STATISTICAL ISSUES AND NOVEL STRATEGIES FOR EXPRESSION QUANTITATIVE TRAIT LOCI MAPPING

By

Shaoyu Li

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Statistics

2011

ABSTRACT

STATISTICAL ISSUES AND NOVEL STRATEGIES FOR EXPRESSION QUANTITATIVE TRAIT LOCI MAPPING

By

Shaoyu Li

Gene regulation is thought to play a pivotal role in determining physiological trait variability by promoting/reducing the expression of functional genes directly or indirectly related to the phenotype. Expression quantitative trait loci (eQTL) mapping studies hold great promise in disentangling gene regulations and have become a popular research area recently. In this dissertation, I explore several statistical strategies, which are applied to eQTL mapping studies, aimed to have a better understanding on the biological mechanism of gene regulation.

The major goal of eQTL studies is to identify genomic regions that are likely to regulate gene expressions. Given that genes function in a network basis, we consider the scenario that a genetic perturbation could lead to a cascade effects on the transcription of multiple genes which belongs to a gene set, e.g., network/pathway. We develop a statistical procedure which incorporates prior biological gene set information (e.g. KEGG pathway, GO terms) into eQTL mapping framework to elucidate gene regulation from a systems biology perspective. Pathway regulators which mediate the expression of genes in a pathway are detected by modeling multiple gene expressions as a multivariate response to test the joint variation changes among different genotype categories. We apply the proposed approach to a yeast eQTL data set. Novel pathway regulators and regulation hotspots are identified.

Currently, most eQTL mapping studies focus on single marker analysis. However, the variability of gene expression may be caused by the regulation of a set of variants that belong to a common genetic system, and individually only with small or moderate effect. To study

the roles of genetic systems in regulating gene expressions, we propose a statistical p-value combination approach to combine individual signals across a pre-defined genetic system to form an overall signal, while considering correlations between genetic variants in the system. Results for simulation studies and the application to the yeast eQTL data are presented.

As part of the DNA sequence variation, gene-gene interaction or epistasis has been ubiquitously observed in nature where its role in shaping the development of an organism has been broadly recognized. Investigating genetic interactions related to mRNA expression is an important step on the path to elucidating the genetic architecture underlying gene expression and could provide valuable functional interpretation of gene regulation. As genes are the functional units in living organisms, we conceptually propose a gene-centric gene-gene interaction framework for genome-wide epistasis detection. Multiple genetic markers (e.g. SNPs) in a gene are modeled simultaneously as a testing unit. We develop a model-based kernel machine approach for detecting pairwise gene-gene interactions. Simulation study and applications of the proposed method to the yeast eQTL data indicate its feasibility to eQTL mapping. We further extend the model-based kernel machine method to binary phenotypic outcomes. Our models provide quantitative and testable framework for assessing the interplay between gene expression and gene regulation,

and will have great implications for elucidating the genetic architecture of gene expression.

I dedicate this work to my parents and my husband.

ACKNOWLEDGMENT

I would like to thank all those people who gave me guidance and help and because of whom the past five years experience of graduate study has become my precious memory.

First and foremost, my sincerest gratitude goes to my advisor Dr. Yuehua Cui, for his encouragement, supervision, patient as well as his continuous support of my graduate study and research. Dr. Cui led me into the fields of statistical genetics and bioinformatics, illuminated my mind and guided me along the way through countless discussions. Dr. Cui's expertise in the subjects of statistics, genetics and biology has been my inspirations to work in the interdisciplinary area. In the past four years, he trained me technical research skills and encouraged me to develop the ability of independent thinking. He is always the one from whom I can get help when I had difficulties or lost directions in my research work.

I appreciate my co-advisor Dr. Shin-Han Shiu from the department of Plant Biology. Dr. Shiu has always been there to listen and give advices. I am deeply thankful for the discussions about my research proposals with Dr. Shiu and for his continuous faith in my research capabilities. I am also grateful to the other two members of my thesis committee, Dr. Marianne Huebner and Dr. Lifeng Wang, for their precious time. My graduate study and research benefited greatly from their statistical wisdom and willingness to share. As the graduate director of the department, Dr. Lijian Yang has been taking very good care of us. I want to thank Dr. Yang for all the attention and encouragement he has given.

I thank all professors and staffs in the department who have taught me and assisted me. My special thanks go to Prof. James Stapleton. He treats us like his children and is always there whenever help is needed.

I would like to thank my husband, Duan Chen, for standing beside me all these years. He is my courage to decide to come to this foreign country for my education. He is the person with whom I am willing to share everything. He cheers for every shining moment of my life and consoles and lifts me up whenever I was down. He makes my happiness doubled and gives me faith to face whatever situation life presents.

My thanks also go to other students in Dr. Cui's group, Dr. Gengxin Li, who is now at Yale University and Mr. Cen Wu, for their valuable help in my daily life and useful discussions about my research. I thank my lovely friends, Dr. Qiongxia Song, Dr. Xiaoqin Tang, Dr. Yun Xue, Ms. Hsiu-Ching Chang and Ms. Shujie Ma for their friendship. We got to know each other in our first year at Michigan State University and the days we spent together have become a beautiful journey of my life.

Last but definitely not least, I thank my parents for their selfless love. They make all possible efforts to help me pursue my dream and respect whatever decision I make. I appreciate the care and support from my brother and sister. I feel so lucky to be in such a big and happy family. My dear family own their credits in every single achievement of mine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 An overview of genetics

1.1.1 Basic concepts

Every single organism on earth has an unique set of chemical blueprints determining how it looks and functions. The chemical blueprints are contained in Deoxyribose Nucleic Acid (DNA). DNA sequence consists four nucleotides, Adenine(A), Thymine(T), Cytosine(C) and Guanine (G) and the sequence of the four bases makes every organism distinguished from others. The bases are paired, A with T and C with G, and the pairs are called base pairs. DNA sequence organized neatly to form *chromosomes*. The number of chromosomes varies widely in different species: for example, Arabidopsis thaliana has 5 pairs, Saccharomyces cerevisiae has 16 pairs, Drosophila melanogaster has 4 pairs and human has 23 pairs. The whole set of chromosomes is defined as *genome* for each species.

Most multicellular organisms are diploid, which means their chromosomes are paired and with one chromosome of each pair inherited from mother and the other chromosome

from father. A *gene* is a segment of DNA that stores the instructions needed to construct components of a cell for making a particular protein (Wikipedia). Genes are known to be the basic functional units in which biological characteristics are inherited from one generation to the next. A site on chromosome is called a *locus*, where one or several genes could reside. At a given locus, there could be multiple forms of DNA sequence. And the total number of forms at the locus depends on the number of copies of each chromosome. For example, diploid organisms have two forms because the chromosomes are paired. Each form of DNA sequence at a given locus is called an *allele*. At a given locus, suppose the two different DNA sequence forms are A and a, then organisms with identical alleles (AA or aa) are homozygous and heterozygous otherwise (Aa). The collection of alleles, AA, Aa and aa are *genotypes* of the locus, which carry unique genetic information for each single subject and contribute to the outcome of certain characteristic traits-*phenotype*, for example, hair color and blood pressure.

1.1.2 Quantitative genetics

Variants in genes or chromosomes may affect the phenotype of a trait. A major goal that lies at the heart of quantitative genetics is to understand the contribution of genetic sequence to the observed variance of a trait of interest. A fundamental idea of classical quantitative genetics is that the phenotypic variance V_P is the sum of genetic variance V_G and environmental variance V_E.

$$V_P = V_G + V_E \tag{1.1}$$

The genetic variance can be further decomposed into the variances of additive, dominance and epistatic components:

$$V_G = V_{GA} + V_{GD} + V_{GI} \tag{1.2}$$

The variance associated with each of these components can be estimated by using the covariance structure for the phenotypic resemblances between groups of relatives. Based on the decomposition of phenotypic variance, the broad-sense heritability is defined as:

$$H^2 = \frac{V_G}{V_G + V_E} \tag{1.3}$$

And the narrow-sense heritability is defined as

$$h^2 = \frac{V_{GA}}{V_G + V_E} \tag{1.4}$$

These two heritability parameters H^2 and h^2 are used to describe the degree of overall genetic contributes for a quantitative trait traditionally [1].

The breakthroughs in genotyping and sequencing technology [2, 3, 4, 5] have accelerated the process toward a complete understanding of the relationship between genotype and phenotype. The emerge of the whole genome sequencing technology in recent years established huge collections of molecular markers for various species, e.g. Yeast, E. Coli, Human, Rice and Soybean [6, 7, 8, 9, 10]. The genome sequence built up detailed maps of chromosomes, showing the precise location of genes and determining the area that can differ from subject to subject. These areas vary among individuals can be used as molecular markers to study the genetic contribution to phenotypes. Different types of molecular markers have being used, including the most commonly used microsatellite markers, also referred as short tandem

repeats (STRs) and single nucleotide polymorphisms (SNPs). STRs happens when a small number of nucleotides (normally, 1-6 base pairs) repeat themselves. The number of repeats differ from one to another and therefore can be used as genetic markers. SNPs are polymorphisms in base pairs throughout the genome. Scientists have been able to genotype up to millions of SNPs throughout the entire genome. Other than these, copy number variation, DNA methylation and histone modification are also applicable as genetic markers.

Based on the genomic maps stands the quantitative trait loci (QTL) mapping methodologies, which aim to identify linkage/association between genomic regions and a quantitative trait of interest [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22].

1.2 Quantitative trait loci (QTL) mapping

1.2.1 Statistical approaches for QTL mapping

Statistical methodologies on QTL mapping have been published through the past several decades. include single-marker mapping [11, 12, 13, 14, 22], interval mapping [15], composite interval mapping [16, 17, 18, 19, 20] and multiple interval mapping [21, 23, 24, 25]. Single marker mapping is the simplest method for QTL mapping studies. At each genotyped marker, one may split phenotypes into groups according to the genotypes at the marker, e.g., a backcross population has two genotype groups (AA, Aa) and an F2 population has three genotype groups (AA, Aa, aa). Then apply either a t-statistic or analysis of variance (ANOVA) or logarithm of the odds (LOD) score to test the variation of phenotypes between different genotype groups, whichever works properly to measure the evidence of linkage between the marker and the QTL. Markers access statistical significance are then identified

to be linked to the QTL of the quantitative trait. Although the implementation of single marker analysis is straightforward and does not require genetic map distance information, the weaknesses of the approach are obvious. By looking at molecular markers only, single marker mapping approaches fail to provide estimates of QTL locations and tend to underestimate the true QTL effect. In most cases, molecular markers are not the true QTL and the recombination between a marker and QTL makes the effect of the marker smaller than the true QTL effect generally. Especially, when markers are widely spaced, a QTL could be far away from a single marker. Then the effect size of the maker could be too small to be detected and therefore mapping power is low.

Interval mapping developed by Lander and Botstein [15] adapted the approach of LOD score analysis to obtain estimates of both the genetic location and phenotypic effect size of a QTL. Rather than focusing on molecular markers only, interval mapping estimates and tests effect size at locations in intervals between markers. By assuming a putative QTL and its location between two markers, the effect of the putative QTL can be expressed as a function of the probabilities of its genotype given the genotypes of the two flanking markers and the genetic distances. Single marker analysis can then be conducted at each putative QTL position between markers, where the test statistic exceeds a threshold level is declared as a QTL. Interval mapping approach overcomes some disadvantages of single marker mapping and has been the most popular mapping approach in the past decade. However, interval mapping still focuses on analyzing one single position at a time. When there are multiple QTLs on a chromosome, the test statistic at the position being tested will be influenced by all the QTLs, consequently leading to biased estimates of QTL effect size and position. For example, a "ghost" QTL between two closely located true QTLs could appear. Besides,

by using only the nearest two flanking markers to estimate the probabilities of the QTL genotype may not be efficient.

To overcome these disadvantages of the interval mapping approaches, Zeng [18] extended the interval mapping to a more efficient and precise mapping approach, termed composite interval mapping (CIM). A similar approach called multiple QTL mapping was also developed by Jansen [19] at the same time. The basic idea about the composite interval mapping is to perform interval mapping through a test statistic which could adjust the effects of linked QTL(s). CIM selects a set of markers as covariates to control the genetic variation of other possible QTLs. Appropriate selection of the set of marker loci serve as covariates is crucial in CIM [26]. Either too many or too few markers being included in the model as cofactors may cause problems. Due to the lack of prior knowledge of the number and positions of underlying QTL, no simple solution for the choice of marker covariates has been developed. And because of the issue, there are recommendations against the use of CIM [27].

All the previous methods were developed to target one single QTL at a time. While for quantitative traits with more than one QTL, especially those with linked or interact QTL, a approach which considers multiple QTL simultaneously will be more accurate and therefore attractive. Multiple interval mapping (MIM) [21, 28, 29] extends interval mapping to multiple QTLs. MIM apply the model selection approach to search for the best model among all possible genetic models that considers the true genetic architecture of the quantitative trait. It is not hard to imagine that the number of all possible models will be huge with large scale molecular markers. Exhaustive search for the best model from the big pool of all possible models incurs huge computational burden. Some strategies that can reduce the burden have been proposed. For example, Sen and Churchill [30] developed a new statistical framework

for quantitative trait loci mapping, which is an approach reside in the middle of pure frequentist and Bayesian. The method splits the multiple QTL problem into two distinct parts: the relationship between the QTL and phenotype and the locations of QTL. A simulation based algorithm was used to estimate the number and genotypes of QTL. With this information, the estimation of QTL effects and interaction becomes easier and computationally less demanding.

1.2.2 Issues and challenges for QTL mapping

How to properly model the underlying poly QTL structure has become the major focus and challenge in the analysis of QTL. There are still issues, such as how many markers should be incorporated as covariates in the CIM method; what is a good estimate of the number of QTL across the genome and how many of them should be include in one model? Besides these issues relevant to modeling, multiple testing issue is a well known problem in the analysis of QTL. Modern QTL study deals with high dimensional molecular marker data. When statistical tests are conducted across the genome, declaring statistical significance becomes a challenging problem. With small number of tests, multiple testing correction can be done by the classical Bonnferroni correction to control the family wise error rate (FWER). When the number of tests increases, especially to a large number, Bonnferroni correction turns out to be too conservative for detecting significant signals. The FDR procedure developed by Benjamini and Hochberg has been proved to be less conservative than approaches controlling the FWER [31]. Motivated by more challenging practical questions with even higher dimension or more complicated correlation structure between tests, many other multiple testing procedures have been developed, including the q-value approach [32]. For more information

on multiple testing issues, readers are referred to some comprehensive reviews [33, 34]. Although the topic has been intensely investigated, no standard has been obtained to evaluate the performance of different correction procedures. FDR procedure and q-value approach are the most widely used ones, which still have limitations with respect to different applications.

1.3 Microarray study and gene expression

1.3.1 Microarray technology

Genes are expressed when they transcript into RNA (transcription). The technology of microarry has been enabled the measurement of thousands of gene expressions simultaneously. In a typical microarray experiment, two mRAN samples are isolated and then fluorescently labeled and hybridized to a platform surface. Microarry platforms is a glass slide or membrane arrayed with DNA fragments or oligonucleotides that represent specific gene coding region in a regular pattern. The relative amounts of transcript RNA can be measured by laser scanning or autoradiographic imaging after thorough washing. cDNA array and Affymetrix Gene Chips are the two widely used microarray platforms. With the data generated by microarray experiment, a number of statistical methods can be applied to infer the underlying dynamic functioning.

Studying the change in expression pattern offers an opportunity, that all technologies aforetime could not provide, to understand the molecular mechanisms underlying biological processes in a cell. While, the original scanned gene expression raw data contains noise, missing values and systematic errors within and between arrays, hence must be pre-processed before any other quantitative tools, e.g. clustering and classification, can be properly applied.

Problems relevant to microarray data pre-processing includes normalization [35, 36, 37], missing values estimation [38, 39, 40, 41] and pre-selection which chooses only those informative genes for later analysis [42].

After the pre-processing, those genes show statistically significant differences across different conditions, such as mutant versus wild type, healthy versus diseased or multiple tissues, can be reported by studying their expression profiles. It is also possible to compare gene expression profile over multiple time points. The analysis of expression data reveals valuable insights of gene function and gene regulation. For example, by comparing gene expression patterns of healthy people (controls) and people with a certain disease (cases), genes over- or under-express in cases may be identified to have association with the disease and the function of an unknown gene could be inferred from genes in the same cluster.

1.3.2 Clustering analysis

There are many ways to analyze gene expression data. Gene expression clustering intends to group "similar" genes into the same cluster and the "dissimilar" ones into different groups. Two crucial questions in clustering analysis need to be answered are: how do we define similarity and how to decide whether two genes are similar enough to be clustered into one group? Depends on the objects of a study or project, investigators may have their own expectation of how the clusters would look like. For example, a study aim to study gene functions would like to have genes share similar function (in a same functional circuit/pathway) clustered together. The expectation leads to project-oriented definition for "similarity" as well as the clustering algorithm. Hundreds of clustering algorithms have been developed and some of these algorithms have been applied to gene expression analysis since the huge amount of

data became available. Inventing a new clustering algorithm is easy; inventing a clustering algorithm which is computationally efficient and able to provide accurate and biologically meaningful gene clusters for studies with specific targets is challenging; and inventing a clustering algorithm that performs well under all kinds of circumstances could be mission impossible. It is possible that a clustering algorithm works perfect for one study leads to disaster for another. No standard criteria that can be used to evaluate the performance of all different algorithms up to date. And there's no single algorithm can perform well uniformly in different studies. Statistical issues raised up in the clustering analysis of gene expression profiles include finding appropriate similarity measures, deciding whether two genes are statistically significantly related and a fair evaluation procedure for the performance of different clustering algorithms.

The microarray expression profiling shows its advances in unraveling valuable biological insights. This high-throughput study of multiple genes appears to be more powerful when integrated with genetics. Jansen and Nap proposed a new manner of study termed genetical genomics [43]. It is the study that treats expression level of each single gene of the thousands of genes as quantitative traits and map genomic regions responsible for the gene expression variation. Genetical genomics combines the studies of gene expression profiles and traditional QTL mapping and therefore also referred as expression QTL (eQTL) mapping study. eQTL mapping has being applied widely to different model organisms because of its great promise in inferring gene regulation. The details of a general framework for eQTL mapping will be discussed in the following section.

1.4 Expression quantitative trait loci (eQTL) mapping

Traditional quantitative trait loci (QTL) mapping has been focused on identifying genetic loci responsible for the phenotypic changes of a quantitative trait. Such studies are designed to detect linkage between genetic markers and the functional (causal) variants responsible for the phenotypic variability, and thus fail to disentangle the functional mechanisms of variants due to the regulation of genes. It has been commonly recognized that gene regulations play pivotal roles in determining trait variations in natural populations by promoting or reducing the expression of functional genes directly related to the trait. The obvious difference in the looks and functions between organisms is not due to the static chemical blueprints but rather to the complexity of dynamic functional mechanisms of gene regulation.

The recent advances in microarray technology open an alternative front for multiple gene discovery by studying thousands of gene expression profiles simultaneously under certain conditions or treatments. As an intermediate molecular phenotypes that associates genetic variants with physiological outcomes, e.g. human diseases, analysis of gene expression holds great promise to infer genes accompanying a disease trait, and serves as an alternative to identify novel relationships among genes. A number of studies have shown repeatedly that gene expressions are inheritable traits, thus can be used for genetic mapping [44, 45, 46]. eQTL mapping study successfully integrates the two endeavors, genetic mapping and gene expression analysis, and allows the systematic insights into the biology of gene regulation.

By assaying genome-wide gene expressions and genotype profiles for individuals in a mapping population (e.g. F2, RIL and natural human population), the traditional statistical QTL mapping tools can be used to map the genomic regions that are responsible for the variation in transcriptional abundance of thousands of genes (Figure 1.1). Single trait-single

marker eQTL mapping analyzes the quantitative level of the transcript of one single gene at a time to map eQTL [46]. By assembling eQTL for all genes across the genome, we can build up a comprehensive two-dimensional diagram in which positions of detected eQTL are plotted against the positions of genes for which eQTL was mapped. The diagram provides an important reference source for characterizing *cis* and *trans* regulations. eQTL whose physical location is close to the actual location of the target gene (e.g. 10kb or 20kb up and down the gene region) are defined as *cis*-acting regulators of the gene. And the linkage between the gene and the eQTL is a *cis*-linkage. In contrast, eQTL located remotely away the target gene is *trans*-acting regulator and the linkage is a *trans*-linkage. For example, if a gene on chromosome 10 transcribes to create a transcription factor which regulates the expression of a gene on chromosome 1, then the gene on chromosome 10 *trans*-regulates the gene on chromosome 1. Generally, *cis*-regulators have bigger effects on gene expression and therefore are easier to be detected. Gene expressions are found to be polygenic [45, 47] and the distribution of eQTL across the genome is not uniform. Some genomic regions linked to multiple genes [47] are regulation *hotspots* or *"master regulator"*.

eQTL mapping study involves analysis of large volume of data. Statistical strategies and methods are needed in every step of eQTL mapping: experimental design, data pre-processing, mapping and down-stream functional analysis. Take the mapping stage as an example, efficient and intuitive approaches to the identifications of eQTL for transcription levels are crucial. Besides, multiple testing issue is another well known challenge in large scale eQTL mapping studies. Multiple testing corrections are implemented differently in various studies and no standard statistical approach addressing the multiplicity issue in eQTL mapping study has been developed so far. There are quite a few review articles for

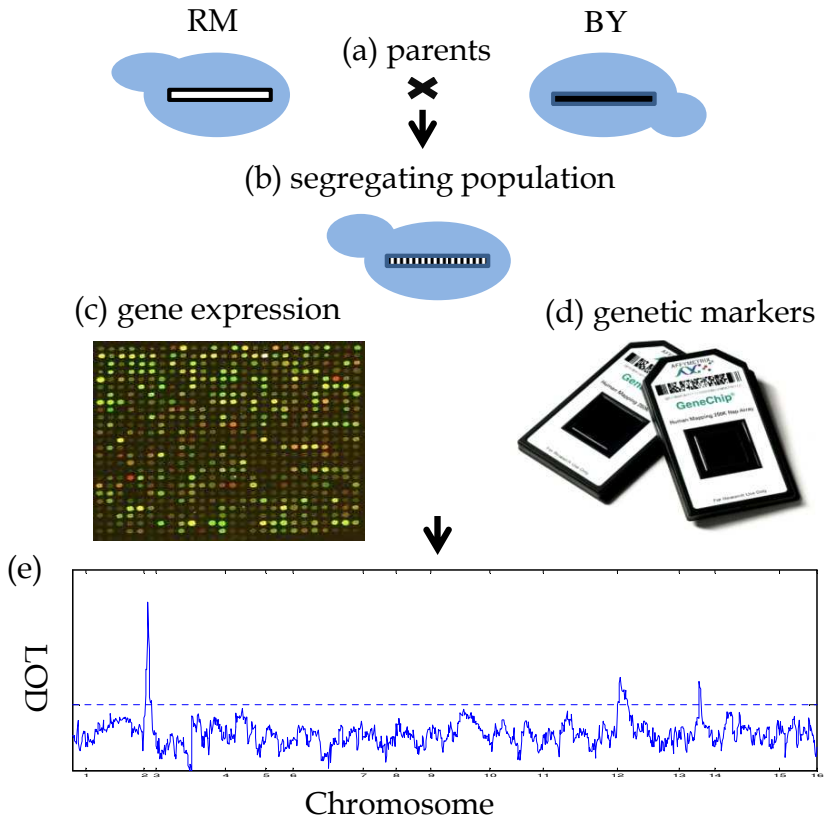


Figure 1.1: A typical design of the eQTL experiment. (a) Two genetically distinct strains (BY and RM) are chosen as parent strains. (b) Segregants are generated by crossing the two parent strains. (c) Gene expression levels of all segregants are measured by microarray. (d) Genetic markers across the genome are genotyped. (e) eQTL mapping results for one gene expression. (For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.)

eQTL mapping in the literature. Readers are referred to [48] for a review of statistical methods in eQTL mapping, and to [49, 50] for a general review of eQTL mapping studies.

1.5 Objectives and organization

Despite the successes of eQTL studies, there are open questions remained to be answered. The classification of *cis* and *trans* regulation relies on how a researcher defines gene regions.

Due to the limited knowledge on the explicit genomic map, no standard definition of gene regions is available. Misclassification of *cis* and *trans* regulation is possible when inappropriate distance rule is applied. For example, a transcription factor near its target gene could be referred to as a *cis* regulator. Moreover, studies of eQTL in a variety of organisms observe that the polygenic basis of transcriptional variation is complex. One single regulator may regulate the expression of multiple genes and the cumulative effects of multiple regulatory elements could function jointly to affect the expression of a single gene or multiple genes in a network. Figure (1.2) shows various gene regulation pattern. Additional hallmarks of the complexity of genetic basis of expression changes includes gene×gene interaction, gene×environment interaction and pleiotropy. A complete understanding of the genetic basis of gene expression is far more than the broad claims of *cis* and *trans* regulations.

One of the major goals of eQTL mapping studies is to elucidate gene regulatory principles and ultimately gain knowledge of the genetic architecture underlying complex physiological phenotypes [43, 49, 51, 52]. It is the aim of this dissertation to develop statistical mapping methods and strategies which can be applied to eQTL studies to obtain novel findings about gene regulation.

The content of the dissertation is organized as follows. In chapter 2, we propose a statistical approach to identify *pathway regulators*, eQTLs that regulate transcriptional variation of pathways. A p-value combination approach which considers correlations between individual p-values to infer pathway regulation is developed in chapter 3. The method combines individual p-values across pre-defined genetic systems to infer the cumulative effect of the whole genetic system in regulating gene expressions. Chapter 4 introduces a statistical framework for detecting gene×gene interactions from a conceptually novel gene-centric perspective. A

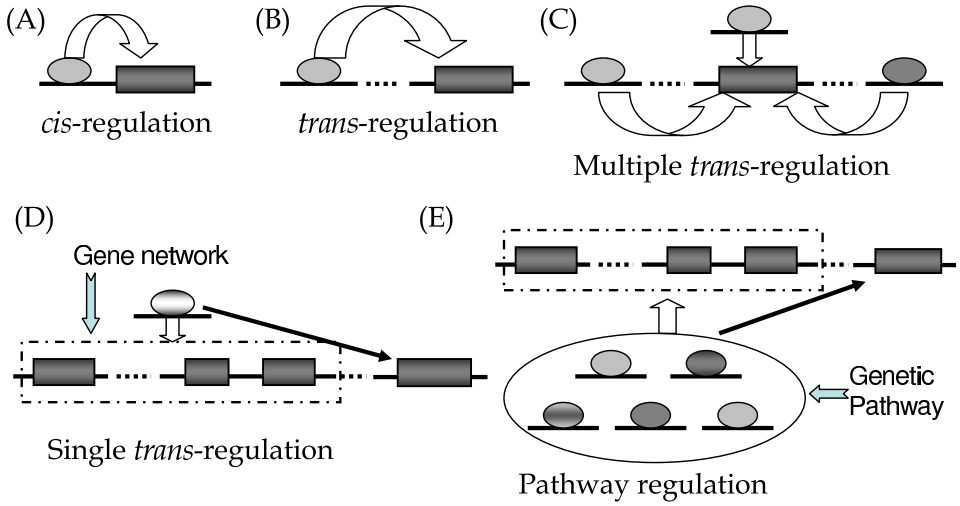


Figure 1.2: Various patterns of gene regulation: (A) *cis*-element regulates its own gene expression; (B) *trans*-element regulates downstream gene expression; (C) multiple *trans*-elements regulate the same gene expression; (D) single *trans*-element regulates single gene expression or multiple gene expressions in a network (i.e., gene network); and (E) multiple regulators in a genetic pathway function jointly to regulate multiple gene expressions in or not in a network. The shaded ovals and rectangular represent regulatory elements and coding genes, respectively. The dotted lines imply that genes are located in different regions.

model based kernel machine approach for investigating gene×gene interaction effects underlying quantitative traits is developed. The approach is extended to studies with binary phenotypic outcomes in chapter 5.

Chapter 2

A systems biology approach for identifying novel pathway regulators in eQTL mapping

2.1 Background and motivation

Most current eQTL mapping studies treat each gene expression as one single trait. The so called single trait analysis may not be powerful enough to identify genetic variants responsible for gene expression changes given that genes function in networks. Wessel et al. found in their eQTL mapping study that many SNPs are responsible for the expression change of genes belonging to a certain pathway [53]. It is commonly recognized that genes in a biological pathway, e.g. metabolic pathway, developmental pathway or signal transduction pathway, "cooperate" with each other and function as a team to fulfill their designated tasks. Different expression of one gene, especially those that play key roles in the pathway, would influence

expression levels of other genes in the same pathway. Thus, a signal perturbation of a particular gene in a pathway would induce a cascade of biochemical events that affect all, or many of the other genes belonging to the same network or pathway. Take this functional mechanism of a pathway into account, the current broad claims of *cis*- or *trans*-regulation detected with the single trait analysis might not be sufficient and efficient enough to capture the relationship between genetic variations and gene expressions.

Mootha et al. [54] have previously showed that focusing on expression data in terms of predefined pathways can provide valuable insights not easily achievable by methods focus on individual genes. Many scientists are thus interested in identifying which genetic variant mediates the expression change of a pathway. The identified regulator, termed *pathway regulator*, provides additional information about the function of gene regulation from a systems biology perspective. A number of eQTL studies have incorporated prior biological pathway information into their analysis [55, 56]. And most of these studies implement a two-stage procedure: perform single trait analysis in the first stage and then conduct gene set enrichment analysis (GSEA) to test whether an expression pathway is enriched at a particular locus. The two-stage approach obviously does not take care the correlation, which is common between genes in a pathway. Moreover, the accuracy and efficiency of the enrichment analysis in the second stage depends heavily on the results of the first stage. When multiple genes function jointly, each with small marginal effects, the approach may fail to identify important pathway regulators. Another disadvantage of the two-stage analysis is the multiplicity issue. With thousands of gene expression profiles, single trait analysis have to adjust for the large number of tests when declaring significance. This may lead to low power in identifying genes with small marginal effects, which again affects the power of the second

stage enrichment analysis.

Considering the limitation of the current single trait-based analysis and motivated by the biological phenomenon, we propose to identify common pathway regulators by treating gene expressions belonging to a common pathway as a multivariate response, and focus our interests in identifying pathway regulators that mediate the expression change of a particular biological pathway or process. More importantly, when multiple gene expressions are jointly considered, the multiple testing burden in a single trait analysis is potentially reduced, and hence leading to increased power. For the illustrations in the chapter we restrict ourselves to one yeast dataset [47]. Our analysis indicates there are potential pathway regulators in regulating pathway expressions. Applying commonly used hotspot detection method, we identified several pathway regulator hotspots. We also performed an enrichment analysis to test which genetic pathway is enriched in regulating the expression change of a certain pathway, and found significantly enriched genetic pathways in regulating other pathway expressions.

2.2 Methods

2.2.1 eQTL dataset

The yeast dataset was generated from 112 meiotic recombinant progeny of two yeast strains: BY4716 (BY; a laboratory strain) and RM11-1a (RM; a vineyard isolate) aimed to understand the genetic architecture of gene expressions. The dataset contains expression profiles of 6216 gene expression traits and 2956 SNP marker genotype profiles. For more details about the dataset, see [47].

In the yeast genotype profiles, genotypes of neighboring markers tend to be very similar and some are even identical. For those SNP markers showing high correlation, we follow the strategy proposed by Sun [57] to construct marker blocks, in order to remove redundancy and reduce the genotype dimension. Specifically,

i) *Merge markers into marker blocks*: Define $u = (u_1, u_2, \cdots, u_n)^T$ and $v = (v_1, v_2, \cdots, v_n)^T$ as vectors of two SNP genotype profiles over n individuals. Each SNP is coded as 0 or 1 depending on whether it is inherited from BY or RM strain. The Manhattan distance between the two SNP genotype vectors is defined as

$$MD = \sum_{i=1}^{n} |u_i - v_i|$$

The value of MD indicates the degree of overlap of the two SNP markers. A small value would indicate much overlap between the two markers. We include a SNP marker into a marker block if the Manhattan distance between a marker and any markers in its neighborhood is less than a predefined value r. In our analysis, we set $r = 1$. Other values like 1.25, 1.5 could also be used, depending on how strict the constrain you want to put on the marker similarity. If either u_i or v_i is missing for any individual i, the term $|u_i - v_i|$ is excluded from the summation, and the MD measure is adjusted by multiplying a factor $\frac{n}{n-m}$ [57], where m is the total number of terms been excluded. By setting $r = 1$, 1168 marker blocks are obtained.

ii) *Define genotype profiles for each marker block*: We find consensus of each marker block and then dichotomize it. An individual genotype is set to 0 or 1 if at least 75% of the markers in a block equals to 0 or 1 for that individual. Otherwise it is set as missing.

Individuals with missing genotypes will be eliminated for further analysis.

In the following presentation, when we see a marker which really means a marker block. Quite often some of the marker blocks only contain one SNP marker and some may contain more than one marker. We interchange the two words "marker" and "marker block" frequently and they do mean the same thing.

2.2.2 Genome-wide pathway regulator identification

We focused our analysis on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and we extracted 99 pathways from the R package: YEAST. Let $Y_i = \{y_{i1}, \cdots, y_{ip}\}^T$ be a vector of gene expressions in a pre-defined pathway for the ith subject, where p is the size of the pathway. Assuming the 1168 SNP marker blocks are the causal genetic variants responsible for the gene expression changes, we test on one marker block at a time, ignoring any variants located at the interval flanking any two marker blocks. Thus, for each marker, there are two genotype categories with each one corresponding to one multivariate expression profiles. To test the differential expression pattern between different genotypes at a locus, a Hotelling's T^2 test can be applied which has the form,

$$T^2 = (\bar{Y}_0 - \bar{Y}_1)^T [(\frac{1}{n_1} + \frac{1}{n_2}) S_{pooled}]^{-1} (\bar{Y}_0 - \bar{Y}_1) \tag{2.1}$$

where $\bar{Y}_j = \sum_{i=1}^{n_j} Y_i$ and n_j are the sample mean expression vector and sample size for genotypes coded as j $(j = 0, 1)$, respectively. Assuming equal variance for expression values in the two genotype categories, a pooled variance estimation S_{pooled} can be used in defining the T^2 statistic. The T^2 test is performed for all 99 KEGG pathways at every marker block across the genome.

Theoretically the T^2 statistic follows a scaled F distribution [58]. To control the family-wise error rate across the whole genome, we perform a permutation test to determine the genomewide cutoff. When doing permutations, each row vector of gene expression is considered as one observation to retain the gene correlation information within a pathway. Then we fix the genotype information and randomly sample expression vectors without replacement. This random reshuffling procedure disturbs the relationship between gene expressions and genotypes. One thousand permutations are conducted to generates a null distribution for the T^2 statistic. For each permutation, T^2 values for all marker blocks are calculated, and the maximum T^2 value is recorded. The 1000 maximum T^2 values represent the genome-wide null distribution of the T^2 statistic in which the 95th percentile is considered as the genome-wide cutoff. A SNP marker block is considered as a pathway regulator if the observed T^2 value is greater than the cutoff value.

The T^2 test is performed when the number of genes in a pathway is less than the sample size. However, in real applications some pathways may contain large number of genes ($p > n$). Given the small sample size (total 112) in the yeast dataset, this does happen (e.g., pathway '04111' and '03010'). When this happens, instead of using the T^2 statistic, we propose to use the F statistic proposed by Zapala and Schork [59]. Consider a multivariate regression model,

$$Y = X\beta + \epsilon \tag{2.2}$$

where Y is a multivariate response (e.g., gene expression in a pathway), X is the design matrix for SNP genotypes. When only one SNP marker is considered, X is an $n \times 2$ matrix with ones in the first column and numerical genotype coding in the second column. The F

statistic proposed by Zapala and Schork has the form,

$$F = \frac{tr(HGH)}{tr[(I-H)G(I-H)]}$$

where H is the hat matrix of the multivariate regression model in (4.1), and

$$G = (I - \frac{1}{n}\mathbf{1}\mathbf{1}')A(I - \frac{1}{n}\mathbf{1}\mathbf{1}')$$

where matrix $A = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$ is a so called distance matrix which measures distance between expression levels of genes in a pathway; $\mathbf{1}$ is a column vector of ones and I is an identity matrix. An easy way to form the distance matrix is to use the correlation matrix and transform them with simple transformation technique, i.e. $d_{ij} = \sqrt{2(1-r_{ij})}$ where r_{ij} is the correlation between genes i and j [59].

The F statistic is specially useful when the number of parameters p is larger than the sample size n [59]. However, it is not trivial to find the theoretical distribution of the F statistic. Here, we still conduct a permutation procedure to assess the statistical significance. When $p = 1$, the F and the T^2 statistics are identical if the distance matrix is computed through the use of the standard Euclidean distance measure. For pathways with small number of genes, results obtained with the two statistics are also very consistent. For pathways with large number of genes, the F method is more time demanding due to large matrix operation. Thus, we only apply this method to pathways with $p > n$.

2.2.3 Pathway regulation hotspot detection

eQTL hotspot is defined as the genetic region where there are a large number of gene expressions are mapped to than by random [60, 61]. For traditional eQTL hotspot detection, genomic regions are generally defined as "bins" with each "bin" covering a genomic interval in a length of, say 5Mb (in humans)[60]. Similar to regular eQTL hotspot detection, we can also identify pathway regulation hotspots. Let N_l ($l = 1, \cdots, L(= 1168)$) be the number of pathways which are significantly mapped to marker block l. Let $N = \sum_{l=1}^{L} N_l$ be the total number of pathways significantly mapped to the whole genome. Then a Poisson distribution can be assumed for each N_l with the mean parameter λ estimated by the empirical mean N/L. Consider each marker block as one potential hotspot, the probability of observing N_l or more significant pathways mapped to a marker block can be considered as the hotspot p-value, denoted as p_l. Take the Bonferroni correction at the 0.05 genome-wide significant level, a marker block is considered as a pathway regulation hotspot if $p_l < 0.05/L$. Alternatively, we can combine neighborhood marker blocks as one synthetic block with a pre-defined length, e.g., 20kb length. Then the total genome can be divided into $K(< L)$ segments. Following the same procedure described above, pathway regulation hotspots can also be tested.

Relaxing the Poisson assumption, we can also use a nonparametric permutation procedure to identify regulation hotspots. Let $Q = (q_{ij})$ be a matrix which contains the mapping results, where $q_{ij} = 1$ if pathway i ($i = 1, \cdots, 99$) is significantly mapped to locus j ($j = 1, \cdots, 1168$), and $q_{ij} = 0$ otherwise. We randomly permute the positions for 1's for each row of matrix Q and generate 1000 permuted matrices $Q_1^*, Q_2^*, \cdots, Q_{1000}^*$ while keeping the row

sums of all these Q_p^*'s the same as the row sum of original observed matrix Q, i.e.,

$$\sum_{j=1}^{1168} q_{p,ij}^* = \sum_{j=1}^{1168} q_{ij},\ i = 1, 2, \cdots, 99, p = 1, 2, \cdots, 1000$$

The distribution of column sums for each permuted matrix is recorded. A locus is declared as a regulation hotspot if the observed count at that locus is larger than the 95th percentile of the permuted distribution.

2.2.4 Genetic pathway enrichment analysis

In a recent genome-wide association study for identifying disease risk variants, Wang et al. [62] first proposed a pathway-based association study to map genetic pathways involving multiple genetic variants functioning together to give rise to a disease phenotype. Motivated by this idea, we expect certain genetic pathways be enriched in responsible for the expression change of an expression pathway. By genetic pathway we mean SNP variants that belong to a common pathway. Here we use GP to denote a genetic pathway and use EP to denote an expression pathway. The purpose of this analysis is to identify which GP is enriched in mediating the expression change of an EP. For an enriched GP corresponding to an EP, we anticipate the expression variation of that EP can be explained by the joint function of SNPs in that GP.

From the genome-wide analysis, we can obtain a list of significant pathway regulators (markers) corresponding to an EP. Total 1149 unique genes (including annotated and non-annotated) are extracted from the whole genome. GPs are then grouped according to the KEGG pathway information. We call a gene is significant if there is at least one marker in this gene is significant. It is possible that there are several markers in a gene are significant.

Similar as the EPs, there are total of 99 GPs retrieved from the KEGG database. Fixing each EP, we can test which GP is enriched to explain the expression variation of an EP. Let n_S be the total number of genes that are significantly associated with an EP. Let n_G be the number of genes that belong to a GP, among which S are significantly associated with an EP. Then we can formulate a 2×2 table shown in Table 2.1. The Fisher's exact test can be applied to calculate the enrichment p-value which is then compared with a significance level α. We use a less conservative α value, i.e., $\alpha = 0.01$ to declare GP enrichment.

Table 2.1: A simple layout for testing genetic pathway enrichment

	No. of genes in a GP	No. of genes not in a GP	Total
No. of sig. genes	S	$n_S - S$	n_S
No. of non-sig. genes	$n_G - S$	$K - n_G - n_S + S$	$K - n_S$
Total	n_G	$K - n_G$	K

$K(= 1149)$ is the total number of unique genes covering the marker blocks across the genome.

2.3 Results

2.3.1 Pathway regulators

We combined prior pathway information (e.g., KEGG) with the proposed pathway mapping approach to detect pathway regulators. There are totally 99 pathways retrieved for the Yeast package in R for this dataset as listed in the supplement Table (A.1). At each marker block position, a T^2 or F statistic was calculated for each gene expression pathway depending on the size of the pathway. We illustrate the idea with one pathway: MAPK signaling pathway. Figure 2.1 shows the T^2 profile plot for the pathway across the 16 yeast chromosome. The horizontal dash-dotted line in the plot indicates the 5% genome-wide threshold by permu-

tation tests. Genomic positions where the T^2 peaks passing the threshold are considered as potential pathway regulators. For this pathway, we identified several pathway regulators on chromosome 2, 3, 5, 8, 14, 15 and 16. A full plot of the T^2 (or F) profiles for all the 99 pathways are listed in a appendix file given by [63].

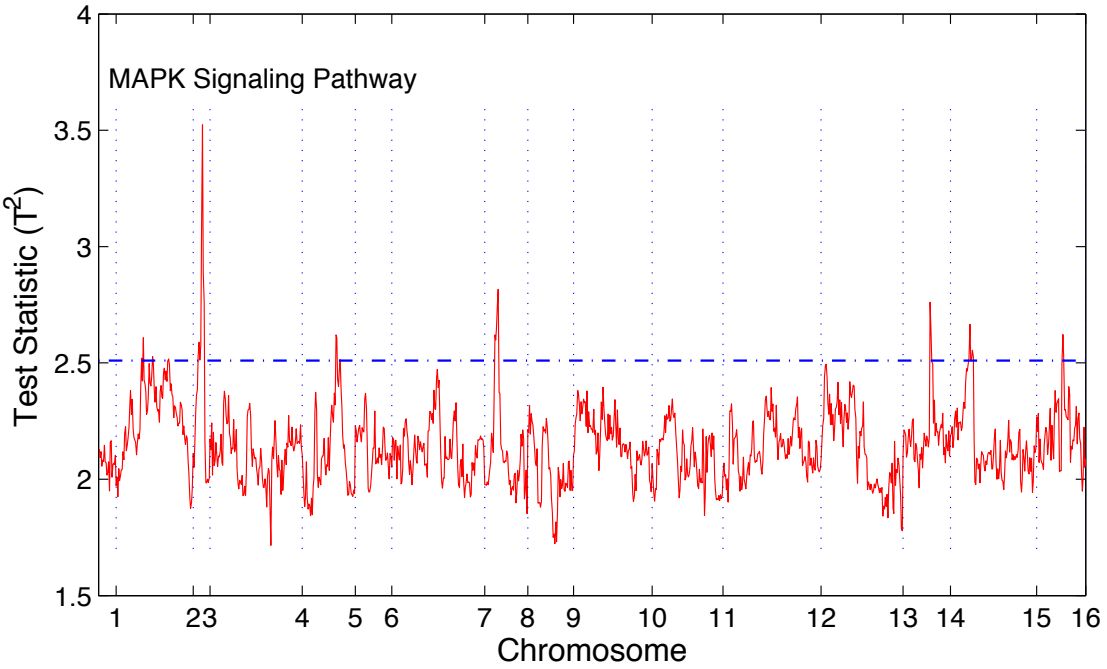


Figure 2.1: The T^2 profile plot across the entire yeast genome (16 chromosomes).The dash-dotted horizontal line is the 5% genome-wide permutation cutoff. The vertical dotted lines separate different chromosome regions. The peaks of the T^2 profiles that pass the cutoff correspond to potential pathway regulation loci (e.g. on chromosome 2, 3, 5, 8, 14, 15 and 16). Both the cutoff and the T^2 values are log10 transformed.

As indicated by the whole genome scan of all the 99 pathways, we can see consistent strong association signals on chromosome 3, 14 and 15 which indicates that there are important pathway regulators located on these three chromosomes. Since large number of EPs are regulated by these regulators, they are potentially "master" pathway regulators. For example, SNP marker YCL009C (ILV6) on chromosome 3 regulates 39 EPs and its neighborhood

genes (e.g., LEU2 and BUD3) also regulate large number of EPs.

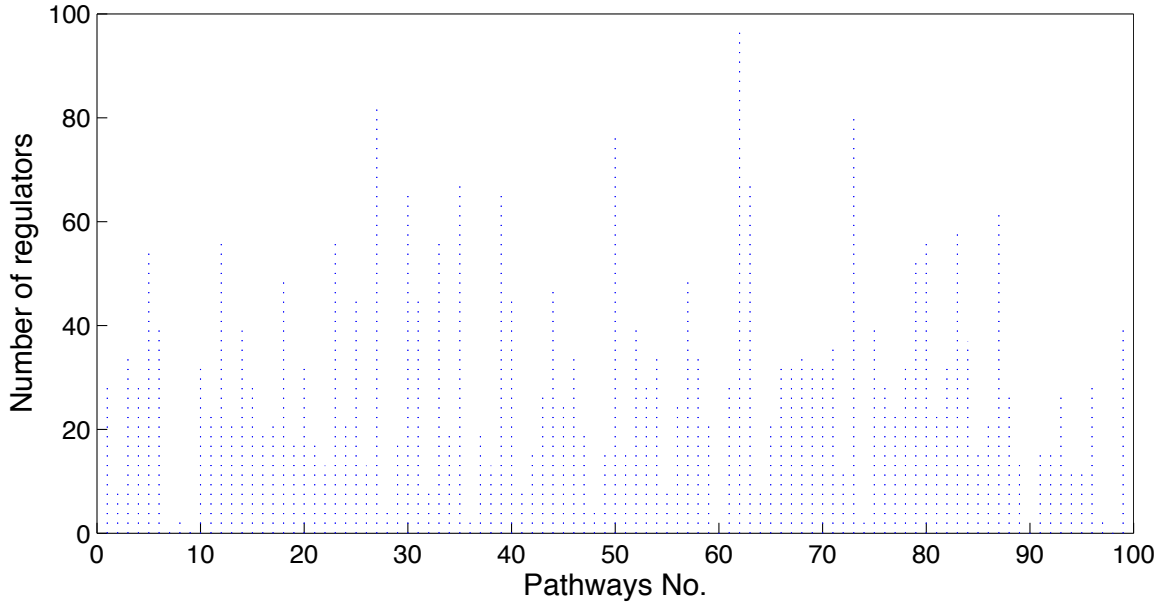


Figure 2.2: Number of regulators for each expression pathway. The horizontal axis denotes the 99 KEGG pathways and the vertical axis denotes the number of marker blocks that are significantly associated with each expression pathway.

Figure 2.2 shows how many regulators each EP has. All the 99 EPs are plotted in the horizontal axis and the vertical axis indicates the number of regulators each pathway has. We can clearly see that the expression of some pathways are affected by many genetic variants. For example, pathway 62 (Pyruvate metabolism pathway) has 98 regulators. Some pathways are not regulated by any variants, for instance, pathway 60 (ABC transporters - General) and 90 (Two-component system - Organism-specific). Note that many markers are highly correlated in this yeast dataset. Even though we merged some markers with large proportion of overlaps, we still expect large number of markers to be highly correlated with neighborhood markers. Thus, Figure 2.2 only gives us a rough idea of how each pathway expressions are affected by many regulators. Whenever there is a causal regulator presented

in a genomic region, due to strong linkage disequilibrium (LD) between neighborhood markers, its neighborhood markers might also show strong association signals. Thus, the true regulators for each EP might be smaller than the reported numbers.

2.3.2 Pathway regulation hotspot

In eQTL mapping study, people are often interested in knowing which genomic region or interval plays important roles in regulating gene expressions. The so identified regions or intervals are called eQTL hotspots [60]. Since we merged some markers to form marker blocks, we simply treat each block as one potential pathway regulation hotspot and assess its significance. We counted the number of pathways being regulated by a marker block across the genome. The average number of association for one marker block is $\hat{\lambda} = 2.52$, and none of the marker block was expected to contain association with more than 10 pathways by chance at the 5% genome-wide significant level after Bonferroni correction (0.05/1168). We detected total 76 pathway regulation hotspots. Figure 2.3 shows the distribution of the identified hotspots. The horizontal dash-dotted line indicates the threshold calculated from the Poisson model. The vertical bars indicates the number of pathways regulated by each marker block. Significant pathways at the hotspots are indicated by red color and all other significant pathways are indicated by cyan color. We identified serval pathway regulation hotspot groups located on chromosome 2, 3, 5, 10, 12, 13, 14 and 15. Chromosome 5 and 15 show two distantly located hotspots.

It is interesting to note that most of the hotspots are clustered together on the genome. Some clusters have narrow band (e.g., the ones on chromosome 5, 10, 14 and 15), and some have wide band (e.g., the ones on chromosome 2, 12 and 13). As we noted from the marker

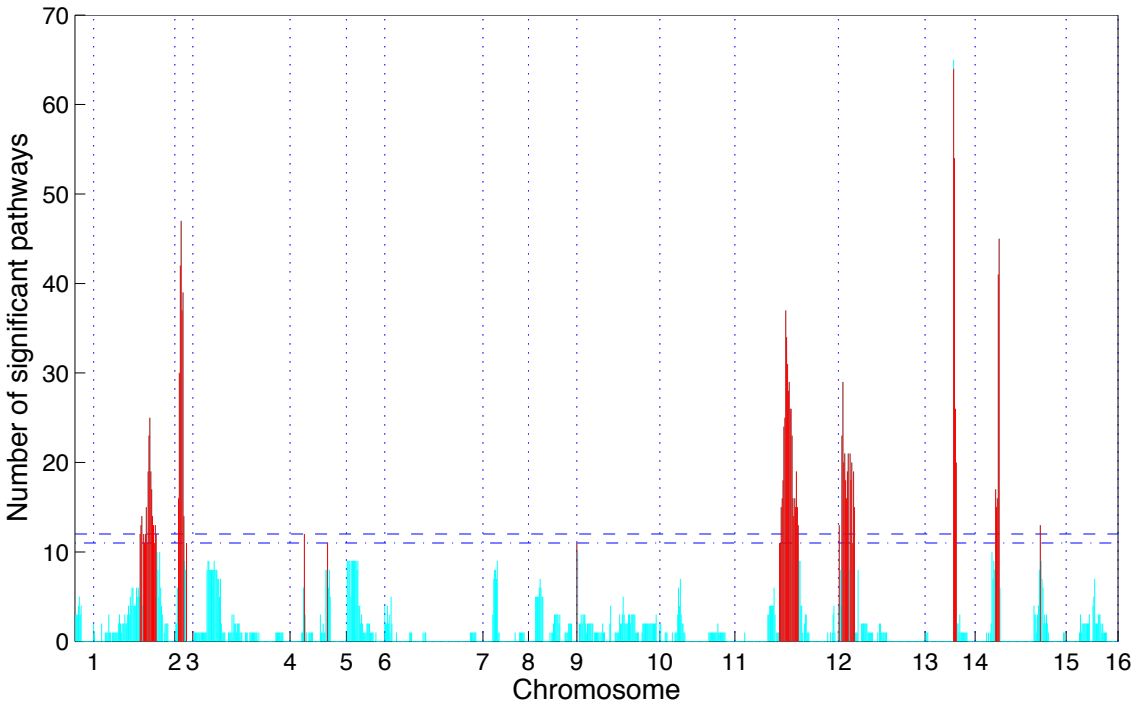


Figure 2.3: The pathway regulator hotspot. The dash-dotted and the dashed lines indicate the threshold calculated using the Poisson distribution and the permutation method, respectively. Red bars indicate the number of EPs regulated by hotspots and cyan bars indicate all other significant EPs not mapped to the hotspots.

data, there are strong correlations (LD) between markers for this yeast dataset. Thus, this kind of pattern is expected. If we increase the hotspot interval size, the hotspot band would become narrower with more sharp peak. Noted also the clustered pattern for other regulation loci due to high LD between neighboring SNP markers.

We also applied the permutation method to detect regulation hotspot. When using the permutation method, the cutoff is changed to 11. The horizontal dashed line in Fig. 2.3 indicates the permutation cutoff. Loci with more than 11 associated pathways were identified as hotspots. With increased threshold, the number of regulation hotspots reduced to 67. Several hotspots including the ones on chromosome 5 and 10 are no longer significant

with the new threshold value. Overall, the two methods for identifying regulation hotspots give quite similar results. A detailed lists of the hotspot regulation are given in appendix, Table (A.2).

In a recent study of genetic basis for small-molecule drug response, Perlstein et al. [64] detected eight QTL hotspots located on chromosome 1, 3, 12, 13, 14, 15 with the same yeast marker data. Five of those (on chromosome 3, 12, 13 and 14) overlap with the hotspots we identified. This information indicates the relative importance of these four genomic regions in regulating gene expressions as well as drug responses. It is possible that the variation in drug response is due to the variation in pathway expressions which are directly related to hotspots regulation. Models can be developed to test this type of causal relationship [65], and will be considered in future work.

2.3.3 Genetic pathway enrichment

In addition to identify pathway regulation hotspots, we also performed a functional enrichment analysis using the Fisher's exact test to assess if a GP is enriched in regulating the expression of an EP. An enrichment p-value was computed to reflect the degree in which a given GP is over-represented. The results are tabulated in Table 2.2. A heatmap of the pathway enrichment analysis is given in the supplemental figure-Figure (A.1). To make the table consistent with the supplemental figure (A.1), we also listed the pathway number (denoted as #) in addition to the pathway identification number (PID). The left column shows the enriched GPs which are responsible for the expression change of the corresponding EPs in the left column. All enriched GPs are claimed at the 1% significance level.

Clearly, pathway 15, 20 and 74 are relatively important in regulating the expression of

other pathways, since each one is enriched for a large number of EPs. It is hypothesized that the signal perturbation of these pathways may have pleiotropic consequences on multiple downstream pathways. Particularly for pathway 20, it may act as potential "master" pathway regulators as it regulates 25 pathway expressions. Also noted that most enriched GPs are metabolism or biosynthesis related pathways, which may indicate that these pathways might play key roles in the yeast genome.

In testing GP enrichment, we found that some GPs are enriched in regulating its own gene expressions. We define these GPs who regulate their own gene expressions as *cis*-pathway regulators. The highlighted bold-font pathways in Table 2.2 are those that show strong *cis*-regulation effects. These six GPs are pathway 13, 15, 20, 27, 43 and 44. All the others show *trans*-regulation effects. Noted that the enriched GPs are claimed at the 0.01 significance level. If we lower the significance level to 0.001, all the *cis*-regulation pathways are gone, indicating that the *cis*-regulation effect is actually weaker than the *trans*-regulation effect in this application.

In a closer look at the enriched GPs, we found that pathway 20 contains two SNP markers (YCL009C in gene ILV6 and YCL018W in gene LEU2) that are located on the hotspot on chromosome 3. LEU2, beta-isopropylmalate dehydrogenase, plays an important role in catalyzing the third step in the leucine biosynthesis pathway. Pathway 78 also contains two SNP markers (YBR176W in gene ECM31 and YCL009C in gene ILV6). In checking the KEGG pathway, we found that ILV6 is on the most upstream in pathway 78. Thus this gene may play a key role in affecting the downstream gene functions and in turn affecting many other pathway expressions.

2.4 Discussion

Understanding the genetic architecture of complex traits is one of the major challenges in modern biology. In a series of recent advances, many efforts have been focused on mapping genetic regions, called QTLs, in responsible for the phenotypic variation of a complex trait. Due to limited mapping resolution and other non-genetic factors contributing to the phenotypic variation, this process has not been very successful in real applications, leaving only a few successful cases being reported in literature [66, 67]. Recent advances in microarray technology allows us to measure the transcription abundance of many organisms and hence open another framework in understanding the genetic basis of gene expression, with an aim to shed new light on the regulation of a genetic system. The initiation of the eQTL mapping with combined genetic mapping and gene expression analysis brings new prospect in understanding the complex process of gene regulation toward the ultimate goal of improving trait quality and disease prevention [68].

Based on the biological assumption that genes function in networks, dissecting the genetic architecture of gene regulation from a systems biology perspective should provide more insights regarding the function of a biological system. In this article, we made an attempt to study gene regulations by combining gene expression and genetic polymorphism data together and proposed a pathway-based systems biology approach that aims to identify genetic variants that regulate pathway gene expressions. We propose to do the analysis by considering gene expressions in a pre-defined pathway as a multivariate response. Since genes in the same biological pathway tend to have similar expression patten, looking at a bunch of expression levels in a pathway as our unit phenotype will give us more information about the differential expression pattern about this pathway, and thereby will give us more

power to steadily detect the association of a genetic variation with the expression changes of a pathway. We focused our application to a real dataset in yeast and identified significant regulation patterns across the 16 yeast chromosomes. The detected pathway regulators tend to cluster together on the genome which might be due to the strong correlations among SNP markers on the genome.

In this study, we identified strong pathway regulation hotspots. Most of the hotspots overlap with the ones tested with single trait analysis [47]. Perlstein et al. recently applied the same yeast data to study individual genetic differences in response to small-molecule drugs and identified eight hotspots in response to multiple compounds [64]. Their hotspots overlap with most of the pathway regulation hotspots identified in our study except the one on chromosome 1. This information indicates that the same polymorphisms may affect both gene expression and compound response. The genetic enrichment test proposed in this work can be applied to their study to understand which genetic pathways are involved in drug response. Noted in our analysis, each hotspot contain either a single pleiotropic polymorphism or several closely linked polymorphisms (marker blocks) affecting the response to multiple pathway expressions. If we group the genomic regions as intervals with 20kb length as a previous work did [47], we may reduce the number of hotspots to a more compact size. On the other hand, to do so we may end up with an interval containing many genes and this may bring difficulties in interpretation.

In a similar analysis of the same yeast dataset by [69], the authors also found out the pattern that gene expressions associated with a common SNP marker are tend to be in the same pathway given that the pathway information is available. In their analysis, for instance, twelve expressions were identified to have strong linkage with the SNP marker at

one locus on chromosome 3. Two out of these 12 traits are included in the same KEGG pathway: MAPK signaling pathway (pathway id "04010"). Seven expression traits were shown to have linkage with the SNP marker at another locus on chromosome 3. Three out of these 7 traits are in the same pathway: Valine, leucine and isoleucine biosynthesis pathway (pathway id "00290"). These two loci were also detected to be pathway regulators in the current study. These results underscore the importance in finding genetic regulators responsible for the joint expression change of a pathway.

From the biological perspective, due to limited knowledge in genome annotation and gene pathways, not all genes can be mapped to a pathway. In the current analysis, only 1193 gene expressions are mapped to the 99 KEGG pathways, leaving a large proportion of genes unmapped. Alternatively, one can also focus the analysis on Gene Ontology (GO) terms which have a more comprehensive coverage of the gene information. As more and more gene information being documented in the public database (e.g., KEGG), this will eventually not be an issue. With limited pathway information, we can also classify gene expressions according to their correlation information to construct gene co-expression networks or modules [70]. These modules can be treated as pseudo pathways for further analysis. When a module is found to be significantly regulated, the function of those unknown genes can thus be inferred from those genes with known function in the same module. Since genes in the same module potentially share the same regulator, this can help generate meaningful biological hypothesis for experimental validation.

Table 2.2: Genetic pathway enrichment analysis results. The right column indicates enriched genetic pathways (GPs) that are responsible for the expression change of the corresponding expression pathways (EPs) in the right column, at the 0.01 significant level. Pathways highlighted with bold faces indicate *cis*-pathway regulation.

#	(PID) Enriched Genetic Pathway	#	(PID) Expression Pathway
1	(04010) MAPK signaling pathway	47	(00480) Glutathione metabolism
13	(03020) RNA polymerase	**13**	**(03020) RNA polymerase**
15	(00051) Fructose and mannose metabolism	**15**	**(00051) Fructose and mannose metabolism**
		16	(00052) Galactose metabolism
		43	(00520) Nucleotide sugars metabolism
		70	(00625) Tetrachloroethene degradation
16	(00052) Galactose metabolism	15	(00051) Fructose and mannose metabolism
		43	(00520) Nucleotide sugars metabolism
17	(03022) Basal transcription factors	83	(00220) Urea cycle and metabolism of amino groups
20	(00290) Valine, leucine and isoleucine biosynthesis	1	(04010) MAPK signaling pathway
		4	(00910) Nitrogen metabolism
		6	(00410) beta-Alanine metabolism
		18	(00053) Ascorbate and aldarate metabolism
		20	**(00290) Valine, leucine and isoleucine biosynthesis**
		25	(00010) Glycolysis / Gluconeogenesis
		26	(00330) Arginine and proline metabolism
		32	(00920) Sulfur metabolism

PID=pathway ID

Table 2.2 (cont'd)

#	(PID) Enriched Genetic Pathway	#	(PID) Expression Pathway
20	(00290) Valine, leucine and isoleucine biosynthesis	38	(00563) GPI-anchor biosynthesis
		45	(00340) Histidine metabolism
		49	(00750) Vitamin B6 metabolism
		52	(00251) Glutamate metabolism
		54	(00252) Alanine and aspartate metabolism
		58	(00670) One carbon pool by folate
		61	(00300) Lysine biosynthesis
		66	(00260) Glycine, serine and threonine metabolism
		73	(00310) Lysine degradation
		75	(00630) Glyoxylate and dicarboxylate metabolism
		77	(00450) Selenoamino acid metabolism
		78	(00770) Pantothenate and CoA biosynthesis
		86	(00360) Phenylalanine metabolism
		88	(00680) Methane metabolism
		92	(00401) Novobiocin biosynthesis
		96	(00272) Cysteine metabolism
		99	(00903) Limonene and pinene degradation
27	(00650) Butanoate metabolism	**27**	**(00650) Butanoate metabolism**

PID=pathway ID

Table 2.2 (cont'd)

#	(PID) Enriched Genetic Pathway	#	(PID) Expression Pathway
35	(00561) Glycerolipid metabolism	51	(00521) Streptomycin biosynthesis
		68	(00530) Aminosugars metabolism
39	(00513) High-mannose type N-glycan biosynthesis	87	(01030) Glycan structures - biosynthesis 1
43	(00520) Nucleotide sugars metabolism	**43**	**(00520) Nucleotide sugars metabolism**
44	(00020) Citrate cycle (TCA cycle)	**44**	**(00020) Citrate cycle (TCA cycle)**
46	(00980) Metabolism of xenobiotics by cytochrome P450	64	(00440) Aminophosphonate metabolism
68	(00530) Aminosugars metabolism	53	(00072) Synthesis and degradation of ketone bodies
74	(03010) Ribosome	45	(00340) Histidine metabolism
		86	(00360) Phenylalanine metabolism
		92	(00401) Novobiocin biosynthesis
78	(00770) Pantothenate and CoA biosynthesis	86	(00360) Phenylalanine metabolism
		92	(00401) Novobiocin biosynthesis

PID=pathway ID

Chapter 3

A combined p-value approach to infer pathway regulations in eQTL mapping

3.1 Introduction

Given that the expression of a gene or a network of genes may be regulated by a group of genetic variants functioning together as a system, studying gene regulations by focusing on the joint function of variants in a system could shed novel light into the complexity of a biological system. It is commonly recognized that genes in a pathway or network act in a coordinated manner to fulfill a joint task. Thus analysis from a systems biology perspective, for instance, focusing on genetic variants in terms of pre-defined pathways/networks, can provide valuable biological insights into gene function and regulation. Moreover, variants in a genetic pathway often confer moderate effects in mediating the expression change of a gene or a gene network, which makes it difficult to detect individual effect and consequently leads to low power in single marker analysis. From a biological point of view, signals in

a genetic system, even though individually are not significant (say p-values of 0.06 for an extreme case), many such values for related genes within a pathway or network when taken together may suggest the relative importance of that particular genetic system in mediating gene expression changes. By a genetic system we mean a group of genes within a genetic functional category which can be obtained from various sources such as Kyoto Encyclopedia of Genes and Genomes (KEGG)[71], Gene Map Annotator and Pathway Profiler [72] and Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) [73], or a category of multiple loci defined by SNP physical locations.

The above thoughts motivated us to consider a joint analysis in which multiple signals are combined together to indicate the contribution of the overall system. Herein, we argue that a joint analysis could provide additional insights into gene function and regulation that otherwise could not be achieved by looking at individual signals alone. We propose to combine individual p-values in a genetic system (e.g., KEGG category) while considering the correlations among them, to form an overall signal for inference of shared gene expression patterns in an eQTL mapping framework.

Methods of combining p-values have been applied to a wide range of problems, including genome-wide association studies [74, 75], multiple endpoints studies in clinical trails and meta-analysis, and detecting differentially expressed genes [76]. There are different p-value combination methods in the literature, for example, the Fisher's combined p-value approach [77]; the truncated product method [78]; the rank truncated product method [79]; and the weighted truncation product method [80]. A commonality among these combining methods is to first take a transformation of individual p-values and then evaluate the distribution of the combined statistic. However, when individual tests are not independent, the distribution

of the combined statistic is difficult to obtain. Moreover, there is no analytical criterion for choosing the truncation threshold for the truncated product methods.

For multiple individual tests in a genetic system, it is known that they are not independent due to linkage disequilibrium (LD) or functional interactions between variants. Regarding the concern of correlations among individual tests, methods that ignore correlations and treat them independently will obviously affect the accuracy of the results and could lead to either inflated false positives or false negatives. Some work has been done to handle correlations when combining individual p-values. For example, one could estimate the empirical null distribution of the combined statistic by a simulation-based procedure [78]; approximate the null distribution based on a known correlation matrix [81]; or apply the most widely used permutation approach. Although, permutation approaches, when performed appropriately, provide an unbiased estimation of the null distribution and are widely considered the gold standard with which other tests are compared, their main disadvantage is the computational cost [82]. For example, to get an empirical p-value of 10^{-5}, at least 10^5 permutations are needed.

When a large number of tests are involved in a study, alternative methods that can provide similar accuracy would be attractive. Brown (1975) proposed to combined dependent tests assuming a multivariate normal distribution of the test statistics with a specified covariance structure [83]. The method later on was extended by Kost and McDermott (2002) assuming a known covariance matrix up to a scaler quantity [81]. The assumption of a known covariance matrix limits their application as in most cases the distribution of the test statistic is unknown with an unknown covariance matrix. In this chapter, we focus our attention on the Fisher's combination statistic and propose to approximate its null distribution with

a scaled chi-square distribution while considering correlations among individual tests. We propose different strategies to estimate the correlation information.

3.2 Statistical Methods

3.2.1 Pattern of gene regulation

It has been commonly recognized that gene regulation plays a pivotal role in determining trait variation in natural populations by promoting or reducing the expression of functional genes that are (in)directly related to a phenotypic trait. Given that genes function in networks, the identification of regulatory elements, as well as the "master regulators" that affect the expression of hundreds of genes, can greatly enrich our knowledge of gene regulatory networks, and ultimately help us gain novel insights into the genetic architecture of complex traits [51, 52].

Figure 1.2 shows several possible gene regulation patterns. Figure 1.2(A) and 1.2(B) show *cis*- and *trans*-regulation patterns, respectively. Figure 1.2(C) indicates that the same gene can be regulated by multiple *trans*-regulatory loci. Each of these regulatory loci are associated with specific genetic variants. In the context of eQTL mapping, we are trying to identify genetic variants that are associated with these regulatory changes and likely regulate gene expression. To map eQTLs as illustrated in Figs. 1.2(A)-(C), single marker–single trait analysis can be applied followed by multiple testing corrections. Figure 1.2(D) shows that a regulatory element can regulate multiple genes, among which some share a common network. When multiple gene expressions are grouped into a network or a pathway, the identified regulators are termed as network or pathway regulators, and methods for this

purpose have been developed [63]. Figure 1.2(E) shows that the expression of a single gene or a network of genes is regulated by the joint function of multiple genetic variants, potentially belonging to a common genetic system (e.g., a genetic pathway). The signal perturbation of a genetic system could cause the expression change of a gene or a network of genes, and consequently result in phenotypic changes such as a disease. In this work we focus our analysis in identifying pathway regulation as shown in Figure 1.2(E). The identification of pathway regulations would help us better understand the genetic architecture of gene expression and regulation from a systems biology perspective.

3.2.2 The Satterthwaite's approximation

As we mentioned in the introduction section, a genetic system can be defined as a genetic pathway from the KEGG database or a GO term, or as a group of variants located physically close to each other. We hypothesize that the signal perturbation of a genetic system could lead to the expression change of a single gene or a network of genes. We assume there are L SNP variants in a given genetic system. For the L SNPs, we conduct L individual tests and obtain L individual test statistics or p-values. Depending on the number of genotype categories at each locus and the expression phenotype distribution, different tests can be applied. For example, a two-sample t-test or Hotelling's T^2 test can be applied depending on whether the response is a single gene expression value or multiple gene expression values, while assuming there are two possible genotype categories at a locus (e.g., in a recombinant inbred line or yeast population). We tried to combine individual signals in a genetic system to determine if it, as a whole system, underlies the expression changes of genes, and hope to gain novel insights into gene regulations from a systems biology perspective.

Let $p_1, p_2, \cdots, p_L$ be the p-values for L individual two-sided tests, $H_{i,0} : \mu_{i1} = \mu_{i0}$ versus $H_{i,1} : \mu_{i1} \neq \mu_{i0}$ $(i = 1, 2, \cdots, L)$ assuming there are two genotype categories (denoted as 1 and 0) at each locus. Define $z_i = -2 \log p_i$. Under the null hypothesis of no genetic effect, each of the L p-values is uniformly distributed and $z_i \sim \chi^2_2$ for $i = 1, \cdots, L$. If we assume the L tests are independent, the Fisher's combined statistic $T = \sum_{i=1}^{L} z_i \sim \chi^2_{2L}$ under the global null hypothesis of no genetic effect.

When multiple genetic variants are considered as a system, they are more or less correlated. Thus the L p-values are not independent and the Fisher's chi-square distribution with $2L$ degrees of freedom (d.f.) does not hold. Here we proposed to approximate T by a scaled chi-square distribution under the null by applying the Satterthwaite's approximation method. We assume that the combined statistic T follows a scaled chi-square distribution, i.e.,

$$T = \sum_{i=1}^{L} z_i \dot{\sim} a\chi^2_g. \tag{3.1}$$

The scale parameter a and the d.f. g are chosen so that the first and second moments of the scaled chi-square distribution and the distribution of T under the null are identical. For correlated p-values, the expectation and variance of the statistic T under the null can be obtained as

$$E(T) = E(\sum_{i=1}^{L} z_i) = 2L,$$

$$\begin{aligned} Var(T) &= Var(\sum_{i=1}^{L} z_i) \\ &= \sum_{i=1}^{L} Var(z_i) + 2\sum_{j<i} Cov(z_i, z_j) \\ &= 4L + 8\sum_{j<i} \rho_{ij}, \end{aligned}$$

where ρ_{ij} is the correlation between the log-transformed p-values z_i and z_j.

By equating the first and the second moments of T and $a\chi^2_g$, we have

$$E(a\chi^2_g) = ag = E(T) = 2L,$$

and

$$Var(a\chi^2_g) = 2a^2g = Var(T) = 4L + 8\sum_{j<i}\rho_{ij}.$$

Solving the two equations, we obtain

$$\hat{a} = \frac{4L + 8\sum_{j<i}\rho_{ij}}{4L} = 1 + \frac{2\sum_{j<i}\rho_{ij}}{L}, \tag{3.2}$$

$$\hat{g} = \frac{2L}{\hat{a}} = \frac{2L^2}{L + 2\sum_{j<i}\rho_{ij}}. \tag{3.3}$$

When the L SNPs are completely independent, i.e., $\rho_{ij} = 0 \; \forall \; i, j$, it can be seen that the approximation is the same as the distribution of the Fisher's combined statistic assuming independence. When the L SNPs are completely dependent, i.e., $\rho_{ij} = 1 \; \forall \; i, j$, then $\hat{a} = L$ and $\hat{g} = 2$. In this case, the statistic T is just a sum of L independent χ^2_2 variables. For $-1 < \rho_{ij} < 1$, parameters a and g approximate the distribution of T, where a and g can be estimated by Equations (3.2) and (3.3). In reality, we rarely see negative correlations for a two-sided test. So the restriction of $2\sum_{j<i}\rho_{ij} > -L$ to get positive estimates of a and g is easily met. The challenge remaining is to estimate the correlation between z_i and z_j from the data. In the following, we illustrate how to estimate the correlation ρ_{ij}.

3.2.3 Estimating the correlation matrix

Let $\boldsymbol{z} = (z_1, \cdots, z_L)$ be a vector of log-transformed p-values and let Γ be the correlation matrix of $\boldsymbol{z}$. From the above approximation we can see that the accuracy of the approximation to the distribution of T depends largely on how well the correlation matrix Γ is estimated. Assuming a multivariate normal distribution of the test statistics, Brown (1975) proposed to estimate Γ with a completely specified covariance matrix [83]. The author argued that the covariance between z_i and z_j is a function of the correlation between the ith and jth variables under the group of affine transformation. This is however not true in a genetic study, and there is no analytically closed form for the structure of Γ. In this paper, we propose two methods to approximate Γ, which are detailed in the follows.

Estimating the correlation matrix by permutation Since we want to approximate the null distribution of T, we need the correlation matrix of the transformed p-value vector $\boldsymbol{z}$ under the null hypothesis. Permutation was applied to generate random samples of $\boldsymbol{z}$ by reshuffling the relationship between the gene expression values and genetic markers, where genetic variants for each individual in a system are maintained as a vector to preserve their correlation structure. For each permutation, we would have a vector of p-values, $p^b = (p_1^b, p_2^b, \cdots, p_L^b)$ and also the transformed p-values $\boldsymbol{z}^b = (z_1^b, z_2^b, \cdots, z_L^b)$. The correlation matrix for $\boldsymbol{z}$ under the null then can be estimated by the sample covariance of the permuted random sample: $\boldsymbol{z}^b (b = 1, 2, \cdots, B)$, and B is the total number of permutations (say 1000). The sample correlation matrix obtained from the permuted samples were used as the estimate of Γ. No assumption is required for the distribution of the test statistics at this step. Generally speaking, the larger the data dimension (L), the more the permutations are required.

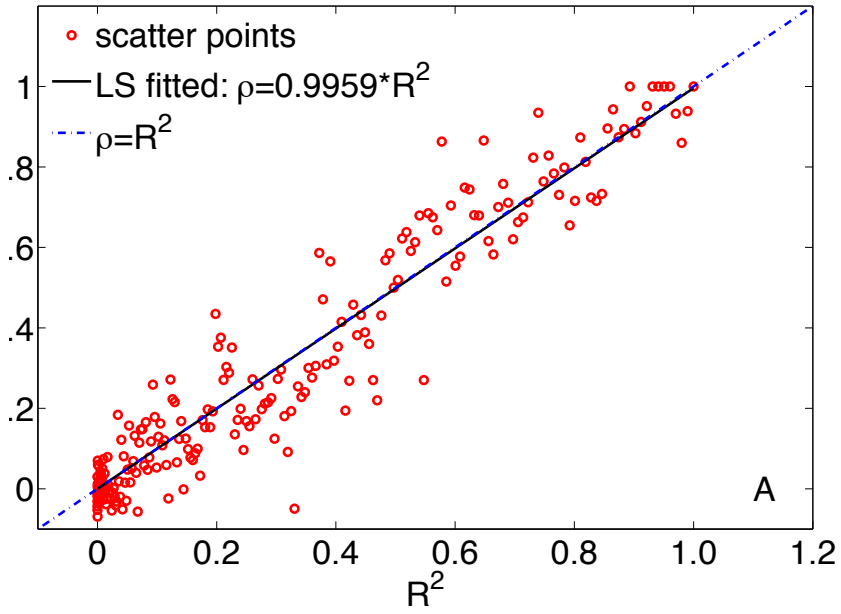


Figure 3.1: Scatter plots of correlation coefficient ρ vs LD R^2. The blue line is $\rho = R^2$, black line is the least square fitted line. (A) $MAF = 0.1$, fitted function: $\rho = 0.996R^2$; (B) $MAF = 0.3$, fitted function: $\rho = 1.006R^2$; and (C) $MAF = 0.5$, fitted function: $\rho = 0.99R^2$

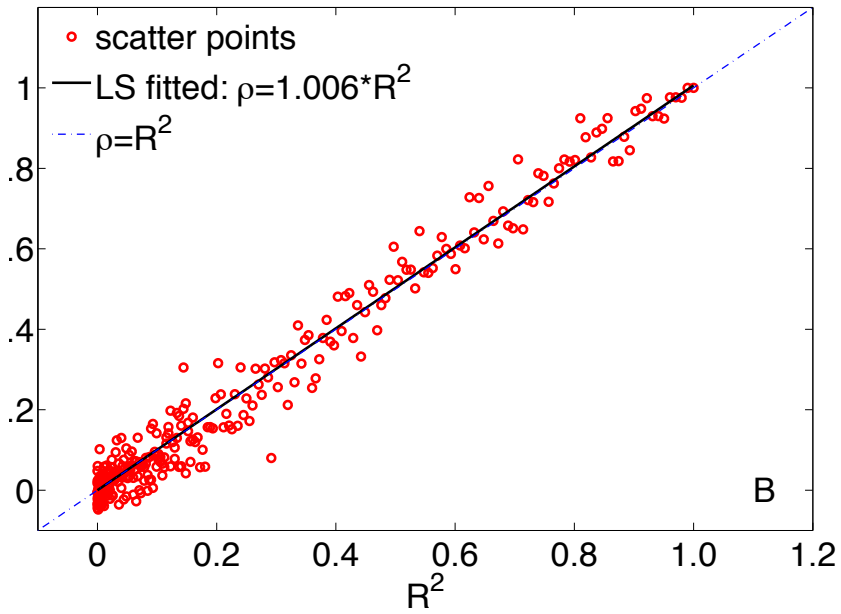


Figure 3.1 (cont'd). $MAF = 0.3$, fitted function: $\rho = 1.006R^2$.

Estimating the correlation matrix by LD approximation Note that multiple variants in a genetic system are either physically close to each other or functionally correlated.

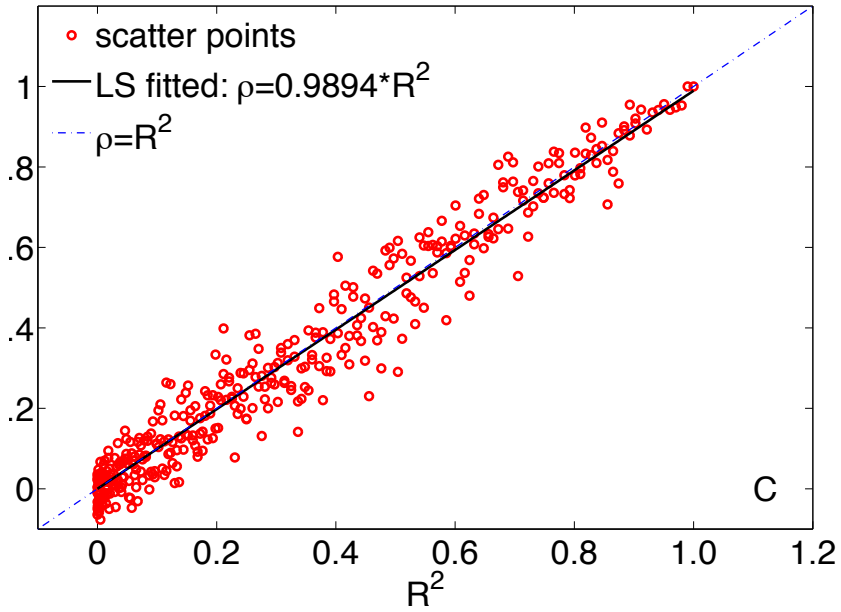


Figure 3.1 (cont'd). $MAF = 0.5$, fitted function: $\rho = 0.99R^2$.

The correlation information is more or less reflected by LDs between the variants. This motivates us to approximate Γ by LDs among SNP variants whose individual p-values are to be combined. Unfortunately there is no analytical solution to assess the relationship between the correlations of z and the LDs. We checked the relationship between the LDs of SNP variants (measured by R^2) and the correlation structure of $\boldsymbol{z}$. To begin with a simple example, we considered two SNP variants, each with a minor allele frequency (MAF) of $q = 0.1$ $(0.3, 0.5)$. For a given MAF, the range of LD denoted by D is given by

$$max\{-q_1q_2, -(1-q_1)(1-q_2)\} \leq D \leq min\{q_1(1-q_2), q_2(1-q_1)\},$$

where q_1 and q_2 denote the MAF for SNPs at two different loci. If we assume the same MAF for both SNPs, the range of D becomes $max\{-q^2, -(1-q)^2\} \leq D \leq q(1-q)$ and the range of $R = \frac{D}{\sqrt{(q_1(1-q_1)q_2(1-q_2))}} = \frac{D}{q(1-q)}$ is $max\{-q/(1-q), -(1-q)/q\} \leq R \leq 1$.

For a fixed MAF, we generated genotypes for two SNPs with different values of D (hence R) in a given range (following the procedure described in the LD-based simulation section). Phenotypes were simulated independent of the SNPs (i.e. under the null distribution) and then tested for association between the phenotype and the two SNP markers with p-values denoted by p_1 and p_2. For a given R value, the correlation coefficient of the two transformed p-values $z_1 = -2\log p_1$ and $z_2 = -2\log p_2$ was calculated from 1000 simulated samples. Scatter plots of the correlation coefficient ρ against R^2 corresponding to MAF 0.1, 0.3 and 0.5 are given in Figures 3.1. The three plots clearly indicate a linear relationship between ρ and R^2. The least squares fitted lines (black) almost perfectly overlap with the $\rho = R^2$ lines (blue). We also tried various allele frequency combinations for the two SNPs and found very similar relationships. Since a two-sided test was performed, even with negatively correlated SNPs, their p-values are still positively correlated. This explains why we rarely see negative correlations between the log-transformed p-values. We assessed the relationship for a real eQTL data set applied in this study (discussed in the real data analysis section). A similar relationship was also observed (Figure 3.2). The assessment in simulation and real data indicates that R^2 provides a good approximation to the correlation between the log-transformed p-values.

3.3 Simulation study

3.3.1 Accuracy of the scaled χ^2 approximation

The accuracy of the scaled chi-square approximation was evaluated by a χ^2-plot. Considering two p-values, p_1 and p_2, which are correlated with $\text{corr}(p_1, p_2) = \rho*$. We generated

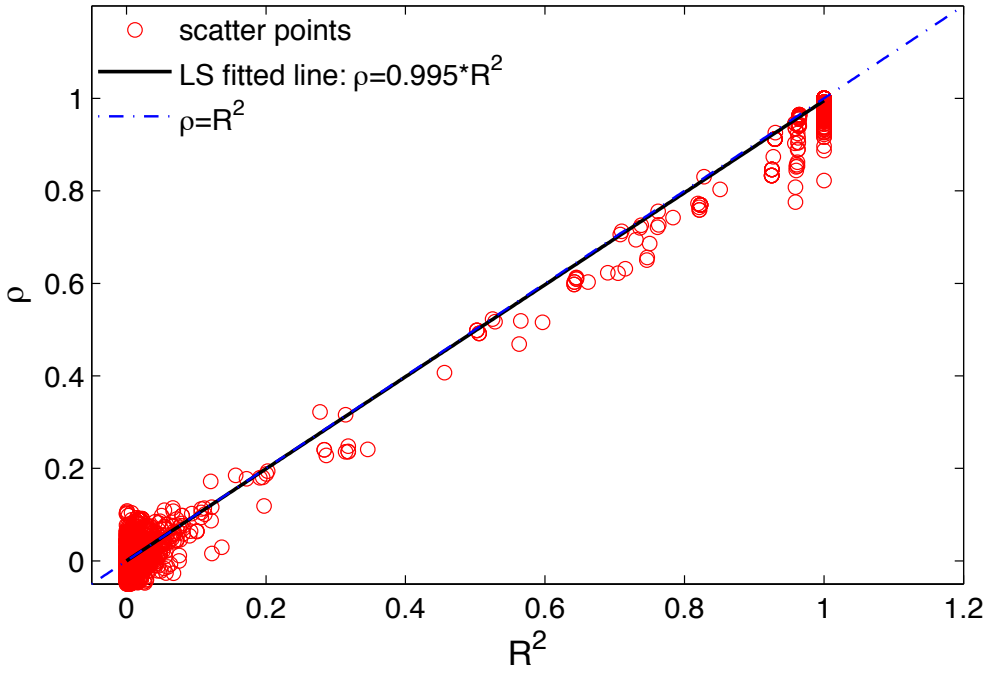


Figure 3.2: Scatter plot of the correlation coefficient ρ and R^2 for the YEAST eQTL data set. The black line is the least square fitted line: $\rho = 0.995R^2$ and the blue line is a straight diagonal line.

1000 random samples of p-values with a given correlation $\rho*$. The corresponding combined statistic T for the 1000 simulated samples were obtained. The estimated correlations between log-transformed p-values were then used to estimate a and g. Figure 3.3 plots the approximated percentiles using $\hat{a}\chi^2_{\hat{g}}$ (right panel) and χ^2_{2L} (left panel) versus the observed empirical percentile of T. As shown in the figure, points of percentiles of scaled chi-square distribution and the empirical percentiles lie roughly on a straight line, while χ^2-plot for the χ^2_{2L} approximation deviates from the straight line, especially at the tail. The plots demonstrate that the scaled chi-square distribution provides a much more accurate approximation to the distribution of T under the null than a regular chi-square distribution does. Simply ignoring the correlations among the test statistics would result in biased approximation and

wrong inference.

3.3.2 Simulation design

Genotype simulation We simulated genotypes for one genetic pathway with multiple SNP variants. These variants function together as a whole system to regulate expression changes of a single gene or a network of genes. Two methods were used to simulate the genotype data. The first method, termed LD-based simulation, generates SNP genotype data based on pairwise LD structure. The second method is a real data-based simulation which mimics gene structure and LD patterns of a real data set by sampling genotypes directly from the data.

LD-based simulation: Let q_A and q_B be the frequencies of two alleles A and B for two adjacent SNPs, with LD denoted by D. The frequencies of four haplotypes can be expressed as $p_{ab} = (1 - q_A)(1 - q_B) + D, p_{AB} = q_A q_B + D, p_{Ab} = q_A(1 - q_B) - D, p_{aB} = (1 - q_A)q_B - D$. Assuming HardyWeinberg equilibrium, the SNP genotype at locus A can be simulated assuming a binomial distribution. Locus B can be simulated conditional on locus A with the conditional probability given by

$$P(B|A) = \frac{P(BA)}{P(A)} = \frac{p_{AB}}{q_A} = \frac{q_A q_B + D}{q_A}. \tag{3.4}$$

This illustration is for simulating a haploid genome (e.g., yeast). The same idea can be applied to simulate a diploid genome. The advantage of this simulation strategy is that we can easily control the pairwise LD pattern between adjacent SNPs. We assume genes in a pathway are in linkage equilibrium (The assumption is not required for the method, but is used only for illustration of the feasibility of the proposed approach to different applications).

SNPs within each gene are in LD and the genotypes for SNPs in each gene were simulated by the LD-based simulation approach. We simulated SNP genotypes for four individual genes, $G1(8)$, $G2(5)$, $G3(3)$ and $G4(4)$, where the number in parenthesis indicates the number of SNP markers in the corresponding genes. The four genes were assumed to belong to one genetic pathway. LDs for SNPs within each gene were set to $R^2 = 0.9$.

Table 3.1: List of data generating models

Model	Gene action
I	$y = \mu + \epsilon$
II	$y = \mu + \beta_1 S_1 + \beta_2 S_2 + \beta_7 S_7 + \beta_8 S_8 + \epsilon$
III	$y = \mu + \beta_1 S_1 + \beta_2 S_2 + \beta_{15} S_1 S_5 + \beta_{38} S_3 S_8 + \epsilon$
IV	$y = \mu + \beta_1 G_{1,1} + \beta_2 G_{1,2} + \beta_3 G_{2,1} + \beta_4 G_{2,2} + \beta_5 G_{3,2} + \beta_6 G_{1,3} G_{3,2} + \epsilon$
V	$y = \mu + \beta_1 G_{1,1} + \beta_2 G_{1,2} + \beta_3 G_{2,1} + \beta_4 G_{2,2} + \beta_5 G_{2,2} G_{2,3} + \beta_6 G_{1,5} G_{4,4} + \epsilon$

Where S_j represents the jth SNP in a genetic pathway; $G_{i,j}$ represents the jth SNP in the ith gene. The effect of β_{ij}'s were considered the same.

Real data-based (RD) simulation: To simulate SNPs which mimic the gene structure and LD patterns among SNPs in a real genetic pathway, we took genotype vectors for SNPs within the #20 genetic pathway ("00290", Valine, leucine and isoleucine biosynthesis) in the yeast data set. Genotype vectors were randomly drawn with replacement from the real data to create a simulation sample. This genetic pathway has four individual genes with 14 SNPs in total. Missing genotypic values were imputed before the random draw. We found that the pairwise LDs in this pathway varies with $D \in (-0.035, 1)$ and $R \in (-0.14, 1)$.

Phenotype simulation Several simulation scenarios assuming different gene actions were considered (Table 3.1). Model I considers the case in which there is no genetic effect at all. So model I is the null model we used to assess the false positive rate. Model II assumes only main SNP effects in a genetic pathway (SNPs 1, 2, 7 and 8). Model III assumes main

SNP effects (SNPs 1 and 2) as well as the interactions between SNPs 1 and 5 and between 3 and 8. Model IV and V simulate phenotypes considering the gene structure in a genetic pathway. Interactions were considered for SNPs in different genes. Model IV considers interactions only when the corresponding gene has a main effect. Model V assumes there is an interaction effect between two genes and one of which has no marginal main effect.

We applied model II and model III to simulate phenotypes with genotype simulated by the RD-based simulation method. The LD-based simulation method were applied for model IV and model V to generate phenotype data. Thus four different simulation scenarios were considered: (A) RD-based genotype + Model II phenotype; (B) RD-based genotype + Model III phenotype; (C) LD-based genotypes + Model IV phenotype; and (D) LD-based genotypes + Model V phenotype. Type I error rate was assessed with phenotypic data simulated by Model I.

3.3.3 Simulation Results

We evaluated the type I error rate and power of the scaled chi-square approximation to infer genetic regulatory patterns. The type I error rate was estimated by simulating 1000 data sets under the null distribution (Model I). Similarly, we estimated power by simulating 1000 data replicates for each model (Model II-V). Two-sided two sample t-tests were applied to test for associations between SNP markers and a quantitative trait y. Individual p-values for all SNP markers within the pathway were then combined to form the test statistic $T = -2\sum_{i=1}^{L} \log p_i$. For each simulated data set, a p-value for the combined statistic T is assessed and is denoted by $p^c_{\chi^2_{2L}}$, $p^c_{a\chi^2_g}(\text{perm})$, $p^c_{a\chi^2_g}(R^2)$ and p^c_{perm}. For $p^c_{\chi^2_{2L}}$, the combined p-value follows a χ^2_{2L} distribution under the null; for $p^c_{a\chi^2_g}(\text{perm})$ and $p^c_{a\chi^2_g}(R^2)$, the combined

p-value follows a scaled $a\chi^2_g$ distribution, where parameters a and g were estimated by using correlations approximated by the permutation-based and the LD-based approximation (i.e., $\rho = R^2$) approaches, respectively; and for p^c_{perm}, the significance of the combined p-values were assessed by permutation tests with 10,000 permutation samples. In all simulations, we treated the results obtained by the p^c_{perm} method as the underlying truth with which the performance of other methods was compared.

Type I error rate Empirical type I error rates at the 0.05 significance level for 1000 replicates are summarized in the third column of Tables 3.2 and 3.3. The results clearly show that the type I error rates are significantly inflated for the χ^2_{2L} approximation under different simulation scenarios. The scaled chi-square approximation and the permutation procedure yield similar type I error rates which are close to the 0.05 nominal level. The two methods for correlation estimation have no significant effect on type I error rate.

Table 3.2: Empirical type I error rate and power for scenarios A and B under different sample sizes. The effects of β_j's are fixed at 0.1.

n	Methods	Model I	Model II	Model III
200	χ^2_{2L}	0.217	0.935	0.942
	$a\chi^2_g$(perm)	0.051	0.787	0.785
	$a\chi^2_g(R^2)$	0.053	0.788	0.786
	Permutation	0.049	0.788	0.787
500	χ^2_{2L}	0.204	1.000	0.999
	$a\chi^2_g$(perm)	0.052	0.994	0.991
	$a\chi^2_g(R^2)$	0.047	0.992	0.990
	Permutation	0.047	0.992	0.991

Power comparison Table 3.2 summarizes the empirical power for scenarios A and B. The results obtained with the permutation method is considered as the underlying truth. It can be seen that the χ^2_{2L} approximation always gives the highest power (see column 3), which is due to its high false positive rate. The results produced by the scaled chi-square

approximation are very close to the permutation-based results, which indicates the good performance of the scaled chi-square approximation. No significant differences in power were observed for the two scaled chi-square approximation methods. However, the calculation with the $a\chi^2_g(R^2)$ method is much faster than the permutation-based $a\chi^2_g$(perm) method. The effect of sample size on power is clear: large sample size always gives large power, as we expected.

The results for scenarios C and D are summarized in Table 3.3. Similar trends as in Table 3.2 were observed. Again, the χ^2_{2L} approximation yields inflated false positive rates and is less attractive than the scaled chi-square approximation does. We also tried other correlations and found that negative or low positive correlations may reduce the overall power for given genetic effects. However, the overall trend as we observed in Tables 3.2 and 3.3 remains unchanged, when comparing the performance of different methods.

Table 3.3: Empirical type I error rate and power for scenarios C and D under different sample sizes. The effects of β_j's are fixed at 0.15.

n	Methods	Model I	Model IV	Model V
200	χ^2_{2L}	0.179	0.882	0.885
	$a\chi^2_g$(perm)	0.056	0.706	0.718
	$a\chi^2_g(R^2)$	0.053	0.703	0.709
	Permutation	0.054	0.704	0.714
500	χ^2_{2L}	0.189	0.998	0.996
	$a\chi^2_g$(perm)	0.057	0.986	0.989
	$a\chi^2_g(R^2)$	0.056	0.984	0.989
	Permutation	0.052	0.986	0.989

3.4 Real data analysis

We applied our method to the yeast eQTL dataset introduced in section (2.2.1). The pathway information was retrieved from the R package: YEAST. There are 99 KEGG pathways in the package, but only 83 pathways were retrieved for follow-up analysis. The genotype profiles of neighboring markers tend to be highly correlated and some are even identical. With this information, markers were first merged to blocks [57]. Then missing genotypes were imputed based on available genotype information in each block. In cases where markers did not belong to any block, missing data were imputed by assuming a Bernoulli distribution with allele frequency estimated based on available data for the corresponding marker. We focused our analysis on the pathway regulation of a network of genes as illustrated in Figure 1.2(E). We first built up gene expression networks using the gene expression traits. Then the method described in this work was applied to identify pathway regulations for each network.

3.4.1 Gene co-expression network

There are many ways to construct gene expression networks. We focused on gene co-expression networks following the method proposed by Zhang and Horvath [84]. Because of the computational burden, only the top 2001 connected genes out of the 4000 most varying genes were considered to build the co-expression networks. The average linkage hierarchical clustering method was applied to group genes with coherent expression profiles based on a topological overlap matrix (TOM) dissimilarity measure. In our study, we obtained six gene modules (Table 3.4). Figure 3.4 shows the six co-expression network modules. For a detailed description of the weighted gene co-expression network approach, the readers are referred to [84].

Table 3.4: Information on gene co-expression networks

Modules	Blue	Brown	Green	Red	Turquoise	Yellow
# of genes	251	153	125	56	325	151
# of eigengenes	12	7	7	1	9	6

3.4.2 Network singular value decomposition

For each network, gene expression values were treated as multivariate responses and tested for association at each SNP marker locus. For the yeast data, there are two possible genotype categories at each locus. So a two sample Hotelling's T^2 test can be applied to test if mean responses are different for the two groups at each locus [63]. A gene co-expression network usually consists of many genes. In this dataset, most co-expression networks contain hundreds of genes. So the dimension of a network is greater than the sample size in most cases. Therefore it is infeasible to use Hotelling's T^2 test for expression profiles of all genes in a network. To reduce the dimension of a network, we applied the singular value decomposition (SVD) method. Because genes in a network are often highly correlated, using SVD could dramatically reduce the data dimensionality with only relatively few "eigengenes" capturing the total variation of a network. In this study, "eigengenes" that account for more than 85% of the total variation of a network of gene expression values were chosen as the response variable for further analysis.

Consider a gene expression network with N genes, all expression profiles can be represented by a matrix X with $N \times n$ dimension where n is the sample size. Each row of X represents the expression of one gene belonging to the network. The SVD of matrix X is given by

$$X = UDV^T,$$

where U is an $N \times L$ matrix; $D = diag\{d_1, d_2, \cdots, d_L\}$ is an $L \times L$ diagonal matrix, $d_1 \geq d_2 \geq \cdots \geq d_L$ are eigenvalues of X; and V^T is an $L \times n$ matrix with $L = min\{N, n\}$. Each row of matrix V^T represents a so-called "eigengene" of the original network. The proportion of "eigengenes" calculated by $v_l = d_l^2 / \sum_{i=1}^{L} d_i^2$ indicates the amount of total variation captured by the lth eigengene. Top K eigengenes will be remained for further analysis if the cumulative variation captured by the top K eigengenes is larger than 85%, i.e., $\sum_{l=1}^{K} v_l \geq 85\%$. The eigengenes are orthogonal to each other and are treated as a multivariate response to represent each co-expression network for further analysis.

3.4.3 Results by the scaled chi-square approximation

Hotelling's T^2 test was applied at each locus for gene expression networks with two or more eigengenes. For the red module with only one eigengene, a two-sided two sample t-test was applied. Individual p-values were then combined for each of the 83 genetic pathways to assess the significance by the scaled chi-square approximation. SNPs in different GPs may overlap which may cause dependence among GPs. The overlap issue was ignored in the current analysis and will be studied in future work. We also did the pathway enrichment analysis (PEA) proposed by Wang et al. [62]. The results are summarized in Table 3.5. Only GPs with p-values less than 0.001 were reported. The last three columns list the p-values for the combined statistic T using different methods to estimate the correlations plus those with the PEA analysis. The overlapped GPs with p-values less than 0.001 are highlighted with bold font. In many cases the enriched GPs identified with the two methods are very similar, except for the Blue module. In terms of the computation time, the combined p-value approach took much less time than the PEA analysis. For example, it took about 5 minutes

to calculate the combined p-value with LD-based correlation approximation, while it took about 8 hours to run 1000 permutations for one network module with the PEA analysis.

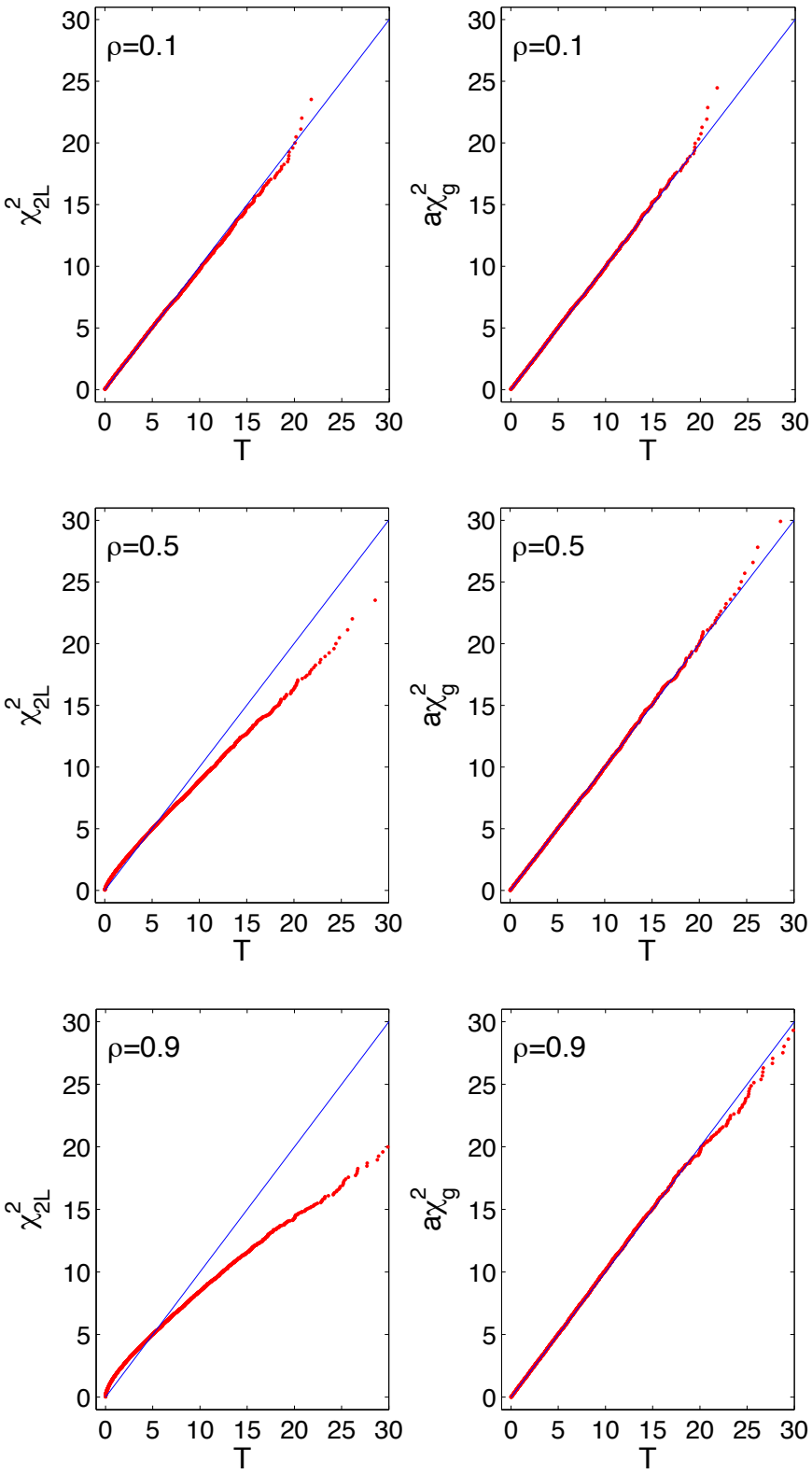


Figure 3.3: χ^2 plot for percentiles of the observed statistic T against the χ^2_{2L} approximation (left panel) and $a\chi^2_g$ approximation (right panel). Two correlations were assumed: $\rho = 0.1$ (upper panel), $\rho = 0.5$ (middle panel) and $\rho = 0.9$ (lower panel).

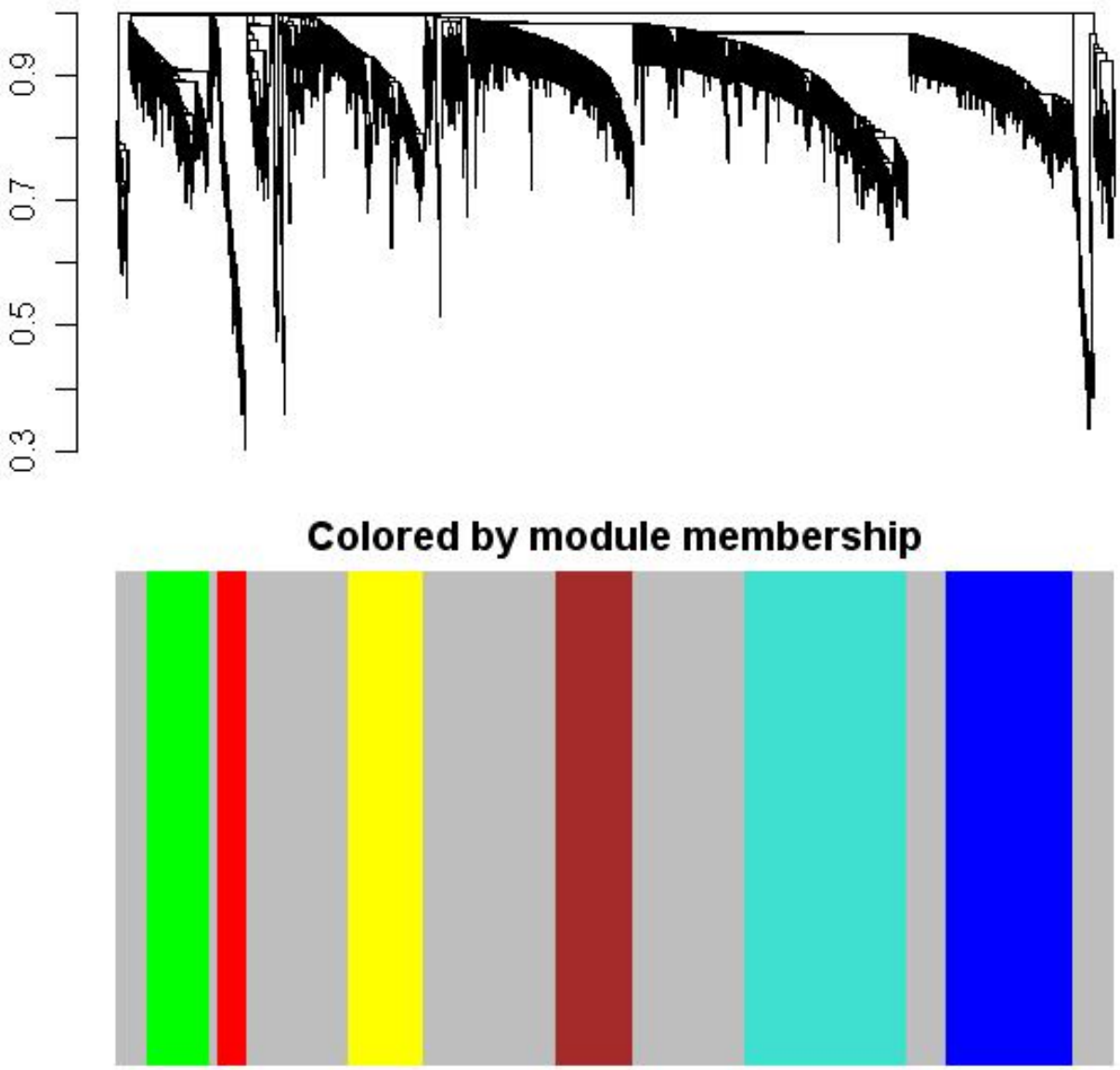


Figure 3.4: Weighted gene co-expression network with hierarchical clustering trees for the yeast gene expression data. See Zhang and Hovath (2005) for details of the algorithm.

Table 3.5: List of enriched genetic pathways (GPs) with the scaled chi-square approximation method and the gene set enrichment analysis. Only GPs with p-values ≤ 0.001 using either the p-value combined method or the PEA method are listed. The middle column is the list of GPs that are associated with the expression change of the corresponding co-expression networks given in the first column. GPs that show enrichment with both methods are highlighted with bold font.

Gene Network (# of genes)	P#	(PID)	Name of enriched GPs	$p_{a\chi^2_g}(R^2)$	$p_{a\chi^2_g}$(perm)	p_{PEA}
Blue	17	(03022)	Basal transcription factors	2.28e-03	1.75e-03	*< 0.001*
(251)	34	(04111)	Cell cycle - yeast	*7.55e-04*	3.03e-03	0.010
	78	(00770)	Pantothenate and CoA biosynthesis	*4.68e-04*	*7.69e-04*	0.011
Brown	10	(00500)	Starch and sucrose metabolism	8.89e-02	8.97e-02	*< 0.001*
(153)	**13**	(03020)	RNA polymerase	*2.53e-04*	*4.39e-04*	*< 0.001*
	17	(03022)	Basal transcription factors	*2.87e-04*	*3.69e-04*	*< 0.001*
	25	(00010)	Glycolysis / Gluconeogenesis	2.66e-02	3.05e-02	*< 0.001*
	32	(00920)	Sulfur metabolism	*7.11e-04*	1.12e-03	0.002
	34	(04111)	Cell cycle - yeast	*4.68e-05*	*2.81e-04*	*0.001*
	78	(00770)	Pantothenate and CoA biosynthesis	*3.97e-05*	*6.08e-05*	0.039
	83	(00220)	Urea cycle and metabolism of amino groups	*4.28e-04*	*6.41e-04*	*< 0.001*
	84	(00860)	Porphyrin and chlorophyll metabolism	*6.92e-04*	1.07e-03	*< 0.001*

P#=pathway number; PID=pathway ID.

Table 3.5 (cont'd)

Gene Network (# of genes)	P#	(PID)	Name of enriched GPs	$p_{a\chi^2_g}(R^2)$	$p_{a\chi^2_g}$(perm)	p_{PEA}
Green(125)	**20**	(00290)	Valine, leucine and isoleucine biosynthesis	*3.50e-05*	*4.19e-05*	*< 0.001*
Red	**1**	(04010)	MAPK signaling pathway	*1.19e-04*	*1.06e-04*	*< 0.001*
(56)	10	(00500)	Starch and sucrose metabolism	1.23e-02	1.56e-02	*< 0.001*
	43	(00520)	Nucleotide sugars metabolism	*2.28e-05*	*3.53e-05*	*< 0.001*
	85	(00040)	Pentose and glucuronate interconversions	*3.04e-04*	*3.86e-04*	*0.001*
Turquoise	**20**	(00290)	Valine, leucine and isoleucine biosynthesis	*5.75e-07*	*3.45e-06*	*< 0.001*
(325)	**27**	(00650)	Butanoate metabolism	*6.40e-04*	1.30e-03	*< 0.001*
	78	(00770)	Pantothenate and CoA biosynthesis	*3.67e-05*	*1.43e-04*	0.002
Yellow	**20**	(00290)	Valine, leucine and isoleucine biosynthesis	*2.91e-39*	*1.05e-35*	*< 0.001*
(151)	**27**	(00650)	Butanoate metabolism	*1.92e-13*	*2.615e-13*	*< 0.001*
	74	(03010)	Ribosome	*2.99e-04*	*3.93e-04*	0.006
	78	(00770)	Pantothenate and CoA biosynthesis	*2.10e-19*	*6.41e-18*	*< 0.001*

P#=pathway number; PID=pathway ID.

We also tried the Fisher's χ^2_{2L} approximation assuming SNPs in a genetic pathway are independent. We found more significant pathways than with the scaled chi-square approximation (data not shown). As indicated by the simulation studies, the additional GPs identified are most likely false positives. From Table 3.5, we can see that pathways 78 (Pantothenate and CoA biosynthesis) and 20 (Valine, leucine and isoleucine biosynthesis) are responsible for several network expression changes. This implies the relative importance of these pathways in the regulation of yeast gene expressions.

In order to understand the biological significance of our findings, it is important that we first describe the origin of strains used in the original yeast crossing design. As mentioned earlier, the parental strains are derived from natural isolates. The first strain, BY4716, is a lab strain whose origin can be traced back to a natural isolate that was found growing on a rotting fig [85]. However, this strain has had a long history of use as a laboratory model and has been selected for many properties that make it more amenable to experimentation [86]. In addition, because it is derived from a haploid segregant of the original heterozygous, diploid natural isolate, and because it has been harbored in the relatively benign lab environment for many generations, several known loss-of-function alleles have been identified in this parental strain [87]. Finally, all yeast strains used in experimental genetic crosses are altered to some degree. Most commonly these alterations include the generation of a null mutation for the HO endonuclease, the loss of which prevents mating type switching and allows for manipulation of ploidy and mating type [88]. In addition, experimental yeast strains also harbor loss-of-function alleles for genes within amino acid biosynthetic pathways, so that nearly all lab strains are auxotrophic for some combination of amino acids (e.g., Uracil, Leucine, Lysine, Histine, Tryptophan, Methionine, Adenine) [88]. Such auxotrophies provide a mechanism

for phenotypic selection on yeast media that lacks specific amino acid supplements. Even though the second parental strain, a haploid derivative of the natural vineyard isolate RM11-1a, was chosen to represent the prototrophic representative of a natural strain, it does carry loss-of-function alleles for HO endonuclease and auxotrophies for the Leucine and Uracil biosynthetic pathways [45].

Strikingly, all of the pathways inferred to influence co-expressed gene groups can be traced to either the engineered or lab selected loss-of-function alleles segregating in the parental stains. For example, in Table 3.5, the Yellow gene co-expression module exhibited the highest statistical significance with respect to the functional categories that explain the observed variation. We did a GO term search and found that 43.7% of genes in this module are mapped to GO cellular amino acid and derivative metabolic process. This represents the highest percentage these genes can be mapped to the GO process category. Also 28.5% of genes are mapped to the GO transferase activity function category, which explains the enrichment of pathway 74 (Ribosome). KEGG genetic pathways 20 (Valine, leucine, and isoleucine biosynthesis), 27 (Butanoate metabolism), and 78 (Pantothenate and CoA biosynthesis) are all either directly requiring or downstream of the Lue2 (YCL018W) and Ilv6 (YCL009C) genes. These genes are both physically and functionally linked in that they are required for leucine and isoleucine biosynthesis and found with 13 kilobases of one another (roughly 3-5 centiMorgans) [89]. Because Leu2 is a complete knock-out, there were several markers all found within this locus, each strongly associated with a given pathway. Similarly, the Ilv6 gene, with only a single marker, is also strongly associated with all three of these KEGG genetic pathways. In addition, all or some combination of these genetic pathways are strongly associated with the Blue, Brown, Green, and Turquoise,

gene co-expresssion networks, and in each case, the association is mediated by the same genetic markers. Hence a single engineered mutation that was known to be segregating in the parental cross explains most of the co-expressed genes in the Yellow module, and these same associations are found in the Blue, Brown, Green, and Turquoise gene networks. All of these effects are likely mediated by a single loss-of-function at Leu2 with direct effect. In addition, the indirect effects of Leu2 on the regulation and activity of Ilv6 as well as the linkage of Ilv6 with Leu2 may also play an important role [90, 89, 91]. Note that pathway 78 is enriched for the Blue, Brown and Turquoise network only by our approach, which indicates the better performance of our method against the PEA analysis in this study. Thus, this systems biology approach has allowed for the elucidation of many interacting gene networks and the genetic pathways through which they are most likely influenced. Importantly, these conceptual linkages derive from a clear biological reason, in this case an engineered mutation with pleiotropic effects.

In addition to the associations mediated via auxotrophic markers, the remaining genetic pathways can be broadly categorized in three groups: mitochondrial function (17 - Basal transcription factors; 13 - RNA polymerase), cell cycle (34 - Cell cycle), and cell signaling, filamentous/invasive growth, and mating (1 - MAPK signaling pathway). All of these effects are in pathways that can be traced to additional alleles of large effect that are known to have been segregating in the cross. Amn1 and Flo8 mutations in the lab strain were selected at some point in the past for reduced flocculation (clumpy growth due to cell-cell adhesion), and the 112 segregants differ in mating type at the MAT locus [45, 85]. All of these selected and engineered alleles are known to be strongly involved in MAPK signaling. In fact, gene Ste20 (YHL007C) in the MAPK signaling pathway in this analysis shows the strongest single

marker associations, and the gene is directly downstream of another well characterized QTL in previous studies, the Gpa1 gene [92]. Perhaps accidentally, the lab strain also is known to exhibit several phenotypes indicative of reduced mitochondrial function [93]. While loss-of-function alleles were known to exist in the lab strain for the Hap1 (YLR256W) and Mkt1 (YNL085W) genes, a recent study mapping variation in mitochondrial function with these same data, identified three additional mitochondrial alleles of strong effect at Sal1 (YNL083W), Cat5 (YOR125C), and Mip1 (YOR330C), respectively [94]. In particular, Mip1 is part of the mitochondrial DNA polymerase and Hap1 is required for cytochrome function [95, 96]. Hence, the many genetic pathways related to mitochondrial function and localization (e.g., 92% genes in the Green module map to mitochondria via Gene Ontology) are likely a downstream pathway that was altered as a result of these known deficient alleles segregating in the cross. In this case, we suspect that given the importance of proper mitochondrial function in the wild, each of these alleles is due to relaxed selection in the lab environment [91].

Finally, the single largest effect size typically observed in studies utilizing data from this cross is at the Ira2 gene (YOL081W) [97]. We observed very strong signals at this gene for all six co-expressed modules. The strongest one (p-value$< 10^{-14}$) corresponds to the Brown module. Even though this gene is not mapped to any KEGG pathways in this analysis, it is located upstream of the RAS/PKA signaling pathway and has strong downstream effects on nutrient signaling, cyclic AMP signaling, cell proliferation, and polymerase II activity [98]. The downstream effects of this polymorphism are apparent in the many genetic pathways related to nutrient metabolism, transcription, and cell cycle. Interestingly, this allele has not been traced to lab engineering or relaxed selection, but is more likely a naturally segregating

difference that is derived in the vineyard isolate [99].

In summary, our analysis has elucidated how a systems biology approach can identify the variation in genetic pathways that control co-expressed gene networks, and nearly all of the effects identified in this cross can be traced back to either engineered mutations or loss-of-function alleles that arose due to relaxed constraint in the benign lab environment.

3.5 Discussion

The integration of gene expression analysis and genetic mapping, termed eQTL mapping, brings great promise in elucidating the genetic architecture of gene expression. Empirical studies have shown that eQTL mapping can shed new light into gene network prediction, provide additional biological insights into gene regulation, and facilitate functional gene identification [100, 101, 102, 103]. Moreover, eQTL mapping results can provide additional directional information in gene regulatory network construction [104, 105]. With more biological data being generated at the sequence, transcriptional, proteomic and metabolic levels, together with the end-point phenotypic data such as a disease status, we are progressively approaching the era where various sources of data information can be integrated to gain novel biological insights from a systems biology perspective.

Our study is driven by the biological fact that genes function in networks or systems. Most biological phenomena occur through the expression of multiple genes which are potentially regulated by a cascade of genetic variants. Mootha et al. previously showed that focusing on expression data in terms of predefined pathways/networks (genetic features) can provide valuable insights into gene function not easily achievable by methods focus on individual genes [54]. This inspired us to focus on features of genetic variants that belong to predefined

pathways/networks in order to understand the genetic basis of gene regulation. Given the complexity of a genetic system, it is very unlikely that the function of a single variant will induce an overt identifiable or physiologically meaningful expression change of a network of genes. Also features defined by groups of genes should be more robust to genetic variation. Thus, we proposed to incorporate pathway (e.g., KEGG pathway) information into an eQTL mapping framework to gain novel insights into pathway regulation of gene expression. By combining evidence of multiple signals in a genetic system, our method addresses the limitation of the traditional single marker–single trait analysis: 1) Without a single encompassing theme, results could be hard to interpret; 2) Moderate changes which were disregard in single marker analysis, may afford more insight into gene regulation mechanisms [54].

As reviewed in the introduction section, there are many ways to combine evidences. It is commonly recognized that variants in a genetic system are often correlated. In this study we proposed to approximate the combined p-values of individual signals with a scaled chi-square approximation considering correlations among variants. Newton et al. proposed a random-set method in assessing gene-set enrichment by averaging gene scores [106]. As discussed by the authors, among-gene dependence was not an issue in their enrichment analysis because factors that caused dependence were excluded from the calculation of a gene score. Instead of averaging, we proposed to combine signals. In addition, correlations among genetic variants preserved a structural relationship due to LD. Our simulation studies indicated that large false positive rates could be observed if correlations were not properly accounted for. We proposed two different methods for an estimation of the correlation information between the log-transformed p-values. The results indicate that using the LD information to approximate the correlation produces similar results as using permutation-based methods. Real data

analysis also confirmed the result (Table 3.5). Thus, LD information could be directly applied in order to save computation time. It is also worth noting that depending on whether it is a one-sided or two-sided test, the relationship between the LD (R) and the correlation (ρ) could be different.

In the real data analysis, we focused our attention on gene expression networks as the response variables. We can also focus the responses on expression pathways extracted from public database such as those from KEGG database or from GO terms. Since only p-values are required, any sophisticated statistical tests can be applied. Even though the LD-based approximation for correlation of the log-transformed p-values may not be valid for a nonlinear model, the correlations can always be evaluated with the proposed permutation-based method. Depending on the interest of an investigator, our method provides a general strategy for regulation inference in a single gene or pathway level [107]. In addition, the method can also be extended to a (genome-wide) genetic association study to identify novel pathways underlying complex disease.

Chapter 4

Gene-centric gene-gene interaction: a model-based kernel machine method

4.1 Introduction

Accumulative evidence shows that much of the genetic variation for a complex trait can be explained by the joint function of multiple genetic factors, as well as environmental contributions. Searching for these contributing genetic factors and further characterizing their effect sizes, is one of the primary goals and challenges for modern genetics. The recent breakthroughs in high-throughput genotyping technologies and the completion of the International Haplotype Mapping (HapMap) project provide unprecedented opportunities to characterize the genetic machinery of living organisms. Genetic association analyses focusing on single nucleotide polymorphisms (SNPs) or haplotypes have led to the identification of many novel genetic determinants of complex traits. However, despite enormous success in genome-wide association studies, single SNP or haplotype based studies still suffer from

low replication rates because of the infeasibility of dealing with the complex patterns of association, e.g. genetic heterogeneity, epistasis and gene-environment interaction. Much of the genetic components of many traits remains unaccounted for and only a small proportion of the heritability has been explained.

It has been broadly recognized that most common human diseases are likely to have complex etiologies [108]. In a recent report, Neale and Sham discussed the choice of the basic genetic components to be considered for association with a complex trait [109]. It is demonstrated that a gene-based approach, in which all variants within a putative gene are considered jointly, have relative advantages over single SNP or haplotype analysis. There are multiple reasons for this. First, it is well known that genes are the functional units of human genome. Variants in genes have high probability of being functionally important than those that occur outside of genes [110]. Because of this characteristic, gene-based association analysis would provide more biologically interpretable results than the single-SNP or haplotype based analysis. Second, the position, sequence and function of genes are highly consistent across diverse human populations, which makes the gene-based studies more powerful in terms of replication [109]. Third, when there are multiple variants within a gene that function in a complicated manner, the gene-based association test can gain additional power compare to a single SNP analysis by capturing the joint function of multiple variants simultaneously [111, 112]. Finally, a gene-based analysis is statistically appealing. By considering multiple SNP markers within a gene as a testing unit, the number of tests would decrease dramatically, hence reducing the multiple testing problem and improving the power of the association testing.

We all know that genes do not function alone, rather they constantly interact with each

other. It has been widely recognized that gene-gene interaction, or epistasis, is an important category that contributes to the unexplained heritability of complex traits [108, 113, 114, 115]. Methods for detecting gene-gene interaction have been historically pursued on a single locus level, either parametrically such as the regression-based tests of interaction [116] and the Bayesian epistasis mapping [117], or non-parametrically such as the entropy-based approaches [118], and some data mining methods such as the multifactor dimensionality reduction (MDR) [119] and random forests [120]. Methods based on interaction of haplotypes have also been developed [121]. However, due to the phase-ambiguity problem, the haplotype-based methods are limited to only small size haplotypes. Extensions to interaction of large size haplotypes are challenged by computational cost. For a comprehensive review of statistical methods developed for detecting gene-gene interactions, readers are referred to [122].

With the relative merits of the gene-based association analysis, the identification of genetic interactions by focusing on genes as functional units should carry the same benefits and gains as it does with single gene analysis. Thus we propose to jointly model the genetic variation of SNPs within a gene, then pairwise gene-gene interactions can be carried out in a genome-wide search. The idea of Gene-centric Gene-Gene (denoted as 3G) interaction would conceptually change the way we model gene-gene interactions and meantime bring statistical challenges. Through the modeling of the joint variation of a gene pair, we argue that a 3G interaction analysis is biologically attractive. In addition, by focusing on genes as testing units, the number of pairwise interaction tests can be dramatically reduced compared to a single SNP-based pairwise interaction analysis. Thus a 3G interaction analysis is also statistically appealing.

In this work, we propose a model-based kernel machine method for the purpose of identifying significant gene-gene interactions under the proposed 3G analysis framework. Kernel based methods have been proposed to evaluate association of genetic variants with complex traits in the past decades [123, 124, 125, 126, 127]. A general kernel machine method can account for complex nonlinear SNP effects within a genetic feature (e.g. a gene or a pathway) by using an appropriately selected kernel function. Generally speaking, a kernel function captures the pairwise genomic similarity between individuals for variants within an appropriately defined feature [126]. The application of kernel-based method in genetic association analysis has been reported in the literature [124, 128, 129]. However, none of these considers interaction of genes. Here, we propose a general 3G interaction framework by applying the smoothing-spline ANOVA model [130] to model gene-gene interaction. The proposed method, termed **G**ene-centric **G**ene-**G**ene interaction with **S**moothing-s**P**line **A**NOVA **M**odel (3G-SPAM), is implemented through a two-step procedure: (1) an exhaustive 2-dimensional genome-wide search for pairwise gene-gene interactions; and (2) significance assessment of pairwise interactions.

The rest of the chapter is organized as follows. In section 4.2, we describe the detailed model derivation of our method. We proposed two score statistics for testing the overall genetic effect and the interaction effect based on the 3G-SPAM. To evaluate the performance of the proposed method, Monte Carlo simulations are performed in section 4.3. The utility of the method are demonstrated by real data analysis in section 4.4 followed by discussions in section 4.5.

4.2 Statistical methods

4.2.1 Smoothing Spline-ANOVA (SS-ANOVA) model

We assume n unrelated individuals sampled from a population, each of which possesses a measurement for certain quantitative trait of interest. The quantitative measurements of the n individuals are denoted as $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$. Traditional approaches for detecting gene-gene interactions, such as MDR or regression type analysis, identify SNP-SNP interactions. In this work, we focus our attention to pairwise gene-gene interactions by considering each gene as a unit. Consider two genes, denoted as G_1 and G_2, with L_1 and L_2 SNP markers respectively. Let $\boldsymbol{x}_i = (x_{i,1}, \cdots, x_{i,L})^T$ be an $L \times 1$ genotype vector of the gene pair for subject i. Here $L = L_1 + L_2$ is the total number of SNP markers in the gene pair. We model the relationship between the genotypes of the gene pair ($\boldsymbol{x}_i$) and the phenotype y_i by the following model

$$y_i = m(\boldsymbol{x}_i) + \epsilon_i, i = 1, 2, \cdots, n \tag{4.1}$$

where m is an unknown function and $\epsilon_i \sim N(0, \sigma^2)$ is a random subject-specific error term which is generally assumed to be normal with mean 0 and variance σ^2 and be independent of $\boldsymbol{x}_i$.

Gu has discussed the ANOVA decomposition of multivariate functions on generic domains of each single coordinate [131]. Actually, the decomposition can also be defined on nested domains; see Appendix B.1. With the prior knowledge of genes, the genotype vector $\boldsymbol{x}_i$ is partitioned as $\boldsymbol{x}_i = [\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}]$, with $\boldsymbol{x}_i^{(j)}$ represents the L_j SNP predictors for gene j ($j = 1, 2$). Let a product domain be $\mathcal{X} = \mathcal{X}^{(1)} \otimes \mathcal{X}^{(2)}$ with $\boldsymbol{x}^{(j)} \in \mathcal{X}^{(j)}$ and A_j be an averaging operator on $\mathcal{X}^{(j)}$, that averages out $\boldsymbol{x}_i^{(j)}$, $j = 1, 2$. Then a function $m(\boldsymbol{x})$ defined

on the product domain has a functional ANOVA decomposition as in the following equation:

$$\begin{aligned} m &= \prod_{j=1}^{2}(I - A_j + A_j)m \\ &= A_1A_2 + (I - A_1)A_2 + A_1(I - A_2) + (I - A_1)(I - A_2)m \\ &= \mu + m_1 + m_2 + m_{12} \end{aligned} \tag{4.2}$$

μ is the overall mean, m_1, m_2 are the main effects of the two genes and m_{12} describes the interaction effect between them.

4.2.2 Reproducing kernel Hilbert space and the dual representation

Based on the ANOVA decomposition, an reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions on $\mathcal{X}$ can be constructed [132, 133]. Let $\mathcal{H}^{(j)}$ be an RKHS of functions on $\mathcal{X}^{(j)}$, $j = 1, 2$ and $\mathbf{1}^{(j)}$ be a space of constant functions on $\mathcal{X}^{(j)}$, then

$$\begin{aligned} \mathcal{H} &= \prod_{j=1}^{2}(\mathbf{1}^{(j)} \oplus \mathcal{H}^{(j)}) \\ &= [\mathbf{1}] \oplus [\mathcal{H}^{(1)} \otimes \mathbf{1}^{(2)}] \oplus [\mathbf{1}^{(1)} \otimes \mathcal{H}^{(2)}] \oplus (\mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}) \\ &= [\mathbf{1}] \oplus \mathcal{H}^1 \oplus \mathcal{H}^2 \oplus \mathcal{H}^3 \end{aligned} \tag{4.3}$$

where $\oplus$ refers to direct sum and $\otimes$ refers to tensor product. Equation (4.3) provides an orthogonal decomposition of the entire functional space $\mathcal{H}$. So $\mathcal{H}$ is an RKHS with the associated reproducing kernel as the sum of the reproducing kernels of these component subspaces. Each functional component in (4.2) lies in a subspace in (4.3), and is estimated in

the corresponding RKHS. The identifiability of the components is assured by side conditions : $\int_{\mathcal{X}^{(j)}} m_j(\boldsymbol{x}^{(j)})d\mu_j = 0,\ j = 1, 2.$

We assume that function m is a member of the RKHS $\mathcal{H}$ and it can be estimated as the minimizer of the following penalized sum of squares.

$$\mathcal{L}(\boldsymbol{y}, m) = \sum_{i=1}^{n}(y_i - m(\boldsymbol{x}_i))^2 + \frac{1}{2}\lambda J(m) \tag{4.4}$$

where J is a roughness penalty. Since the orthogonal decomposition of spcase $\mathcal{H}$, the penalty can also be decomposed, then

$$\mathcal{L}(\boldsymbol{y}, m) = \sum_{i=1}^{n}(y_i - m(\boldsymbol{x}_i))^2 + \frac{1}{2}\sum_{l=1}^{3}\lambda_l \parallel P^l m(.) \parallel^2_{\mathcal{H}} \tag{4.5}$$

where P^l is the orthogonal projector in $\mathcal{H}$ onto $\mathcal{H}^l$, λ_l are the tuning parameters which balance the goodness of fit and complexity of the model. The minimizer of (4.5) is known to have a representation [130] in terms of a constant and the associated reproducing kernels $\{k_l(s,t)\}$ of the $\mathcal{H}^l, l = 1, 2, 3$.

$$\begin{aligned} m(\boldsymbol{x}) &= \mu\mathbf{1} + \sum_{i=1}^{n} c_i \sum_{l=1}^{3} \theta_l k_l(\boldsymbol{x}_i, \boldsymbol{x}) \\ &= \mu\mathbf{1} + \sum_{l=1}^{3} K_l^T(\boldsymbol{x}) C_l \end{aligned} \tag{4.6}$$

where $K_l^T(\boldsymbol{x}) = (k_l(\boldsymbol{x}_1, \boldsymbol{x}), \cdots, k_l(\boldsymbol{x}_n, \boldsymbol{x}))$, $C_l = (c_1, \cdots, c_n)^T\theta_l$. Details on the choice of the reproducing kernel functions corresponding to the three subspaces will be discussed in a later section.

Substitute the representation of $m(\cdot)$ into equation (4.5)

$$\begin{aligned}\mathcal{L}(\boldsymbol{y}, m) &= \sum_{i=1}^{n}(y_i - m(\boldsymbol{x}_i))^2 + \frac{1}{2}\sum_{l=1}^{3}\lambda_l \parallel P^l m(\cdot) \parallel_{\mathcal{H}}^2 \\ &= (\boldsymbol{y} - m(\boldsymbol{X}))^T(\boldsymbol{y} - m(\boldsymbol{X})) + \frac{1}{2}\sum_{l=1}^{3}\lambda_l C_l^T \mathbf{K}_l C_l \\ &= (\boldsymbol{y} - \mu\mathbf{1} - \sum_{l=1}^{3}\mathbf{K}_l C_l)^T(\boldsymbol{y} - \mu\mathbf{1} - \sum_{l=1}^{3}\mathbf{K}_l C_l) + \frac{1}{2}\sum_{l=1}^{3}\lambda_l C_l^T \mathbf{K}_l C_l\end{aligned} \tag{4.7}$$

where

$$\mathbf{K}_l = \begin{bmatrix} K_l^T(\boldsymbol{x}_1) \\ K_l^T(\boldsymbol{x}_2) \\ \vdots \\ K_l^T(\boldsymbol{x}_n) \end{bmatrix}$$

The gradients of $\mathcal{L}$ with respect to the coefficients (μ, $C_l : l = 1, 2, 3$) are

$$\frac{\partial \mathcal{L}}{\partial \mu} = \mathbf{1}^T(\boldsymbol{y} - \mu\mathbf{1} - \sum_{l=1}^{3}\mathbf{K}_l C_l)$$

and

$$\frac{\partial \mathcal{L}}{\partial C_l} = \mathbf{K}_l^T(\boldsymbol{y} - \mu\mathbf{1} - \sum_{l=1}^{3}\mathbf{K}_l C_l) + \lambda_l \mathbf{K}_l C_l$$

Therefore, the first order condition is satisfied by the system

$$\begin{bmatrix} n & \mathbf{1}^T\mathbf{K}_1 & \mathbf{1}^T\mathbf{K}_2 & \mathbf{1}^T\mathbf{K}_3 \\ \mathbf{K}_1^T\mathbf{1} & \mathbf{K}_1^T\mathbf{K}_1 + \lambda_1\mathbf{K}_1 & \mathbf{K}_1^T\mathbf{K}_2 & \mathbf{K}_1^T\mathbf{K}_3 \\ \mathbf{K}_2^T\mathbf{1} & \mathbf{K}_2^T\mathbf{K}_1 & \mathbf{K}_2^T\mathbf{K}_2 + \lambda_2\mathbf{K}_2 & \mathbf{K}_2^T\mathbf{K}_3 \\ \mathbf{K}_3^T\mathbf{1} & \mathbf{K}_3^T\mathbf{K}_1 & \mathbf{K}_3^T\mathbf{K}_2 & \mathbf{K}_3^T\mathbf{K}_3 + \lambda_3\mathbf{K}_3 \end{bmatrix} \begin{bmatrix} \mu \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{K}_1^T \\ \mathbf{K}_2^T \\ \mathbf{K}_3^T \end{bmatrix} \boldsymbol{y} \tag{4.8}$$

The connection between smoothing splines and linear mixed effects model has been previously established [130, 134]. For the two-way ANOVA decomposition model considered in this paper, we show that the above first order system is equivalent to the Henderson's normal equation the following linear mixed effects model; see Appendix B.2.

$$y = \mu\mathbf{1} + \tilde{m}_1 + \tilde{m}_2 + \tilde{m}_{12} + \epsilon \tag{4.9}$$

where $\tilde{m}_1, \tilde{m}_2, \tilde{m}_{12}$ are independent $n \times 1$ vector of random effects; $\tilde{m}_1 \sim N(\mathbf{0}, \tau_1^2\mathbf{K}_1)$, $\tilde{m}_2 \sim N(\mathbf{0}, \tau_2^2\mathbf{K}_2)$, $\tilde{m}_{12} \sim N(\mathbf{0}, \tau_3^2\mathbf{K}_3)$, and $\epsilon \sim N(0, \sigma^2 I)$ is independent of $\tilde{m}_1, \tilde{m}_2, \tilde{m}_{12}$. This connection indicates the estimators of functions m_1, m_2, m_{12} are just the BLUPs of the random effects in the linear mixed effects model. Tuning parameters $\lambda_l, l = 1, 2, 3$ are functions of the variance components, which can be estimated either by maximum likelihood method or by restricted maximum likelihood (REML) method. Since REML method gives unbiased estimates for the variance components, we adopt the REML estimation in this work. The obtained dual representation of the linear mixed effects model for the SS-ANOVA model makes it feasible to do inferences about the main and interaction components under the mixed effects model framework.

4.2.3 Choice of kernel function for genotype similarity

The choice of reproducing kernel is not arbitrary in the sense that the kernel function must be non-negative definite. By theorem 2.3 [131], given a non-negative definite function k on $\mathcal{X}$, we can construct a unique RKHS of real-valued functions on $\mathcal{X}$ with k as its reproducing kernel. In a genetic association study, kernel function captures the pairwise genomic similarities between two individuals across multiple SNPs in a gene. It projects the genotype data from the original space, which can be high dimensional and nonlinear, to a one-dimensional

linear space. The Allele Matching (AM) kernel is one of the most popularly used kernels for measuring genotype similarity. This type of kernel measure has been used in linkage analysis [135] and in association studies [123, 124, 125, 128, 136]. For a comprehensive review of genomic similarity and kernel methods, readers are referred to [126, 127]. With the notable strength of not requiring knowledge of the risk allele for each SNP, AM kernel has been chosen as the kernel function in this study. This similarity kernel counts the number of matches among the four comparisons between two genotypes $g_{i,s}$ (with two alleles A and B) and $g_{j,s}$ (with two alleles C and D) of two individuals i and j at locus s, and can be expressed as

$$AM(g_{i,s} = A/B, g_{j,s} = C/D) = I(A \equiv C) + I(A \equiv D) + I(B \equiv C) + I(B \equiv D) \quad (4.10)$$

where I is the indicator function and "$\equiv$" means the two alleles are in identical-by-state (IBS). The kernel function based on AM similarity measure then takes the form

$$f(g_i, g_j) = \frac{\sum_{s=1}^{S} AM(g_{i,s}, g_{j,s})}{4S} \quad (4.11)$$

where S is the number of SNPs considered.

To incorporate valuable SNP-specific information into analysis to potentially improve performance, a weighted-AM kernel can be applied which has the form

$$f(g_i, g_j) = \frac{\sum_{s=1}^{S} w_s AM(g_{i,s}, g_{j,s})}{4\sum_{s=1}^{S} w_s} \quad (4.12)$$

where w_s is the weighting function which can be adopted to incorporate prior knowledge to

gain extra power. For example, when a study is trying to identify the effect of rare variants, the weight function can be taken as the inverse of the minor allele frequency to boost the signal for rare variants [127].

We use the AM kernel as the reproducing kernel for the two subspaces $\mathcal{H}^1$ and $\mathcal{H}^2$ corresponding to the main effects of the two genes. Utilizing the fact that the reproducing kernel for a tensor product of two reproducing kernel spaces is the product of the two reproducing kernels [137], the associated reproducing kernel for $\mathcal{H}^3$ can be taken as the product of the reproducing kernels of the two subspaces: $\mathcal{H}^1$ and $\mathcal{H}^2$.

4.2.4 Hypothesis testing

Testing overall genetic effect In a gene-based genetic association study, one is interested in whether a gene as a system is associated with a disease trait. In the proposed 3G interaction study, we are interested in the association of each gene with a quantitative trait as well as the interaction between genes if any. The analysis starts with a two-dimensional pairwise search for gene pairs with overall contribution to the phenotypic variation and then test those contributing gene pairs for interaction effect. In the SS-ANOVA framework, testing the overall contribution of a gene pair to a phenotypic trait is to test $H_0 : m_1(\boldsymbol{x}^{(1)}) = m_2(\boldsymbol{x}^{(2)}) = m_{12}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = 0$. Similarly, testing for interaction effect can be formulated as $H_0 : m_{12}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = 0$ With the linear mixed effects model representation, the aforementioned two tests are equivalent to

$$(I) \quad H_0^1 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$$

and

$$(II) \quad H_0^2 : \tau_3^2 = 0$$

respectively. Here, $\tau_1^2, \tau_2^2, \tau_3^2, \sigma^2$ are the variance components in model (4.9).

A well-known issue in testing variance component is that the parameters under the null hypotheses are on the boundary of the parameter space. Moreover, the kernel matrices $\mathbf{K}_s$'s are not block-diagonal. Thus, the asymptotic distribution of the likelihood ratio test (LRT) statistic does not follow a central chi-square distribution under the null hypothesis. The mixture chi-square distribution proposed by Self and Liang [138] under irregular conditions does not apply in our case either. In this work, we construct score test statistics based on the restricted likelihood. Consider the linear mixed model in (4.9), $\boldsymbol{y} \sim N(\mu\mathbf{1}, V(\beta))$, and the restricted log-likelihood function can be written as

$$\ell_R \propto -\frac{1}{2} ln(|V(\beta)|) - \frac{1}{2} ln(|\mathbf{1}^T V^{-1}(\beta)\mathbf{1}|) - \frac{1}{2}(\boldsymbol{y} - \hat{\mu}\mathbf{1})^T V(\beta)^{-1}(\boldsymbol{y} - \hat{\mu}\mathbf{1}) \tag{4.13}$$

where $\beta = (\sigma^2, \tau_1^2, \tau_2^2, \tau_3^2)$, $V(\beta) = \sigma^2 I + \tau_1^2\mathbf{K}_1 + \tau_2^2\mathbf{K}_2 + \tau_3^2\mathbf{K}_3$. The first order derivative of the restricted log-likelihood function with respect to each variance component:

$$\frac{\partial \ell_R}{\partial \beta_i} = -\frac{1}{2} tr(RV_i) + \frac{1}{2}(\boldsymbol{y} - \hat{\mu}\mathbf{1})^T V^{-1}(\beta) V_i V^{-1}(\beta)(\boldsymbol{y} - \hat{\mu}\mathbf{1}) \tag{4.14}$$

where $V_i = \frac{\partial V(\beta)}{\partial \beta_i}, i = 1, \cdots, 4$, so $V_1 = I, V_2 = \mathbf{K}_1, V_3 = \mathbf{K}_2, V_4 = \mathbf{K}_3$ and $R = V^{-1} - V^{-1}\mathbf{1}(\mathbf{1}^T V^{-1}\mathbf{1})^{-1}\mathbf{1}^T V^{-1}$.

The restricted score function under the null hypothesis: $H_0^1 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ is

$$\frac{\partial \ell_R}{\partial \beta_i}|_{\tau_1^2=\tau_2^2=\tau_3^2=0} = -\frac{1}{2\sigma^2} tr(P_0 V_i) + \frac{1}{2\sigma^4}(\boldsymbol{y} - \hat{\mu}\mathbf{1})^T V_i (\boldsymbol{y} - \hat{\mu}\mathbf{1})$$

where $P_0 = \mathbf{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$ is the projection matrix under the null. Thus, H_0^1 can be tested by the following score statistic

$$S(\sigma^2) = \frac{1}{2\sigma^2}(\boldsymbol{y} - \hat{\mu}_0\mathbf{1})^T \sum_{l=1}^{3} \mathbf{K}_l(\boldsymbol{y} - \hat{\mu}_0\mathbf{1})$$

where $\hat{\mu}_0 = (\mathbf{I} - P_0)y$ is the MLE of μ under the null. This leads to

$$S(\sigma^2) = \frac{1}{2\sigma^2}\boldsymbol{y}^T P_0 \sum_{l=1}^{3} \mathbf{K}_l P_0 \boldsymbol{y} \tag{4.15}$$

Denoting σ_0^2 as the true value of σ^2 under the null, then $S(\sigma_0^2)$ is a quadratic form in $\boldsymbol{y}$. Following Liu and Lin [139], we use the satterthwaite method to approximate the distribution of $S(\sigma_0^2)$ by a scaled chi-square distribution, i.e., $S(\sigma_0^2) \sim a\chi_g^2$, where the scale parameter a and the degrees of freedom g can be estimated by the method of moments (MOM). By equating the mean and variance of the test statistic $S(\sigma_0^2)$ with those of $a\chi_g^2$, we have

$$\begin{cases} \delta = E[S(\sigma_0^2)] = tr(P_0 \sum_{i=1}^{3} \mathbf{K}_i)/2 = E[a\chi_g^2] = ag \\ \nu = Var[S(\sigma_0^2)] = tr(\sum_{i=1}^{3}(P_0\mathbf{K}_i) \sum_{i=1}^{3}(P_0\mathbf{K}_i))/2 = Var[a\chi_g^2] = 2a^2g \end{cases} \tag{4.16}$$

Solving for the two equations leads to $\hat{a} = \nu/2\delta$ and $\hat{g} = 2\delta^2/\nu$.

In practice, we do not know the true value σ_0^2 and we usually replace it by its MLE under the null model, denoted as $\hat{\sigma}_0^2$. The asymptotic distribution of $S(\hat{\sigma_0}^2)$ can still be approximated by the scaled chi-square distribution because the MLE is $\sqrt{n}$ consistent. To account for this substitution, we estimate a and g by replacing ν with $\tilde{\nu}$ based on the efficient

information. The Fisher's information matrix of $\boldsymbol{\tau} = (\tau_1^2, \tau_2^2, \tau_3^2)$ is given by

$$I_{\boldsymbol{\tau}\boldsymbol{\tau}} = \frac{1}{2}\begin{bmatrix} tr(P_0\mathbf{K}_1P_0\mathbf{K}_1) & tr(P_0\mathbf{K}_1P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_1P_0\mathbf{K}_3) \\ tr(P_0\mathbf{K}_2P_0\mathbf{K}_1) & tr(P_0\mathbf{K}_2P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_2P_0\mathbf{K}_3) \\ tr(P_0\mathbf{K}_3P_0\mathbf{K}_1) & tr(P_0\mathbf{K}_3P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_3P_0\mathbf{K}_3) \end{bmatrix}$$

$$I_{\boldsymbol{\tau}\sigma^2} = \frac{1}{2}\begin{pmatrix} tr(P_0\mathbf{K}_1) & tr(P_0\mathbf{K}_2) & tr(P_0\mathbf{K}_3) \end{pmatrix}^T$$

and $I_{\sigma^2\sigma^2} = \frac{1}{2}tr(P_0P_0)$. Then the efficient information $\tilde{I}_{\boldsymbol{\tau}\boldsymbol{\tau}} = I_{\boldsymbol{\tau}\boldsymbol{\tau}} - I^T_{\boldsymbol{\tau}\sigma^2}I^{-1}_{\sigma^2\sigma^2}I_{\boldsymbol{\tau}\sigma^2}$ and

$$\tilde{\nu} = Var[S(\hat{\sigma}^2)] \approx SUM[\tilde{I}_{\boldsymbol{\tau}\boldsymbol{\tau}}] \tag{4.17}$$

where operator "SUM" indicates the sum of every elements of the matrix.

Testing G×G interaction For testing interaction effect, i.e., testing $H_0^2 : \tau_3^2 = 0$, we also apply a score test. Denote $\Sigma = \sigma^2 I + \tau_1^2\mathbf{K}_1 + \tau_2^2\mathbf{K}_2$. The score function (4.14) under this null hypothesis becomes:

$$\begin{aligned} \frac{\partial \ell_R}{\partial \tau_3^2}|_{\tau_3^2=0} &= -\frac{1}{2}[tr(P_{01}\mathbf{K}_3) - (\boldsymbol{y} - \hat{\mu}\mathbf{1})^T\Sigma^{-1}\mathbf{K}_3\Sigma^{-1}(\boldsymbol{y} - \hat{\mu}\mathbf{1})] \\ &= -\frac{1}{2}(tr(P_{01}\mathbf{K}_3) - \boldsymbol{y}^TP_{01}\mathbf{K}_3P\boldsymbol{y}) \end{aligned} \tag{4.18}$$

where $P_{01} = \Sigma^{-1} - \Sigma^{-1}\mathbf{1}(\mathbf{1}^T\Sigma^{-1}\mathbf{1})^{-1}\mathbf{1}^T\Sigma^{-1}$ is the projection matrix under the null, then

$$S_I = \frac{1}{2}\boldsymbol{y}^TP_{01}\mathbf{K}_3P_{01}\boldsymbol{y} \tag{4.19}$$

Similarly, Satterthwaite method is used to approximate the distribution of S_I by $a_I\chi^2_{g_I}$.

Parameters a_I and g_I are estimated by MOM. Specifically, $\hat{a}_I = \nu_I/2\delta_I$ and $\hat{g}_I = 2\delta_I^2/\nu_I$, where $\delta_I = \frac{1}{2}tr(P_{01}\mathbf{K}_3)$ and $\nu_I = \frac{1}{2}tr(P_{01}\mathbf{K}_3P_{01}\mathbf{K}_3) - \frac{1}{2}\Phi^T\Delta^{-1}\Phi$,

$$\Phi = [tr(P_{01}^2\mathbf{K}_3), tr(P_{01}\mathbf{K}_3P_{01}\mathbf{K}_1), tr(P_{01}\mathbf{K}_3P_{01}\mathbf{K}_2)]^T$$

and

$$\Delta = \begin{bmatrix} tr(P_{01}^2) & tr(P_{01}^2\mathbf{K}_1) & tr(P_{01}^2\mathbf{K}_2) \\ tr(P_{01}^2\mathbf{K}_1) & tr(P_{01}\mathbf{K}_1P_{01}\mathbf{K}_1) & tr(P_{01}\mathbf{K}_1P_{01}\mathbf{K}_2) \\ tr(P_{01}^2\mathbf{K}_2) & tr(P_{01}\mathbf{K}_2P_{01}\mathbf{K}_1) & tr(P_{01}\mathbf{K}_2P_{01}\mathbf{K}_2) \end{bmatrix}$$

4.3 Simulation study

4.3.1 Simulation design

Monte Carlo simulations were conducted to evaluate the performance of the proposed approach for detecting genetic effects as well as gene-gene interaction in an association study. The genotype data were simulated using two approaches introduced in [111]. In the following, we describe the details of the two genotype generating methods: MS program and LD-based simulation.

MS program: The MS program developed by Hudson [140] generates haplotype samples by using the standard coalescent approach in which the random genealogy of a sample is first generated and the mutations are randomly placed on the Genealogy. We first simulated two independent samples of haplotypes by using MS program. Parameters of the coalescent model were set as following: (1) The diploid population size $N_0 = 10,000$; (2) The mutation parameter $\theta = 4N_0\mu = 5.610 \times 10^{-4}/bp$; and (3) The cross-over rate parameters are $\rho =$

$4N_0r = 4.0 \times 10^{-3}/bp$ and $\rho = 8 \times 10^{-3}/bp$ for the two independent samples respectively. In each sample, 100 haplotypes were simulated for a locus with 10kb long and the number of SNP sequences were set to be 100. Two haplotypes were then randomly drawn within each simulated haplotype pool and paired to form the genotype on the locus for an individual. For each individual, we randomly selected 10 adjacent SNPs with minor allele frequency (MAF) greater than 5% to form a gene. This was done separately for each simulated haplotype pool and finally we had genotypes for n individuals for two separate genes with 10 SNPs each, and the two genes were independent.

LD-based simulation: Under this scenario, SNP genotypes were simulated by controlling pairwise LD values. Let p_A be the MAF for SNP1. Assuming Hardy-Weinberg equilibrium (HWE), the first SNP marker can be simulated according to a multinomial distribution with frequencies p_A^2, $2p_A(1-p_A)$ and $(1-p_A)^2$ for genotype AA, Aa and aa, respectively. Let the MAF of the next simulated marker (SNP2) as p_B and the LD between SNP1 and SNP2 be D. Assuming HWE, the four haplotype frequencies can be calculated as $p_{AB} = p_Ap_B+D$, $p_{Ab} = p_A(1-p_B)-D$, $p_{aB} = (1-p_A)p_B-D$ and $p_{ab} = (1-p_A)(1-p_B)+D$ for haplotype AB, Ab, aB and ab, respectively. The conditional genotype distribution of SNP2 given on SNP1 can be derived as

$$P(BB|AA) = \frac{P(AABB)}{P(AA)} = \frac{p_{AB}^2}{p_A^2} = \frac{(p_Ap_B + D)^2}{p_A^2} \tag{4.20}$$

Similarly we can get the other 8 conditional genotype distributions (see Table 1 in [111] for more details). Two genes with 10 SNPs each were simulated by applying the LD-based simulation method. For gene 1, we assume MAF=0.3 and pairwise SNP correlation $r^2 = 0.5$ ($r^2 = \frac{D^2}{p_Ap_B(1-p_A)(1-p_B)}$). For gene 2, we assume MAF=0.2, and r^2=0.8.

Phenotype simulation: Four simulation scenarios were considered in simulating the phenotype (Table 4.1). In Scenario I, the three genetic effects were set as zero, with which we can assess the false positive control of different methods. In Scenario II, we considered the main effects for the two genes, but set the interaction effect as zero. In Scenarios III and IV, both main effects and interaction effect were considered. The difference between the scenario III and IV is that the interaction effect in Scenario III is smaller than the main effect, while in Scenario IV it is larger than both main effects. Quantitative trait of interest were simulated from a multivariate normal distribution with mean $\mu\mathbf{1}_{n\times 1}$ and variance-covariance matrix $V = \sigma^2\mathbf{I} + \tau_1^2\mathbf{K}_1 + \tau_2^2\mathbf{K}_2 + \tau_3^2\mathbf{K}_3$, where $\tau_1^2, \tau_2^2, \tau_3^2$ took different values under different scenarios; $\mathbf{K}_i, i = 1, 2, 3$ are the kernel matrices using the allele matching method described before. Different sample sizes ($n = 200$ and 500) and different heritability (H^2=0.1, 0.2, 0.4) were assumed. Let $\sigma_G^2 = \tau_1^2 + \tau_2^2 + \tau_3^2$, then the heritability is defined as $H^2 = \sigma_G^2/(\sigma_G^2 + \sigma^2)$. For a given value of residual variance σ^2, the main effects of the two genes were set equal. When the interaction effect was considered, it was set as either half of the main effect (Scenario III) or double the main effect (Scenario IV). Thus for a given heritability level, the parameter values were different under different scenarios. Specific values for $\sigma^2, \tau_1^2, \tau_2^2, \tau_3^2$ were given in the first column of Table 4.1.

4.3.2 Model comparison

We mainly compared our simulation results with two other methods described in the follows. Wang et al. proposed an interaction method using a partial least square approach which is developed specifically for binary disease traits [141]. The method cannot be applied for quantitative traits. However, in [141], the authors compared their method with a regression-

based principle component analysis method. Specifically, assuming an additive model for each marker in which genotypes AA, Aa and aa are coded as 2,1,0, respectively, the singular value decomposition (SVD) can be applied to both gene matrices. Let G_j be an $n \times L_j$ SNP matrix for gene j $(= 1, 2)$. The SVD for G_j can be expressed as $G_j = U_j D_j V_j^T$, where D_j is a diagonal matrix of singular values, and the elements of the column vector U_j are the principal components $U_j^1, U_j^2, \cdots, U_j^{m_j}$ ($m_j \leq L_j$ is the rank for G_j). An interaction model can be expressed as

$$y = \mu + \sum_{l_1=1}^{L_1} \beta_{l_1} x_{l_1} + \sum_{l_2=1}^{L_2} \beta_{l_2} x_{l_2} + \gamma U_1^1 U_1^2 \tag{4.21}$$

where γ represents the interaction effect between the first pair of PCs corresponding to the largest eigenvalues in the two genes. The main effect of the each gene is modeled through the sum of all single marker effects. For simplicity, only one interaction effect between the first PC corresponding to the largest eigenvalues in each gene was considered in [141]. We followed their way and compared the performance of our model with this model.

In principle, one can select PCs for each gene based on the proportion of variation explained (say $> 85\%$). Then, pairwise interactions can be considered for all selected PCs in model (4.21). Thus, we replaced the main effect of each gene in model (4.21) with PCs rather than single SNPs to reduce the model degrees of freedom, model (4.21) then becomes

$$y = \mu + \sum_{k_1=1}^{K_1} \beta_{k_1} U_{k_1} + \sum_{k_2=1}^{K_2} \beta_{k_2} U_{k_2} + \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \gamma_{k_1 k_2} U_{k_1}^1 U_{k_2}^2 \tag{4.22}$$

where $U_{k_j}, j = 1, 2$ represents the PCs for gene j, and $K_j, j = 1, 2$ is chosen based on the proportion of variation explained by the number of PCs in gene j. With this regression model, we considered all possible pairwise PC interactions between the two genes and G×G

interaction was done by testing $H_0 : \gamma_{k_1 k_2} = 0$, for all k_1 and k_2. This model was applied by [142], in their gene-based interaction analysis.

In addition to the above two models, we also compared our gene-centric approach to a pairwise SNP interaction model. Details of the comparison is given in section 4.3.3. For a given simulation scenario, 1000 simulation runs were conducted. Type I error rates and power were examined at the nominal level $\alpha = 0.05$.

4.3.3 Simulation results

Table 4.1 summarizes the comparison results between our kernel method and model (4.21) and (4.22). The power of an association test was denoted by $P_{\dot{1}}$, $P_{\dot{2}}$ and $P_{\dot{3}}$ which correspond to the power by using the proposed gene-centric interaction method, model (4.21) and (4.22), respectively. The superscript letters o and i denote the power for testing the significance of the overall genetic effects and the interaction effect, respectively. Noted that the power for the interaction test was calculated only when the overall test showed significance. Thus, the power and the false positive rate for the interaction test are smaller than the ones obtained without this constraint.

Comparisons of the the proposed method with the two PCA-based methods: The results for Scenario I indicate that our method has type I error rate reasonably controlled for the overall genetic effect tests under the two genotype simulation scenarios (see Scenario I in Table 4.1). The two PCA-based interaction models produced slightly conservative results when the genotypes were simulated with the MS program. For example, the type I error rates were 0.033 and 0.023 for the two methods when sample size is 500.

In Scenarios II-IV, we fixed the residual variance σ^2 to 0.8, and varied the three genetic

Table 4.1: List of empirical type I error and power based on 1000 simulation runs.

Parameter values			LD-based					
$(\sigma^2, \tau_1^2, \tau_2^2, \tau_3^2)$	H^2	n	P_1^{o*}	P_1^{i*}	P_2^{o*}	P_2^{i*}	P_3^{o*}	P_3^{i*}
Scenario I								
(1,0,0,0)	0	200	0.049	0.004	0.045	0.016	0.095	0.025
		500	0.061	0.002	0.044	0.021	0.055	0.012
Scenario II								
(0.8, 0.044, 0.044,0)	0.1	200	0.285	0.019	0.212	0.032	0.209	0.042
		500	0.531	0.026	0.420	0.052	0.374	0.043
(0.8, 0.1, 0.1,0)	0.2	200	0.459	0.029	0.386	0.058	0.387	0.055
		500	0.776	0.048	0.636	0.045	0.686	0.042
(0.8, 0.267, 0.267,0)	0.4	200	0.734	0.072	0.661	0.058	0.684	0.072
		500	0.927	0.065	0.862	0.069	0.939	0.071
Scenario III								
(0.8, 0.036, 0.036, 0.018)	0.1	200	0.289	0.025	0.234	0.051	0.238	0.037
		500	0.565	0.054	0.415	0.059	0.414	0.062
(0.8, 0.08, 0.08, 0.04)	0.2	200	0.486	0.053	0.389	0.065	0.389	0.046
		500	0.806	0.086	0.686	0.085	0.746	0.074
(0.8, 0.21, 0.21, 0.11)	0.4	200	0.765	0.109	0.654	0.087	0.740	0.107
		500	0.946	0.163	0.881	0.131	0.956	0.140
Scenario IV								
(0.8, 0.022, 0.022, 0.044)	0.1	200	0.318	0.047	0.245	0.048	0.253	0.051
		500	0.571	0.064	0.466	0.090	0.432	0.062
(0.8, 0.05, 0.05, 0.1)	0.2	200	0.500	0.074	0.409	0.076	0.443	0.087
		500	0.805	0.141	0.720	0.117	0.755	0.119
(0.8, 0.133, 0.133, 0.266)	0.4	200	0.771	0.172	0.694	0.115	0.750	0.136
		500	0.938	0.304	0.881	0.230	0.961	0.244

*$P.^o$ and $P.^i$ refer to the power for testing the overall genetic effects (i.e., $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$) and for testing interaction effect (i.e., $H_0 : \tau_3^2 = 0$), respectively. $P_{\dot{1}}$, $P_{\dot{2}}$ and $P_{\dot{3}}$ refer to powers by using the proposed gene-centric method, the full PCA-based interaction with model (4.22) and the partial PCA-based interaction analysis with model (4.21), respectively.

effects to get different heritability levels. As we expected, the testing power increases as the heritability level and sample size increase. For example, under the LD-based simulation, the overall power increases from 0.565 to 0.946 when H^2 increases from 0.1 to 0.4 with fixed sample size 500 in Scenario III. Under the same Scenario, the overall power increases from

Table 4.1 (cont'd)

Parameter values			MS program					
$(\sigma^2, \tau_1^2, \tau_2^2, \tau_3^2)$	H^2	n	P_1^{o*}	P_1^{i*}	P_2^{o*}	P_2^{i*}	P_3^{o*}	P_3^{i*}
Scenario I								
(1,0,0,0)	0	200	0.070	0.002	0.048	0.025	0.034	0.011
		500	0.052	0.001	0.033	0.019	0.023	0.008
Scenario II								
(0.8, 0.044, 0.044,0)	0.1	200	0.255	0.016	0.186	0.057	0.115	0.019
		500	0.525	0.036	0.339	0.045	0.254	0.030
(0.8, 0.1, 0.1,0)	0.2	200	0.485	0.044	0.324	0.058	0.253	0.041
		500	0.755	0.041	0.615	0.071	0.594	0.050
(0.8, 0.267, 0.267,0)	0.4	200	0.758	0.080	0.611	0.066	0.604	0.052
		500	0.946	0.066	0.842	0.066	0.917	0.048
Scenario III								
(0.8, 0.036, 0.036, 0.018)	0.1	200	0.299	0.019	0.164	0.041	0.126	0.027
		500	0.548	0.030	0.399	0.069	0.298	0.034
(0.8, 0.08, 0.08, 0.04)	0.2	200	0.491	0.069	0.346	0.056	0.279	0.046
		500	0.752	0.061	0.640	0.089	0.632	0.045
(0.8, 0.21, 0.21, 0.11)	0.4	200	0.766	0.100	0.629	0.091	0.616	0.069
		500	0.941	0.131	0.872	0.128	0.914	0.097
Scenario IV								
(0.8, 0.022, 0.022, 0.044)	0.1	200	0.280	0.027	0.189	0.051	0.136	0.032
		500	0.571	0.038	0.449	0.089	0.325	0.045
(0.8, 0.05, 0.05, 0.1)	0.2	200	0.514	0.053	0.377	0.062	0.291	0.043
		500	0.787	0.111	0.669	0.119	0.667	0.105
(0.8, 0.133, 0.133, 0.266)	0.4	200	0.779	0.153	0.619	0.103	0.680	0.092
		500	0.963	0.256	0.874	0.211	0.955	0.194

* $P.^o$ and $P.^i$ refer to the power for testing the overall genetic effects (i.e., $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$) and for testing interaction effect (i.e., $H_0 : \tau_3^2 = 0$), respectively. $P_1^{\cdot}$, $P_2^{\cdot}$ and $P_3^{\cdot}$ refer to powers by using the proposed gene-centric method, the full PCA-based interaction with model (4.22) and the partial PCA-based interaction analysis with model (4.21), respectively.

0.486 to 0.806 when sample size increases from 200 to 500 under fixed H^2. We observed a similar trend for genotypes simulated with the MS program (Table 4.1).

Relatively little power to detect interactions was observed for the three methods (partly due to the way we calculated the interaction power). As sample size or heritability increase, the interaction power also increases. Larger interaction effect (Scenario IV) results in larger

interaction power compared to the one obtained with smaller interaction effect (Scenario III). For example, for fixed sample size ($n = 500$) and fixed heritability ($H^2 = 0.4$), the interaction power increases from 16% to 30% under the LD-based simulation when the interaction effect was doubled. We did additional simulation by increasing the sample size to 1000 and achieved reasonable interaction power (data not shown). The simulation results indicate that large sample size is needed in order to obtain reasonable power to detect the interaction effect.

Model performance under different interaction effect sizes: Interactions may be caused by a variety of underlying mechanisms. Some genes might have both significant main and interaction effects, while others might only incur epistatic effects without main effects. Simulation studies were designed to evaluate the performance of the proposed kernel machine approach in discovering gene $\times$ gene interaction under different epistasis effect sizes. We defined the proportion of the epistatic variance among the total genetic variance as $\rho = \tau_3^2/(\tau_1^2+\tau_2^2+\tau_3^2)$, which gave us an indication of the strength of the epistatic effect between two genes for a fixed total genetic variance.

Two genes each with 10 SNPs were considered as in previous simulation studies. Genotype data and phenotype data were generated as described in Section 3.1, but with different values for the variance components. For a given heritability level ($H^2 = 0.4$) and a fixed residual error variance ($\sigma^2 = 0.6$), the total genetic variance is calculated as 0.4. We then assumed the same effect size for the two main components, and varied the proportion ρ. For example, we had $(\tau_1^2, \tau_2^2, \tau_3^2) = (0.16, 0.16, 0.08)$ when $\rho = 0.2$, and $(\tau_1^2, \tau_2^2, \tau_3^2) = (0.04, 0.04, 0.32)$ when $\rho = 0.8$. Six values of proportion $\rho = (0, 0.2, 0.4, 0.6, 0.8, 1.0)$ were considered, including the two extreme cases: no epistatic effect at all ($\rho = 0$) and pure epistasis ($\rho = 1$). Comparisons with the other two PCA-based interaction analyses were considered under two

different sample sizes, 500 and 1000. Empirical powers was calculated based on testing the interaction effect only.

Results based on 1000 replicates were summarized in Figure 4.1. All the three methods can reasonably control the type I error ($\rho = 0$). As we expected that the empirical interaction power increases as the interaction effect size increases. When SNPs are correlated (Figure 4.1B), small number of PCs might be enough to capture the variation of each gene. So the power is larger than MS-based simulation (Figure 4.1A). Among the three methods, our kernel-based method has the highest power. Model (4.21) has the lowest power, which implies that only considering one pair of PC interaction is not enough to capture the interaction effect between two genes. The effect of sample size on the interaction power is also significant. Larger sample size always leads to larger power. The results also confirm that detecting gene $\times$ gene interactions generally requires relatively larger sample size than it does for detecting main genetic effects.

Comparison with the single SNP interaction model: In a regression-based analysis for interaction, the commonly used approach is the single SNP interaction model with the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \tag{4.23}$$

where β_0 is the intercept; β_1, β_2 and β_{12} represent the effects of SNP x_1 in gene 1, SNP x_2 in gene 2 and the interaction effect between the two; and $\varepsilon \sim N(0, \sigma^2)$. We simulated data according to model (4.23) assuming a MAF $p_A = 0.3$. Different heritabilities and different sample sizes were assumed. Obviously it is unfair to compare the two since the single SNP interaction model is the true analytical model and it should have the best performance. However, it is worth to evaluate the performance of our kernel method when there is only

one functional pair of SNPs in two genes. For simplicity, we assumed the same effect size for the three coefficients which are calculated under specific heritability ($H^2 = 0.2$ and 0.4) when generating the data. We considered an extreme case in which each gene only contains one single SNP. Data generated with model (4.23) are subject to both the single SNP interaction and the proposed kernel interaction analysis. The results are summarized in Table 4.2.

Table 4.2: List of empirical type I error and power based on 1000 simulation runs (single SNP interaction model).

Heritability	Coefficients	Sample size	Single SNP		Kernel	
(H^2)	($\beta_0, \beta_1, \beta_2, \beta_{12}$)	(n)	P_o	P_i	P_o	P_i
		200	0.055	0.019	0.059	0.003
	(0.19, 0, 0, 0)	500	0.058	0.019	0.057	0.003
		1000	0.052	0.017	0.059	0.003
		200	0.497	0.03	0.534	0.032
0.2	(0.19, 0.19, 0.19, 0)	500	0.923	0.045	0.911	0.046
		1000	0.999	0.048	0.997	0.053
		200	1	0.221	1	0.183
	(0.19, 0.19, 0.19, 0.19)	500	1	0.419	1	0.349
		1000	1	0.714	1	0.635
		200	0.053	0.022	0.053	0.003
	(0.51, 0, 0, 0)	500	0.049	0.016	0.062	0.001
		1000	0.054	0.024	0.057	0.008
		200	1	0.051	1	0.058
0.4	(0.51, 0.51, 0.51, 0)	500	1	0.062	1	0.067
		1000	1	0.054	1	0.058
		200	1	0.850	1	0.648
	(0.51, 0.51, 0.51, 0.51)	500	1	0.996	1	0.964
		1000	1	1	1	1

* P_o and P_i refer to the power for testing the overall genetic effects (i.e., $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$ for the kernel approach and $H_0 : \beta_1 = \beta_2 = \beta_{12} = 0$ for the pairwise SNP interaction analysis) and for testing interaction effect (i.e., $H_0 : \tau_3^2 = 0$ for the kernel approach and $H_0 : \beta_{12} = 0$ for the pairwise SNP interaction analysis), respectively.

Both models show comparable type I error control for the overall genetic test (see P_o in the table). For the interaction test, it looks like that the kernel approach generates more

conservative results. Here the interaction test is nested within the overall genetic test. If we aggregate the results by dividing P_i by P_o, the single SNP analysis actually produces more inflated false positives compared to our kernel approach when no genetic effect is involved at all. When data were simulated assuming only main effects but no interaction (case $\beta_{12} = 0$), the two approaches yield very similar false positive rate, indicating reasonable performance of the kernel approach for false positive control.

For the power analysis, we found little difference between the two methods for the overall genetic test (P_o), especially under large sample size and high heritability level. For the interaction test (P_i), we found the power increase as sample size and heritability level increase. For example, P_i increases from 0.183 to 0.635 for the kernel approach when sample size increases from 200 to 1000, a 2.5 fold increase in power under a fixed heritability level (H^2=0.2). When heritability level increases from 0.2 to 0.4 under a fixed sample size (say 500), we saw a dramatic power increase from 0.349 to 0.964 for the kernel approach. A similar trend is also observed for the single SNP interaction analysis. The information implies that from a practical point of view, large sample is always preferred, especially when environmental noise is large. Overall, the single SNP interaction model (4.23) yields slightly higher power than the kernel approach, although the difference is diminished under large sample size ($n = 1000$) and high heritability level ($H^2 = 0.4$). This is not surprising since one would expect to see large power when data are analyzed with the true model. We did additional simulations in which more than one functional SNPs within each gene were involved in interacting with each other to affect a trait variation. Results showed that the kernel method consistently outperformed the single SNP interaction model (data not shown).

In summary, our model performs reasonably well in different scenarios compared to the

other methods. Even when there is only one single SNP pair interacting with each other in two genes, our analysis produces results as good as the ones analyzed with the true model, especially under large sample size and high heritability (Table 4.2). For the powers obtained under the two genotype simulation methods, the difference is not remarkable. To achieve high power, large sample size (say $n > 500$) is always encouraged.

4.4 Applications to real data

4.4.1 Analysis of yeast eQTL mapping data

The real data set we analyzed with our model is the well studied yeast eQTL mapping dataset generated to understand the genetic architecture of gene expression. The details of the dataset is described in section (2.2.1). As an example to show the utility of our approach to an eQTL mapping study, we picked the expression profile of one gene (BAT2) as the quantitative response to identify potential genes or epistasis that regulate the expression of this gene. Noted that one of the parental strain RM11-1a is a LEU2 knockout strain. We expect strong segregation of this gene in the mapping population. Thus we picked this gene which is in the downstream of Leucine Biosynthesis Pathway (see Fig. 5(a) in [57]) as the response. A two-dimensional pairwise interaction search was done. Due to strong signals, Bonferroni correction was applied to adjust multiple testings for the 1072380 gene pairs. Overall test for pairs of gene effects was conducted followed by the score test for interaction if the overall test is significant.

There are total 1465 genes with some containing a single SNP marker. All the genes were subject to the proposed kernel interaction analysis. Figure 4.2A shows the pairwise inter-

action plot for -log10 transformed p-values associated with the overall genetic test (I). The yellow hyperplane indicates the Bonferroni correction threshold. Data points with p-values larger than 10^{-4} were masked. The plot indicates a strong genetic effect at chromosome 3 and 13, which implies that the two locations are potential regulation hotspots. In checking the recent literature, we found that the two positions were reported as eQTL hotspots in a number of studies [45, 64, 63].

Out of the 1072380 gene pairs, 87 pairs were found to have significant interactions at an experimental wide level of 0.05. Figure 4.2B plots the pairwise significant interactions. Circles corresponds to significant interaction pairs with the darkness of the color indicating the strength of the interaction. We saw a strong interaction pattern on chromosome 13. One or several genes at this location interact with many other genes to affect the transcription of gene BAT2. Another interaction "hotspot" is at chromosome 3 where genes (containing LUE2 and its neighborhood genes) interact with genes at chromosome 5, 13 and 15 to regulate BAT2 expression. We used Cytospace [143] to generate an interaction network (see Fig. 4.3). Each node represents a gene and the thickness of the connection line indicates the strength of the interaction effect. Genes at the same chromosome location are clustered together in the plot. Light nodes with oval shapes indicates weak or no marginal effects. We found strong marginal effects for genes on chromosome 3 and 13. The most strongest interaction effect is between genes on chromosome 3 and chromosome 13. We also highlighted (red lines) the interaction between genes on chromosome 3 and others. Among the genes with no marginal effects (light oval nodes), is URA3 which is a known transcription factor [144]. Even though it does not show any main effect, it interacts with several genes on chromosome 3 to regulate the expression of BAT2. The results also imply the important role of several

loci on chromosome 13. Since their functions are unknown, they can be potential candidate genes for further lab validation.

4.5 Discussion

The importance of gene-gene interaction in complex traits has stimulated enormous discussion and fundamental works in statistical methodology development have been broadly pursued (reviewed in [122]). Previous investigations have demonstrated the importance of a gene-centric approach in genetic association studies by simultaneously considering all markers in a gene to boost association power and reduce the number of tests [111, 112, 145]. This motivates us to develop a gene-centric approach to understand gene-gene interaction associated with complex traits.

In this work, we have proposed a gene-centric kernel machine framework for gene-gene interaction analysis. Our model considers all variants in a gene as a system and adopts a kernel function to model the genomic similarity between SNP variants. The kernel machine method was previous developed for an association test and has been shown to be powerful in association studies [128, 129]. Motivated by these work, we propose a spline-smoothing ANOVA decomposition method to decompose the genetic effects of two genes into separate main and interaction effects, and further model and test the genetic effects in the reproducing kernel Hilbert space. The joint variation of SNP variants within a gene is captured by a properly defined kernel function, which enables one to model the interaction of two genes in a linear reproducing Hilbert space by a cross-product of two kernel functions. Following rigorous derivations, the kernel machine method is shown to be equivalent to a linear mixed effects model. Thus, testing main and interaction effects can be done by testing the signif-

icance of different variance components. Extensive simulations under various settings and the analysis of two real data sets demonstrate the advantage of the gene-centric analysis.

He et al. [142] previously proposed a gene-based interaction method in which each gene is summarized by several principle components and interaction was tested through the modeling of the PC terms rather than single SNPs. The authors proposed a weighted genotype scoring method using pairwise LD information to test gene-gene interaction. Their method is similar to several other methods which jointly consider information contributed by multiple markers [146, 147]. Our method is fundamentally different from their approach in which we capture the joint variation of SNP variants within and between genes by kernel functions (see [126] for more discussion of the advantage of the kernel methods). Our method can also be extended to test interaction of variants by incorporating various weighting functions to define a kernel measure. Simulation studies demonstrate the advantage of the method over the PC-based regression analysis.

The advantage of the gene-centric gene-gene interaction analysis was previously discussed in [142], such as reducing the number of hypothesis tests in a genome-wide scan. However, we should not over-emphasize the role of gene-centric analysis. Our simulation study indicates that when the underlying truth is that interaction only occurs between two single SNPs in two genes, single-SNP interaction analysis performs better. This result agrees with the conclusion made by He et al. [142]. Therefore, we recommend investigators conduct both types of analysis (single SNP and gene-centric) in real applications, especially when no prior knowledge is available on how SNPs function within a gene as well as between genes. For a large-scale genome-wide or candidate gene study, one can also use the gene-centric approach as a screening tool, then further target which SNPs in different genes interact with each

other.

The choice of kernel function may have potential effects on the testing power [126, 127]. In this paper, we consider the allele matching (AM) kernel. Choice of other kernel functions can also be applied such as the identical-by-state (IBS) kernel and many others. Schaid gave a very nice summary of various choices of kernel functions and their applications in genetic association studies [126, 127]. It is not the purpose of this paper to compare the performance of difference kernel choices on the power of an association test. A comparison study of different kernel functions on the power of the interaction test will be considered in future investigation.

The proposed method considers two genes as two units to test their interaction. It is easy to extend the idea to incorporate other genomic features such as pathways as testing units to assess pathway-pathway interaction under the proposed framework. The mapping results can then be visualized by some network graphical tools such as the Cytospace software [143] which can help investigators generate important biological hypotheses for further lab validation.

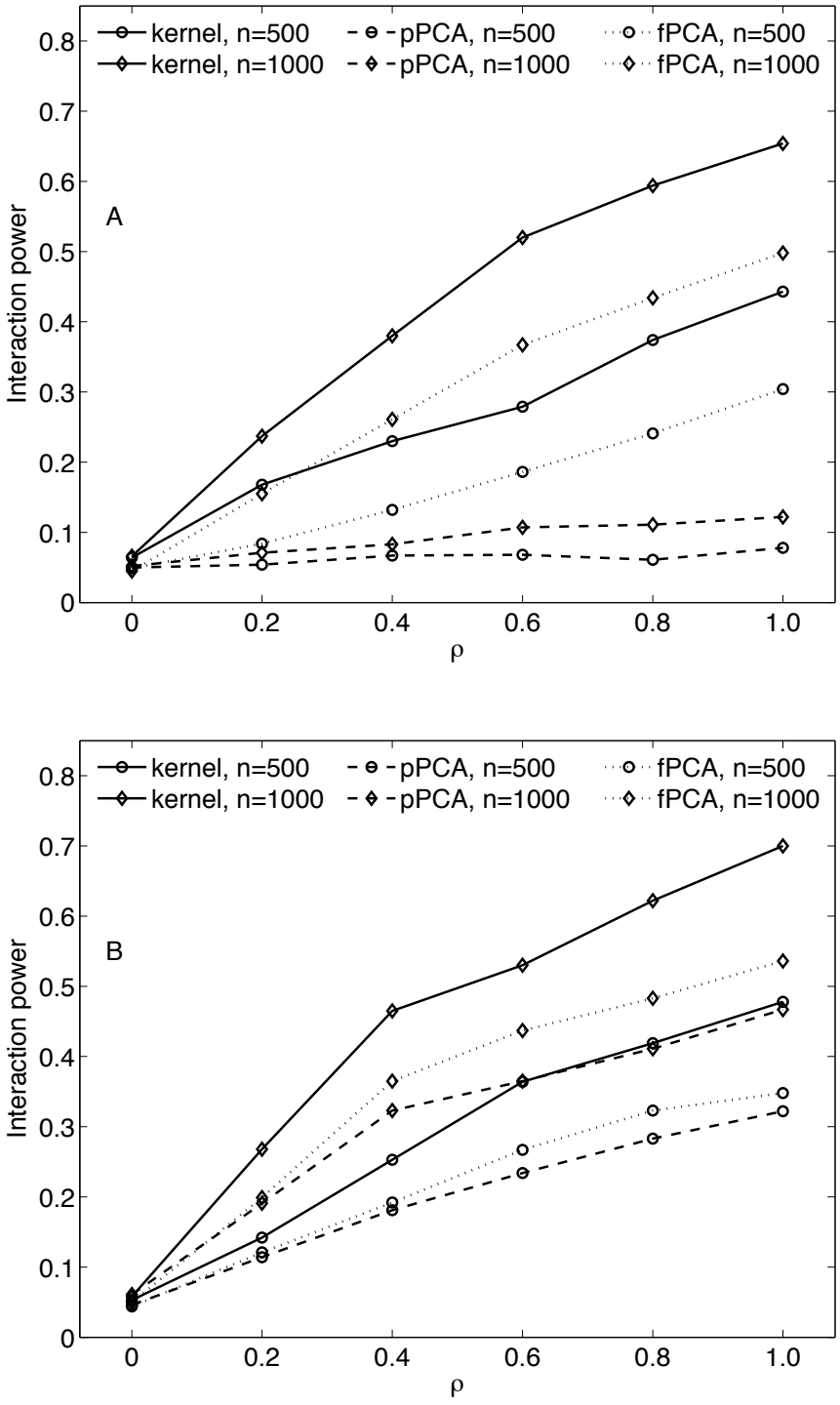
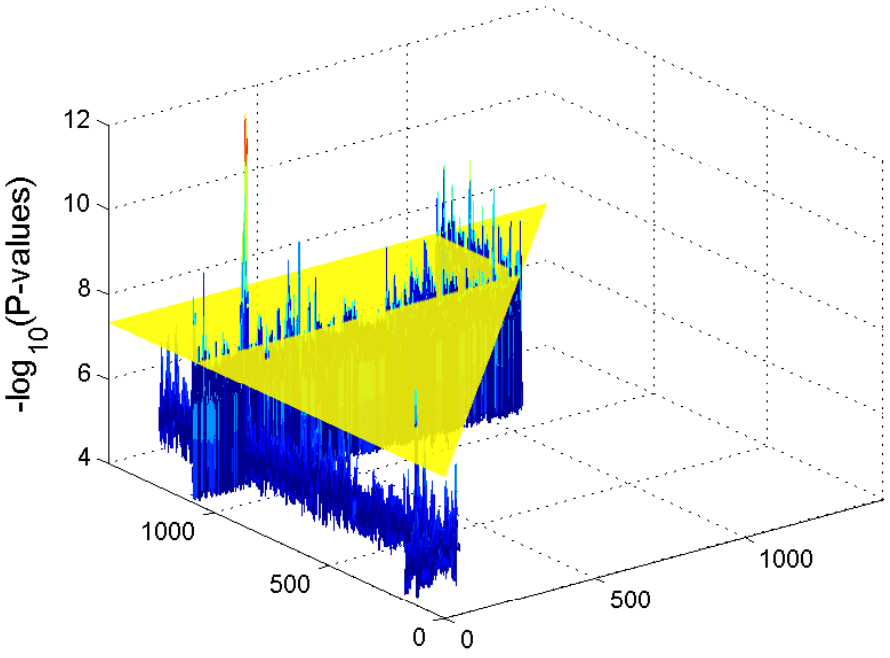
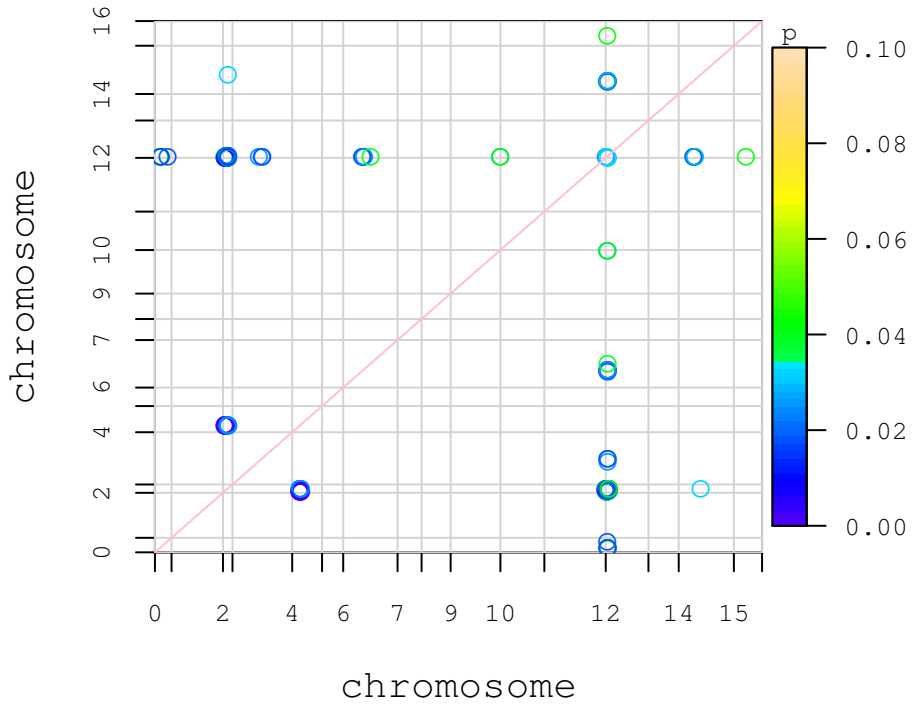


Figure 4.1: Power comparison of the proposed kernel approach (solid line), the partial PCA-based interaction model (4.21) (dashed line, denoted as pPCA) and the full PCA-based interaction model (4.22) (dotted line, denoted as fPCA) under different sample sizes and different proportions (ρ) of epistasis variance. Genotypes were simulated with the MS program (A) and the LD-based algorithm (B).



A



B

Figure 4.2: The -log10 transformed p-value profile plot of all gene pairs for the overall test (A) and the interaction test (B). The yellow hyperplane in A represents the Bonferroni cutoff.

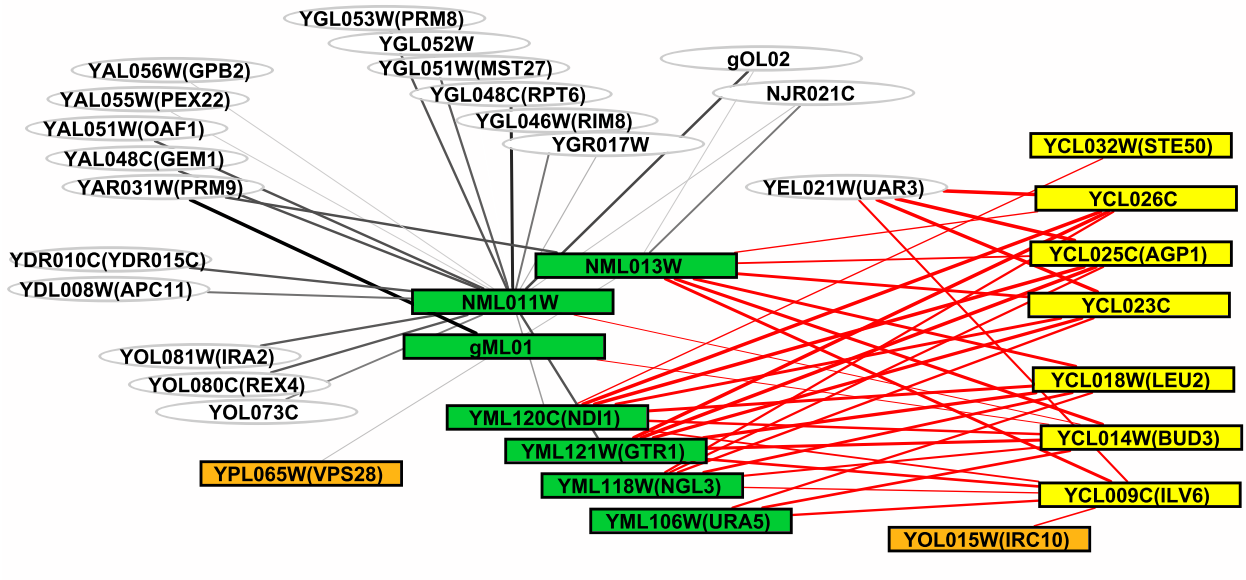


Figure 4.3: The network graph of interacting genes generated with Cytoscape (Shannon et al. 2003). The thickness of the connection line indicates the strength of the interaction. Nodes with light oval shapes indicate no marginal effect.

Chapter 5

An extension of the kernel-based gene-centric gene×gene interaction to binary phenotypes

5.1 Introduction

In chapter 4, we have illustrated the importance of incorporating gene×gene interactions in association studies. We proposed a model-based kernel machine method to detect gene×gene interaction for continuous quantitative trait from a gene-centric point of view. Simulation and real data studies indicate the promise and utility of the method as a statistical tool, which can be applied to various studies including eQTL mapping and disease association studies. However, in reality, many research problems have dichotomous phenotypic traits of interest, such as, human diseases, which are normally classified as diseased and healthy status. Gene×gene interaction is ubiquitous and believed to be an important contributor

to the "missing" part of the heritability of complex human diseases as argued in chapter 4. The aim of this chapter is to extend the model-based kernel machine approach to binary phenotypic outcomes.

5.2 Methods

Statistical model Suppose we have binary disease outcomes of n unrelated subjects, denoted as $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^T$, in which $y_i = 1$ if the i^{th} subject is affected and $y_i = 0$ otherwise. Let $\boldsymbol{x}_i$ denote the genetic vector of a gene pair which can be partitioned as $\boldsymbol{x}_i = (\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)})$; $\boldsymbol{x}_i^{(j)}, j = 1, 2$ represents the genotypes of SNPs in the two genes, respectively. Let $E(y_i|\boldsymbol{x}_i) = \pi_i$, we model the relationship between the disease status and the gene pair with a smooth function m.

$$log(\frac{\pi_i}{1-\pi_i}) = m(\boldsymbol{x}_i) \tag{5.1}$$

By functional ANOVA decomposition [130, 133, 131], m can be decomposed as the sum of several components.

$$log(\frac{\pi_i}{1-\pi_i}) = \mu + m_1(\boldsymbol{x}_i^{(1)}) + m_2(\boldsymbol{x}_i^{(2)}) + m_{12}(\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}) \tag{5.2}$$

where m_1, m_2 model the main effects of the two genes and m_{12} captures the interaction between them. The above equation defines a logistic two-factor interaction model. Assume function $m \in \mathcal{H}$, then $m_1 \in \mathcal{H}^1, m_2 \in \mathcal{H}^2$ and $m_{12} \in \mathcal{H}^3$ are orthogonal components, where $\mathcal{H}, \mathcal{H}^l, l = 1, 2, 3$ are defined the same as in chapter 4. We estimate function m as the

maximizer of the following penalized log-likelihood function.

$$\mathcal{L}(\boldsymbol{y}|m(\boldsymbol{x}_i)) = \sum_{i=1}^{n}(y_i m(\boldsymbol{x}_i) - log(1 + exp(m(\boldsymbol{x}_i))) - \frac{1}{2}\sum_{l=1}^{3}\lambda_l \|P^l m\|^2_{\mathcal{H}} \tag{5.3}$$

Function $m \in \mathcal{H}$ which maximizes (5.3) is known to have the following representation [130, 133]

$$m(\boldsymbol{x}) = \mu\mathbf{1} + \sum_{i=1}^{n} c_i \sum_{l=1}^{3}\theta_l k_l(\boldsymbol{x}_i, \boldsymbol{x}) \tag{5.4}$$

Plug in the representation of the maximizer into the penalized log-likelihood function, we have

$$\mathcal{L}(\boldsymbol{y}|\mu, C_l) = \sum_{i=1}^{n}(y_i(\mu + \sum_{l=1}^{3} K_l^T(\boldsymbol{x}_i)C_l) - log(1 + e^{(\mu+\sum_{l=1}^{3} K_l^T(\boldsymbol{x}_i)C_l)})) + \frac{1}{2}\sum_{l=1}^{3}\lambda_l C_l^T \mathbf{K}_l C_l \tag{5.5}$$

where $K_l(\boldsymbol{x})$, C_l and $\mathbf{K}$ are defined as in (4.6) and (4.7).

We show that the penalized log-likelihood function is identical to the penalized quasi-likelihood function of the following logistic mixed effects model for $\tilde{\pi} = E(\boldsymbol{y}|\tilde{m}_1, \tilde{m}_2, \tilde{m}_{12})$.

$$logit(\tilde{\pi}) = \mu\mathbf{1} + \tilde{m}_1 + \tilde{m}_2 + \tilde{m}_{12} \tag{5.6}$$

with independent random effects $\tilde{m}_1 \sim N(0, \tau_1^2\mathbf{K}_1)$, $\tilde{m}_2 \sim N(0, \tau_2^2\mathbf{K}_2)$ and $\tilde{m}_{12} \sim N(0, \tau_3^2\mathbf{K}_3)$.

The penalized quasi-likelihood function [148] of model (5.6) is

$$L = \sum_{i=1}^{n}(y_i(\mu + \sum_{l=1}^{3}\tilde{m}_{l,i}) - log(1 + e^{\mu+\sum_{l=1}^{3}\tilde{m}_{l,i}}) - \frac{1}{2}\sum_{l=1}^{3}\frac{1}{\tau_l^2}\tilde{m}_l^T\mathbf{K}_l^{-1}\tilde{m}_l \tag{5.7}$$

where $\tilde{m}_l = (\tilde{m}_l(\boldsymbol{x}_1), \tilde{m}_l(\boldsymbol{x}_2), \cdots, \tilde{m}_l(\boldsymbol{x}_n))^T$ and $\tilde{m}_{l,i} = \tilde{m}_l(\boldsymbol{x}_i)$. Letting $\tau_l^2 = 1/\lambda_l$, $\tilde{m}_l =$

$K_l^T C_l$, then

$$L = \sum_{i=1}^{n}(y_i(\mu + \sum_{l=1}^{3} K_l^T(\boldsymbol{x}_i)C_l) - log(1 + e^{\mu+\sum_{l=1}^{3} K_l^T(\boldsymbol{x}_i)C_l}) - \frac{1}{2}\sum_{l=1}^{3}\lambda_l C_l^T \mathbf{K}_l C_l \qquad (5.8)$$

The above expression is identical to the penalized log-likelihood function given in (5.5). The identity indicates the connection between the two models.

By the dual representation theorem of the logistic mixed effect model, (μ, m_1, m_2, m_{12}) and the penalty parameters $(\lambda_\ell, \ell = 1, 2, 3)$ can be estimated by the BLUPs and REML estimates under the logistic mixed effect model framework.

Parameter estimation Take the first order derivatives of function (5.7) with respect to μ and C_l $(l = 1, 2, 3)$

$$\frac{\partial L}{\partial \mu} = \mathbf{1}^T(\mathbf{y} - \tilde{\pi}) \qquad (5.9)$$

$$\frac{\partial L}{\partial C_l} = \mathbf{K}_l(\boldsymbol{y} - \tilde{\pi}) - \lambda_l \mathbf{K}_l C_l, l = 1, 2, 3 \qquad (5.10)$$

Then the Hessian matrix H is

$$H = -\begin{bmatrix} \mathbf{1}^T D\mathbf{1} & \mathbf{1}^T D\mathbf{K}_1 & \mathbf{1}^T D\mathbf{K}_2 & \mathbf{1}^T D\mathbf{K}_3 \\ \mathbf{K}_1^T D\mathbf{1} & \mathbf{K}_1^T D\mathbf{K}_1 + \lambda_1\mathbf{K}_1 & \mathbf{K}_1^T D\mathbf{K}_2 & \mathbf{K}_1^T D\mathbf{K}_3 \\ \mathbf{K}_2^T D\mathbf{1} & \mathbf{K}_2^T D\mathbf{K}_1 & \mathbf{K}_2^T D\mathbf{K}_2 + \lambda_2\mathbf{K}_2 & \mathbf{K}_2^T D\mathbf{K}_3 \\ \mathbf{K}_3^T D\mathbf{1} & \mathbf{K}_3^T D\mathbf{K}_1 & \mathbf{K}_3^T D\mathbf{K}_2 & \mathbf{K}_3^T D\mathbf{K}_3 + \lambda_3\mathbf{K}_3 \end{bmatrix}$$

where matrix $D = diag\{\tilde{\pi}_1(1-\tilde{\pi}_1), \tilde{\pi}_2(1-\tilde{\pi}_2), \cdots \tilde{\pi}_n(1-\tilde{\pi}_n)\}$. Let $q = (\frac{\partial L}{\partial \mu}, \frac{\partial L}{\partial C_1}, \frac{\partial L}{\partial C_2}, \frac{\partial L}{\partial C_3})^T$ and $\alpha = (\mu, C_1, C_2, C_3)^T$, assume λ_l $(l = 1, 2, 3)$ are known, then α can be estimated by the

Newton-Raphson iteration as

$$\alpha^{(k+1)} = \alpha^{(k)} - (H^{(k)})^{-1} q^{(k)}$$

The α value at convergence is the BLUPs of the model (5.6) when the penalty parameters are known. In reality, we do not know the values for λ_l $(l = 1, 2, 3)$ and need to estimate them. Substitute the Hessian matrix and the first order derivatives into the Newton-Raphason iteration, we arrive at

$$\begin{aligned} H^{(k)}\alpha^{(k+1)} &= H^{(k)}\alpha^{(k)} - q^{(k)} \\ &= \begin{bmatrix} \mathbf{1}D^{(k)}(\mathbf{1}\mu^{(k)} + \sum_{l=1}^{3} \mathbf{K}_l C_l + (D^{(k)})^{-1}(\boldsymbol{y} - \tilde{\pi}^{(k)})) \\ \mathbf{K}_1 D^{(k)}(\mathbf{1}\mu^{(k)} + \sum_{l=1}^{3} \mathbf{K}_l C_l + (D^{(k)})^{-1}(\boldsymbol{y} - \tilde{\pi}^{(k)})) \\ \mathbf{K}_2 D^{(k)}(\mathbf{1}\mu^{(k)} + \sum_{l=1}^{3} \mathbf{K}_l C_l + (D^{(k)})^{-1}(\boldsymbol{y} - \tilde{\pi}^{(k)})) \\ \mathbf{K}_3 D^{(k)}(\mathbf{1}\mu^{(k)} + \sum_{l=1}^{3} \mathbf{K}_l C_l + (D^{(k)})^{-1}(\boldsymbol{y} - \tilde{\pi}^{(k)})) \end{bmatrix} \end{aligned} \tag{5.11}$$

$\tilde{\boldsymbol{y}} = \mathbf{1}\mu^{(k)} + \sum_{l=1}^{3} \mathbf{K}_l C_l + (D^{(k)})^{-1}(\boldsymbol{y} - \tilde{\pi}^{(k)})$ is known as the working response vector in the generalized linear model context. Breslow [148] proposed to estimate the variance components of a generalized mixed effect model by assuming normality for the working vector $\tilde{\boldsymbol{y}}$ at convergence. Then parameters $(\tau_1^2, \tau_2^2, \tau_3^2)$ are estimated as the REML estimates of variance components in a linear mixed effects model with the response variable $\tilde{\boldsymbol{y}}$ expressed as

$$\tilde{\boldsymbol{y}} = \mu\mathbf{1} + \tilde{m}_1 + \tilde{m}_2 + \tilde{m}_{12} + \epsilon \tag{5.12}$$

where the $\tilde{m}'s$ are the same independent effects in model (5.6) and $\epsilon \sim N(0, D^{-1})$. The detailed algorithm of the estimation procedure is summarized as in the following algorithm.

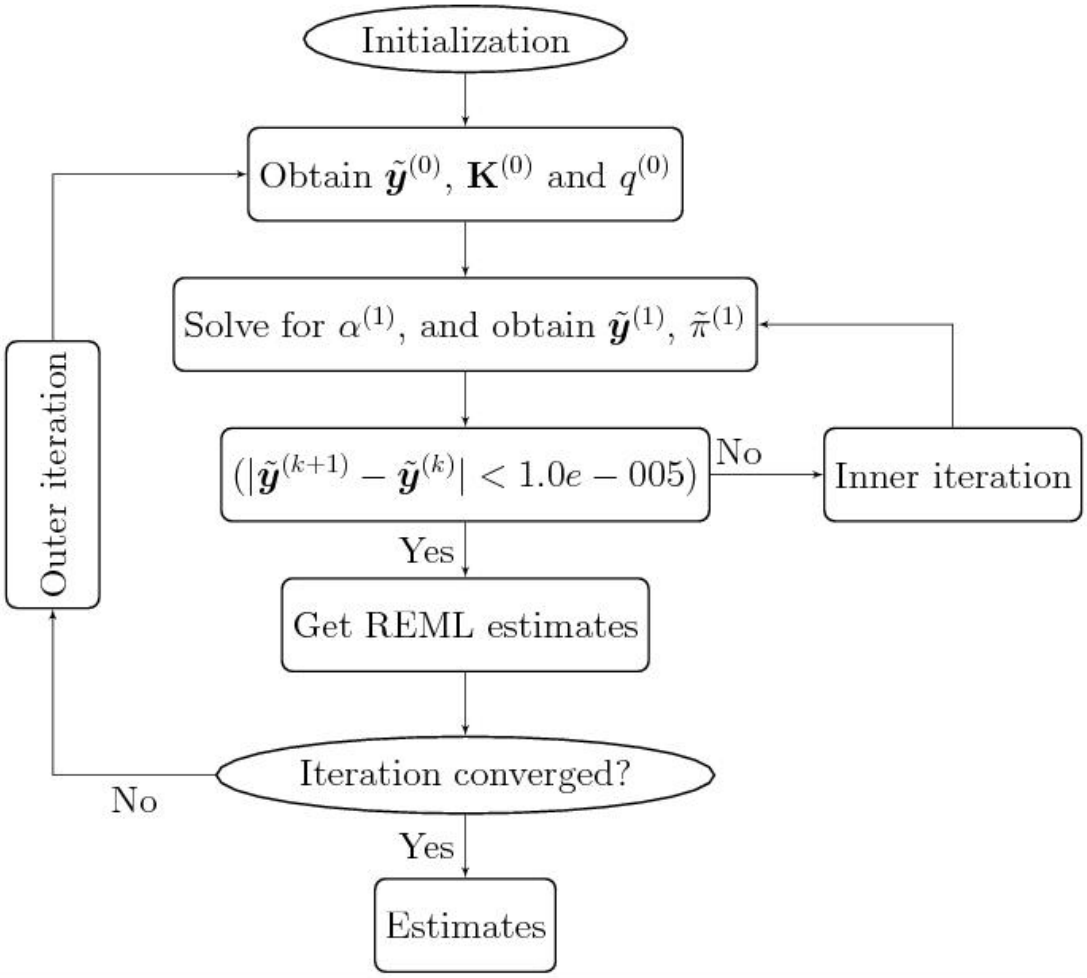


Figure 5.1: Work flow of the estimating algorithm

Figure 5.1 gives the work flow of the present estimating procedure.

- Step 0: Initialization: $(\tau_1^2, \tau_2^2, \tau_3^2) = (\tau_1^2, \tau_2^2, \tau_3^2)^{(0)}$, $\tilde{\pi} = \tilde{\pi}^{(0)}$
- Step 1: Calculate the values of $\tilde{\boldsymbol{y}}^{(0)} = logist(\tilde{\pi}^{(0)}) + (D^{(0)})^{-1}(\mathbf{y} - \tilde{\pi}^{(\mathbf{0})})$ and $H^{(0)}$ and $q^{(0)}$ by corresponding equations.
- Step 2: Obtain $\alpha^{(1)}$ by solving the system (5.11). Get the corresponding $\tilde{\boldsymbol{y}}^{(1)}$ and $\tilde{\pi}^{(1)}$.
- Step 3: Remain $(\tau_1^2, \tau_2^2, \tau_3^2)^{(0)}$ unchanged, iterate the process until convergence (e.g. $|\tilde{\boldsymbol{y}}^{(k+1)} - \tilde{\boldsymbol{y}}^{(k)}| < 1.0e-005$) and denote the value of $\tilde{\boldsymbol{y}}$ at convergence as $\tilde{\boldsymbol{y}}^{(c)}$.

- Step 4: Assume normality for $\tilde{\boldsymbol{y}}^{(c)}$ and solve for REML estimates of $(\tau_1^2, \tau_2^2, \tau_3^2)^{(1)}$ based on the approximate linear mixed effect model of (5.12).

- Step 5: Iterate step 1 through 4 until convergence (e.g. $|\tilde{\pi}^{(t+1)} - \tilde{\pi}^{(t)}| < 1.0e-005$).

- Step 6: The values of the parameters at convergence are used as estimates.

Hypothesis tests We constructed the following test statistics for the two hypotheses we are interested in. (1) Testing the overall genetic effect of a gene pair, i.e., $H_0 : m_1 = m_2 = m_{12} = 0$; and (2) testing the interaction effect between a gene pair, i.e., $H_0 : m_{12} = 0$, which are equivalent to the following hypotheses with the logistic mixed effect model representation: (I) $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$; and (II) $H_0 : \tau_3^2 = 0$. Based on the approximate linear mixed effects model for the working vector $\tilde{\boldsymbol{y}}^{(c)}$ at convergence [149], two score test statistics are obtained similarly as in chapter 4. The restricted log-likelihood function of the approximate linear mixed effects model (5.12) is

$$l_R = -\frac{1}{2}log(|V|) - \frac{1}{2}|\mathbf{1}^T V^{-1}\mathbf{1}| - \frac{1}{2}(\tilde{\boldsymbol{y}} - \mathbf{1}\hat{\mu})^T V^{-1}(\tilde{\boldsymbol{y}} - \mathbf{1}\hat{\mu}) \tag{5.13}$$

where $\hat{\mu}$ is the MLE of μ. We use the score test statistic to test the overall genetic effect of a pair of genes by

$$S_{binary} = \frac{1}{2}(\tilde{\boldsymbol{y}} - \mathbf{1}\hat{\mu})^T D \sum_{l=1}^{3} \mathbf{K}_l D(\tilde{\boldsymbol{y}} - \mathbf{1}\hat{\mu}) = (\boldsymbol{y} - \mathbf{1}\hat{\mu})^T \sum_{l=1}^{3} \mathbf{K}_l (\boldsymbol{y} - \mathbf{1}\hat{\mu}) \tag{5.14}$$

where $\hat{\mu}$ is the MLE of μ under the null hypothesis $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$.

The score test statistic for testing the interaction between the two genes is derived as

$$S_{binary,I} = \frac{1}{2}(\tilde{\boldsymbol{y}} - \mathbf{1}\hat{\mu})^T V_1^{-1}\mathbf{K}_3 V_1^{-1}(\tilde{\boldsymbol{y}} - \mathbf{1}\hat{\mu}) \tag{5.15}$$

where $\hat{\mu}$ is the MLE of μ under the null hypothesis $H_0 : \tau_3^2 = 0$, $\tilde{\boldsymbol{y}}$ is the working vector at convergence and $V_1 = D^{-1} + \tau_1^2\mathbf{K}_1 + \tau_2^2\mathbf{K}_2$ with τ_1^2, τ_2^2 and D estimated under the null hypothesis.

We approximate the distribution of both score statistics by a scaled χ^2 distribution. The scale parameter and degrees of freedom are estimated by the method of moments (see Appendix C).

To identify gene×gene interactions associated with a disease, a two-step strategy is implemented: (1) an exhaustive search for gene pairs with significant overall genetic effects; and (2) an interaction test at the position where there is significant overall genetic effect. A general concern about an exhaustive pair-wise screen is the potential demand of computation time. While, for the gene-centric approach, the total number of pairs to be tested has been dropped dramatically by treating multiple SNPs in a gene simultaneously as one testing unit. What's more, under the null hypothesis, i.e., $H_0 : \tau_1^2 = \tau_2^2 = \tau_3^2 = 0$, the logistic mixed effect model becomes a general logistic regression model. Hence, the genome-wide scan in the first step should not be time consuming. In practice, to avoid losing potential interaction effects, a relatively nonconservative cutoff value (e.g., $p - value < 0.1$) can be used in the pre-selection step. For those selected gene pairs (i.e., with the overall testing p-value< 0.1), we continue to test interaction effects.

The methodology development of the model-based kernel machine method for detecting gene×gene interactions for binary phenotypes has been completed. We have applied the

approach to a candidate gene association study and derived interesting findings. But the application doesn't fit in the theme of this dissertation and has been excluded.

Chapter 6

Conclusion and future work

6.1 Concluding remarks

The total number of human genes is estimated to be around 23,000. This number is not much greater than the numbers found in species look very different from humans, such as mouse and fruit fly (around 13,000), and even smaller than the number of genes in rice ($> 46,000$). Many wondered how human complexity could be explained by such few genes. It has become common knowledge that the complexity is not due the static number of genes but rather to the dynamic regulation of the transcription of these gene. Understanding mechanistic principles of gene regulation is important for understanding the functions of a living organism.

Data generated by genome research coupled with the recent advancements in microarray technology have made it possible to measure thousands of gene expression profiles simultaneously and genotype up to millions of genetic markers. The combination of traditional QTL mapping and the microarray technology, i.e., expression quantitative trait loci (eQTL)

mapping, has been a powerful paradigm in the past decade which holds great promise in elucidating the genetic architecture of gene expression as well as inferring gene regulation. Current single trait-single marker eQTL mapping studies have achieved valuable insights into gene regulation, for example, identifying *cis* and *trans* regulatory elements for genes across the genome. However, there are still many open questions in regard to how genes are being regulated.

It is postulated that a regulatory gene alters its own gene expression level and could consequently alters expression levels of relevant genes through cellular signaling pathways. And like most phenotypic traits, gene expression levels are multifactorial and complex genetically. It is very likely that genetic variants belonging to one functional group interact with each other in a complicated manner to affect transcript levels of genes. Hence, traditional approaches focusing on single gene analysis could have limited power and lead to results that are difficult to be interpreted biologically. In this dissertation, we proposed to integrate the biological pathway/gene set information into eQTL mapping studies. We considered two levels of pathway/gene set based analysis. One level is the pathway analysis across thousands of genes. Expression levels of genes in a pre-defined pathway are modeled as multivariate response to test the association with any given genomic locus. The study identifies regulators for the whole pathway, which are called pathway regulators. Based on the mapping results, we also detected regulation hotspots which regulate the transcription of genes in more pathways than by random. Another level is the marker set analysis across the whole genome. Genetic features were defined by grouping genetic markers in a pre-defined pathway or a region on the genome. Then association test is conducted between expression of a given gene or a gene set with the genetic feature. Association studies based on genetic

features is potentially more robust and could lead to biologically meaningful results. Statistical strategies for the two levels of pathway based eQTL mapping were described in chapter 2 and 3.

Chapter 4 and 5 of the dissertation introduced a statistical association approach for detecting gene×gene interactions. By treating genes as testing units, the testing dimension is reduced and the analysis may be more robust in the context of reproducibility. We modeled the relationship between genotype and phenotype by a smooth function. Functional ANOVA decomposed the function into additive main effects and interaction effect between a pair of genes. Score test statistics were constructed to test for genetic factors associated with the quantitative traits of interest. This model-based kernel machine method for detecting associated gene ×gene interactions was developed in chapter 4. The mapping method was applied to the yeast eQTL data to find epistases which control the transcription of genes. In chapter 5, we extended the model-based method to binary phenotypic outcomes.

Our overall aim in this dissertation is to develop novel statistical strategies for pathway-based eQTL mapping study, which incorporates prior biological pathway information into the eQTL mapping framework to disentangle gene regulations from the systems biology perspective. eQTL mapping analysis at pathway level can shed new lights into the functional interpretation of gene regulation. Besides, this dissertation presents a statistical gene-centric procedure for detecting gene×gene interactions underlying complex traits. It is our expectation that results obtained by these novel statistical strategies will help experimental scientists to generate informative biological hypothesis for further functional evaluation of any genes related to complex traits.

6.2 Future work

Research in eQTL mapping has making great progress in understanding the genetics of gene expression. However, substantial multidisciplinary efforts are still required including efficient statistical strategies for the analysis and interpretation of the analysis results in the functional context. Even though eQTL mapping only differs from the traditional QTL mapping in the number of phenotypes, the challenge lies on the multiplicity issue of multiple gene expression profiles and the correlations among them. Statistical approaches which model the functional relationship between gene expressions is thus biologically appealing. For example, the pathway-based eQTL mapping strategies proposed in this dissertation could provide additional biological insights by integrating additional gene set information into the eQTL mapping framework. However, genes as components in a pathway have their own particular responsibilities. For example, up-stream genes in the toll-like receptor signaling pathway recognize pathogens and pass the signal through down-stream genes in the pathway to activate immunity. How to statistically model the complicated functional relationship is a challenging and interesting research problem for future investigation. What's more, pathway information retrieved from public databases only represent the most conserved parts of pathways. Transcriptional pathway can be plastic; it depends on a large numbers of factors, such as environment, developmental stage and different tissues. Therefore, pathway regulators can be extremely context dependent. What biological information should be incorporated into a statistical model is another question that needs to be considered.

Recent achievements of the next generation sequencing (NGS) technology generates extremely large volume of data with an unprecedented speed. A Roche 454 machine can complete sequencing a genome in about 8 hours. The new technology presents both challenges

and opportunities in turning the massive NGS data into biological information. Sophisticated statistical methodologies and strategies are in demand for base-calling, sequence alignment and the down-stream functional interpretation analysis. The parallel RNA-sequencing (RNA-seq) allows accurate measurement of transcript complexity, including rare and common transcripts, novel gene structure, alternative splicing (AS) and allele-specific expression (ASE). To use RNA-seq data in the context of eQTL mapping calls for novel statistical method due to the new characteristics of the RNA-seq data. We are especially interested in developing efficient statistical mapping approaches which combines the AS and ASE information to infer eQTL and therefore derive novel insights of gene regulation, such as regulators of transcriptional, cotranscriptional and post-transcriptional levels, respectively.

The DNA-sequencing (DNA-seq) technology is able to provide an entire spectrum of genetic variations, including a large proportion of low frequency polymorphisms (rare variants). Genome-wide association study (GWAS) has been the primary approach for detecting genetic variants associated with complex diseases. And hundreds of related genes have been identified for human diseases in the past decade. However, researchers started realizing that only a small proportion of heritability is explained by those genes. Rare variants is thought to be a potential source that contributes to the missing heritability. Traditional association methods based on common variants common disease hypothesis have little power in detecting rare variants due to the low appearance frequency and modest effect size. Novel statistical approaches are needed to detect the association between rare variants and complex traits. Recently, several association tests for rare variants based on grouping and collapsing were proposed [150, 151, 152]. The common idea of these approaches is to pool rare variants within a given genomic region to investigate the cumulative evidence of association of the

region. The current statistical approaches for rare variants detection are intuitive and easy to implement, but still have limitations because of the simplified biological background between the genetic variants, for example, linkage disequilibrium, functional correlation and uneven effect size. Besides, systematic errors could also cause high false positive rate. More work is required to refine the various approaches for association study with rare variants and ultimately determine their properties and performance under different scenarios.

The development of efficient and useful statistical methods for eQTL mapping studies and functional interpretation of analysis results are two interwoven issues. Regulators identified by statistical approaches can be distinguished between true positives and false positives by looking for consistent supports from a different data source. Conversely, richer biological knowledge helps to develop more suitable statistical models. With the support of highly improving biotechnologies, a multi-field integrative analysis which combines the advantages of different sources, for example, DNA-Seq, RAN-Seq and ChIP-Seq could be a direction of future work. The importance of integrating multiple sources of data sets has been recognized. The marriage of eQTL mapping and gene networks has led to the oriented gene regulatory network, which provides valuable information about gene regulation [153]. We anticipate the analysis from multiple data sources and in an integrative manner is the path that could lead us to a full understanding of life functions.

APPENDIX

Appendix A

Supplementary materials for chapter 2

Table A.1: List of KEGG pathways and their ID numbers

#	PID	Function
1	04010	MAPK signaling pathway
2	00460	Cyanoamino acid metabolism
3	00780	Biotin metabolism
4	00910	Nitrogen metabolism
5	00280	Valine, leucine and isoleucine degradation
6	00410	beta-Alanine metabolism
7	00730	Thiamine metabolism
8	00230	Purine metabolism
9	00550	Peptidoglycan biosynthesis
10	00500	Starch and sucrose metabolism
11	00190	Oxidative phosphorylation
12	00640	Propanoate metabolism
13	03020	RNA polymerase
14	00960	Alkaloid biosynthesis II
15	00051	Fructose and mannose metabolism
16	00052	Galactose metabolism
17	03022	Basal transcription factors
18	00053	Ascorbate and aldarate metabolism
19	04070	Phosphatidylinositol signaling system
20	00290	Valine, leucine and isoleucine biosynthesis
21	00740	Riboflavin metabolism
22	00240	Pyrimidine metabolism
23	00380	Tryptophan metabolism
24	00510	N-Glycan biosynthesis
25	00010	Glycolysis / Gluconeogenesis

Table A.1: List of KEGG pathways and their ID numbers (cont'd)

#	PID	Function
26	00330	Arginine and proline metabolism
27	00650	Butanoate metabolism
28	03030	DNA replication
29	00970	Aminoacyl-tRNA biosynthesis
30	00600	Sphingolipid metabolism
31	00790	Folate biosynthesis
32	00920	Sulfur metabolism
33	00100	Biosynthesis of steroids
34	04111	Cell cycle - yeast
35	00561	Glycerolipid metabolism
36	00061	Fatty acid biosynthesis
37	00562	Inositol phosphate metabolism
38	00563	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
39	00513	High-mannose type N-glycan biosynthesis
40	00564	Glycerophospholipid metabolism
41	00565	Ether lipid metabolism
42	04120	Ubiquitin mediated proteolysis
43	00520	Nucleotide sugars metabolism
44	00020	Citrate cycle (TCA cycle)
45	00340	Histidine metabolism
46	00980	Metabolism of xenobiotics by cytochrome P450
47	00480	Glutathione metabolism
48	00430	Taurine and hypotaurine metabolism
49	00750	Vitamin B6 metabolism
50	00071	Fatty acid metabolism

Table A.1: List of KEGG pathways and their ID numbers (cont'd)

#	PID	Function
51	00521	Streptomycin biosynthesis
52	00251	Glutamate metabolism
53	00072	Synthesis and degradation of ketone bodies
54	00252	Alanine and aspartate metabolism
55	04130	SNARE interactions in vesicular transport
56	00030	Pentose phosphate pathway
57	00350	Tyrosine metabolism
58	00670	One carbon pool by folate
59	03050	Proteasome
60	02010	ABC transporters - General
61	00300	Lysine biosynthesis
62	00620	Pyruvate metabolism
63	00120	Bile acid biosynthesis
64	00440	Aminophosphonate metabolism
65	00760	Nicotinate and nicotinamide metabolism
66	00260	Glycine, serine and threonine metabolism
67	00710	Carbon fixation
68	00530	Aminosugars metabolism
69	00624	1- and 2-Methylnaphthalene degradation
70	00625	Tetrachloroethene degradation
71	00627	1,4-Dichlorobenzene degradation
72	04140	Regulation of autophagy
73	00310	Lysine degradation
74	03010	Ribosome
75	00630	Glyoxylate and dicarboxylate metabolism

Table A.1: List of KEGG pathways and their ID numbers (cont'd)

#	PID	Function
76	00130	Ubiquinone biosynthesis
77	00450	Selenoamino acid metabolism
78	00770	Pantothenate and CoA biosynthesis
79	00900	Terpenoid biosynthesis
80	00400	Phenylalanine, tyrosine and tryptophan biosynthesis
81	00590	Arachidonic acid metabolism
82	00720	Reductive carboxylate cycle (CO2 fixation)
83	00220	Urea cycle and metabolism of amino groups
84	00860	Porphyrin and chlorophyll metabolism
85	00040	Pentose and glucuronate interconversions
86	00360	Phenylalanine metabolism
87	01030	Glycan structures - biosynthesis 1
88	00680	Methane metabolism
89	03060	Protein export
90	02021	Two-component system - Organism-specific
91	00271	Methionine metabolism
92	00401	Novobiocin biosynthesis
93	00361	gamma-Hexachlorocyclohexane degradation
94	01031	Glycan structures - biosynthesis 2
95	00632	Benzoate degradation via CoA ligation
96	00272	Cysteine metabolism
97	00362	Benzoate degradation via hydroxylation
98	01032	Glycan structures - degradation
99	00903	Limonene and pinene degradation

Table A.2: Detailed list of hotspot regulations

Chr	Hotspot(Gene)	Start	Stop	Regulated EPs
2	YBR131W(CCZ1)	499012	499012	“03020” “00010” “00600” “00790” “00100” “00020” “00300” “00620” “00310” “00900” “00720” “00220”
2	YBR132C(AGP2)	499889	499895	“03020” “00740” “00010” “00600” “00790” “00100” “00300” “00620” “00310” “00900” “00400” “00590” “00220”
2	gBR07	506661	508843	“00500” “03020” “00740” “00010” “00790” “00100” “04111” “00020” “00300” “00620” “00900” “00400” “00590” “03060”
2	YBR139W	516889	517123	“00500” “04070” “00740” “00790” “00100” “04111” “00564” “00620” “00760” “00900” “00400” “00220”
2	gBR08	519049	521415	”00500” ”03020” ”00740” ”00790” ”00100” ”00564” ”00020” ”00620” ”00900” ”00400” ”00220” ”03060”
2	YBR142W(MAK5)	530481	530481	“00500” “03020” “00740” “00970” “00600” “00790” “00100” “00020” “00620” “00450” “00900” “00400” “00590” “00220” “03060”
2	YBR147W	537314	537314	“00500” “03020” “00740” “00510” “00970” “00600” “00790” “00100” “04111” “00564” “00020” “00620” “00310” “00450” “00900” “00400” “00220” “03060” “00361”
2	YBR154C(RPB5)	548401	548401	“04010” “00500” “03020” “04070” “00510” “00970” “00600” “00790” “00100” “04111” “00562” “00513” “00564” “00020” “00620” “00760” “00450” “00900” “00400” “00590” “00220” “01030” “03060”
2	YBR156C(SLI15) NBR031C NBR034W NBR035W	551299 553812 555575 555778	551299 553812 555596 555787	“00500” “00640” “03020” “04070” “00740” “00510” “00970” “00600” “00790” “00100” “04111” “00562” “00513” “00564” “00020” “00030” “00620” “00760” “00450” “00900” “00400” “00590” “00220” “01030” ”03060”

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
2	YBR161W	562409	562415	"00500" "03020" "00960" "00650" "00600" "00790" "00100" "04111" "00513" "00564" "00020" "00620" "00530" "00450" "00900" "00400" "00220" "00360" "03060"
2	YBR162W(YSY6)	565216	565216	"00500" "03020" "00960" "00650" "00600" "00790" "00100" "04111" "00564" "00030" "00620" "00530" "00450" "00900" "00400" "00220" "00360"
2	YBR163W(DEM1) YBR165W(UBS1)	567221 569414	567221 569414	"00500" "03020" "00960" "00650" "00600" "00790" "00100" "00564" "00030" "00620" "00900" "00400" "00220" "01030"
2	YBR165W(UBS1) YBR166C(TYR1)	569420 570229	569420 570229	"00500" "03020" "00650" "00600" "00790" "00100" "00564" "00030" "00620" "00900" "00400" "00220" "01030"
2	YBR172C(SMY2) YBR174C	579459 582419	579459 582419	"00500" "00960" "04070" "00650" "00600" "00790" "00100" "04111" "00562" "00564" "00900" "00400" "00220"
2	YBR176W(ECM31) NBR038W NBR041W	584351 592863 592989	584357 592863 592989	"00500" "00960" "00650" "00600" "00790" "0100" "00564" "00620" "00530" "00900" "00400" "00220"
3	YCL026C(FRM2)	75021	75021	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00330" "00650" "00750" "00670" "00630" "00770" "00680" "00401" "00903"
3	YCL025C(AGP1)	76127	76127	"00910" "00280" "00410" "00640" "00053" "00290" "00740" "00380" "00010" "00330" "00650" "00561" "00020" "00340" "00750" "00071" "00252" "00670" "00300" "00620" "00120" "00310" "00630" "00770" "00220" "00360" "00680" "00401" "00272" "00903"

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
3	YCL023C	79091	79091	"04010" "00910" "00280" "00410" "00640" "00053" "00290" "00740" "00380" "00010" "00330" "00650" "00600" "00920" "04111" "00561" "00563" "00564" "00020" "00340" "00980" "00750" "00071" "00251" "00252" "00350" "00670" "00300" "00620" "00120" "00624" "00310" "00630" "00450" "00770" "00720" "00220" "00360" "00680" "00401" "00272" "00903"
3	YCL022C gCL01 YCL018W(LEU2)	81832 90412 91977	81832 91496 92391	"04010" "00910" "00280" "00410" "00640" "00051" "00053" "00290" "00380" "00010" "00330" "00650" "00970" "00600" "00920" "04111" "00561" "00563" "00564" "00020" "00340" "00980" "00750" "00071" "00251" "00252" "00350" "00670" "00300" "00620" "00120" "00440" "00260" "00624" "00310" "00630" "00450" "00770" "00720" "00220" "00360" "00680" "00271" "00401" "00632" "00272" "00903"
3	YCL014W(BUD3)	100213	100213	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00010" "00330" "00650" "00600" "00920" "00561" "00563" "00564" "00020" "00340" "00980" "00750" "00071" "00251" "00252" "00350" "00670" "00300" "00620" "00120" "00260" "00310" "00630" "00450" "00770" "00220" "00680" "00401" "00272" "00903"
3	YCL009C(ILV6)	105042	105042	"04010" "00910" "00280" "00410" "00640" "00053" "00290" "00380" "00010" "00330" "00650" "00920" "00561" "00563" "00564" "00020" "00340" "00750" "00071" "00251" "00252" "00350" "00670" "00300" "00620" "00120" "00260" "00710" "00310" "00630" "00450" "00770" "00220" "00360" "00680" "00401" "00272" "00903" "00600"

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways					
3	NCR015C	175799	175808	"04010"	"00280"	"00410"	"00290"	"00380"	"00600"
	gCR02	177850	177850	"00563"	"00340"	"00350"	"00440"	"00630"	"00450"
				"00770"	"00903"				
5	gEL02	109310	109310						
	YEL021W(URA3)	116530	116830	"00740"	"00240"	"00510"	"00020"	"00251"	"00252"
	NEL011C	117046	117056	"00030"	"00620"	"00710"	"00630"	"00720"	"01030"
	YLR014C(PPR1)	117705	117705						
3	YLR236C	611967	611997	"00410"	"00380"	"00650"	"00600"	"00100"	"00513"
				"00071"	"00072"	"00620"	"00530"	"00627"	"00900"
				"00720"	"00361"	"00903"			
12	gLR07	634225	634226	"00280"	"00640"	"00380"	"00650"	"00600"	"00100"
				"00072"	"04130"	"00620"	"00530"	"00627"	"00900"
				"00720"	"01030"	"00361"	"00903"		
12	gLR07	634227	634227	"00280"	"00640"	"00380"	"00650"	"00600"	"00100"
				"00513"	"00072"	"04130"	"00620"	"00530"	"00627"
				"00310"	"00900"	"00720"	"01030"	"00361"	"00903"
12	gLR07	635380	635380	"00280"	"00410"	"00190"	"00640"	"00240"	"00380"
				"00010"	"00650"	"00600"	"00100"	"00564"	"00071"
				"00072"	"04130"	"00620"	"00530"	"00627"	"00310"
				"00900"	"00720"	"00860"	"01030"	"00361"	"00903"
12	YLR252W	642137	642137	"00280"	"00410"	"00190"	"00640"	"00380"	"00010"
				"00650"	"00600"	"00100"	"00513"	"00564"	"00071"
	YLR253W	644082	644136	"00072"	"04130"	"00620"	"00530"	"00627"	"00310"
				"00130"	"00900"	"00720"	"00860"	"01030"	"00361"
				"00903"					

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
12	YLR257W YLR258W(GSY2)	659357 662627	659357 662627	"00280" "00410" "00190" "00640" "00960" "00240" "00380" "00510" "00010" "00650" "00600" "00100" "00561" "00513" "00564" "00020" "00071" "00072" "00252" "04130" "00030" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00220" "00860" "01030" "03060" "00361" "00632" "00903"
12	YLR261C(VPS63) YLR263W(RED1)	668249 672779	668249 672785	"00280" "00410" "00190" "00640" "00240" "00380" "00510" "00010" "00650" "00600" "00100" "00513" "00564" "00020" "00071" "00251" "00072" "00252" "04130" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00220" "00860" "01030" "03060" "00271" "00361" "00903"
12	YLR265C(NEJ1)	674651	674651	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00020" "00071" "00251" "00072" "00252" "04130" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00860" "01030" "03060" "00361" "00903"
12	NLR116W	677957	677957	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00251" "00072" "00252" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00220" "00860" "01030" "00361" "00903"
12	YLR267W(BOP2) YLR269C YLR271W	679808 681096 683457	679808 681096 683463	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00251" "00072" "00252" "00620" "00120" "00530" "00627" "00310" "00130" "00900" "00720" "00220" "00860" "01030" "00361" "00903"

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
12	YLR273C(PIG1) YLR274W(CDC46)	689211 693610	689217 693616	"00280" "00410" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00251" "00072" "00620" "00530" "00627" "00310" "00130" "00900" "00720" "00860" "01030" "00271" "00361" "00903"
12	YLR274W(CDC46)	693790	693790	"00280" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00020" "00071" "00251" "00072" "00620" "00530" "00627" "00310" "00130" "00900" "00720" "00860" "01030" "00271" "00361" "00903"
12	YLR277C(YSH1)	697260	697260	"00280" "00190" "00640" "00380" "00010" "00650" "00600" "00100" "00513" "00564" "00071" "00251" "00072" "00620" "00530" "00627" "00130" "00900" "00720" "00860" "00271" "00361" "00903"
12	YLR281C YLR282C	704828 705088	704828 705220	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00564" "00071" "00072" "00252" "00620" "00627" "00900" "00720" "00271" "00361"
12	YLR282C	705226	705226	"00280" "00640" "00380" "00650" "00600" "00100" "00072" "00252" "00620" "00627" "00130" "00900" "00720" "00361"
12	YLR285W	708035	708041	"00280" "00640" "00380" "00010" "00650" "00600"
12	NLR118C	708258	708260	"00100" "00020" "00071" "00072" "00620" "00627"
12	YLR286C(CTS1)	708594	708594	"00130" "00900" "00720" "00361"
12	NLR121W	710924	710924	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00020" "00072" "00620" "00627" "00130" "00900" "00720" "00361" "00361"

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
12	YLR288C(MEC3)	713638	713644	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00564" "00071" "00251" "00072" "00252" "00620" "00627" "00130" "00900" "00720" "00271" "00361"
12	YLR288C(MEC3)	713686	713686	"00280" "00640" "00380" "00010" "00650" "00600" "00100" "00071" "00072" "00620" "00627" "00130" "00900" "00720" "00361"
12	YLR292C(SEC72)	719857	719857	"00280" "00640" "00010" "00650" "00600" "00100" "00072" "00627" "00450" "00900" "00720" "00271" "00361"
13	YML120C(NDI1)	28622	28694	"00280" "00410" "00640" "00053" "00290" "00380" "00561" "00071" "00710" "00770" "00220" "00272" "00903"
13	NML013W	46070	46084	"00280" "00410" "00640" "00053" "04070" "00290" "00380" "00330" "00650" "00970" "00561" "00020" "00071" "00251" "00670" "00120" "00260" "00710" "00310" "00770" "00220" "00272" "00903"
13	NML011W	49894	49903	"00910" "00280" "00410" "00500" "00640" "00053" "04070" "00290" "00380" "00010" "00330" "00650" "00970" "00561" "00020" "00071" "00251" "00252" "00670" "00620" "00120" "00260" "00710" "00310" "00770" "00720" "00220" "00272" "00903"
13	YML108w YML106W(URA5)	54913 57145	54913 57145	"00280" "00410" "00640" "00053" "00290" "00380" "00650" "00970" "00561" "00020" "00071" "00251" "00670" "00120" "00260" "00710" "00770" "00220" "00272" "00903"

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
13	gML01	64970	69122	"00280" "00410" "00640" "00053" "00290" "00380" "00650" "00970" "00561" "00071" "00251" "00252" "00670" "00120" "00260" "00310" "00770" "00220" "00040" "00272" "00903"
13	YML098W(TAF13)	77684	77684	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00251" "00670" "00120" "00310" "00770" "00220" "00272" "00903"
13	YML097C(VPS9)	78655	78655	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00120" "00770" "00220" "00272" "00903"
13	NML009C YML096W	79760 81250	79786 81358	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00710" "00310" "00770" "00220" "00272" "00903"
13	YML091C(RPM2)	87587	87587	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00260" "00710" "00310" "00770" "00220" "00360" "00272" "00903"
13	YML086C(ALO1)	96015	96015	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00260" "00710" "00310" "00770" "00220" "00360" "00272" "00903"
13	YML084W	99585	99720	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00310" "00770" "00360" "00272" "00903"

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
13	YML083C	100048	100048	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00650" "00561" "00071" "00670" "00620" "00120" "00710" "00310" "00770" "00220" "00360" "00272" "00903"
13	YML072C YML071C(DOR1)	124876 129925	124876 130069	"00910" "00280" "00410" "00640" "00053" "00290" "00380" "00010" "00650" "00561" "00340" "00071" "00670" "00620" "00120" "00710" "00220" "00272" "00903"
13	YML061C(PIF1)	149075	149075	"00280" "00410" "00640" "00053" "00650" "00561" "00340" "00071" "00670" "00120" "00710" "00220" "00272" "00903"
14	NNL035W	449639	449639	"04010" "00410" "00500" "00190" "03020" "00960" "00051" "00052" "03022" "00053" "04070" "00290" "00240" "00380" "00510" "00010" "00330" "00650" "03030" "00970" "00600" "00790" "00061" "00562" "00563" "00564" "00565" "04120" "00520" "00340" "00480" "00750" "00071" "00521" "00251" "00252" "04130" "00030" "00350" "00670" "03050" "00300" "00620" "00120" "00440" "00760" "00260" "00710" "00625" "00310" "03010" "00450" "00400" "00590" "00220" "00860" "00040" "00360" "01030" "03060" "00271" "01031" "00632" "00272"

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
14	YNL074C(MLF3)	486861	486861	"04010" "00500" "00190" "00640" "03020" "03022" "04070" "00290" "00240" "00380" "00510" "00010" "00330" "00650" "03030" "00970" "00790" "00920" "00061" "00562" "00563" "00513" "04120" "00020" "00480" "00750" "00521" "00251" "00252" "04130" "00030" "00350" "00670" "03050" "00300" "00620" "00440" "00760" "00260" "00710" "04140" "00310" "00450" "00400" "00590" "00220" "00860" "01030" "00680" "03060" "00271" "01031" "00632" "00272"
14	YNL066W(SUN4)	502316	502316	"03022" "04070" "00290" "00330" "00970" "00790" "00920" "00562" "00563" "00750" "00251" "00252" "00300" "00620" "00440" "00760" "00260" "00310" "00450" "00400" "00590" "00220" "00860" "00271" "00632" "00272"
14	gNL07	525061	525064	"03022" "04070" "00290" "00330" "00970" "00562" "00563" "00750" "00251" "00252" "00300" "00620" "00760" "00260" "00630" "00400" "00590" "00220" "00860" "00271"
15	YOL094C(RFC4)	141621	141633	"04010" "00052" "04111" "00561" "00513" "00564" "00565" "00980" "00071" "00030" "00350" "00120" "00710" "00624" "04140" "00130" "00040"
15	YOL093W YOL092W YOL089C(HAL9)	143597 144659 150651	143597 144959 150651	"00561" "00513" "00565" "00980" "00071" "00251" "00252" "00030" "00350" "00120" "00710" "00624" "00625" "04140" "00040"
15	YOL088C(MPD2)	154177	154309	"04010" "00561" "00513" "00564" "00565" "00520" "00980" "00071" "00251" "00350" "00120" "00710" "00624" "00625" "04140" "00040"

Table A.2: Detailed list of hotspot regulations (cont'd)

Chr	Hotspots(Gene)	Start	Stop	Regulated pathways
15	gOL02	170945	174364	"04010" "00910" "00280" "00410" "00500" "00640" "00053" "00380" "00010" "00790" "00920" "00561" "00513" "00564" "00565" "00520" "00340" "00980" "00430" "00750" "00071" "00521" "00251" "00252" "00030" "00350" "00670" "00120" "00760" "00260" "00710" "00624" "00625" "00590" "00860" "00040" "00360" "00680" "00272" "00362" "00903"
15	gOL02 YOL081W(IRA2) YOL080C(REX4)	175594 180180 180961	179289 180222 180961	"04010" "00910" "00280" "00410" "00640" "00053" "00380" "00010" "00790" "00920" "00561" "00513" "00564" "00565" "00520" "00340" "00980" "00430" "00750" "00071" "00521" "00251" "00252" "00030" "00350" "00670" "00120" "00760" "00260" "00710" "00530" "00624" "00625" "04140" "00310" "00590" "00220" "00860" "00040" "00360" "00680" "00271" "00272" "00362" "00903"
15	YOR127W(RGA1)	563943	563943	"00052" "00380" "00010" "00561" "00020" "00980" "00071" "00350" "03050" "00120" "00710" "00624" "00130"

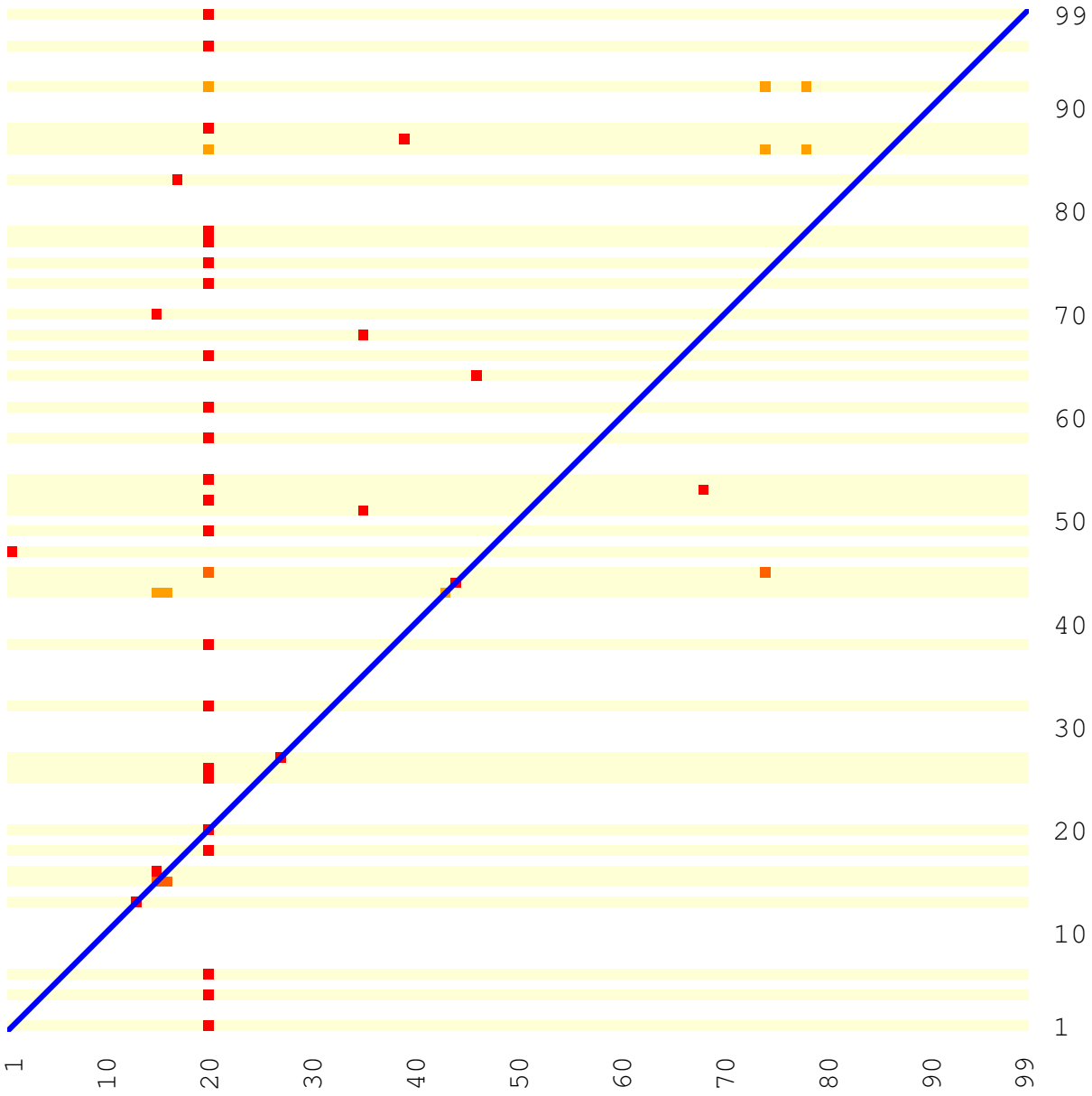


Figure A.1: A heatmap of enriched pathways. Only significantly enriched pathways are shown in the plot (indicated by squares). The darker the color of each square, the smaller the enrichment p-value and hence the strong the association. Squares on the diagonal line indicate *cis*-pathway regulation and those on off-diagonals indicate *trans*-regulation. The horizontal and vertical axes denote the genetic pathway (GP) and the gene expression pathway (EP), respectively. Strong *trans*-pathway regulations are detected.

Appendix B

Smoothing spline ANOVA decomposition and the dual representation

B.1 SS-ANOVA decomposition

Suppose a function f is on domain $\Gamma = \Gamma_1 \otimes \Gamma_2 \otimes \Gamma_3$. Define corresponding averaging operator A_γ on each generic domain $\Gamma_\gamma, \gamma = 1, 2, 3$. An ANOVA decomposition of function f can be obtained:

$$\begin{aligned} f &= \{\prod_{\gamma=1}^{3}(I - A_\gamma + A_\gamma)\}f \\ &= (I - A_1)(I - A_2)(I - A_3) + (I - A_1)(I - A_2)A_3 + (I - A_1)A_2(I - A_3) \\ &+ A_1(I - A_2)(I - A_3) + (I - A_1)A_2A_3 + A_1A_2(I - A_3) + A_1(I - A_2)A_3 + A_1A_2A_3 \end{aligned}$$

For a nested domain $(\Gamma_1 \otimes \Gamma_2) \otimes \Gamma_3$, let A_{12} be the averaging operator on domain $(\Gamma_1 \otimes \Gamma_2)$. Then the ANOVA decomposition becomes

$$f = \{(I - A_{12})(I - A_3) + A_{12}(I - A_3) + (I - A_{12})A_3 + A_{12}A_3\}f$$

Since

$$(I - A_1)(I - A_2) + (I - A_1)A_2 + A_1(I - A_2) = I - A_1A_2$$

By letting $A_{12} = A_1A_2$,

$$\{(I - A_{12})(I - A_3) + A_{12}(I - A_3) + (I - A_{12})A_3 + A_{12}A_3\}f = \{\prod_{\gamma=1}^{3}(I - A_\gamma + A_\gamma)\}f$$

Recursively, it shows that the ANOVA decomposition can also be conducted on product of nested domains.

B.2 The dual representation

Consider the linear mixed effect model

$$y = \mu\mathbf{1} + \tilde{m}_1 + \tilde{m}_2 + \tilde{m}_{12} + \epsilon$$

with $\tilde{m}_1, \tilde{m}_2, \tilde{m}_{12}$ are independent $n \times 1$ vector of random effects; $\tilde{m}_1 \sim N(\mathbf{0}, \tau_1^2\mathbf{K}_1)$, $\tilde{m}_2 \sim N(\mathbf{0}, \tau_2^2\mathbf{K}_2)$, $\tilde{m}_{12} \sim N(\mathbf{0}, \tau_3^2\mathbf{K}_3)$, and $\epsilon \sim N(0, \sigma^2 I)$ is independent of $\tilde{m}_1, \tilde{m}_2$ and $\tilde{m}_{12}$. The Henderson's normal equation for obtaining the BLUPs of the random effects is

$$\begin{bmatrix} n & \mathbf{1}^T & \mathbf{1}^T & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} + \frac{\sigma^2}{\tau_1^2}\mathbf{K}_1^{-1} & \mathbf{I} & \mathbf{I} \\ \mathbf{1} & \mathbf{I} & \mathbf{I} + \frac{\sigma^2}{\tau_2^2}\mathbf{K}_2^{-1} & \mathbf{I} \\ \mathbf{1} & \mathbf{I} & \mathbf{I} & \mathbf{I} + \frac{\sigma^2}{\tau_3^2}\mathbf{K}_3^{-1} \end{bmatrix} \begin{bmatrix} \mu \\ \tilde{m}_1 \\ \tilde{m}_2 \\ \tilde{m}_{12} \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{I} \\ \mathbf{I} \\ \mathbf{I} \end{bmatrix} \boldsymbol{y} \tag{B.1}$$

It can be shown this normal equation is equivalent to the first order condition for estimating function m, equation (4.8). Multiply both sides of equation (4.8) by the following matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \mathbf{K}_1^{-1} & 0 & 0 \\ 0 & 0 & \mathbf{K}_2^{-1} & 0 \\ 0 & 0 & 0 & \mathbf{K}_3^{-1} \end{bmatrix}$$

then

$$\begin{bmatrix} n & \mathbf{1}^T\mathbf{K}_1 & \mathbf{1}^T\mathbf{K}_2 & \mathbf{1}^T\mathbf{K}_3 \\ \mathbf{1} & \mathbf{K}_1+\lambda_1\mathbf{I} & \mathbf{K}_2 & \mathbf{K}_3 \\ \mathbf{1} & \mathbf{K}_1 & \mathbf{K}_2+\lambda_2\mathbf{I} & \mathbf{K}_3 \\ \mathbf{1} & \mathbf{K}_1 & \mathbf{K}_2 & \mathbf{K}_3+\lambda_3\mathbf{I} \end{bmatrix} \begin{bmatrix} \mu \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{I} \\ \mathbf{I} \\ \mathbf{I} \end{bmatrix} \boldsymbol{y}$$

Letting $\tilde{m}_l = \mathbf{K}_l C_l, l = 1, 2$, $\tilde{m}_{12} = \mathbf{K}_3 C_3$ and $\tau_l^2 = \sigma^2/\lambda_l, l = 1, 2, 3$, the system is exactly the equation (B.1), which is the Henderson's normal equation of linear mixed effects model (4.9).

Appendix C

Scaled χ^2 approximation

We approximate the distribution of statistics S_{binary} and $S_{binary,I}$ with scaled χ^2 distributions.

$$S_{binary} = (\boldsymbol{y} - \mathbf{1}\hat{\mu})^T \sum_{l=1}^{3} \mathbf{K}_l (\boldsymbol{y} - \mathbf{1}\hat{\mu}) \sim a\chi_g^2 \tag{C.1}$$

$$S_{binary,I} = \frac{1}{2}(\tilde{\boldsymbol{y}} - \mathbf{1}\hat{\mu})^T V_1^{-1} \mathbf{K}_3 V_1^{-1} (\tilde{\boldsymbol{y}} - \mathbf{1}\hat{\mu}) \sim a_I \chi_{g_I}^2 \tag{C.2}$$

The parameters a, g, a_I and g_I are estimated by the method of moments.

$$E(S_{binary}) = \sum_{l=1}^{3} \frac{1}{2} tr(P_0 \mathbf{K}_l) = ag \tag{C.3}$$

$$Var(S_{binary}) = tr(\sum_{l=1}^{3}(P_0\mathbf{K}_l) \sum_{l=1}^{3}(P_0\mathbf{K}_l))/2 = 2a^2 g \tag{C.4}$$

Solve the equations,

$$\hat{a} = \frac{Var(S_{binary})}{2E(S_{binary})} = \frac{tr(\sum_{l=1}^{3}(P_0\mathbf{K}_l) \sum_{l=1}^{3}(P_0\mathbf{K}_l))}{2\sum_{l=1}^{3} tr(P_0\mathbf{K}_l)} \tag{C.5}$$

and

$$\hat{g} = \frac{(\sum_{l=1}^{3} tr(P_0\mathbf{K}_l))^2}{tr(\sum_{l=1}^{3}(P_0\mathbf{K}_l) \sum_{l=1}^{3}(P_0\mathbf{K}_l))} \tag{C.6}$$

Since τ_1^2, τ_2^2 in the statistic $S_{binary,I}$ are replaced with their MLE in practice, corresponding corrections (based on the efficient information) are needed when estimate the mean and variance of the statistic [154].

$$e = E(S_{binary,I}) = \frac{1}{2} tr(P_1 \mathbf{K}_3) = a_I g_I \tag{C.7}$$

$$I_{\tau\tau} = Var(S_{binary,I}) \approx \frac{1}{2} tr((P_1\mathbf{K}_3 P_1 \mathbf{K}_3)) - \frac{1}{2}\Phi^T \Delta^{-1} \Phi = 2a_I^2 g_I \tag{C.8}$$

where $\Phi = (tr(P_1\mathbf{K}_3P_1\mathbf{K}_1), tr(P_1\mathbf{K}_3P_1\mathbf{K}_2))^T$ and

$$\Delta = \begin{bmatrix} tr(P_1\mathbf{K}_1P_1\mathbf{K}_1) & tr(P_1\mathbf{K}_1P_1\mathbf{K}_2) \\ tr(P_1\mathbf{K}_2P_1\mathbf{K}_1) & tr(P_1\mathbf{K}_2P_1\mathbf{K}_2) \end{bmatrix} \tag{C.9}$$

Therefore,

$$\hat{a}_I = \frac{I_{\tau\tau}}{2e} \tag{C.10}$$

$$\hat{g}_I = \frac{2e^2}{I_{\tau\tau}} \tag{C.11}$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA, 1998.

[2] J. Watson and F. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[3] E. M. Southern. Detection of specific sequences among DNA fragments separated by gene electrophoresis. *J. Mol. Biol.*, 98:503–517, 1975.

[4] F. Sanger, S. Niiken, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.*, 74:5463–5468, 1980.

[5] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. *Science*, 230:1350–1354, 1985.

[6] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver. Life with 6000 genes. *Science*, 274(5287):546–563, 1996.

[7] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. The complete genome sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, 1997.

[8] J. C. Venter, M. D. Adams, and et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[9] J. Yu, S. Hu, J. Wang, and et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, 296(5565):79–92, 2002.

[10] J. Schmutz, S. B. Cannon, and et al. Genome sequence of the palaeopolyploid soybean. *Nature*, 463(7278):178–183, 2010.

[11] M. Soller and T. Brody. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theor. Appl. Genet.*, 47:35–39, 1976.

[12] M. D. Edwards, C. W. Stuber, and J. F. Wendel. Molecularmarker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics*, 116:113–125, 1987.

[13] J. S. Beckmann and M. Soller. Detection of linkage between marker loci and loci affecting quantitative traits in crosses between segregating populations. *Theor. Appl. Genet.*, 76:228236, 1988.

[14] Z. W. Luo and M. J. Kearsey. Maximum likelihood estimation of linkage between a marker gene and a quantitative trait locus. *Heredity*, 63:401–408, 1989.

[15] E. S. Lander and D. Botstein. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121:185–199, 1989.

[16] R. C. Jansen. A general mixture model for mapping quantitative trait loci by using molecular markers. *Theor. Appl. Genet.*, 85:252–260, 1992.

[17] Z.-B. Zeng. Theoretical basis of precision mapping of quantitative trait loci. *Proc. Natl. Acad. Sci. USA*, 90:10972–10976, 1993.

[18] Z.-B. Zeng. Precision mapping of quantitative trait loci. *Genetics*, 236:1457–1468, 1994.

[19] R. C. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135:205–211, 1993.

[20] R. C. Jansen and P. Stam. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136:1447–1455, 1994.

[21] C. H. Kao, Z.-B. Zeng, and R. D. Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152:1203–1216, 1999.

[22] S. Xu and N. Yi. Mixed model analysis of quantitative trait loci. *Proc. Natl Acad. Sci. USA*, 97:14542–14547, 2000.

[23] J. M. Satagopan, B. S. Yandell, M. A. Newton, and T. C. Osborn. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, 144:805–816, 1996.

[24] P. Unimari and I. Hoeschele. Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics*, 146:735–743, 1997.

[25] M. J. Sillanpää and E. Arjas. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics*, 15:1605–1619, 1999.

[26] H. P. Piepho and H. G. Gauch. Marker pair selection for mapping quantitative trait loci. *Genetics*, 157:433–444, 2001.

[27] K. W. Broman. Review of statistical methods for QTL mapping in experimental crosses. *Lab Anim. (NY)*, 30(7):44–52, 2001.

[28] C.-H. Kao and Z.-B. Zeng. General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics*, 53(2):653–665, 1997.

[29] Z.-B. Zeng, C.-H. Kao, and C. J. Basten. Estimating the genetic architecture of quantitative traits. *Genetical Research*, 74(3):279–289, 1999.

[30] S. Sen and G. A. Churchill. A statistical framework for quantitative trait mapping. *Genetics*, 159:371–387, 2001.

[31] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Sco.*, 57(1):289–300, 1995.

[32] J. D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):289–300, 2003.

[33] W. Zou and Z.-B. Zeng. Statistical methods for mapping multiple QTL. *Int. J. Plant. Genomics.*, 2008:286561, 2008.

[34] A. Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Stat. Methods. Med. Res.*, 17(4):347–388, 2008.

[35] J. Schuchhardt, D. Beule, A. Malik, E. Wolski, H. Eickhoff, H. Lehrach, and H. Herzel. Normalization strategies for cDNA microarrays. *Nucleic Acids Res.*, 28(10):E47, 2000.

[36] A. Hill, E. Brown, M. Whitley, G. Tucker-Kellogg, C. Hunter, and D. Slonim. Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.*, 2(12):research 0055.10055.13, 2001.

[37] M. Bilban, L. K. Buehler, S. Head, G. Desoye, and V. Quaranta. Normalizing DNA microarray data. *Curr. Issues Mol. Biol.*, 4(2):57–64, 2002.

[38] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[39] S. Oba, M. A. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

[40] X. Wang, A. Li, Z. Jiang, and H. Feng. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, 22:7–32, 2006.

[41] M. K. Choong, M. Charbit, and H. Yan. Autoregressive-model-based missing value estimation for DNA microarray time series data. *IEEE Trans. Inf. Technol. Biomed.*, 13(1):131–137, 2009.

[42] D.-U. Ramón and A. Sara. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3), 2006.

[43] R. C. Jansen and J. P. Nap. Genetical genomics: the added value from segregation. *Trends Genet.*, 17:388–391, 2001.

[44] Vivian G. Cheung, Laura K. Conlin, Teresa M. Weber, Melissa Arcaro, Kuang-Yu Jen, Michael Morley, and Richard S. Spielman. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.*, 33(3):422–425, 2003.

[45] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296:752–755, 2002.

[46] E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and

S. H. Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, 2003.

[47] R. B. Brem and L. Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA*, 102(5):1572–1577, 2005.

[48] C. Kendziorski and P. Wang. A review of statistical methods for expression quantitative trait loci mapping. *Mamm. Genome.*, 17(6):509–517, 2005.

[49] Y. Gilad, S. A. Rifkin, and J. K. Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, 24(8):408–415, 2008.

[50] J. Li and M. Burmeister. Genetical genomics: combining genetics with gene expression analysis. *Hum. Mol. Genet.*, 14(2):R163–169, 2005.

[51] G. Yvert, R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, R. Mackelprang, and L. Kruglyak. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.*, 35(1):57–64, 2003.

[52] E. Petretto, J. Mangion, M. Pravanec, N. Hubner, and T. J. Aitman. Integrated gene expression profiling and linkage analysis in the rat. *Mamm. Genome.*, 17(6):480–489, 2006.

[53] J. Wessel, M. A. Zapala, and N. J. Schork. Accommodating pathway information in expression quantitative trait locus analysis. *Genomics*, 90(1):132–142, 2007.

[54] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstràle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, 34(3):267–273, 2003.

[55] E. Lee, J. H. Woo, J. W. Park, and T. Park. Finding pathway regulators: gene set approach using peak identification algorithms. *BMC Proc.*, 1(Suppl 1):S90, 2007.

[56] C. Wu, D. L. Delano, N. Mitro, S. V. Su, J. Janes, P. McClurg, S. Batalov, G. L. Welch, J. Zhang, A. P. Orth, J. R. Walker, R. J. Glynne, M. P. Cooke, J. S. Takahashi, K. Shimomura, A. Kohsaka, J. Bass, E. Saez, T. Wiltshire, and A. I. Su. Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genet.*, 4(5):e1000070, 2008.

[57] W. Sun, S. Yuan, and K. Li. Trait-trait dynamic interaction: 2D-trait eQTL mapping for genetic variation study. *BMC Genomics*, 9:242, 2008.

[58] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis.* New Jersey: Prentice Hall, 2007.

[59] M. A. Zapala and N. J. Schork. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Natl. Acad. Sci. USA*, 103(51):19430–19435, 2006.

[60] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747, 2004.

[61] R. Breitling, Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. de Haan, A. I. Su, and R. C. Jansen. Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.*, 4(10):e1000232, 2008.

[62] K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, 81(6):1278–1283, 2007.

[63] S. Li, Q. Lu, and Y. Cui. A systems biology approach for identifying novel pathway regulators in eqtl mapping. *J. Biopharm. Stat.*, 20(2):373–400, 2010.

[64] E. O. Perlstein, D. M. Ruderfer, D. C. Roberts, S. L. Schreiber, and L. Kruglyak. Geneticbasis of individual differences in the response to small-moledule drugs in yeast. *Nat. Genet.*, 39(4):496–502, 2007.

[65] E. E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S. K. Sieberts, S. Monks, M. Reitman, C. Zhang, P. Y. Lum, A. Leonardson, R. Thieringer, J. M. Metzger, L. Yang, J. Castle, H. Zhu, S. F. Kash, T. A. Drake, A. Sachs, and A. J. Lusis. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, 37(7):710–717, 2005.

[66] A. Frary, T. C. Nesbitt, S. Grandillo, E. Knaap, B. Cong, J. Liu, J. Meller, R. Elber, K. B. Alpert, and S. D. Tanksley. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science*, 289(5476):85–88, 2000.

[67] C. Li, A. Zhou, and T. Sang. Rice domestication by reducing shattering. *Science*, 311(5769):1936–1939, 2006.

[68] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, 10(3):184–194, 2009.

[69] J. D. Storey, J. M. Akey, and L. Kruglyak. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.*, 3(8):e267, 2005.

[70] J. Dong and S. Horvath. Understanding network concepts in modules. *BMC Syst. Biol.*, 4:1–24, 2007.

[71] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, 32:277–280, 2004.

[72] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.*, 31(1):19–20, 2002.

[73] Klein T. E. Thorn, C. F. and R. B. Altman. PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. *Methods Mol. Biol.*, 311:179–191, 2005.

[74] G. Peng, L. Luo, H. Siu, Y. Zhu, P. Hu, J. Hong, S.and Zhao, X. Zhou, J. D. Reveille, L. Jin, C. I. Amos, and M. Xiong. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.*, 18(1):111–117, 2010.

[75] K. Yu, Q. Li, A. W. Bergen, R. M. Pfeiffer, P. S. Rosenberg, N. Caporaso, P. Kraft, and N. Chatterjee. Pathway analysis by adaptive combination of p-values. *Genet. Epidemiol.*, 33(8):700–709, 2009.

[76] A. Hess and H. Iyer. Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. *BMC Genomics.*, 8:96, 2007.

[77] R. A. Fisher. *Statistical methods for research workers.* London: Oliver and Boyd., 1932.

[78] D. V. Zaykin, L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir. Truncated product mehtod for combining P-values. *Genet. Epidemiol.*, 22(2):170–185, 2002.

[79] F. Dudbridge and B. P. Koeleman. Rank truncated product of p-values, with application to genomewide association scans. *Genet. Epidemiol.*, 25(4):360–366, 2003.

[80] O. De la Cruz, X. Wen, B. Ke, M. Song, and D. L. Nicolae. Gene, region and pathway level analyses in whole-genome studies. *Genet. Epidemiol.*, 34(3):222–231, 2010.

[81] J. T. Kost and M. P. McDermott. Combining dependent p-values. *Stat. Prob. Lett.*, 60(2):183–190, 2002.

[82] K. N. Conneely and M. Boehnke. So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *Am. J. Hum. Genet.*, 81(6):1158–1168, 2007.

[83] Brown M. 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, S1:987–992, 1975.

[84] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, 4:Article17, 2005.

[85] R. K. Mortimer and J. R. Johnston. Genealogy of principal strains of the yeast genetic stock center. *Genetics*, 13:35–43, 1986.

[86] J. A. Barnett. A history of research on yeasts 10: foundations of yeast genetics. *Yeast*, 24:799–845, 2007.

[87] Z. Gu, L. David, D. Petrov, T. Jones, R. W. Davis, and L. M. Steinmetz. Elevated evolutionary rates in the laboratory strain of Saccharomyces cerevisiae. *Proc, Natl. Acad. Sci. USA*, 102:1092–1097, 2005.

[88] *Methods in yeast genetics: A cold spring harbor laboratory course manual, publisher = CSHL Press, year = 2000, author = Burke, D. and Dawson, D. and Stearns, T.*

[89] J. M. Cherry, C. Ball, S. Weng, R. Juvik, G. amd Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R. K. Mortimer, and D. Botstein. Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387(6632 Suppl):67–73, 1997.

[90] C. Cullin, A. Baudin-Baillieu, E. Guillemet, and O. Ozier-Kalogeropoulos. Functional analysis of YCL09C: evidence for a role as the regulatory subunit of acetolactate synthase. *Yeast*, 12(15):1511–1518, 1996.

[91] J. Ronald and J. M. Akey. The evolution of gene expression QTL in Saccharomyces cerevisiae. *PLoS One*, 2(7):e678, 2007.

[92] Y. Wang and H. G. Dohlman. Pheromone signaling mechanisms in yeast: a prototypical sex machine. *Science*, 306(5701):1508–1509, 2004.

[93] M. Gaisne, Verdire J. Bcam, A. M., and C.J. Herbert. A "natural" mutation in Saccharomyces cerevisiae strains derived from S288c affects the complex regulatory gene HAP1 (CYP1). *Curr. Genet.*, 36(4):195–200, 1999.

[94] L. N. Dimitrov, R. B. Brem, L. Kruglyak, and D. E. Gottschling. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of Saccharomyces cerevisiae S288C strains. *Genetics*, 183(1):365–383, 2009.

[95] F. J. Foury. Cloning and sequencing of the nuclear gene MIP1 encoding the catalytic subunit of the yeast mitochondrial DNA polymerase. *J. Biol. Chem.*, 264(34):20552–20560, 1989.

[96] K. Pfeifer, K. S. Kim, S. Kogan, and L. Guarente. Functional dissection and sequence of yeast HAP1 activator. *Cell*, 56(2):291–301, 1989.

[97] I. M. Ehrenreich, J. P. Gerke, and L. Kruglyak. Genetic dissection of complex traits in yeast: insights from studies of gene expression and other phenotypes in the BY×RM cross. *Cold Spring Harb. Symp. Quant. Biol.*, 74:145–153, 2009.

[98] J. R. Broach. Ras-regulated signaling processes in Saccharomyces cerevisiae. *Cold Spring Harb. Symp. Quant. Biol.*, 1(3):370–377, 1991.

[99] E. N. Smith and L. Kruglyak. Gene-environment interaction in yeast gene expression. *PLoS Biol.*, 6(4):e83, 2008.

[100] L. Bao, J. L. Peirce, M. Zhou, H. Li, D. Goldowitz, R. W. Williams, L. Lu, and Y. Cui. An integrative genomics strategy for systematic characterization of genetic loci modulating phenotypes. *Hum. Mol. Genet.*, 16(11):1381–1390, 2007.

[101] Y. Chen, J. Zhu, P. Y. Lum, X. Yang, S. Pinto, D. J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S. K. Sieberts, A. Leonardson, L. W. Castellini, S. Wang, M. F. Champy, B. Zhang, V. Emilsson, S. Doss, A. Ghazalpour, S. Horvath, T. A. Drake, A. J. Lusis, and E. E. Schadt. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452(7186):429–435, 2008.

[102] E. E. Schadt, C. Molony, E. Chudin, K. Hao, X. Yang, P. Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, J. Zhu, J. Millstein, S. Sieberts, J. Lamb, D. GuhaThakurta, J. Derry, J. D. Storey, I. Avila-Campillo, M. J. Kruger, J. M. Johnson, C. A. Rohl, A. van Nas, M. Mehrabian, T. A. Drake, A. J. Lusis, R. C. Smith, F. P. Guengerich, S. C. Strom, E. Schuetz, T. H. Rushmore, and R. Ulrich. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, 6(5):e107, 2008.

[103] X. Yang, J. L. Deignan, H. Qi, J. Zhu, S. Qian, J. Zhong, G. Torosyan, S. Majid, B. Falkard, R. R. Kleinhanz, J. Karlsson, L. W. Castellani, S. Mumick, K. Wang, T. Xie, M. Coon, C. Zhang, D. Estrada-Smith, C. R. Farber, S. S. Wang, A. van Nas, A. Ghazalpour, B. Zhang, D. J. Macneil, J. R. Lamb, K. M. Dipple, M. L. Reitman, M. Mehrabian, P. Y. Lum, E. E. Schadt, A. J. Lusis, and T. A. Drake. Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat. Genet.*, 41(4):415–423, 2009.

[104] R. Alberts, J. Fu, M. A. Swertz, L. A. Lubbers, C. J. Albers, and R. C. Jansen. Combining microarrays and genetic analysis. *Brief. Bioinform*, 6(2):135–145, 2005.

[105] J. J. Keurentjes, J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, A. J. Snoek, L. B .and Peeters, D. Vreugdenhil, M. Koornneef, and R. C. Jansen. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. USA*, 104(5):1708–1713, 2007.

[106] M. A. Newton, F. A. Quintana, J. A. den Boon, S. Sengupta, and P. Ahlquist. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.*, 1:85–106, 2007.

[107] H. Zhong, X. Yang, L. M. Kaplan, C. Molony, and E. E. Schadt. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.*, 84(4):581–591, 2010.

[108] T. A. Thornton-Wells, J. H. Moore, and J. L. Haines. Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet.*, 20(12):640–647, 2004.

[109] B. M. Neale and P. C. Sham. The future of association studies: Gene-based analysis and replication. *Am. J. Hum. Genet.*, 75(3):353–362, 2004.

[110] E. Jorgenson and J. S. Witte. A gene-centric approach to genome-wide association studies. *Nat. Rev. Genet.*, 7:885–891, 2006.

[111] Y. H. Cui, G. L Kang, K. L Sun, M. P. Qian, R. Romero, and W. J. Fu. Gene-centric genomewide association study via entropy. *Genetics*, 179:637–650, 2008.

[112] A. Buil, A. Martinez-Perez, A. Perera-Lluna, L. Rib, P. Caminal, and J. M. Soria. A new gene-based association test for genome-wide association studies. *BMC Proc.*, 3:S130, 2009.

[113] B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456:18–21, 2008.

[114] J. H. Moore and S. M. Williams. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.*, 85:309–320, 2009.

[115] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, 11:446–450, 2010.

[116] W. W. Piegorsch, C. R. Weinberg, and J. A. Taylor. Non-hierarchical logistic models and case-only designs for accessing susceptibility in population-based case-control studies. *Stat. Med.*, 13(2):153–162, 1994.

[117] Y. Zhang and J. S. Liu. Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.*, 39(9):1167–1173, 2007.

[118] G. Kang, W. Yue, J. Zhang, Y. H. Cui, Y. Zuo, and D. Zhang. An entropy-based approach for testing genetic epistasis underlying complex diseases. *J. Theor. Biol.*, 250(2):362–374, 2008.

[119] Ritchie M. D., L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.*, 69(1):138–147, 2001.

[120] L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, 2001.

[121] M. Li, R. Romero, W. J. Fu, and Y. H. Cui. Mapping haplotype-haplotype interactions with adaptive lasso. *BMC Genet.*, 11:79, 2010.

[122] H. J. Cordell. Detecting gene-gene interactions that underlie human disease. *Nat. Rev. Genet.*, 10(6):392–404, 2009.

[123] J. Y. Tzeng, B. Devlin, L. Wasserman, and K. Roeder. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.*, 72(4):891–902, 2003.

[124] D. J. Schaid, S. K. McDonnell, S. J. Hebbring, J. M. Cunningham, and S. N. Thibodeau. Nonparametric tests of association of multiple genes with human disease. *Am. J. Hum. Genet.*, 76(5):780–793, 2005.

[125] J. Wessel and N. J. Schork. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.*, 79(5):792–806, 2006.

[126] D. J. Schaid. Genomic similarity and kernel methods i: Advancements by building on mathematical and statistical foundations. *Hum. Hered.*, 70(2):109–131, 2010.

[127] D. J. Schaid. Genomic similarity and kernel methods ii: Methods for genomic information. *Hum. Hered.*, 70(2):132–140, 2010.

[128] L. Kwee, D. Liu, X. Lin, D. Ghosh, and M. Epstein. A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.*, 82(2):386–397, 2008.

[129] M. C. Wu, P. Kraft, M. P. Epstein, D. M. Taylor, S. J. Chanock, D. J. Hunter, and X. Lin. Powerful snp-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.*, 86(6):929–942, 2010.

[130] G. Wahba. *Spline models for observational data: CBMS-NSF regional conference series in applied mathematics*. Philadelphia: Society of Industrial and Applied Mathematics, 1990.

[131] C. Gu. *Smoothing spline ANOVA models*. Springer-Verlag, 2002.

[132] C. Gu and G. Wahba. Smoothing spline anova with component-wise bayesian "confidence intervals". *J. Comput. Graph. Statist.*, 2(1):97–117, 1993.

[133] G. Wahba, Y. D. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.*, 23(6):1865–1895, 1995.

[134] T. Speed. [That BLUP is a good thing: The estimation of random effects]: Comment. *Statist. Sci.*, 6:42–44, 1991.

[135] D. E. Weeks and K. Lange. The affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.*, 42(2):315–326, 1988.

[136] I. Mukhopadhyay, E. Feingold, D. E. Weeks, and A. Thalamuthu. Association tests using kernel-based measures of multi-locus genotypes similaritybetween individuals. *Genet. Epidemiol.*, 34(3):213–221, 2010.

[137] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[138] S. G. Self and K. Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, 82(398):605–610, 1987.

[139] D. Liu, X. Lin, and D. Ghosh. Semiparametric regression of multi-dimensional genetic pathway data: least squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, 2007.

[140] R. R. Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[141] T. Wang, G. Ho, K. Ye, and R. Elston. A partial least square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Bioinformatics*, 33(1):6–15, 2009.

[142] J. He, K. Wang, A. C. Edmondson, D. J. Rader, C. Li, and M. Li. Gene-based interaction analysis by incorporating external linkage disequilibrium information. *Eur. J. Hum. Genet.*, 19(2):164–172, 2011.

[143] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, 2003.

[144] A. Roy, F. Exinger, and R. Losson. cis- and trans-acting regulatory elements of the yeast URA3 promoter. *Mol. Cell Biol.*, 10(10):5257–5270, 1990.

[145] S. Li, Q. Lu, W. Fu, R. Romero, and Y. Cui. A regularized regression approach for dissecting genetic conflicts that increase disease risk in pregnancy. *Stat. Appl. Genet. Mol. Biol.*, 8(1):Article 45, 2009.

[146] N. Chatterjee, Z. Kalaylioglu, R. Moslehi, U. Peters, and S. Wacholder. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am. J. Hum. Genet.*, 79(6):10021016, 2006.

[147] J. Chapman and D. Clayton. Detecting association using epistatic information. *Genet. Epidemiol.*, 31(8):894909, 2007.

[148] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *J. Amer. Stat. Assoc.*, 88(421):9–25, 1993.

[149] D. Liu, D. Ghosh, and X. Lin. 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, 2008.

[150] B. Li and Leal S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, 83(3):311–321, 2008.

[151] B. E. Madsen and S. R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, 5(2):e1000384, 2009.

[152] A. L. Price, G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, and S. R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, 86(6):832–838, 2010.

[153] J. Zhu, P. Y. Lum, J. Lamb, D. GuhaThakurta, S. W. Edwards, R. Thieringer, J. P. Berger, M. S. Wu, J. Thompson, A. B. Sachs, and E. E. Schadt. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome. Res.*, 105(2-4):363–374, 2004.

[154] D. Zhang and X. Lin. Hypothesis testing in semeparametric additive mixed models. *Biostatistics*, 4:57–74, 2003.