Title
Design of efficient and accurate statistical approaches to correct for confounding effects and identify true signals in genetic association studies

Author
JOO, JONG WHA JOANNE

Publication Date
2015

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Design of efficient and accurate statistical approaches to correct for confounding effects and identify true signals in genetic association studies

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Bioinformatics Program

by

Jong Wha Joanne Joo

2015

Abstract of the Dissertation

Design of efficient and accurate statistical approaches to correct for confounding effects and identify true signals in genetic association studies

by

Jong Wha Joanne Joo

Doctor of Philosophy in Bioinformatics Program

University of California, Los Angeles, 2015

Professor Eleazar Eskin, Chair

Over the past decades, genome-wide association studies have dramatically improved especially with the advent of the hight-throughput technologies such as microarray and next generation sequencing. Although genome-wide association studies have been extremely successful in identifying tens of thousands of variants associated with various disease or traits, many studies have reported that some of the associations are spurious induced by various confounding factors such as population structure or technical artifacts. In this dissertation, I focus on effectively and accurately identifying true signals in genome-wide association studies in the presence of confounding effects. First, I introduce a method that effectively identifying regulatory hotspots while correcting for false signals induced by technical confounding effects in expression quantitative loci studies. Technical confounding factors such as a batch effect complicates the expression quantitative loci analysis by inducing heterogeneity in gene expressions. This creates correlations between the samples and may cause spurious associations leading to spurious regulatory hotspots. By formulating the problem of identifying genetic signals in a linear mixed model framework, I show how we can identify regulatory hotspots while capturing het-

erogeneity in expression quantitative loci studies. Second, I introduce an efficient and accurate multiple-phenotype analysis method for high-dimensional data in the presence of population structure. Recently, large amounts of genomic data such as expression data have been collected from genome-wide association studies cohorts and in many cases it is preferable to analyze more than thousands of phenotypes simultaneously than analyze each phenotype one at a time. However, when confounding factors, such as population structure, exit in the data, even a small bias is induced by the confounding effects, the bias accumulates for each phenotype and may cause serious problems in multiple-phenotype analysis. By incorporating linear mixed model in the statistics of multivariate regression, I show we can increase the accuracy of multiple phenotype analysis dramatically in high-dimensional data. Lastly, I introduce an efficient multiple testing correction method in linear mixed model. The significance threshold differs as a function of species, marker densities, genetic relatedness, and trait heritability. However, none of the previous multiple testing correction methods can comprehensively account for these factors. I show that the significant threshold changes with the dosage of genetic relatedness and introduce a novel multiple testing correction approach that utilizes linear mixed model to account for the confounding effects in the data.

The dissertation of Jong Wha Joanne Joo is approved.

Matteo Pellegrini

Jason Ernst

Bogdan Pasaniuc

Aldons J. Lusis

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2015

This dissertation is dedicated to my husband Sang Eon Bak who supported my studies, my daughter Jooin Bak who was born during the course of my studies for being such a joy, my second baby who is not born yet, and my parents Seung Ki Joo and Seong Ae Kim for all of their love and support. Lastly, I thank to the God.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Professor Eleazar Eskin for his continuous guidance and patience throughout the course of my research and thesis. I also thank my committee members Jason Ernst, Matteo Pellegrini, Bogdan Pasaniuc, and Aldons J. Lusis for their support and encouragement. Finally, I would like to thank my lab colleagues Farhad Hormozdiari, Jaehoon Sul, Buhm Han, Eun Yong Kang, Nick Furlotte, and Emrah Kostem for helpful discussion and advice. This work was supported by NSF grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448 and 1320589, and NIH grants K25-HL080079, U01- DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-MH101782 and R01-ES022282. I acknowledge the support of the NINDS Informatics Center for Neurogenetics and Neurogenomics (P30 NS062691).

Vita

2000-2005	B.Sc., Computer Science and Engineering, Seoul National Uni., Seoul, Rep. of Korea
2004-2005	Research Intern, Neo poly Inc., Seoul, Rep. of Korea
2005-2007	M.Sc., Computer Science, Brown Uni., Providence, RI, U.S.A.
2007-2010	Researcher Assistant and Software Developer, Functional Proteomics Center of KIST, Seoul, Rep. of Korea
2015	Teaching Assistant, Uni. of California Los Angeles, LA, U.S.A.

Publications

JW Joo, S Na, JH Baek, C Lee and E Paek, "Target-Decoy with Mass Binning: A Simple and effective validation method for shotgun proteomics using high resolution mass spectrometry" in *Journal of Proteome research*, 2010.

PY Chen, S Feng, JW Joo, S Jacobsen and M.Pellegrini, "A comparative analysis of DNA methylation across human embryonic stem cell lines" in *Genome Biology*, 2011.

A Ghazalpour, C Rau, C Farber, B Bennett, L Orozco, A Nas, C Pan, H Allayee, S Beaven, M Civelek, R Davis, T Drake, R Friedman, N Furlotte, S Hui, J Jentsch, E Kostem, HM Kang, EY Kang, JW Joo, V Korshunov, R Laughlin, L Martin, J Ohmen, B Parks, M Pellegrini, K Reue, D Smith, S Tetradis, J Wang, Y Wang, J Weiss, T Kirchgessner, P Gargalovic, E Eskin, A Lusis, R Leboeuf,"Hybrid mouse diversity panel: a panel of inbred mouse strains suitable for analysis of complex genetic traits," in *Mammalian*

Genome, 2012.

EY Kang*, B Han*, N Furlotte*, JW Joo, Diana, S Davis, J Lusis, E Eskin, "Meta-analysis identifies gene-by-environment interactions as demonstrated in a study of 4,323 mouse samples" in *PLOS genetics*, 2014.

JW Joo, JH Sul, B Han, C Ye, E Eskin, "Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies" in *Genome Biology*. 2014.

D He, N Furlaotte, F Hormozdiari, JW Joo, R Ostrovsky*, A Sahai*, E Eskin*. "Identifying Genetic Relatives without Compromising Privacy" in *Genome Research*. 2014.

J Ohmen, EY Kang, X Li, JW Joo, F Hormozdiari, Qing Y Zheng, R Davis, A Lusis, E Eskin, R Friedman, "Genome-Wide Association Study for Age-Related Hearing Loss (AHL) in the Mouse: A Meta-Analysis." in *J Assoc Res Otolaryngol*, 2014.

F Hormozdiari*, JW Joo*, F Guan, R Ostrosky, A Sahai, E Eskin**. "Privacy Preserving Protocol for Detecting Genetic Relatives Using Rare Variants" in *Bioinformatics*, 2014.

JW Joo, EY Kang, E Org, N Furlotte, B Parks, A Lusis, E Eskin. "Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure" in *RECOMB2015*, 2015.

E Org, B Parks, JW Joo, M Mehrabian, Y Blum, EY Kang, B Emert, R Knight, T Drake, E Eskin, A Lusis., "Genetic and environmental control of host-gut microbiota interaction" in *Genome Research*, 2015.

S Na, JJ Lee, J Jeong, JW Joo, KJ Lee, and E Paek, "DDD:Detecting Deuterated Distribution from Heterogeneous Protein States via HDX-MS", *Under review.*

EY Kang, N Furlotte, JW Joo, E Kostem, N Zaitlen, B Han, E Eskin, "Accounting for genetic architecture differences between the sexes in association studies", *Under review.*

JW Joo, F Hormozdiari, B Han*, E Eskin*, "Multiple Testing Correction in Linear Mixed Models,", *Under review.*

F Hormozdiari, JW Joo, B Pasaniuc, E Eskin, "Joint Fine mapping of GWAS and eQTL detects Target Gene and Relevant Tissue", *Under review.*

CHAPTER 1

Introduction

1.1 Motivation

Elucidating the content of a DNA sequence is critical to deeper understand and decode the genetic information for any biological system. With the advent of high-throughput technologies such as microarray and next generation sequencing, nowadays it is feasible to scan millions of variants and tens of thousands of gene expressions. In addition, systematic tools have been developed to analyze the massive amount of data. As a result, genome-wide association studies (GWAS) have successfully identified numerous loci at which variants influence disease risk or quantitative traits.

However, it has been found that many of these identifications are from false positive signals. Many GWAS have reported the existence of various hidden confounding factors, such as unobserved covariates, batch effects, genetic relatedness, environmental perturbations, and so on. These confounders may cause spurious associations by inducing complex dependencies among the individuals and lead to serious problems in various fields of GWAS. For example, it has been reported in many studies that genetic relatedness or population structure, which is one of the representative confounding factors in GWAS, creates many spurious associations in GWAS [KCP02, FRP04, MCP04, COL05, HYH05, RZL05, VP05,BSK06,SSV06,FG06,FE12]. Another representative confounding factor is technical artifacts such as a batch effect. This creates heterogeneity in measurements of gene expres-

sions and may create spurious associations in eQTL studies [LS07,KYE08,LKS10,FSL12].

These confounders may cause even more serious problem in multiple-phenotype analysis. Typical GWAS test correlation between a single phenotype and each genotype one at a time, referred to as "single-phenotype analysis". However, analyzing multiple phenotypes simultaneously, referred to as "multiple-phenotype analysis", often has advantages over single-phenotype analysis. For example, multiple-phenotype analysis can increase the statistical power and may allow to detect signals which is not detectable in single-phenotype analysis due to its small effect sizes. However, when confounding effects exist in the data, it may cause serious problems in the multiple-phenotype analysis. This is because even a small bias is induced by the confounding effects, this bias accumulates for each phenotype in the multiple-phenotype analysis.

Morover, these confounding factors affect multiple testing correction which is an essential step in GWAS analysis and small change in significance threshold may cause many false positives or false negatives. As the number of SNPs genotyped by current association studies is dramatically increasing, the large number of correlated markers brings to the forefront the multiple hypothesis testing correction problem and has motivated much recent activity to address it [WY93,Lin05,SM05,CB07,HKE09]. Unfortunately, in the case when confoundings are present, these cause a violation of the basic assumption necessary for previous approaches which is that the individuals in the sample are i.i.d.. How to correct for multiple testing in the presence of confounding is a fundamental problem and a promising avenue for GWAS.

In this dissertation, I would like to introduce my works on correcting for false signals induced by various confounding factors such as population structure and technical artifacts. By correcting the confounding effects I was able to remove false positive signals and

increase statistical power to identify true signals. In addition, I worked on multiple testing correction to estimate accurate significance threshold in the presence of confounding effects to identify true signals in GWAS.

1.2 Organization

The main focus of my research has been developing efficient and accurate statistical approaches to correct for the false signals due to various confounding factors such as technical artifacts or genetic relatedness, and identify true genetic signals in GWAS.

Linear mixed model has been established as a standard analysis tool for GWAS as it can correct for these hidden factors [KSS10, LLK12, LLH13, YZG14, LTB15]. I applied the linear mixed model to various fields of GWAS to explicitly model these hidden factors and avoid false positives and increase statistical power to detect the true signals.

First, in Chapter 2, I introduce some basic concepts of GWAS and backgrounds of my research such as confounders that cause problems in GWAS, Linear Mixed Models that can fix the confounding effects, and multiple hypothesis testing correction which is an essential step in GWAS that perform up to millions of statistical tests.

Then in Chapter 3, I introduce my work in eQTL studies that tries to correct for confounding factors such as a batch effect to identify genetic regulatory hotspots. I have shown that previous methods that attempt to correct for confounding effects, either fail to correct for spurious signals or remove both spurious signals and true signals. I introduce a novel approach that accurately and effectively remove the false signals while preserving true signals in eQTL studies to identify genetic regulatory hotspots.

In Chapter 4, I introduce my work in multiple-phenotype analysis that tries to efficiently and accurately perform multiple phenotype analysis for high dimensional data in the presence of population structure. From experiments, I have shown that in the multiple-phenotype analysis, the bias due to the genetic relatedness accumulates for each phenotype and may cause serious problems in the multiple-phenotype analysis even if it is ignorable in the single-phenotype analysis. I introduce an approach that can correct for confounding effects induced by population structure which dramatically increases the accuracy of multiple-phenotype analysis in high-dimensional data.

In Chapter 5, I introduce my work in multiple testing correction in linear mixed models. In this work, I have shown that the p-value threshold for significance, referred to as the per-marker threshold, differs as a function of genetic relatedness. As well as I have shown that none of the previous approaches, including permutation test, give accurate per-marker thresholds as their underlying i.i.d. assumption no longer holds under the linear mixed model. I introduced an approach based on the multivariate normal distribution that can efficiently estimate per-marker threshold correcting the effects of population structure.

Lastly, in Chapter 6, I talk about the conclusions of my research.

CHAPTER 2

Background

2.1 Genetic association studies

SNP is a single letter change in DNA, part of the natural genetic variation within a population that are known to associated with complex disease or traits. Genetic association studies seek for SNPs potentially causing phenotypic changes. For the purpose, genetic studies look for the association between a SNP's minor allele frequency and phenotype values or disease status by testing whether the data is more likely under the null hypothesis which assumes there is no association or under the alternative hypothesis which assumes there exists an association. Often a test statistic that reflects how likely the data is under the null model is used to compute a p-value and a predetermined p-value threshold is used to determine whether the association is statistically significant or not.

With the advent of high throughput technologies such as microarray and next generation sequencing, now it is possible to perform the genetic association studies in Genome-wide scale, referred to as Genome-wise association studies (GWAS) and association study is in a new era of big data. As the data size grows, efficient analytical tools for analyzing large datasets are required and many efficient systematical tools for analyzing a large amount of data have been emerged.

2.2 Expression quantitative trait loci studies

Given the rapid increase of the available data on genetic variants, GWAS have successfully identified tens of thousands of SNPs associated with many traits for the past years. However, none of the biological knowledge was encoded in standard GWAS analysis and the functional relevance of most discovered loci remains unclear. In has been reported that a large portion of phenotypic variability in disease risk can be explained by regulatory variants, referred to as expression quantitative trait loci (eQTL), that is, genetic variants that regulate the expression levels of genes [SMD03, NGZ10, GHC10, DYK13, ND13]. eQTL studies treat gene expression as a molecular phenotypic trait and seek for SNPs that are associated with the gene expression. Along with GWAS, eQTL studies became popular in genetic research not only to characterize functional sequence variation but also for understanding the basic processes of gene regulation and interpretation of GWAS.

2.3 Confounding factors in genetic studies

GWAS have reported the existence of various hidden factors, such as unobserved covariates, genetic relatedness, environmental perturbations, and so on. Those shared confounding factors induces complex dependencies among the individuals and complicates the analysis of GWAS leading to spurious associations.

One of the representative confounding factors in GWAS is the genetic relatedness or population structure. GWAS predict an association by looking at the association between a SNP's minor allele frequency and phenotype values or disease status. However, not only disease-causing SNPs cause allele frequency differences but also SNPs influenced by ancestry may cause the differences [KCP02, FRP04, MCP04, COL05, HYH05, RZL05, VP05, BSK06, SSV06, FG06]. This is because allele frequencies vary from population to

population due to each populations unique genetic and social history. Figure 2.1 is a figure from nature reviews genetics [FE12] and it shows an example of how serious problems population structure can causes in mice GWAS. The panel C shows a graphical model that many SNPs and a phenotype are both correlated with population structure. Panel b and c show associations detected before and after correcting for population structure, accordingly. In human studies, the genetic relatedness may not be as extreme as the one in mice studies, however, without correcting for population structure, they may result many false positives as well.

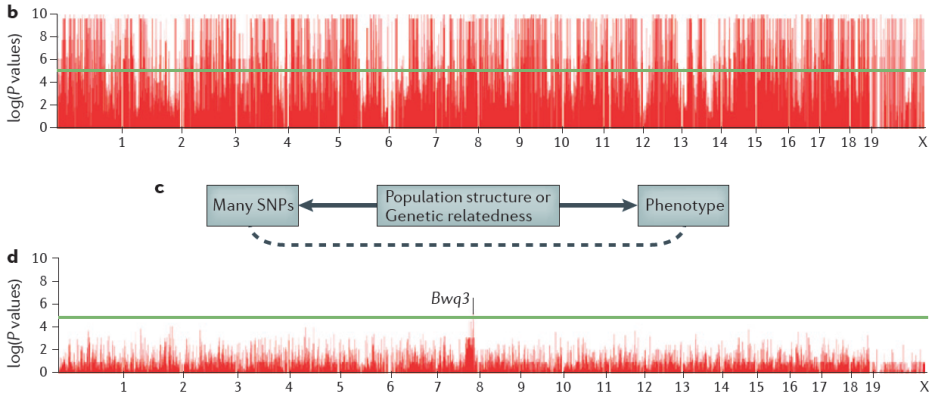


Figure 2.1. An example that population structure causes false signals in GWAS. This figure and caption is from nature reviews genetics [FE12]. Panel b shows a Manhattan plot with the association results for 140,000 SNPs and body weight for 38 inbred strains from the Mouse Phenome Database [BGM07, HAA07, GMB09]. Almost every locus appears to be associated with body weight as each of the many SNPs that differentiate the wild-derived and classical inbred strains appears to be associated with body weight. A visualization of the cause of the spurious associations is shown panel c. Many SNPs and the phenotype are both correlated with the genetic relatedness or population structure among the strains. Statistical techniques can take into account the genetic relationships between the strains to correct for population structure, thus minimizing spurious associations. In this example, EMMA [KZW08] was applied to the data (panel d).

Another popular confounding factor that may induce false identifications is the technical confounding factors such as a batch effect that affects the global correlation structure of

the expression levels of the genes in eQTL studies. For example, let's say the gene expressions of the first half of the sample are measured in the first day of the experiment and gene expressions of the other half of the samples are measured in the second day of the experiment. And let's say, by a technical artifact, all the measurements in the first day are higher than those of the second day. This induces spurious correlations between the first half of the samples and between the other half of the samples in terms of gene expressions. As GWAS test tens of thousands of SNPs, by chance, some SNPs can be correlated with the genes due to the spurious correlation structure induced by the confounding effects and may result false positive identifications. This phenomenon has been reported by many previous studies [FE12, LS07, LKS10] including an eQTL study of recombinant inbred mice, where the regulatory hotspots are inconsistent between replicates [KYE08] (Figure 2.2).

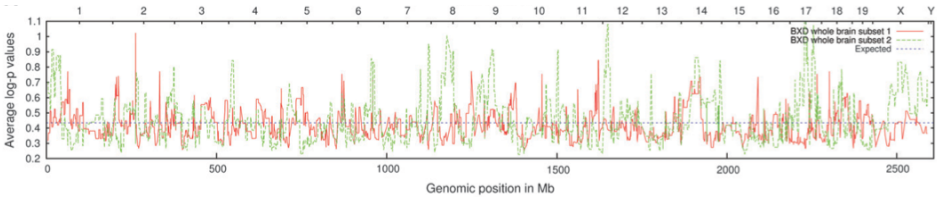


Figure 2.2. An example that the hotspots are inconsistent between replicates in eQTL study. This figure and caption is from genetics [KYE08]. Comparison of the strength of trans-regulatory bands between replicated subsets. The horizontal axis is the genomic positions of the markers in megabases, and the vertical axis is the strength of regulatory hotspots quantified as the average log P-values at each marker across all genes. Each peak represents the strength of the trans-regulatory band at a particular marker SNP. The taller the peak, the more pronounced a trans-regulatory band is in an eQTL map.

2.4 Linear Mixed Model

To explicitly model the hidden factors in order to avoid false positives and increase statistical power, nowadays the linear mixed model (LMM) has been established as a standard analysis tool for GWAS [KSS10, LLK12, LLH13, YZG14, LTB15].

The typical GWAS test the association by estimating the maximum likelihood parameter of the following linear model;

$$y = \mu 1_n + \beta_k x_k + e \tag{2.1}$$

Let n bet the number of samples analyzing, then y is a vector of length n that contains phenotype values, μ is the model mean and 1_n is a vector of length n that contains 1s, x_k is a vector of length n that contains genotype values of kth SNP that we are testing, β_k is the effect size of the kth SNP, and e is a random vector of length n drawn from multivariate normal distribution with mean 0 and covariance matrix $\sigma_e^2 I$, denoted as $e \sim N(0, \sigma_e^2 I)$, where I is the identity matrix.

However, when there exist confounding factors, the equation (5.1) does no longer fit to the generative model of the data [Esk15]. To accommodate the effects of confounding factors, LMM adds an extra term u to the standard linear model (Equation (5.1)) and the statistical model for LMM is as follows:

$$y = \mu 1_n + \beta_k x_k + u + e \tag{2.2}$$

Here, u contains the random variables that accommodates the effects of confounding factors, where we assume $u \sim N(0, \sigma_g^2 K)$ and K is the spurious correlation structure between the samples induced by confounding factors that we want to correct. For example, if we are intended to correct for the spurious correlation structure induced by population structure, K will be a matrix that contains the correlation between the samples in terms of the SNPs, referred to as the "kinship matrix". However, if we are intended to correct for the spurious correlation structure induced by the confounding effects that affects gene

expressions, the K will be a matrix that contains the correlation between the samples in terms of the gene expressions.

2.5 Multiple hypothesis testing correction

The multiple hypothesis testing problem is the situation when we consider many hypotheses simultaneously. When we perform one hypothesis test, the probability of making an error with a standard p-value cut-off of α is α. However, when we perform m hypothesis test, if we use the p-value cut-off of α, the probability of making at least one error increases as a function of m, $1-(1-\alpha)^m$. Thus, we want to control this type I error by adjusting the p-value threshold for each test, α_p, thus the probability of making at least one error to be α; $1-(1-p_{\alpha_p})^m = \alpha$.

Multiple hypothesis testing is an essential step in GWAS analysis, especially, as the number of SNPs genotyped by current association studies is dramatically increasing. A challenge in multiple hypothesis testing in GWAS is a large number of markers are correlated thus the typical multiple hypothesis correction such as Bonferroni correction may cause many false negatives. To address this problem, many statistical methods have been proposed based on permutation test [WY93] or multivariate normal distribution framework to speed up the process [Lin05, SM05, CB07, HKE09]. Another challenge in multiple testing correction in GWAS is that the correct per-marker threshold differs as a function of species, marker densities, genetic relatedness, and trait heritability. However, none of the previous multiple testing correction methods can comprehensively account for these factors; therefore, these methods are not applicable for linear mixed models.

CHAPTER 3

Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies

3.1 Introduction

Understanding the relationship between genetic variation and gene regulation has recently received a lot of interest. The most common approach to study this relationship is through an expression quantitative trait (eQTL) study where both genetic variation and expression levels are collected from a set of individuals and associations between genetic variation and expression are estimated [BYC02, BK05, KFT07, CLS05, BWD05, CSE05, SNF07, ETZ08, SBB07]. Any identified association, or eQTL, suggests the presence of a region harboring genetic variation that affects expression levels.

In eQTL mapping, two types of eQTLs are analyzed: *cis*-eQTLs that are in close proximity to the gene locus and *trans*-eQTLs that occur at a greater distances from the gene locus [MLB09]. Previous eQTL studies in multiple organisms [BK05,KFT07,CLS05,CSE05] have shown that many genes are *trans*-regulated by a small number of genomic regions, known as "regulatory hotspots". Although several eQTL studies have successfully identified regulatory hotspots [CLE05,HWW05,WKH04], it has been reported in studies of recombinant inbred (RI) mice that regulatory hotspots replicate poorly [PLW06]. Previous

studies have discovered that these regulatory hotspots are spurious associations caused by various confounding factors, such as batch effects or other technical artifacts which induce noise during sample preparation or expression measurements [Chu02, FCD03, BMH07]. Confounding factors create heterogeneity in expression data and may induce spurious associations between SNPs and gene expressions, leading to the identification of "spurious regulatory hotspots" [KYE08]. In these spurious hotspots, SNPs appear to be associated with gene expression levels, although they do not have genetic effects on the genes.

Several computational methods have been developed to correct for confounding effects using various statistical methods such as singular value decomposition or linear mixed models [LS07, KYE08, LKS10, FSL12]. The main assumption behind most of these methods is that the confounding factors influence the global correlation structure between the gene expression values. Hence, the methods, such as ICE [KYE08] and SVA [LS07], attempt to estimate the global correlation structure and use it as a covariate in the association to remove confounding effects from the association statistic. Although these methods effectively remove spurious regulatory hotspots, they may also remove true hotspots caused by genetic factors. This is because global correlation structure contains genetic effects, so by correcting for the global structure, genetic effects are removed as well. For example, in a well studied yeast dataset several hotspots are known to be true genetic effects since they have been validated by additional data such as protein measurements [FRS07, PRR07]. Unfortunately, these hotspots are removed in addition to the spurious ones. Other methods [LKS10, FSL12] also do not explicitly remove true genetic signals and either eliminate true hotspots or fail to remove spurious hotspots in our experiments.

In this chapter, we introduce a new method called Next-generation Intersample Correlation Emended (NICE) eQTL mapping that attempts to eliminate spurious regulatory hotspots while retaining hotspots caused by genetic effects utilizing a novel statistical

framework. Our method leverages an insight that confounding factors would affect the majority of genes, while genetic effects would only affect a subset. This insight allows us to distinguish between confounding and genetic effects. We used a recently developed statistic [HE12] to differentiate between genes that are affected by both genetic effects and confounding effects versus genes that are affected by only confounding effects. Using genes only affected by confounding, we are able to correct for the confounding effects but preserve the genetic effects. We first show by simulations that NICE successfully eliminates spurious regulatory hotspots while preserving regulatory hotspots corresponding to real genetic effects. On the other hand, previous methods either fail to eliminate confounding effects or fail to retain the genetic effects.

We demonstrate the utility of NICE with a yeast dataset. Two versions of a yeast dataset were generated in 2005 [BK05] and 2008 [SK08]. Since they were generated 3 years apart in different locations, the hotspots that are shared between the datasets are likely to be the real genetic effects, while hotspots that are different between the datasets are likely to be spurious hotspots. We utilize our method on only the first dataset to see if we can discriminate which hotspots are real and spurious as determined by the second dataset. Applied to the yeast dataset, NICE identifies 83% of the putative regulatory hotspots which are consistent between the two versions of a yeast dataset. Previous methods applied to this dataset either eliminate many of the putative hotspots or predict many spurious hotspots. In addition, NICE eQTL mapping identifies either more or a comparable number of *cis* associations relative to previous methods. Furthermore, applied to a yeast dataset grown in different conditions, NICE identifies genes that are related to gene-by-environment interactions and discovers novel yeast regulatory hotspots that are likely to have a true biological mechanism.

3.2 Results

3.2.1 NICE eQTL mapping

Our goal is to identify true genetic associations in an eQTL mapping study without predicting spurious associations due to confounding factors. Consistent with previous approaches that correct for the confounding factors based on singular value decomposition or linear mixed models, we assume that the confounding factors affect the global correlation structure of expressions. That is, we assume that confounding factors affects the expression levels of most of the genes. On the other hand, we assume that genetic factors only affect the expression levels of a subset of the genes related to the regulatory pathways. Figure 3.1 (a) shows a graphical model that contains both genetic and spurious associations due to a confounding factor. SNP 1 has a genetic effect on multiple genes and thus, is a regulatory hotspot. Unlike SNP 1, SNP 2 has no direct genetic effects on any of the gene. However, SNP 2 has spurious associations with many of the genes because by chance, it happens to be correlated with the confounder and this results in a spurious regulatory hotspot.

To eliminate spurious associations, ICE [KYE08] models the confounding effects by estimating the global correlation structure of the expression levels of all genes and using this structure as a covariate in the association statistic. This has the same effect as if the confounding factor itself is included as a covariate in the association statistic removing its effect. Unfortunately, any regulatory hotspots, as in the case of the first three genes of Figure 3.1 (b), will also be captured in the global correlation structure and be eliminated. For this reason, ICE [KYE08] tends to eliminate true regulatory hotspots in addition to confounding effects.

In contract to ICE [KYE08], NICE uses only a subset of genes to model the global correlation structure between expression levels used to correct for confounding factors. Since most genes are affected by confounding effects, any subset of genes will likely capture the confounding effects and utilizing only those genes to estimate the global correlation structure is enough to correct for the confounding factor. Theoretically, if we can correct for confounding effects using the genes that are not involved in true regulatory hotspots, we would eliminate spurious associations while preserving true genetic associations. Unfortunately, we do not know in advance which genes are involved in the regulatory hotspots which complicates choosing which genes to use to correct for the confounding. Practically, almost all genes are affected by confounding effects while only some genes have both confounding and genetic effects. We expect this second group of genes to show stronger associations than the others. Thus we use the weakly associated genes to model the confounding factors. For example, in Figure 3.1 (c), by using the four most weakly associated genes, we may correct for the confounding effects but preserve the genetic effects.

3.2.2 NICE eliminates spurious regulatory hotspots while preserving true genetic effects

To validate that our method eliminates spurious regulatory hotspots while preserving regulatory hotspots corresponding to real genetic effects, we generated a simulated dataset with both true regulatory hotspots and a batch effect that creates spurious hotspots. We create a dataset that has 1000 samples with 1000 SNPs and 1000 gene expression levels. We added 5 *trans*-regulatory hotspots and *cis* effects. For each of the *trans*-regulatory hotspots, 20% of the genes have *trans* effects. SNPs are randomly generated with minor allele frequencies of 30%. A batch effect is simulated where expression levels in the first half of samples are correlated with each other, but not correlated with the second half of the samples, and vice versa.

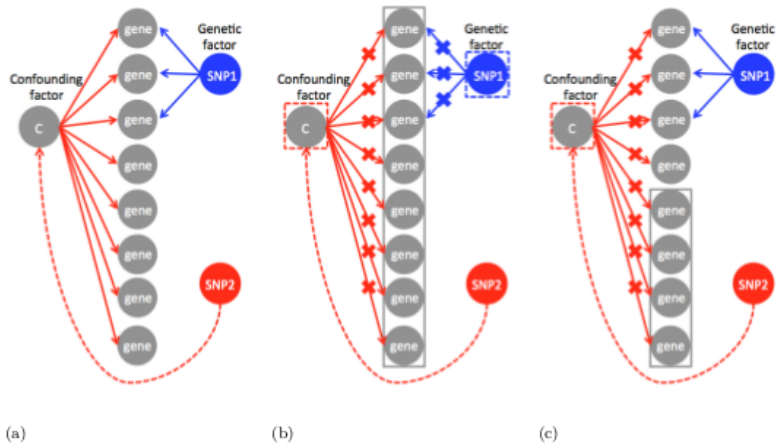


Figure 3.1. (a) Graphical model of genetic and spurious associations. SNP 1 has genetic effect on the first three genes (blue arrows). SNP 2 has no genetic effect on any of the genes, however, it has spurious associations with many of the genes, because by chance, it happens to be correlated with a confounding factor (red arrows). (b) ICE [KYE08] models the confounding effects by estimating the global correlation structure of the expression levels of all genes (grey block). However, this eliminates genetic associations in addition to confounding effects as any regulatory hotspots, the first three genes, will also be captured in the global correlation structure and be eliminated. (c) NICE uses only a subset of genes to model the global correlation structure between expression levels used to correct for confounding factors. Since any subset of genes will capture the confounding effects, in practice, using the bottom four genes (grey block), we can eliminate the confounding effects but preserve the genetic effects.

We visualize the results of an eQTL study through use of an eQTL plot such as those shown in Figure 3.2. The x-axis corresponds to SNP positions and the y-axis corresponds to the gene positions. The intensity of a point on the plot represents the significance of the association. The diagonal band represents the cis-effects and the vertical bands represent hotspots. On the eQTL plot, we mark successfully identified regulatory hotspots with blue arrows, missed regulatory hotspot with green arrows and spurious hotspots with red arrows. In the simulated data, the eQTL plot shows 5 regulatory hotspots (Figure 3.2 (a), blue arrows) and 8 spurious hotspots (Figure 3.2 (a), red arrows) induced by the batch

effect we simulated.

We compared our method to several methods including SVA [LS07], ICE [KYE08], LMM-EH [LKS10], and PANAMA [FSL12], that try to correct expression heterogeneity on simulated data and showed the results on eQTL plots (Figure 3.2). NICE successfully identifies 5 real regulatory hotspots and eliminates spurious ones (Figure 3.2 (f)). SVA also identifies 5 real regulatory hotspots and eliminated spurious ones (Figure 3.2 (b)), which is expected as SVA meant to capture only the broad signal and our simulated data contains only one large batch effect. In the next section, we show that SVA does not perform as well on a real dataset which contains more realistic confounding effects. On the other hand, ICE and LMM-EH remove not only spurious hotspots but also real hotspots (Figure 3.2 (c) and (d)). PANAMA fails to remove the spurious hotspots and does not show a big difference with the standard t-test (Figure 3.2 (a) and (e))). The results are summarized in Table 3.1. All methods successfully identify *cis* effects. We further study the test statistics of p-values by estimating the genomic control inflation factor λ [DR99] to check if the p-values are either inflated ($\lambda > 1$) or deflated ($\lambda < 1$). Figure 3.3 shows $\Delta\lambda$ which is defined as $1 - \lambda$. $\Delta\lambda$ of SVA and NICE are close to zero. On the other hand, the standard t-test and PANAMA show inflation ($\Delta\lambda > 0$) and ICE and LMM-EH show deflation ($\Delta\lambda < 0$).

3.2.3 NICE eliminates spurious hotspots while preserving genetic hotspots in Yeast dataset

We take advantage of a unique dataset consisting of two versions of a yeast dataset generated in 2005 [BK05] and 2008 [SK08] to validate our method. The two datasets contain similar strains, but were generated 3 years apart in different locations. For this reason, the hotspots that are shared between the datasets are likely real genetic effects, while

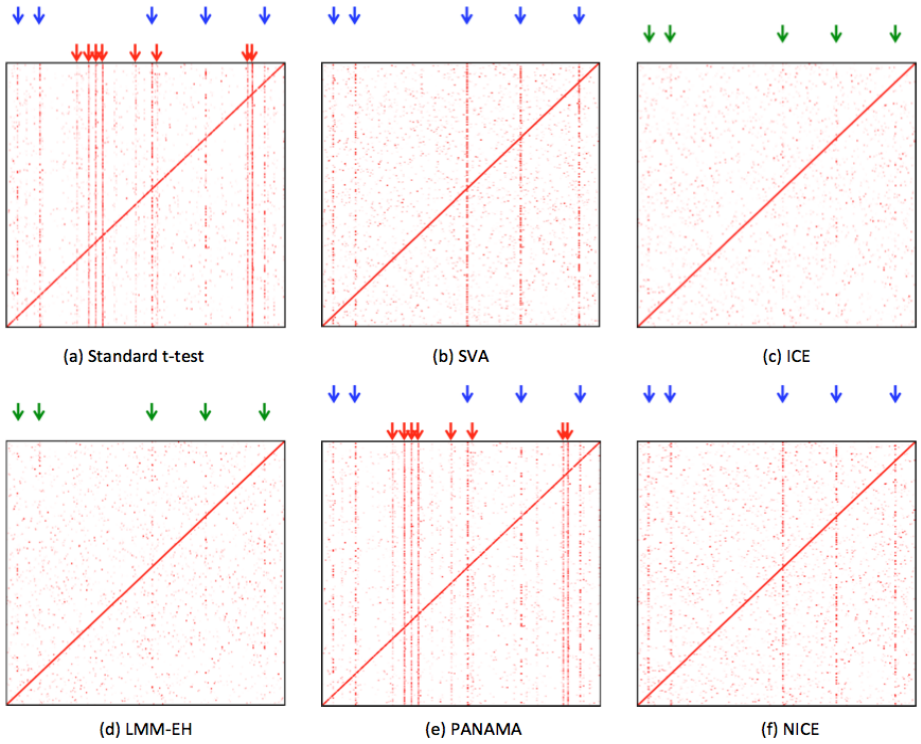


Figure 3.2. eQTL maps of different methods applied to a simulated data. The x-axis corresponds to SNP positions and the y-axis corresponds to the gene positions. The intensity of a point on the plot represents the significance of the association. The diagonal band represents the cis-effects and the vertical bands represent hotspots. Blue arrows show the locations of real genetic regulatory hotspots, green arrows show those of missing hotspots, and red arrows show those of spurious hotspots. (a)-(f) shows the eQTL map applying the standard t-test, SVA, ICE, LMM-EH, PANAMA, and NICE, respectively.

hotspots that are different between the datasets may be spurious hotspots caused by technical confounding factors present at the time of generation of one of the datasets. In addition, some of these hotspots were further validated by other experimental data such as protein levels [FRS07, PRR07].

To determine which hotspots in the two datasets are regulatory hotspots due to genetic effects, we use the following approach. We first compute a p-value for each gene-SNP pair in both datasets using the standard t-test (Figure 3.4 (a) and Figure 3.4 (b)). We then

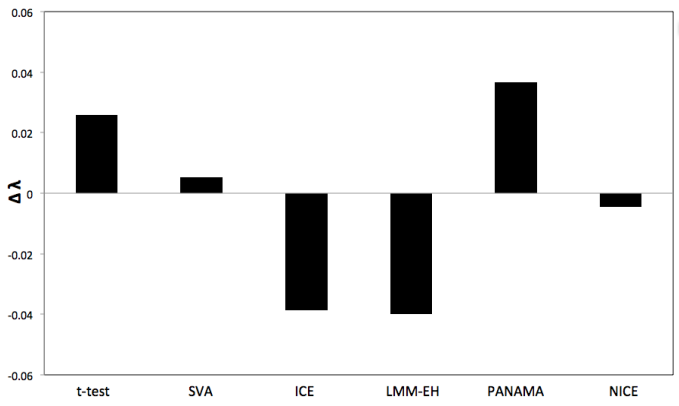


Figure 3.3. Inflation factors of different methods, t-test, SVA, ICE, LMM-EH, PANAMA, and NICE applied to a simulated data. $\Delta\lambda$ is defined as $1 - \lambda$.

Method	Real hotspots	Missing hotspots	Spurious hotspots	Sensitivity	FDR
t-test	5	0	8	1.0	0.62
SVA	5	0	0	1.0	0
ICE	0	5	0	0	NA
LMM-EH	0	5	0	0	NA
PANAMA	5	0	8	1.0	0.62
NICE	5	0	0	1.0	0

Table 3.1. The number of real, missing and spurious hotspots identified by different methods applied to a simulated data and their sensitivities and false discovery rates. Given the number of real (R), missing (M) and spurious (S) hotspots identified, we calculate sensitivity and false discovery rate as $R/(R+M)$ and $S/(R+S)$, respectively.

merge the p-values of two datasets by taking a maximum p-value between the two (Figure 3.4 (c)). The idea is that associations due to true genetic effects are likely to have significant p-values in both datasets while associations due to the confounding effects tend to have a significant p-value in only one of datasets. Thus, by taking the maximum p-value, we can identify the associations that are significant in both datasets. From the merged p-values of two datasets, we identify the top 12 hotspots in terms of their association strength, hotspot level which is described in the following paragraph, and consider them as regulatory hotspots. We call the top 12 hotspots as "putative hotspots". We are in-

terested in the number of hotspots each method recovers among these putative hotspots. The 2008 dataset is of higher quality than the 2005 dataset since it uses a newer version of the array technology. We can verify the relative quality of the datasets by comparing the number of *cis*-eQTLs identified in each dataset which demonstrates that the 2008 dataset is of higher quality. For this reason, we expect that all hotspots originally found in the 2005 dataset that are true effects will be found in the 2008 dataset. Thus hotspots identified in the 2005 dataset but not in the 2008 dataset are likely spurious.

We measure the presence of a hotspot by computing the sum of the log p-values of all of associations of a single marker with the expression level of each gene. This measure called "hotspot level" identifies the hotspots since it captures which SNPs are associated with many gene expression levels. We visualize the hotspot level at the top of our eQTL plots (Figure 3.4). We use the hotspot level to identify the putative hotspots from the merged p-values of two datasets (blue asterisks in Figure 3.5 (a)).

Our goal in this experiment is to identify the true regulatory hotspots and eliminate the spurious hotspots using 2005 yeast dataset [BK05]. We apply each approach to the 2005 data and evaluate the results using knowledge of which hotspots are true hotspots obtained using both the 2005 and the 2008 datasets [BK05,SK08]. We compute the hotspot levels for following methods; the standard t-test, SVA, ICE, LMM-EH, PANAMA, and NICE (Figure 3.5). We consider a SNP as a hotspot if its hotspot level is two standard deviations above the mean. We use this criteria because it provides a reasonable threshold to separate hotspots and noisy peaks. Other thresholds are shown to be inappropriate since using lower thresholds, all methods identify not only most of the putative hotspots but also many spurious hotspots and using higher thresholds, all methods miss most of the putative hotspots. Figure 3.6 shows a plot similar to ROC curve that shows sensitivity and the number of spurious hotspots of different thresholds. The x-axis shows the

number of spurious hotspots and the y-axis shows the sensitivity. For different thresholds, NICE performs the best. We annotate results of each method with blue asterisks, green and red arrows which indicate putative genetic regulatory hotspots predicted from the p-values of merged datasets, missing hotspots and false positive hotspots, respectively. The results show that our method identifies all but two of the putative hotspots while only predicting one spurious hotspots (Figure 3.5 (g)). ICE and LMM-EH makes several false positive predictions and SVA, LMM-EH, and PANAMA miss many hotspots (Figure 3.5 (c) $\sim$ (f)). We note that the t-test has a distinct advantage in this evaluation because p-values from the t-test are used to determine the gold standard (Figure 3.5 (b)) and it is inappropriate to evaluate the t-test in terms of Sensitivity and FDR estimated from the gold standard. Table 3.2 summarizes the results. Figure 3.7 shows the inflation factors of the methods. The standard t-test, SVA and PANAMA show inflation. NICE shows deflation but not as much as ICE and comparable to LMM-EH. Supplementary Figure 3.1 and Supplementary Table 3.1 show the results of the same analysis from the point of view of analyzing the 2008 data and comparing to the hotspots found in the intersection of the 2005 and 2008 datasets. We note that NICE discovers several additional hotspots not identified in the 2005 data which is expected because the 2008 data is of higher quality in general. Below we show that several of these additional hotspots are likely real genetic effects.

Consistent with previous analyses [KYE08, LKS10, FSL12], to compare the statistical power of the methods, we compared the number of *cis* associations reported by the different methods (Figure **??**). NICE is able to identify more *cis* associations than the t-test, SVA, ICE, and PANAMA and identify a comparable number of *cis* associations to LMM-EH. This suggest that NICE is not only able to identify true regulatory hotspots but also increases the general sensitivity of the eQTL detection.

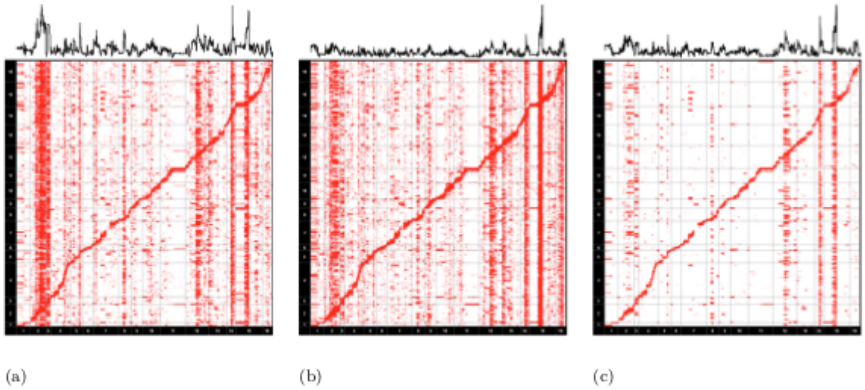


Figure 3.4. eQTL maps of two versions of a yeast dataset that were generated 3 years apart in different locations. The standard t-test is used to generate the p-values. (a) An eQTL map of the yeast dataset generated in 2005 [BK05]. (b) An eQTL map of the yeast dataset generated in 2008 [SK08]. (c) The eQTL map using the maximum p-value of two datasets to show the putative genetic associations in the yeast dataset. Graph on the top of each eQTL map shows the strength of regulatory hotspots by taking the average over all genes of the -log p-values for a given SNP.

Method	Putative hotspots	Missing hotspots	Spurious hotspots	Sensitivity	FDR
t-test	9	3	0	0.75	0
SVA	5	7	0	0.42	0
ICE	8	4	3	0.67	0.27
LMM-EH	2	10	5	0.17	0.71
PANAMA	5	7	0	0.42	0
NICE	10	2	1	0.83	0.09

Table 3.2. The number of putative, missing and spurious hotspots identified by different methods applied to the 2005 yeast data [BK05] and their sensitivities and false discovery rates. Given the number of putative (R), missing (M) and spurious (S) hotspots identified, we calculate sensitivity and false discovery rate as $R/(R+M)$ and $S/(R+S)$, respectively. Note that it is inappropriate to evaluate the t-test in terms of Sensitivity and FDR estimated from putative regulatory hotspots because p-values from the t-test are used to determine the putative hotspots.

3.2.4 NICE discovers novel yeast regulatory hotspots

We reanalyze the 2008 yeast dataset described above using NICE to demonstrate the utility of our approach. The dataset contains expression for yeast strains grown in both

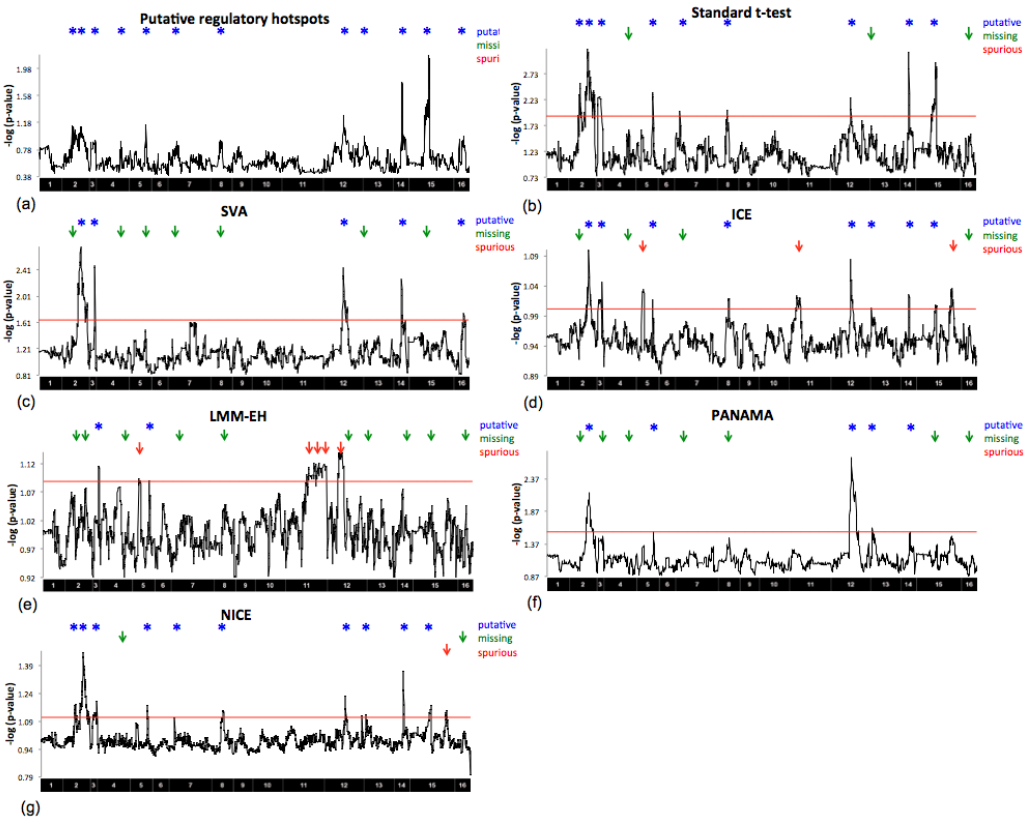


Figure 3.5. Putative, missing and spurious hotspots for the standard t-test, SVA, ICE, LMM-EH, PANAMA and NICE applied to the yeast dataset generated in 2005 [BK05] (a) The average over all genes of the -log of the maximum p-value of the two yeast datasets for each SNP. (b)-(g) The average over all genes of the -log p-value for each SNP for several methods, the standard t-test, SVA, ICE, LMM-EH, PANAMA, and NICE. Blue asterisks show putative genetic regulatory hotspots predicted from merged dataset, green arrows show missing hotspots, and red arrows show spurious hotspots identified by each method. Red horizontal lines show thresholds to select significant peaks which is two standard deviations above the mean. Note that the t-test has a distinct advantage in this evaluation because p-values from the t-test are used to determine the putative regulatory hotspots.

glucose and ethanol media. In our experiments above, we compared the consistency between the 2005 data and the 2008 data both grown in glucose. Here we analyze both conditions in the 2008 data in order to identify both hotspots in each condition as well as hotspots involved in gene by environment interactions consistent with the previous anal-

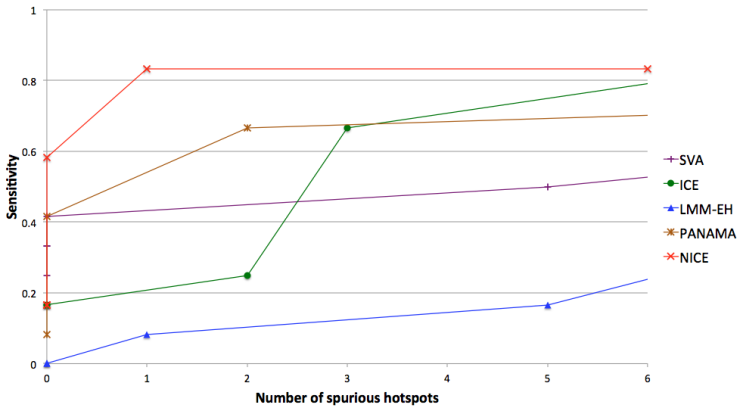


Figure 3.6. Sensitivity and the number of spurious hotspots of different thresholds applied to the yeast dataset generated in 2005 [BK05]. x-axis corresponds to the number of spurious hotspots and y-axis corresponds to the sensitivity. Threshold of the mean, 1 standard deviation above the mean, 2 standard deviation above the mean, and etc. applied.

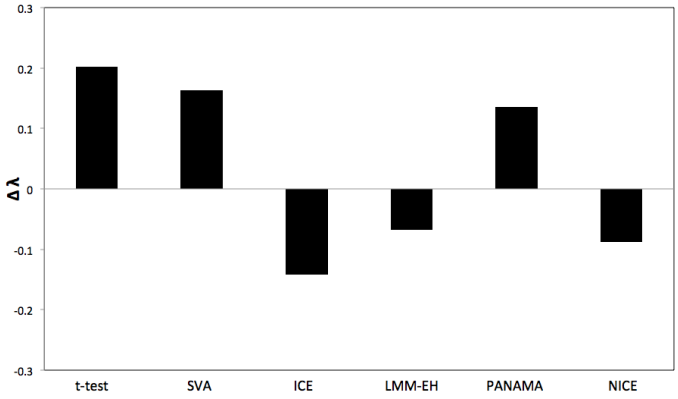


Figure 3.7. Inflation factors of different methods, t-test, SVA, ICE, LMM-EH, PANAMA, and NICE applied to the yeast dataset generated in 2005 [BK05]. $\Delta\lambda$ is defined as $1 - \lambda$.

yses of this data [SK08]. In order to be consistent with the previous analyses, we utilize the method for determining the presence of a hotspot defined in Smith and Kruglyak (2008) instead of the metric we use above. We begin by dividing the yeast genome into 611 20kb bins. For each bin, we count the number of significant trans linkages in the bin. Assuming a Poisson process, the number of expected linkages in each bin is the ratio between the number of trans linkages and the number of bins. For simplicity, we used

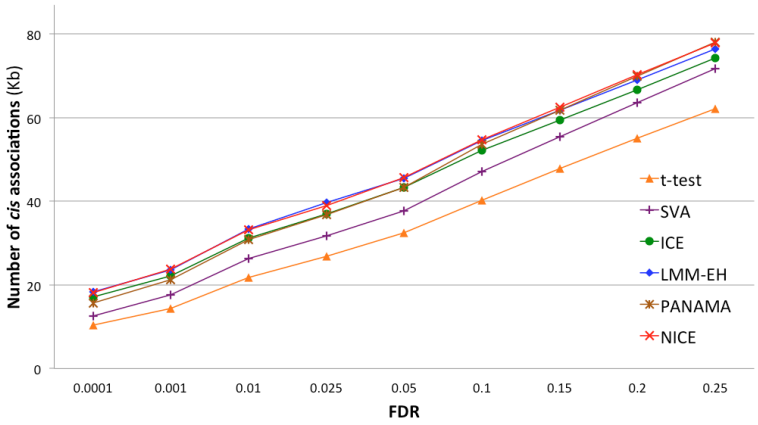


Figure 3.8. The number of *cis* associations in the yeast dataset [BK05] applying the standard t-test, SVA, ICE, LMM-EH, PANAMA, and NICE.

the top 3000 trans associations identified by each method yielding a lambda of 4.9. After adjusting for the number of bins using a Bonferroni correction, a bin is considered to have a statistically significant ($p < 0.05$) if it has > 13 linkages. When we identify significant linkages using a p-value cutoff ($p < 5 \times 10^{-5}$), we achieved almost the same result.

To compare the regulatory hotspots found by various methods, we first define eleven putative regulatory hotspots from a collection of independent experiments using the same parental strains grown in glucose [BYC02, YBW03](Supplementary Table 3.2). Some of these hotspots are expected because of deletions in one of the strains including a hotspot at Chr3:90000 for LEU2 and a hotspot at chr5:110000 for URA3.

We first analyze the glucose data from Simith and Kruglyak (2008). Table 3.3 shows the number of known hotspots captured by each method. We see that both the t-test and NICE capture 9 of the putative regulatory hotspots while ICE captures only 8. Both NICE and ICE capture many more hotspots than the t-test and we wanted to know whether these are spurious or real. For this, we analyzed the 2005 dataset using the same definition of regulatory hotspots as in Smith and Kruglyak (2008) and found that 2/5,

5/14 and 8/17 of the additional hotspots found by the t-test, ICE and NICE replicated, respectively. Those additional hotspots that do not overlap with the 2005 dataset could be specific to the 2008 experiment so we further compared them to the ethanol dataset. We found that 3/5, 13/14 and 15/17 of the additional hotspots found by the t-test, ICE and NICE respectively, replicated in the ethanol experiment. These two results suggest that not only does NICE control for spurious regulatory hotspots, it also discovers more regulatory hotspots that are likely to have a true biological mechanism.

One of the additional hotspots NICE finds to be shared between ethanol and glucose is at Chr7:380000. However, from the t-test results, this hotspot appears to be ethanol specific. We further confirm that this hotspot was also found by NICE in the 2005 data suggesting that it is likely to be a real hotspot that is not condition specific. The two possible candidate genes are RPB9 and MNP1 since the regulatory hotspot is linked in cis to the expression of both of these genes at p-values of 6.8×10^{-6} and 2.2×10^{-4} respectively. RPB9 is a RNA polymerase II subunit that is crucial for transcription fidelity while MNP1 is a putative mitochondrial ribosomal protein that is required for respiratory growth. NICE also finds an additional hotspot at Chr14:1360000 in the glucose data from both 2005 and 2008 but is absent from the ethanol data suggesting that is a glucose specific hotspot. The t-test does not find this hotspot in either glucose datasets. The closest gene is APT2 which is an apparent pseudogene that is not expressed in normal conditions. Interestingly, in our data, we find a strong association between Chr14:1360000 and APT2 in cis at a p-value of 2.3×10^{-17} suggesting that this gene might be functional in a glucose dependent way.

Method	Putative	Missing	Additional	Glucose shared	Ethanol Shared
t-test	9	2	5	2	3
ICE	8	3	14	5	13
NICE	9	2	17	8	15

Table 3.3. The number of putative, missing, and additional hotspots identified by different methods applied to the 2008 yeast dataset grown in glucose media [SK08]. The last two columns show the number of glucose shared and ethanol shared hotspots of the additional hotspots compared to glucose dataset generated in 2005 [BK05] and ethanol dataset generated in 2008 [SK08].

3.3 Discussion

In this chapter, we present a novel approach, NICE, for identifying true genetic regulatory hotspots while eliminating spurious hotspots caused by confounding factors. We leverage the insight that confounding factors are likely to affect the majority of genes, while genetic effects are likely to affect only a smaller subset. This insight allows our approach to distinguish between true and spurious regulatory hotspots. Our approach is related to previous methods that correct for confounding factors such as ICE or SVA which model the global correlation structure and use this structure to correct the association statistics to eliminate the effect of any confounding factors affecting the association statistic. NICE uses only a subset of genes predicted not to be part of the true genetic hotspot to model the global correlation structure between expression levels to correct for confounding factors which eliminates the confounding factors but preserves true hotspots.

We compared to several previous approaches [LS07, KYE08, LKS10, FSL12] on both simulated and real dataset, and demonstrated that our method also achieves higher or comparable statistical power to identify associations while correcting for confounding factors.

While our approach NICE extends the mixed model approach that ICE [KYE08] pro-

posed, in principal, the basic idea behind our approach can be applied to other approaches for correcting for expression heterogeneity such as singular value decomposition (SVD) based SVA approach [LS07]. In that case, for each SNP, a separate SVD would be computed only taking into account genes that are predicted not to be part of a genetic hotspot. Similarly, the techniques presented in LMM-EH [LKS10] can also be adapted in this framework to incorporate multiple variance components to correct for population structure and also correct for bias in estimating the global correlation structure in the presence of population structure.

Our method is based on the assumption that confounding factors are likely to affect the majority of genes, while genetic effects are likely to affect only a subset of the genes. While our approach is an improvement over current methods, in some cases this assumption may be violated, for example, slightly different growth temperatures between batches may result in a specific subset of genes being differentially expressed (e.g. heat shock, cell cycle regulators, etc.). In these cases, our approach would be unable to distinguish those confounding effects from real genetic effects. An additional challenge in eQTL studies is correcting for multiple testing. Possible approaches for multiple testing correction is either applying permutation tests or false discovery rates. Unfortunately, in the case when confounding is present, the confounding causes a violation of the basic assumption necessary for these approaches which is that the individuals in the sample are i.i.d. Shared confounding factors induces complex dependencies among the gene expression patterns of individuals and complicates multiple testing. How to correct for multiple testing in the presence of confounding is a fundamental problem and a promising avenue of future work which is beyond the scope of this dissertation.

3.4 Materials and Methods

3.4.1 Generative model

We assume the following linear mixed model as the generative model of the expression levels,

$$Y = \mu + X\beta + u + e \tag{3.1}$$

Let n be the number of individuals, m be the number of genes and l be the number of SNPs. Y is a $n \times m$ matrix with the gene expression values, μ is a $n \times l$ matrix with the means of expression levels of individuals, X is a $n \times l$ matrix with SNPs encoded by 0 and 1 for haploid and 0, 1, and 2 for diploid, β is a $l \times m$ matrix with their coefficients, and u and e are $n \times m$ matrix with multivariate normal random variables sampled from $N(0, \sigma_g^2 H)$ and $N(0, \sigma_e^2 I)$ accounting for the confounding effects and random errors. Here, H is $n \times n$ covariance matrix that explains intersample correlation structure induced by confounders and I is an $n \times n$ identity matrix. σ_g^2 and σ_e^2 are coefficients of the two variance components.

3.4.2 eQTL mapping

Based on our generative model, equation (3.1), we map eQTL as follows. To test the effect of SNP j on the expression level of gene i, we assume the model

$$y_i = \mu_i + x_j \beta_{ij} + u_i + \epsilon_{ij} \tag{3.2}$$

where y_i is a size n vector denoting gene expression levels of individuals, μ_i is a size n vector denoting the mean of expression levels of individuals, x_j is a size n binary vector denoting SNPs of individuals, $u_i \sim N(0, \sigma_g^2 H)$ is confounding effects, and $\epsilon_{ij} \sim N(0, \sigma_e^2 I)$

is residual errors. The null hypothesis that we want to test is $\beta_{ij} = 0$. Typically, H is defined or estimated before the eQTL mapping. Given estimated $\widehat{H}$, we use the efficient mixed-model association (EMMA) C package [KZW08] to efficiently estimate the variance components (σ_g^2 and σ_e^2). We use the F test as previously suggested on the basis of REML estimates of variance components [KZW08, YPB06, ZAK07]. The challenge in this model is how to estimate $\widehat{H}$ that is close to the true covariance structure of confounding, H.

3.4.3 ICE eQTL mapping

ICE (Intersample Correlation Emended) eQTL mapping approach [KYE08] utilizes global intersample correlation generated from all genes to estimate H. Global intersample correlation matrix from an expression dataset is generated as follows. Let Y be an $m \times n$ expression matrix with n individuals for m genes. Let μ_i, σ_i be the mean and standard deviation of expression values of the ith genes $(Y_{i1}, Y_{i2}, ..., Y_{in})$. Let Z be an $m \times n$ matrix with each element $Z_{ij} = (Y_{ij} - \mu_i)/\sigma_i$. The intersample correlation matrix is defined as the covariance matrix of Z, $\widehat{H} = Cov(Z)$. The estimated intersample correlation matrix $\widehat{H}$ is then used in the linear mixed model in equation (3.2) to correct for the confounding effects.

3.4.4 NICE eQTL mapping

We propose a new eQTL mapping approach called NICE (Next-generation Intersample Correlation Emended) eQTL mapping. NICE builds upon the framework of ICE eQTL mapping but uses a more refined strategy to estimate H, the covariance matrix of confounding effects. The primary limitation of ICE is that it uses the global intersample correlation generated from all genes. If there exists a regulatory hotspot that affects many genes, ICE will overly correct for the confounding and remove the associations to the regulatory hotspot. To overcome this challenge, we must use the genes that are only

affected by the confounding but not by the regulatory hotspots to estimate H. It turns out that segregating these two groups of genes is a highly challenging computational problem.

3.4.4.1 Assumptions

We assume that confounding affects the global correlation structure of the gene expressions thereby affecting most of the genes. This is the standard assumption consistent to previous approaches. We then assume that true regulatory hotspots affect only a subset of the genes. This assumption will be invalid only if a hotspot affects most of the genes, which will be unlikely in practice. Our goal is to separate the genes affected by the true genetic effects from the genes affected only by the confounding. If we can successfully separate them, we will be able to more accurately estimate H using the genes affected only by the confounding. To this end, we make an assumption that the effect size of genetic effects is greater than the magnitude by which the confounding affects the expression levels. That is, we assume that the genes with true genetic effects tend to have more significant results than the other genes affected only by the confounding. This assumption may not be true if the genetic effects are small and the confounding is severe, but in such cases, the noisy data will be highly challenging and in this dissertation we will ignore such cases. Additionally, we assume that the true genetic effects of regulatory hotspot may have a structure. For example, the hotspot may be related to an enhancer element up-regulating many genes, in which case the mean of the effect will be nonzero. Ideally, we would want to use such structures to discriminate the true genetic effects from confounding.

3.4.4.2 Bayesian framework

For our purpose of separating the genes with true genetic effects from the genes with only confounding effects, Bayesian framework fits well because it gives for each gene a posterior probability that the genetic effect will exist or not. Given a SNP that we want to test,

we first apply the standard t-test between the SNP and all genes to obtain the effect sizes and standard errors of the SNP effect with respect to all genes. Let β_i be the estimated effect size of the SNP to gene i and let V_i be the variance of it. We assume a model that

$$P(\beta_i|\text{no genetic effect}) = N(\beta_i; 0, V_i)$$

and

$$P(\beta_i|\text{genetic effect}) = N(\beta_i; \mu, V_i)$$

Note that a few simplifications are employed in this model. First, based on our assumption that the confounding effects are sufficiently smaller than the genetic effects, we approximated the confounding effects as zero. Second, in order to capture the possible structure within genetic effects, we employed the mean term μ. Although this is a simplified model, we found that this approach can capture the majority of the genes affected by the genetic effects, which turns out to be sufficient for our purpose of finding accurate H.
We assume a prior for the effect size

$$\mu \sim N(0, \sigma^2) \ .$$

Let T_i be a random variable which has a value 1 if gene i is affected by the genetic effect of the SNP of interest and a value 0 otherwise. Let π be the prior probability that each gene is affected by the genetic effect such that

$$P(T_i = 1) = \pi, \qquad i = 1, ..., m \ .$$

Then we assume a beta prior on π

$$\pi \sim Beta(\alpha_1, \alpha_2) \ .$$

Let $T = (T_1, ..., T_m)$ be the vector indicating the existence of genetic effect in all genes. Let $\vec{\beta} = (\beta_1, ..., \beta_m)$. Our goal is to estimate the posterior probability that the genetic effect exists for each gene i, namely

$$P(T_i = 1|\vec{\beta}) \ .$$

Notice that T can have 2^m different values. Let $U = \{t_1, ..., t_{2^m}\}$ be the set of those values. By the Bayes' theorem,

$$\begin{aligned} P(T_i = 1|\vec{\beta}) &= \frac{P(\vec{\beta}|T_i = 1)P(T_i = 1)}{P(\vec{\beta}|T_i = 0)P(T_i = 0) + P(\vec{\beta}|T_i = 1)P(T_i = 1)} \\ &= \frac{\sum_{t \in U_i} P(\vec{\beta}|T = t)P(T = t)}{\sum_{t \in U} P(\vec{\beta}|T = t)P(T = t)} \end{aligned} \tag{3.3}$$

where U_i is a subset of U whose elements' ith value is 1. Thus, we should calculate for each t the posterior probability of T,

$$g(t) = P(\vec{\beta}|T = t)P(T = t) \propto P(T = t|\vec{\beta}) \ ,$$

consisting of the probability of $\vec{\beta}$ given T and the prior probability of T.

3.4.4.3 Connection to meta-analytic approach

It turns out that our Bayesian model for eQTL mapping is equivalent to a meta-analysis model although their contexts are different. In a meta-analysis of genetic association studies that combines multiple independent studies, if there exists heterogeneity which refers to the differences in effect sizes of studies [HE11], it is challenging to predict which study has an effect and which study does not. Thus, the problem of finding studies having effect is essentially equivalent to the problem of finding genes having genetic effects in our context. Recently, we have developed an efficient method to solve this problem in the

context of meta-analysis [HE12]. Here we adapt this approach to calculate the posterior probability of T. We briefly describe below how we can calculate $g(t)$.

$g(t)$ consists of the prior probability of T and the probability of $\vec{\beta}$ given T. The prior probability of T is

$$\begin{aligned} P(T=t) &= \int_{-\infty}^{\infty} P(T=t|\pi)p(\pi)d\pi \\ &= \int_{-\infty}^{\infty} \pi^{|t|}(1-\pi)^{m-|t|}p(\pi)d\pi \\ &= \int_{-\infty}^{\infty} \pi^{|t|}(1-\pi)^{m-|t|}\frac{1}{B(\alpha_1,\alpha_2)}\pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}d\pi \\ &= \frac{B(|t|+\alpha_1, m-|t|+\alpha_2)}{B(\alpha_1,\alpha_2)} \end{aligned}$$

where $|t|$ is the number of 1's in t and B is the beta function.

The probability of $\vec{\beta}$ given T is

$$\begin{aligned} P(\vec{\beta}|T=t) &= \int_{-\infty}^{\infty} \prod_{i\in t_0} N(\beta_i;0,V_i) \prod_{i\in t_1} N(\beta_i;\mu,V_i)p(\mu)d\mu \\ &= \prod_{i\in t_0} N(\beta_i;0,V_i) \int_{-\infty}^{\infty} \prod_{i\in t_1} N(\beta_i;\mu,V_i)p(\mu)d\mu \end{aligned} \tag{3.4}$$

where t_0 is the indices of 0 in t and t_1 is the indices of 1 in t. We can analytically work on the integration to obtain

$$\int_{-\infty}^{\infty} \prod_{i\in t_1} N(\beta_i;\mu,V_i)p(\mu)d\mu = \bar{C}\cdot N(\bar{\beta};0,\bar{V}+\sigma^2)$$

where

$$\bar{\beta} = \frac{\sum_i W_i\beta_i}{\sum_i W_i} \qquad \text{and} \qquad \bar{V} = \frac{1}{\sum_i W_i}$$

where $W_i = V_i^{-1}$ is the inverse variance or precision. The summations are all with respect to $i \in t_1$. $\bar{C}$ is a scaling factor such that

$$\bar{C} = \frac{1}{(\sqrt{2\pi})^{m-1}} \sqrt{\frac{\prod_i W_i}{\sum_i W_i}} \exp\left\{ -\frac{1}{2} \left(\sum_i W_i \beta_i^2 - \frac{(\sum_i W_i \beta_i)^2}{\sum_i W_i} \right) \right\} .$$

See Han and Eskin [HE12] for the details of the derivation. As a result, we can calculate $g(t)$ for every t.

3.4.4.4 MCMC

Although we calculate $g(t)$ for each t, it is impractical to perform an exact calculation of $P(T_i = 1|\vec{\beta})$ in equation (3.3) since we have a large number of genes. Thus, we use the following Markov Chain Monte Carlo (MCMC) method [HE12].

1. Start from a random t.
2. Choose a next t, t', based on the moves defined below.
3. If $g(t) < g(t')$, move to t'. Otherwise, move to t' with probability $g(t')/g(t)$.
4. Repeat from step 2.

The set of moves we use for choosing t' is $\{M_1, M_2, ..., M_m\} \cup \{M_{shuffle}\}$. M_i is a simple flipping move of T_i between 0 and 1. $M_{shuffle}$ is a move that shuffles the values of T. At each step, we randomly choose a move from this set assuming a uniform distribution. Other moves can also be used such as moves based on the Bayes factors. We allow n_B burn-in and sample n_S times. After sampling, n_S samples gives us an approximation of the distribution over $g(t)$, which subsequently gives the approximations of equation (3.3). Calculating the posterior probability is the most computationally intensive part of NICE relative to ICE [KYE08] with respect to the running time. By using MCMC, we make dramatic reductions in computational cost which allows NICE to scale to large datasets.

3.4.4.5 NICE intersample correlation matrix

After we calculate the posterior probability that the gene is affected by the true genetic effect of the SNP for each gene, we select genes with probability less than a threshold $\eta = 0.5$. This set of genes represents the genes that are putatively affected only by the confounding. Thus, we use this set of genes to build the intersample correlation matrix $\widehat{H}_{\text{NICE}}$. Then we apply $\widehat{H}_{\text{NICE}}$ to the linear mixed model (3.2) to correct for the confounding in our eQTL mapping. The reason why we choose $\eta = 0.5$ threshold is because we want to approximately find a subset of genes without genetic effects. Although it is ideal to select all the genes without genetic effects, any subset of those genes are likely to capture the global correlation structure as shown in Figure 3.1 (c), and is enough to correct for confounding effects. We choose genes which have an effect with less than 50% chance to select genes that are putatively affected only by the confounding effect. However, we find that unless the threshold is too extreme (e.g. $\eta \leq 0.1$ and $\eta \geq 0.9$), all thresholds yield similar results and the result is robust to the parameter (Supplementary Figure 3.2). We count the number of genes selected by NICE using the posterior probability with threshold $\eta = 0.5$ applied to the yeast data generated in 2005 [BK05] (blue dots in Supplementary Figure 3.3). Except for the putative hotspots, mostly, NICE uses majority of the genes to build the intersample correlation matrix $\widehat{H}_{\text{NICE}}$ similar to ICE [KYE08].

3.4.4.6 Implementation

To calculate the posterior probability in equation (3.3), we used METASOFT [HE11] with prior parameters, $\sigma = 0.05$, $\alpha_1 = 1$, and $\alpha_2 = 5$. We used $\sigma = 0.05$ assuming a small effect size but the choice of the effect size up to 0.4, which is a possible choice of a large effect size in GWAS [**?**, **?**], did not affect the results significantly (Supplementary Figure 3.4). We assume that confounding affects most of the genes while true regulatory hotspots affect only a subset of the genes. Based on the assumption we assume that 20%

of the genes have trans effects for our prior ($\alpha_1/\alpha_2 = 0.2$). We used $\alpha_1 = 1$ and $\alpha_2 = 5$ for diffuse distribution. In practice, changing α_1 and α_2 priors results in similar results as changing the threshold η (data not shown). As shown in the Supplementary Figure 3.2, the results are robust unless the threshold/priors are too extreme. We suggest to use these default priors as they are based on our model assumption if one does not have prior information about a data. We used $n_B = 1,000$ burn-in and $n_S = 1,000,000$ sampling in MCMC. We selected genes with posterior probability less than $\eta = 0.5$. If less than 1% of the genes were selected to calculate the covariance matrix, we used the standard t-test instead of NICE.

3.4.5 *p*-value based approach

Instead of using posterior probability described in previous sections, here we show if a more standard test statistics, such as p-value from the standard t-test could be used for selecting genes without genetic effects to estimate the intersample correlation matrix H. We apply p-value for selecting genes without genetic effects in the following approach. For each SNP, we first order genes based on the p-value obtained using the standard t-test, $\{g_1, g_2, ..., g_m\}$, where g_1 is a gene with the largest p-value and g_m is a gene with the smallest p-value when there are m number of genes. Then we select the first x % of the ordered genes $\{g_1, g_2, ..., g_{\frac{xm}{100}}\}$ as the genes without genetic effects and use expression levels of those genes to estimate H. The following processes are the same as those of NICE. Let's say α is the percentage of genes that have *trans* effects on a *trans*-regulatory hotspot. When we apply this approach to various simulated datasets with different α, $x = 100 - \alpha$ gives the best estimation of H to correct for the confounding effects but retain the true genetic effects (here we show only the case when $\alpha = 20$). However, when we use less ($x < 100 - \alpha$) or more ($x > 100 - \alpha$) genes, we fail to remove confounding effects or fail to retain the true genetic effects, respectively. Supplementary Figure 3.5 (a)$\sim$(c)

show eQTL maps when this approach is applied to our simulated data. Our simulated data has *trans* effects on 20 % of the genes ($\alpha = 20$), in other words, 80 % of genes do not have *trans* effects. Therefore, when we use 80 % ($x = 80$) of the genes to build the intersample correlation matrix H, we are able to correct for the confounding effects but retain the genetic effects. However, when we use less, e.g. 60 % ($x = 60$), of the genes, or more, e.g. 99 % ($x = 99$), of the genes, we either fail to remove confounding effects or fail to retain the true genetic effects, respectively. Unfortunately, we do not know how many genes have *trans* effects for each marker in advance. Moreover, we note that this approach creates many spurious associations other than the ones induced by confounding effects. For example, many horizontal lines appear in the eQTL map (Supplementary Figure 3.5 (a)). This is because when we select x % of genes with the largest p-value, some of the selected genes are shared between many SNPs and this creates spurious associations between the shared genes and the SNPs. We also applied this approach to the yeast dataset generated in 2005 [BK05] using 10 %, 30 %, 50 %, 70 %, and 90 % of the genes. As a result, it misses many putative hotspots as well as makes many false positive predictions (Supplementary Figure 3.6).

Thus, we conclude that p-value is an ineffective approach for selecting genes without genetic effects. On the other hand, the posterior probability that we use for NICE is robust as the value of η neither has significant influence on the results nor is specific to the datasets.

3.4.6 Simulated dataset

We generated a simulated dataset for 1000 genes, 1000 SNPs, over 100 samples based on our generative model, equation (3.1), with $\sigma_g = 0.9$ and $\sigma_e = 0.1$. Assuming haploid, SNPs are encoded by 0 and 1 and randomly generated with minor allele frequency of 30%. A batch effect is simulated as a confounding effect where expression levels in the first half

of samples are correlated with each other, but not correlated with the second half of the samples, and vice versa. 5 randomly selected *trans*-regulatory hotspots are simulated and for each of them, 20% of the genes have *trans* effects of size 0.4 where half have positive effects and the other half have negative effects. *Cis* effect is simulated with size of 0.5.

3.4.7 Yeast datasets

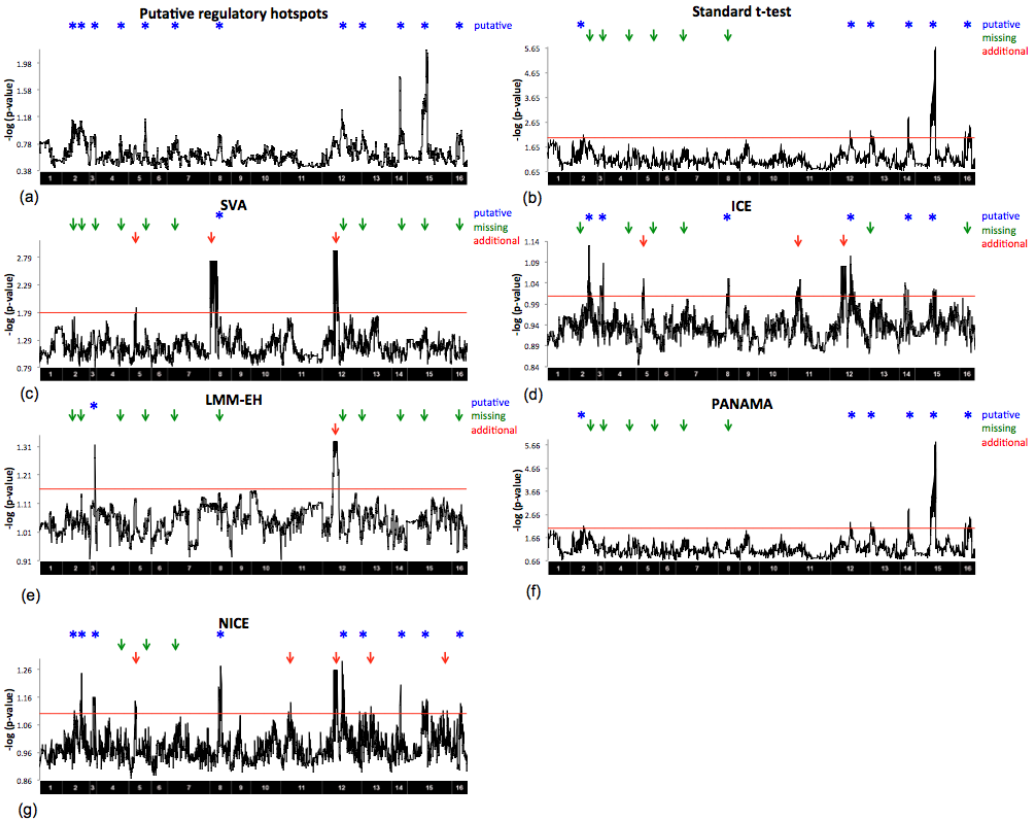
We evaluate our method by utilizing replicate gene expression datasets. We used two versions of a yeast dataset produced 3 years apart at different locations using different microarray platforms. The first data [BK05] was generated in 2005 of which 6,138 probes and 2,956 genotyped loci in 112 segregants are used. The second data [SK08] was generated in 2008 of which 6,138 probes and 2,956 genotyped loci in 109 segregants are used. We classified the eQTL as *cis*-acting when the location of the SNP and the location of the probe are within 50 kb. We showed the number of *cis*-eQTLs for different FDRs where FDRs were calculated using q-value function of R.

3.4.8 Running previous methods

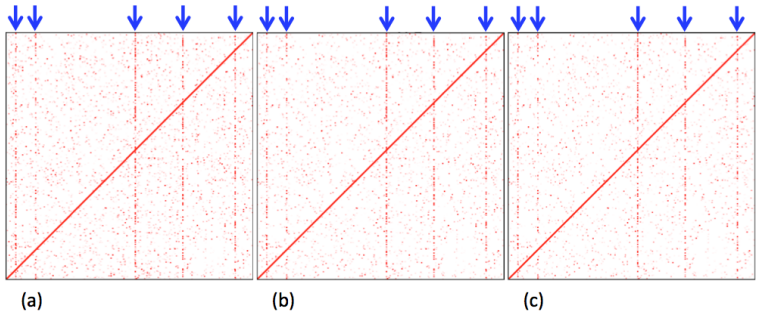
For running previous methods, SVA [LS07], ICE [KYE08], LMM-EH [LKS10] and PANAMA [FSL12], we downloaded the program available from the authors and run the program using default options. For running SVA, 'two-step' method is used. For running LMM-EH, eLMM v1.2 is used for generating covariance matrix K_{EH} and FaST-LMM v2.05 [LLL11] is used for calculating the associations. For running eLMM, ICE covariance matrix is used for initial K_{EH}. REM number of EM steps for each full iteration is set to 3 and REM number of total iterations is set to 10$\sim$20. eLMM provides LMM-EH-PS which corrects for confounding factors as well as population structure. We use LMM-EH instead of LMM-EH-PS because neither of our simulated nor yeast datasets contain population structure. In addition, LMM-EH-PS fails to run in our windows machine with 1.73GHz

Intel Core i7 CPU and 4G RAM.; For running PANAMA, Nicolo Fusi who is the author of PANAMA helped running the program.

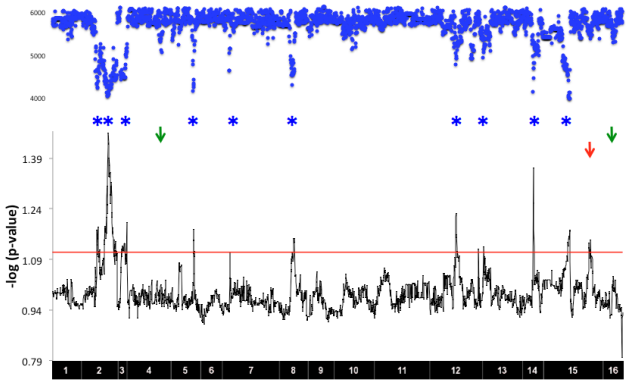
3.5 Supplementary Figure



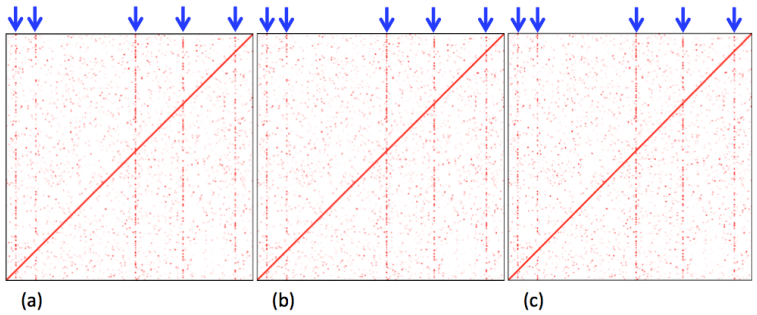
Supplementary Figure 3.1. Putative, missing and additional hotspots for the standard t-test, SVA, ICE, LMM-EH, PANAMA, and NICE applied to the yeast dataset generated in 2008 [SK08]. (a) The average over all genes of the -log of the maximum p-value of the two yeast datasets for each SNP. (b)-(g) The average over all genes of the -log p-value for each SNP for the standard t-test, SVA, ICE, PANAMA and NICE. Blue asterisks show putative genetic regulatory hotspots predicted from merged dataset, green arrows show missing hotspots, and red arrows show additional hotspots. Red horizontal lines show thresholds to select significant peaks which is two standard deviations above the mean. Note that the t-test has a distinct advantage in this evaluation because p-values from the t-test are used to determine the putative regulatory hotspots.



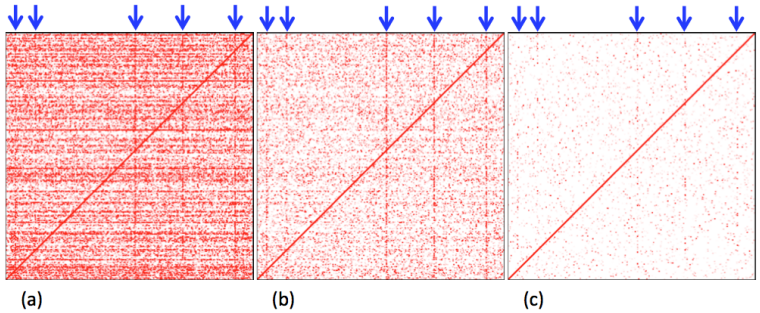
Supplementary Figure 3.2. eQTL maps of NICE using different thresholds applied to a simulated data. (a)-(c) Threshold of $\eta = 0.3$, $\eta = 0.5$, and $\eta = 0.7$ applied, respectively. Blue arrows show the locations of real genetic regulatory hotspots.



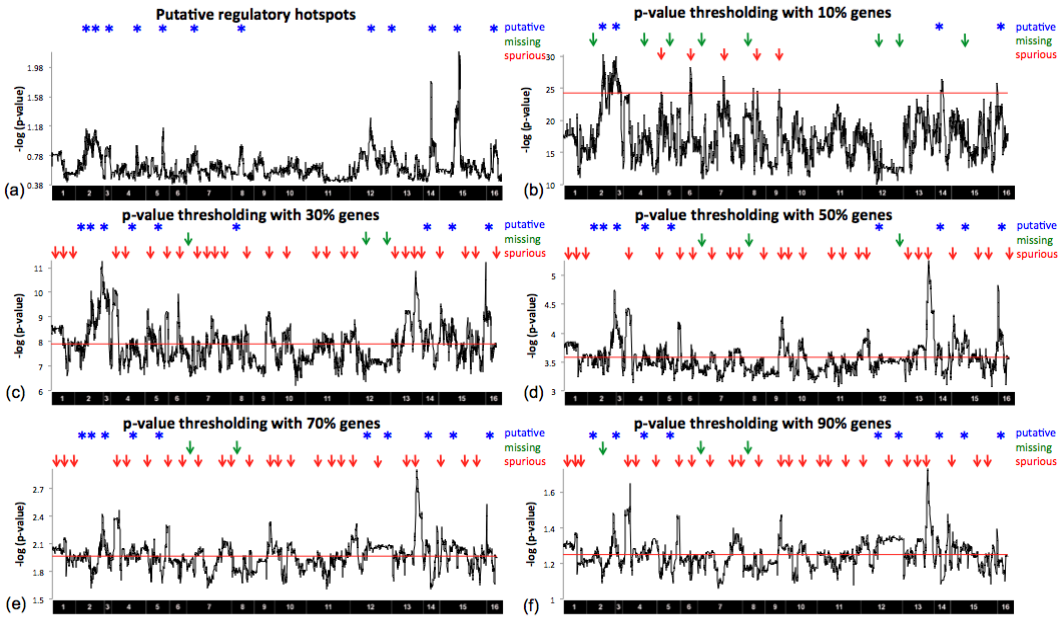
Supplementary Figure 3.3. The number of genes selected by NICE to build the intersample correlation matrix $\widehat{H}_{\text{NICE}}$ applied to the yeast dataset generated in 2005 [BK05]. The bottom plot shows hotspot levels of NICE as in the Figure 3.5 (g). The blue dots on the top of the hotspot levels show the number of genes selected by NICE using the posterior probability less than a threshold $\eta = 0.5$.



Supplementary Figure 3.4. eQTL maps of NICE using different σ values applied to a simulated data. (a)-(c) $\sigma = 0.05$, $\sigma = 0.2$, and $\sigma = 0.4$ applied, respectively. Blue arrows show the locations of real genetic regulatory hotspots. The results of NICE are robust to the prior σ.



Supplementary Figure 3.5. eQTL maps when p-value from the standard t-test is used for selecting genes without genetic effects to build the intersample correlation matrix H applied to a simulated data. (a)$\sim$(c) eQTL maps when 60% ($x = 60$), 80% ($x = 80$), and 99% ($x = 99$) of the genes with the largest p-value are selected, respectively. The simulated data has *trans* effects on 20% of the genes for each *trans*-regulatory hotspot. Blue arrows show the locations of real genetic regulatory hotspots.



Supplementary Figure 3.6. Putative, missing and spurious hotspots when p-value from the standard t-test is used for selecting genes without genetic effects to build the intersample correlation matrix H applied to the yeast dataset generated in 2005 [BK05]. (a) Putative hotspots as in the Figure 3.5 (a). (b)$\sim$(e) eQTL maps when 10% ($x = 10$), 30% ($x = 30$), 50% ($x = 50$), 70% ($x = 70$), and 90% ($x = 90$) of the genes with the largest p-value are selected, respectively.

3.6 Supplementary Table

Method	Putative hotspots	Missing hotspots	Additional hotspots
t-test	6	6	0
SVA	1	11	3
ICE	6	6	3
LMM-EH	1	11	1
PANAMA	6	6	0
NICE	9	3	5

Supplementary Table 3.1. The number of putative, missing, and additional hotspots identified by different methods applied to yeast data generated in 2008 [SK08].

Chromosome	Position	Gene
2	380000	unknown
2	550000	AMN1
3	90000	LEU2
3	200000	MAT
5	110000	URA3
8	90000	GPA1
12	670000	HAP1
12	1000000	SIR3
14	480000	unknown
15	180000	unknown
15	580000	CAT5

Supplementary Table 3.2. List of putative hotspots. We define eleven putative regulatory hotspots from a collection of independent experiments using the same parental strains grown in glucose [BYC02, YBW03].

CHAPTER 4

Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure

4.1 Introduction

Over the past few years, genome-wide association studies (GWAS) have been used to find genetic variants that are involved in disease and other traits by testing for correlations between these traits and genetic variants across the genome. A typical GWAS examines the correlation of a single phenotype and each genotype one at a time. Recently, large amounts of genomic data such as expression data have been collected from GWAS cohorts. This data often contains thousands of phenotypes per individual. The standard approach to analyze this type of data is to perform a GWAS on each phenotype individually, a single-phenotype analysis.

The genomic loci that are of the most interest are the loci that affect many phenotypes. For example, researchers may want to detect variants that affect the profile of gut microbiota, which encompasses tens of thousands of species [LDB96, GRG99]. Another example is the detection of regulatory hotspots in eQTL (expression quantitative trait loci) studies. Many genes are known to be regulated by a small number of genomic regions called *trans*-regulatory hotspots [CLE05, HWW05, WKH04], and these are very

important evidence of the presence of master regulators of transcription. Moreover, a major flaw of the analysis strategy of analyzing phenotypes independently is that this strategy is underpowered. For example, unmeasured aspects of complex biological networks, such as protein mediators, could be captured with many phenotypes together that might be missed with a single phenotype or a few phenotypes [OHP12].

Many multivariate methods have been proposed that are designed to or could be applied to jointly analyze large numbers of genomic phenotypes. Most of the methods perform some form of data reduction, such as cluster analysis and factor analysis [ABB00, Qua01]. However, these data-reduction methods have many issues such as the difficulty of determining the number of principal components, doubts about the generalizability of principal components, etc. [NLS07]. Alternatively, Zapala and Schork proposed a way of analyzing high-dimensional data using multivariate distance matrix regression called Multivariate Distance Matrix Regression (MDMR) analysis [ZS12a]. MDMR constructs a distance or dissimilarity matrix whose elements are tested for association with independent variable of interest. Then, based on the traditional linear models, it tests for the association between a set of independent variables. The method is simple and directly applicable to high dimensional multiple phenotype analysis. In addition, users can flexibly choose appropriate distance matrices [WS06] depending on their experiments.

Each of the previous methods is based on the assumption that the phenotypes of the individuals are independently and identically distributed (i.i.d.). Unfortunately, as has been shown in GWAS studies, this assumption is not valid due to a phenomenon referred to as population structure. Allele frequencies are known to vary widely from population to population, due to each population's unique genetic and social history. These differences in allele frequencies along with the correlation of the phenotype with the populations may cause spurious correlation between genotypes and phenotypes and may induce spurious

associations [KCP02,FRP04,MCP04,COL05,HYH05,RZL05,VP05,BSK06,SSV06,FG06, FE12]. This problem is even more serious when analyzing multiple-phenotypes because this bias in test statistics accumulates from each phenotype which we show in our experiments. Unfortunately, none of the previously mentioned multivariate methods are able to correct for the population structure and may cause a significant amount of false positive results. Recently, multiple-phenotypes analysis methods considering population structure [ZS14,KVS12] have been developed but these methods and related methods are not applicable for large numbers of phenotypes because their computational costs scale quadratically with the number of phenotypes which is impractical.

In this chapter, we propose a method, called GAMMA (Generalized Analysis of Molecular variance for Mixed model Analysis), that efficiently analyzes large numbers of phenotypes while simultaneously considering population structure. Recently, the linear mixed model (LMM) has become a popular approach for GWAS as it can correct for population structure [KYE08, KSS10, LLL11, SAB12, ZS12b, SVP12]. The LMM incorporates genetic similarities between all pairs of individuals, known as the kinship, into their model and corrects for population structure. We take the idea of MDMR [NLS07, ZS12a] that performs multivariate regression using distance matrices to form a statistic to test the effect of covariates on multiple phenotypes and extend it to incorporate linear mixed model in the statistics to correct for population structure.

To demonstrate the utility of GAMMA, using both simulated and real datasets, we compared our method with some of representative previous methods, the standard t-test; one of the standard and the simplest method for GWAS, EMMA [KYE08]; a representative single-phenotype analysis method that implements LMM and corrects for population structure [LLL11, ZS12b], and MDMR [ZS12a]; a multiple-phenotypes analysis method. In a simulated study, GAMMA corrects for population structure and accurately identifies

genetic variants associated with phenotypes. However, previous methods that analyze each phenotype individually do not have enough power to detect the associations and are not able to detect the variants. MDMR [ZS12a] predicts many spurious associations due to population structure. We further applied GAMMA to real datasets. Applied to a yeast dataset, GAMMA could identify most of the regulatory hotspots that are known to be related to regulatory elements in previous study [JSH14], while previous methods fail to detect those hotspots. Applied to a gut microbiome dataset from mouse, GAMMA could correct for population structure and identify biological meaningful variants that are likely to be correlated with taxa. While previous methods either result in significant number of false positives or fail to find any of the variants.

4.2 Results

4.2.1 Correcting for population structure in multivariate analysis

Unlike the traditional univariate analysis that tests an association between each phenotype and each genotype, our goal is to identify SNPs that are associated with multiple phenotypes. Let's say n is the number of samples, m is the number of phenotypes, and we are analyzing an association between ith SNP and m phenotypes. The standard multivariate regression analysis assumes a linear model as follows:

$$\mathbf{Y} = X_i\beta + \mathbf{E}$$

where $\mathbf{Y}$ is an $n \times m$ matrix, where each column vector y_j contains jth phenotype values, X_i is a vector of length n containing genotypes of ith SNP, β is a vector of length m, where each entry β_j contains an effect of ith SNP on jth phenotype, and $\mathbf{E}$ is a $n \times m$ matrix, where each column vector e_j contains i.i.d. residual errors of jth phenotype. Here, we assume that each column of the random effect $\mathbf{E}$ follows a multivariate normal

distribution, $e_j \sim N(0, \sigma_{ej}^2 I)$, where I is an $n \times n$ identity matrix with unknown magnitude σ_{ej}^2.

To test an association between ith SNP and m phenotypes, we test whether any of β_j is 0 or not from the linear model. The standard least-squares solution for $\hat{\beta}_j$ is $(X_i' X_i)^{-1} X_i' y_j$, However, this is problematic when $n \ll m$, which is often the case in genomics data, as there could be many solutions when there are more unknown variables than observations. Alternatively, MDMR [ZS12a] form a statistic to test an effect of a variable on multiple phenotypes by utilizing the fact that the sums of squares associated with the linear model can be calculated directly from a $n \times n$ distance matrix D estimated from $\mathbf{Y}$, where each element d_{ij} reflects the distance between sample i and j. This is because the standard multivariate analysis proceeds through a partitioning of the total sum of squares and cross products (SSCP) matrix, and the relevant information contained in required inner product matrices could be achieved by an $n \times n$ outer product matrix $\mathbf{Y}\mathbf{Y}'$, which could be obtained from any $n \times n$ distance matrix estimated from $\mathbf{Y}$.

However, in GWAS, it has been widely known that genetic relatedness, referred to as population structure, complicates the analysis by creating spurious associations. The linear model does not account for population structure and assuming the linear model may induce many false positive identifications. Moreover, this could cause even more significant problem in multiple-phenotypes analysis because the bias accumulates for each phenotype as their test statistics are summed over the phenotypes (See details in Material and Methods.). Recently, the linear mixed model has emerged as a powerful tool for GWAS as it could correct for population structure. To incorporate effects of population structure, GAMMA assumes a linear mixed model instead of the linear model as follows:

$$\mathbf{Y} = X_i \beta + \mathbf{U} + \mathbf{E}$$

which has an extra $n \times m$ matrix term $\mathbf{U}$, where each column vector u_j contains effects

of population structure of jth phenotype. This is an extension of the following widely utilized linear mixed model for univariate analysis:

$$y_j = X_i\beta_j + u_j + e_j$$

where $u_j \sim N(0, \sigma^2_{gj}K)$ and K is the kinship matrix which encodes the relatedness between individuals and σ^2_{gj} is the variance of the phenotype accounted for by the genetic variation in the sample. Based on the linear mixed model, we perform a multivariate regression analysis through partitioning of the total SSCP matrix to estimate a test statistic for the multiple phenotype analysis. Details of how we perform the inference are described in Materials and Methods.

4.2.2 GAMMA corrects for population structure and accurately identifies genetic variances in a simulated study

Our goal is to detect an association between a variant and multiple phenotypes. A *trans*-regulatory hotspot is a variant that regulates many genes, thus, detecting *trans*-regulatory hotspots is a good applications for GAMMA. To validate that our method eliminates effects of population structure and accurately identifies true *trans*-regulatory hotspots, we generated a simulated dataset that contains true *trans*-regulatory hotspots as well as a complicated population structure. We created a dataset that has 96 samples with 100 SNPs and 1000 gene expression levels. To accommodate population structure, we took SNPs from a subset of a Hybrid Mouse Diversity Panel (HMDP) [BFO10] that contains significant amounts of population structure. To accommodate the *trans*-regulatory hotspots, we simulated 5 *trans*-regulatory hotspots on the gene expression. For each of the trans-regulatory hotspot, we added *trans* effects to 20% of the genes. In addition, we added *cis* effects [MLB09], which are associations between SNPs and genes in close proximity, as they are well-known eQTLs that exist in real organisms.

We applied the standard t-test, EMMA [KYE08], MDMR [ZS12a], and GAMMA on the simulated dataset. We visualized the results of a study in a plot (Figure 4.1), where the x-axis shows SNP locations and the y-axis shows -$\log_{10}$ p-values. As the t-test and EMMA give a p-value for each phenotype, we averaged the p-values over all of the phenotypes for each SNP. On the top of each plot, we marked the locations of the true *trans*-regulatory hotspots with blue arrows. As a result, from the plot we could clearly see that GAMMA successfully identifies the true *trans*-regulatory hotspots without any false positive identifications (Figure 4.1 (d)). However, the standard t-test and EMMA fail to identify the true *trans*-regulatory hotspots as they do not have enough power to detect the associations (Figure 4.1 (a) and (b)). MDMR results many false positive identifications induced by spurious associations due to population structure (Figure 4.1 (c)).

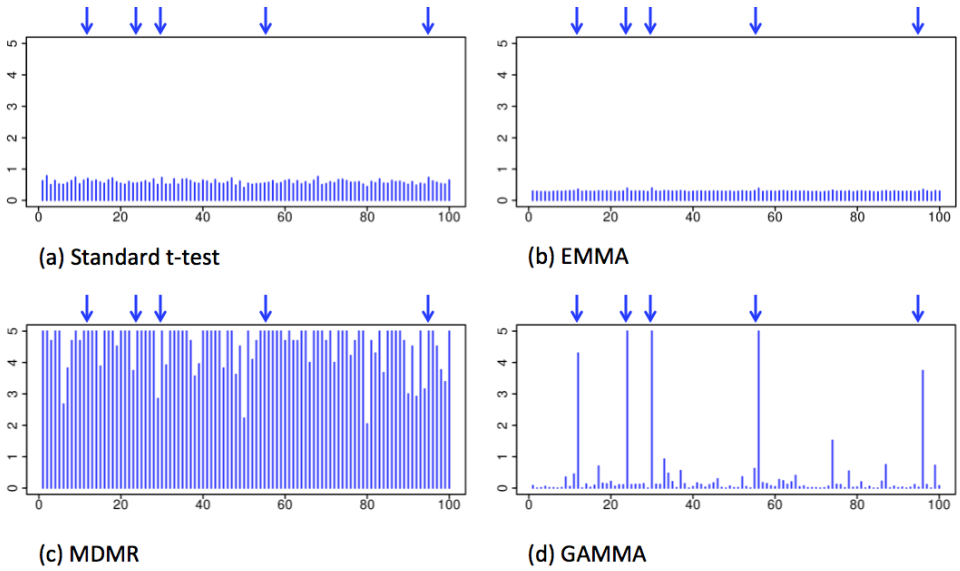


Figure 4.1. The results of different methods applied to a simulated dataset. The x-axis shows SNP locations and the y-axis shows $\log_{10}p$-value of associations between each SNP and all the genes. Blue arrows show the location of the true *trans*-regulatory hotspots. (a) The result of the standard t-test. (b) The result of EMMA. For (a) and (b), we averaged the $\log_{10}p$-values over all of the genes for each SNP. (c) The result of MDMR. (d) The result of GAMMA.

4.2.3 GAMMA identifies regulatory hotspots related to regulatory elements of a yeast dataset

Yeast is one of the model organisms that are known to contain several *trans*-regulatory hotspots. For example, in a well-studied yeast dataset, several hotspots are known to be true genetic effects since they have been validated by additional data such as protein measurements [FRS07, PRR07]. Unfortunately, expression data are known to contain significant amounts of confounding effects from various technical artifacts such as batch effects. To correct for these confounding effects, we applied NICE [JSH14], a recently developed method that corrects for the heterogeneity in expression data, to the yeast dataset and drew an eQTL map shown in Figure 4.2. On the map, the x-axis corresponds to SNP locations and the y-axis corresponds to gene locations. The intensity of each point on the map represents the significance of the association between a gene and a SNP. There are some vertical bands in the eQTL map which represent *trans*-regulatory hotspots. However, it is not easy to tell exactly which ones are the *trans*-regulatory hotspots as the map does not show associations between each SNP and all the genes but only shows associations between each SNP and a single gene.

We applied the standard t-test, EMMA [KYE08], MDMR [ZS12a], and GAMMA to the yeast dataset to detect the *trans*-regulatory hotspots. To remove the confounding effects and other effects from various technical artifacts, we applied genomic control λ which is a standard way of removing unknown plausible effects [DRW01]. The inflation factor λ tells how much the statistics of obtained p-values are departed from a uniform distribution; $\lambda > 1$ indicates an inflation and $\lambda < 1$ indicates a deflation. The λ values are 1.20, 0.86, 3.64 and 0.98 for the t-test, EMMA, MDMR and GAMMA, accordingly. As the yeast dataset does not contain a significant amount of population structure, the λ value is not very big even for the t-test. However, λ value is very big for MDMR which shows that even a small amount of bias could cause significant problem in multiple-phenotypes

analysis. GAMMA could successfully correct for the bias and the λ value for GAMMA is close to 1. Figure 4.3 (a) and (b) show the results of MDMR and GAMMA, accordingly. The x-axis shows locations of the SNPs and the y-axis shows -$\log_{10}$ p-values. The blue stars above each plot show hotspots that were reported as putative *trans*-regulatory hotspots in a previous study [JSH14] for the yeast data. As a result, GAMMA (Figure 4.3 (b)) shows significant signals on most of the putative hotspots. However, MDMR (Figure 4.3 (a)) does not show significant signals on those sites. The t-test and EMMA fail to identify the *trans*-regulatory hotspots as each phenotype are expected to have too small effect that is hard to be detected with a single-phenotype analysis (Supplementary Figure 4.1 in Supplementary).

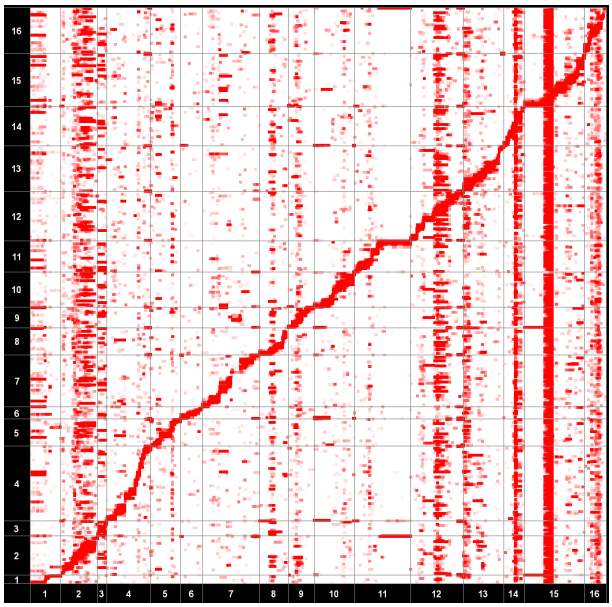


Figure 4.2. An eQTL map of a real yeast dataset. P values are estimated from NICE [JSH14]. The x-axis corresponds to SNP locations and the y-axis corresponds to the gene locations. The intensity of each point on the map represents the significance of the association. The diagonal band represents the *cis* effects and the vertical bands represent *trans*-regulatory hotspots.

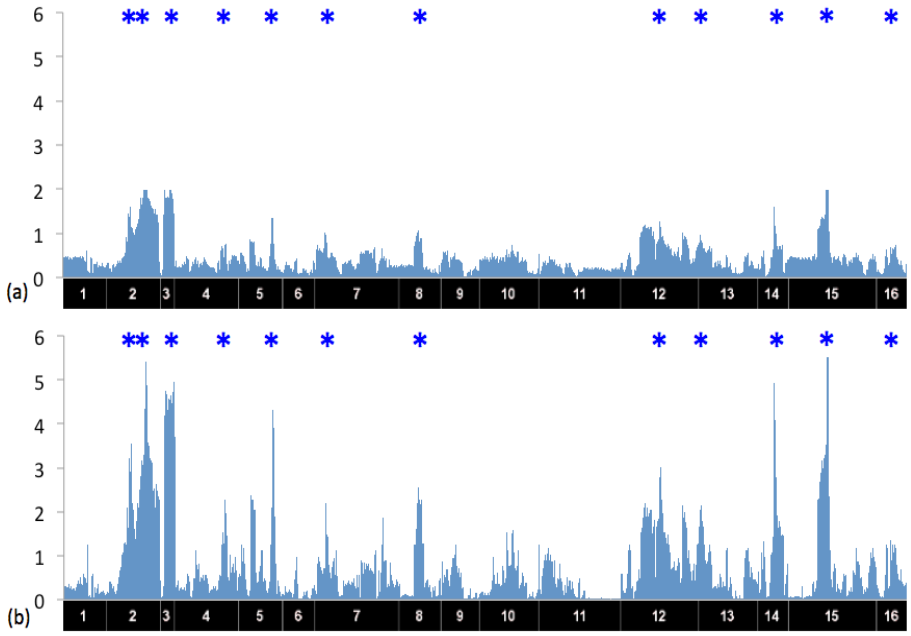


Figure 4.3. The results of MDMR and GAMMA applied to a yeast dataset. The x-axis corresponds to SNP locations and the y-axis corresponds to gene locations. The y-axis corresponds to -$\log_{10}$ of p value. Blue stars above each plot show putative hotspots that were reported in a previous study [JSH14] for the yeast data. (a) The result of MDMR. (b) The result of GAMMA.

4.2.4 GAMMA identifies variants that associated with a gut microbiome

There is an increasing body of evidence that diet and host genetics both affect the composition of gut microbiota, and that shifts in microbial communities can lead to cardio-metabolic diseases such as obesity [LBT05], diabetes [LBT05] and metabolic diseases [KTN13]. Bacteria in a gut constitute a complex ecosystem where most of the interactions in this system are still unknown. There could be clinical overlap between the taxa and some taxa could be co-expressed. The networks between the taxa are complicated and unclear that it is hard to tell a SNP affects a specific taxon but affects jointly many taxa in a profile of the microbiome. For the reason, it would be very useful to perform a multiple-phenotypes analysis for the microbiome data. We applied the standard t-test, EMMA [KYE08], MDMR [ZS12a], and GAMMA on a gut microbiome dataset from HMDP which contains 26 common genus level taxa identified from 592 mice samples, and 197,885 SNPs. Because of the nature of meta-genomics data, the distributions of abundances of species are often highly aggregated or skewed, and there are also usually rare species that contribute many zeros. For the reason, the data is not normally distributed and contains lots of noises for many unknown reasons and we did not apply the genomic control as the λ values are very high except for EMMA which is known to have a deflation problem [LLL11, JSH14].

We applied GAMMA on the dataset (Supplementary Figure 2 in Supplementary). We defined the peaks with p value $\leq 5 \times 10^{-6}$ as the significant ones and we found 9 loci in mouse genome that are likely to be associated with the genus level taxa. Table 4.1 shows the list of the loci and many of these loci contain a number of strong candidate genes based on the literature, overlapping signals with clinical traits and functional variations such as cis-expression quantitative trait loci. For example, chr 1 and 2 loci are the same regions detected with obesity traits in our previous study using the same mice [PNO13]. In addition, global gene expression in epididymal adipose tissue and liver showed a significant

cis-eQTL with genes reside in six out of nine detected loci. On the other hand, MDMR predicts many false positives as mouse data are known to contain significant amounts of population structure. We applied MDMR on one of the smallest chromosome, chr19, and even in the small region, MDMR results 1989 significant peaks out of 5621 loci which shows that MDMR is not applicable for any dataset with population structure (Supplementary Figure 3 in Supplementary). The t-test and EMMA fail to detect significant signals due to the low power (Supplementary Figure 4 in Supplementary).

4.3 Discussion

In this chapter, we present an accurate and efficient method, GAMMA, for identifying genetic variants associated with multiple phenotypes considering population structure. Population structure is a widespread confounding factor that creates genetic relatedness between the samples. This may create many spurious associations between genotypes and phenotypes and results in false identifications. This makes not only the genotypes but also the phenotypes dependent on each other and breaks the i.i.d. assumption of the standard multivariate approaches and makes it inappropriate to apply previous multivariate methods. Moreover, the bias accumulates for each phenotype so when there is a small amount of population structure and even it does not make a big problem in single-phenotype analysis, it could results a serious problem in multiple-phenotypes analysis.

Applied to both a simulated and real datasets including a yeast and a gut microbiome from mouse, GAMMA successfully identifies the variants associated with multiple phenotypes. However [KYE08, ZS12a], other methods either results in many false positives or fail to identify true signals. We applied a pseudo-F statistic that was introduced by Brian H.M. *et al.* (2011), as it provides fast and clear way of estimating a test statistic, especially applicable when the number of phenotypes is much larger than the number of samples, which is often the case in genomics data. However, other appropriate multivariate methods could be applied to GAMMA as well.

There are some complications in comparing results of single-phenotype analysis with those of multiple-phenotypes analysis. We use the average p-value of all the phenotypes for each SNP for the single-phenotype analysis, which could be somewhat a naive way of comparing the results of a single-phenotype analysis and multiple-phenotypes analysis. GAMMA only provides information of whether a set of phenotypes is or is not associated with a SNP but does not provide the information of which phenotypes in a set are asso-

ciated with the SNP. There exist methods for determining which phenotype the SNP is associated with such as the m-values of Han *et al.* [HE12].

4.4 Materials and Methods

4.4.1 Linear Mixed Models

For analyzing ith SNP, we assume the following linear mixed model as the generative model:

$$\mathbf{Y} = X_i\beta + \mathbf{U} + \mathbf{E} \tag{4.1}$$

Let n be the number of individuals and m be the number of genes. Here, $\mathbf{Y}$ is an $n \times m$ matrix, where each column vector y_j contains jth phenotype values, X_i is a vector of length n with genotypes of ith SNP, and β is a vector of length m, where each entry β_j contains an effect of ith SNP on jth phenotype. $\mathbf{U}$ is an $n \times m$ matrix, where each column vector u_j contains the effect of population structure of jth phenotype. $\mathbf{E}$ is a $n \times m$ matrix, where each column vector e_j contains i.i.d. residual errors of jth phenotype. We assume the random effects, u_j and e_j, follow multivariate normal distribution, $u_j \sim N(0, \sigma^2_{gj}K)$ and $e_j \sim N(0, \sigma^2_{ej}I)$, where K is a known $n \times n$ genetic similarity matrix and I is an $n \times n$ identity matrix with unknown magnitudes σ^2_{gj} and σ^2_{ej}, accordingly.

4.4.2 Multiple-phenotypes analysis

Let's say we are analyzing associations between the ith SNP and the jth phenotype. Traditional univariate analysis is based on the following linear model:

$$y_j = X_i\beta_j + e_j \tag{4.2}$$

Here, y_j is a vector of length n with jth phenotype values, X_i is a vector of length n with ith SNP values, β_j is a value contains an effect of ith SNP on jth phenotype and e_j is a vector of length n with i.i.d. residual errors of jth phenotype. To test associations, we test the null hypothesis $H_0 : \beta_j = 0$ against the alternative hypothesis $H_A : \beta_j \neq 0$. We can perform a F-test for the analysis by comparing two models, model 1: $y_j = e_j$ and model 2: $y_j = X_i\beta_j + e_j$. The standard F-statistic is given as follows:

$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(n - p_2)} \tag{4.3}$$

where RSS_1 and RSS_2 are the residual sum of squares (RSS) of model 1 and model 2, accordingly, and p_1 and p_2 are the number of parameters in model 1 and model 2, accordingly. Applying this statistic (Eq. 4.3) to our case, we find the following:

$$\begin{aligned} RSS_1 = y_j'y_j, RSS_2 = (y_j - X_i\hat{\beta}_j)'(y_j - X_i\hat{\beta}_j) = y_j'(I - H_i)y_j = \hat{r}_j'\hat{r}_j \\ RSS_1 - RSS_2 = y_j'y_j - y_j'(I - H_i)y_j = y_j'H_iy_j = \hat{y}_j'\hat{y}_j, p_1 = 1, p_2 = 2 \end{aligned} \tag{4.4}$$

where $\hat{\beta}_j = (X_i'X_i)^{-1}X_i'y_j, H_i = X_i(X_i'X_i)^{-1}X_i'$ and $\hat{r}_j = y_j - \hat{y}_j = y_j - X_i(X_i'X_i)^{-1}X_i'y_j = (I - H_i)y_j$. Applying Eq. 4.4 to Eq. 4.3, we find the following F-statistic:

$$F = \frac{\hat{y}_j'\hat{y}_j/(2-1)}{\hat{r}_j'\hat{r}_j/(n-2)} \tag{4.5}$$

Utilizing the fact that the RSS statistics follow χ^2, we could extend the univariate case into a multivariate case in the following:

$$\mathbf{Y} = X_i\beta + \mathbf{E} \tag{4.6}$$

where $\mathbf{Y}$ is a $n \times m$ matrix, where each column vector y_j contains jth phenotype values, β is a vector of length m, where each entry β_j contains an effect of ith SNP on jth

phenotype, and $\mathbf{E}$ is a $n \times m$ matrix, where each column vector e_j contains i.i.d. residual errors of jth phenotype. Here, we assume that the random effect e_j follows multivariate normal distribution, $e_j \sim N(0, \sigma_{ej}^2 I)$, where I is an $n \times n$ identity matrix with unknown magnitudes σ_{ej}^2. In the multivariate case, both RSS_1 and RSS_2 are $m \times m$ matrices, where the diagonal element $RSS^{j,j}$ is RSS for jth phenotype as calculated in the univariate case. Given this, if we take the trace of this matrix, we obtain a sum of χ^2 statistics. Thus in the multivariate case (Eq. 4.6), we can estimate a pseudo-F statistic as follows:

$$F = \frac{\mathrm{tr}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})/(2-1)}{\mathrm{tr}(\hat{\mathbf{R}}'\hat{\mathbf{R}})/(n-2)} \tag{4.7}$$

where $\hat{\mathbf{R}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - X_i(X_i'X_i)^{-1}X_i'\mathbf{Y} = (I - H_i)\mathbf{Y}$. The reason why we call this a "pseudo" F statistic is because it is not guaranteed that we are summing independent χ^2 statistics, and when they are not independent we do not expect that the result is also χ^2.

Here we note that the trace of an inner product matrix is the same as the trace of an outer product matrix; $\mathrm{tr}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) = \mathrm{tr}(\hat{\mathbf{Y}}\hat{\mathbf{Y}}')$ and $\mathrm{tr}(\hat{\mathbf{R}}'\hat{\mathbf{R}}) = \mathrm{tr}(\hat{\mathbf{R}}\hat{\mathbf{R}}')$. The advantage of this duality is that we can estimate the trace of $\hat{\mathbf{Y}}\hat{\mathbf{Y}}'$ and $\hat{\mathbf{R}}\hat{\mathbf{R}}'$ from the outer product matrix $\mathbf{Y}\mathbf{Y}'$ by utilizing the fact that $\hat{\mathbf{Y}}\hat{\mathbf{Y}}' = H_i(\mathbf{Y}\mathbf{Y}')H_i$ and $\hat{\mathbf{R}}\hat{\mathbf{R}}' = (I - H_i)(\mathbf{Y}\mathbf{Y}')(I - H_i)$. The outer product matrix $\mathbf{Y}\mathbf{Y}'$ could be obtained from any $n \times n$ symmetric matrix of distances or dissimilarities [Gow66, MA01]. Let's say we have a distance matrix D with each element d_{ij}. Let A be a matrix where each element $a_{ij} = (-1/2)d_{ij}$, and we can center the matrix by taking Gower's centered matrix G [Gow66, MA01]:

$$G = (I - \frac{1}{n}11')A(I - \frac{1}{n}11') \tag{4.8}$$

where 1 is a column of 1's of length n. Then this matrix G is an outer product matrix

and we can generate a pseudo F statistic from a distance matrix as follows:

$$F = \frac{\mathrm{tr}(H_i G H_i)/(2-1)}{\mathrm{tr}[(I-H_i)G(I-H_i)]/(n-2)} \tag{4.9}$$

4.4.3 Correcting for population structure

In GWAS, it has been widely known that genetic relatedness, referred to as population structure, complicates the analysis by creating spurious associations. The linear model (Eq. 4.6) does not account for the population structure and applying the model to the multiple-phenotypes analysis may induce false positive identifications. Recently, the linear mixed model has emerged as a powerful tool for GWAS as it could correct for the population structure. To incorporate the effect of population structure, instead of a linear model (Eq. 4.6), GAMMA assumes a linear mixed model (Eq. 4.1) which has an extra term $\mathbf{U}$ accounting for the effects of population structure. This is an extension of the following widely utilized linear mixed model for an univariate analysis:

$$y_j = X_i \beta_j + u_j + e_j$$

Based on the linear mixed model (Eq. 4.1), each phenotype follows a multivariate normal distribution with mean and variance as follows:

$$y_j \sim N(X_i \beta_j, \Sigma_j)$$

where $\Sigma_j = \sigma^2_{g_j} K + \sigma^2_{ej} I$ is the variance of jth phenotype. We compute a covariance matrix, $\hat{\Sigma} = \hat{\sigma}^2_g K + \hat{\sigma}^2_e I$ as described in Implementation and the alternate model is transformed by the inverse square root of this matrix as follows:

$$\hat{\Sigma}^{-1/2} y_j \sim N(\hat{\Sigma}^{-1/2} X_i \beta_j, \sigma^2 I)$$

Thus, to incorporate population structure, we transform genotypes and phenotypes; $\tilde{X}_i = \hat{\Sigma}^{-1/2} X_i$ and $\tilde{y_j} = \hat{\Sigma}^{-1/2} y_j$, and apply them to Eq. 4.9 to get an alternative pseudo-F statistic as follows:

$$F = \frac{\mathrm{tr}(\tilde{H}_i \tilde{G} \tilde{H}_i)/(2-1)}{\mathrm{tr}[(I - \tilde{H}_i)\tilde{G}(I - \tilde{H}_i)]/(n-2)}$$

where $\tilde{H}_i = \tilde{X}_i(\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'$ and $\tilde{G}$ is a Gower's centered matrix estimated from $\tilde{D}$ in turn estimated from $\tilde{\mathbf{Y}}$, where each column vector of $\tilde{\mathbf{Y}}$ is $\tilde{y_j}$.

4.4.4 Implementation

For running GAMMA, we need to compute the covariance matrix $\hat{\Sigma} = \hat{\sigma}_g^2 K + \hat{\sigma}_e^2 I$ and for that we need the estimates of $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$. Let σ_{gj}^2 and σ_{ej}^2 be the two variance components of jth phenotype, where $j = 1, ..., m$. We follow the approach taken in EMMAX [KSS10] or FaST-LMM [LLL11] and estimated σ_{gj}^2 and σ_{ej}^2 in the null model, with no SNP effect. As we take into account multiple phenotypes, a median value of $\hat{\sigma}_{gj}^2$ is used for $\hat{\sigma}_g^2$ and a median value of $\hat{\sigma}_{ej}^2$ is used for $\hat{\sigma}_e^2$ which practically worked well in both of our real datasets. Bray-Curtis measure [BC57, Gow66] is used to calculate the dissimilarity matrix for MDMR and GAMMA. R package vegan is used to estimate the pseudo-F statistics for MDMR and GAMMA. As the distribution of pseudo-F statistic is complicated and does not exactly follows χ^2 distribution as described in the section 4.2., we performed an adaptive permutation for estimating the p-values for MDMR and GAMMA; up to 10^5 permutations for the simulated dataset and 10^6 permutations for the yeast and the microbiome datasets. For running EMMA [KYE08], efficient mixed-model association (EMMA) C package is used.

4.4.5 Simulated dataset

We generated a simulated dataset for 1000 genes, 100 SNPs, over 96 samples based on our generative model (Eq. 4.1) by sampling from a multivariate normal distribution. SNPs are extracted from a HMDP [BFO10] which is a mouse association study panel with significant amounts of population structure. Five randomly selected *trans*-regulatory hotspots are simulated and for each of these, 20% of the genes have *trans* effects of size 1. *Cis* effect is simulated with the size of 2. $\sigma_g^2 = 0.8$ and $\sigma_e^2 = 0.2$ is used.

4.4.6 Real datasets

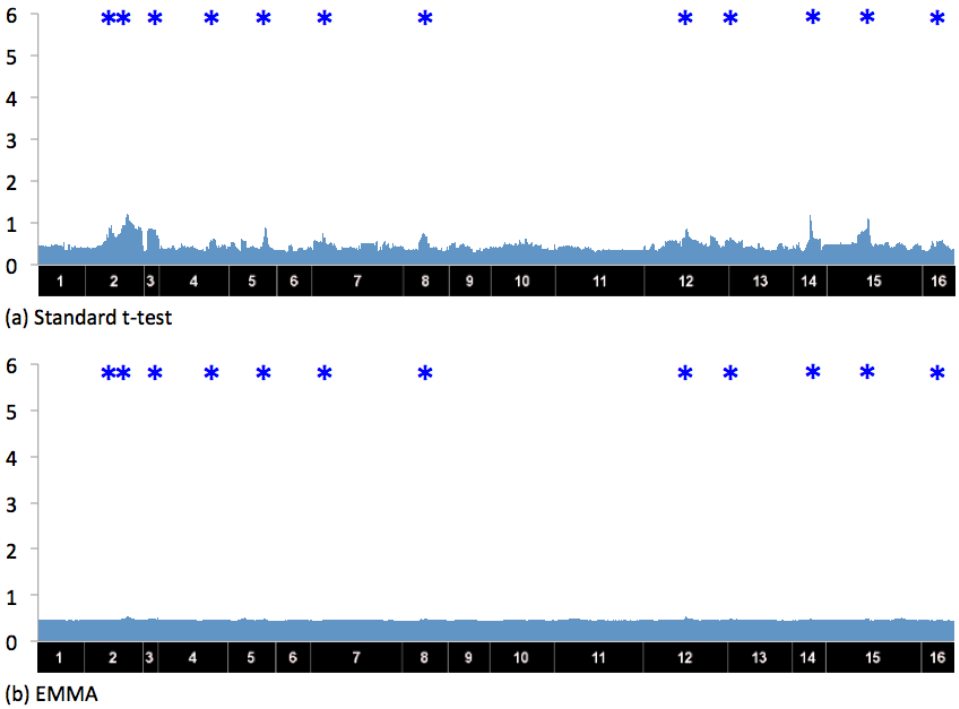
We evaluated our method using a yeast dataset [BK05]. The dataset contains 6,138 probes and 2,956 genotyped loci in 112 segregants. In addition we evaluated our method using a gut microbiome dataset from 592 mice from 110 HMDP strains. The study protocol has been described in detail elsewhere [PNO13]. Bacterial 16S rRNA gene V4 region was sequenced using Illumina MiSeq platform and data was analyzed using established guidelines [BSF13]. The relative abundance of each taxon was calculated by dividing the sequences pertaining to a specific taxon by the total number of bacterial sequences for that sample. We focused on abundant microbes, OTUs with at least 0.01% relative abundance and for genome-wide association study we used 197885 SNPs and 26 genus level taxa. Minor Allele Frequency less than 5% and missing values more than 10% are filtered out. We expect the dataset contains population structure as mouse dataset is one of the known model organism that contains significant amount of population structure. We applied Arcsign transformation on the phenotype values.

Table 4.1. The list of significant associations with a gut microbiome dataset

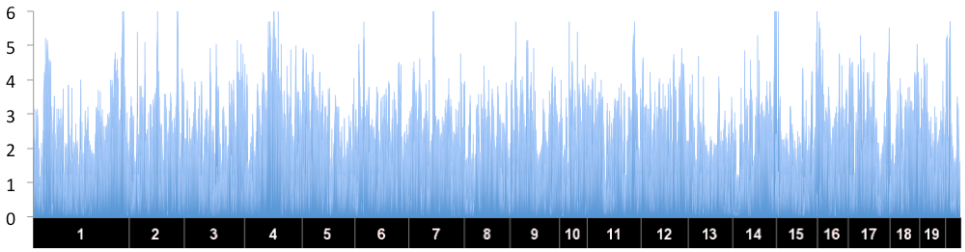
Chr	Peak SNP	Position (Mb)	Associated Region (Mb)	Number of Genes	Clinical QTL	cis eQTL	Overlapping with single Genus GWAS
1	rs31797108	182072111	18.1-18.2	21	body fat % increase		
2	rs27323290	157697578	11.4-15.8	7	food intake, weight	Ctnnbl1	Akkermansia muciniphila
4	rs28319212	95462396	82.1-10.5	74	food intake	Caap1, Ift74	Oscillospira spp.
6	rs50368681	38026365	37.5-38.0	16		Atp6v0a4, Replin1, Zfp467	Sarcina spp.
7	rs33129247	68944648	68.5-71.4	3	TG Gonadal Fat	Nr2f2, Igf1r	Akkermansia muciniphila
11	rs3680824	104011091	10.2-10.4	47		Ccdc85a, Efemp1	
14	rs30384023	120051254	11.9-12.1	5		Dnajc3, Uggt2, Farp1	
16	rs4154709	6236151	62.3-75.0	1			
x	rs29064137	87504122	87.2-88.6	1			

Ctnnbl1, catenin, beta like 1; Capp1, caspase activity and apoptpsis inhibitor; Ift74, intraflagellar transport 74; Atp6v0a4, ATPase, H+ Transporting, Lysosomal V0 Subunit A4; Zfp467, Zinc Finger Protein 467; TG, thyroglobulin; Nr2f2, Nuclear Receptor Subfamily 2, Group F, Member 2; Igf1r, Insulin-Like Growth Factor 1 Receptor; Ccdc85a, Coiled-Coil Domain Containing 85A; Efemp1,EGF Containing Fibulin-Like Extracellular Matrix Protein 1; Dnajc3, DnaJ (Hsp40) Homolog, Subfamily C, Member 3;Uggt2, UDP-Glucose Glycoprotein Glucosyltransferase 2;Farp1, FERM, RhoGEF (ARHGEF) And Pleckstrin Domain Protein 1. Factored Spectrally Transformed Linear Mixed Models (FaST-LMM) [LLL11] is used for single Genus GWAS.

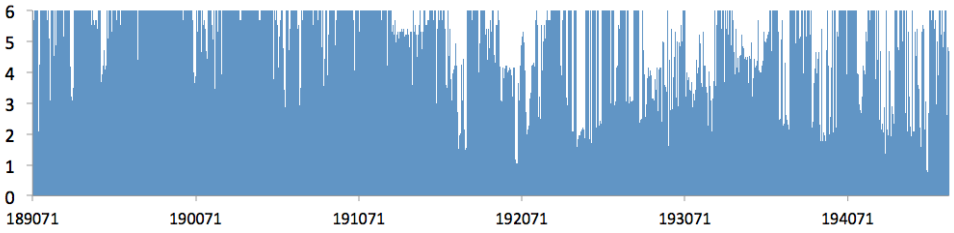
4.5 Supplementary Figure



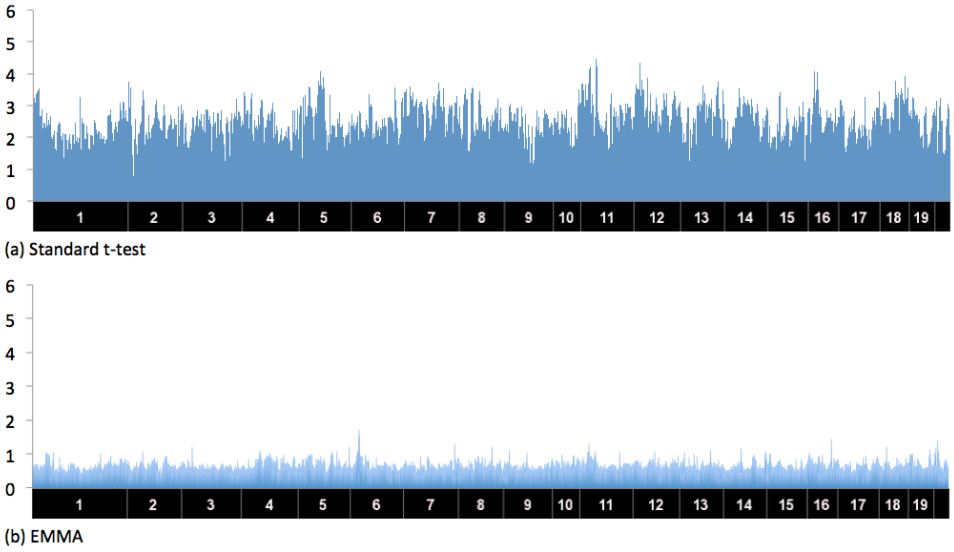
Supplementary Figure 4.1. The results of the standard t-test and EMMA applied to a yeast dataset. The x-axis corresponds to SNP locations and the y-axis corresponds to gene locations. The y-axis corresponds to sum of -$\log_{10}$ of p value over the genes. Blue stars above each plot show putative hotspots that were reported in a previous study [JSH14] in the yeast data. (a) The result of the standard t-test. (b) The result of EMMA.



Supplementary Figure 4.2. The result of GAMMA applied to a gut microbiome dataset. The x-axis corresponds to SNP locations and the y-axis corresponds to gene locations. The y-axis corresponds to -$\log_{10}$ of p value.



Supplementary Figure 4.3. The result of MDMR applied to chromosome 19 of a gut microbiome dataset. The x-axis corresponds to SNP locations and the y-axis corresponds to gene locations. The y-axis corresponds to -$\log_{10}$ of p value.



Supplementary Figure 4.4. The results of the standard t-test and EMMA applied to a gut microbiome dataset. The x-axis corresponds to SNP locations and the y-axis corresponds to gene locations. The y-axis corresponds to sum of -$\log_{10}$ of p value over the genus. (a) The result of the standard t-test. (b) The result of EMMA.

CHAPTER 5

Multiple Testing Correction in Linear Mixed Models

5.1 Introduction

Genome-wide association studies (GWAS) have discovered many variants implicated in complex traits in studies of both humans [HGB07,SRR07,ZWL07,ADL08,MAC08,KAT13, LVB13,ROC13] and model organisms [BK05,SK08,BFO10,FBO11,PGJ11,AVF11,ZKT12, FE12]. In GWAS, both genetic information on variants spread throughout the genome and phenotypic information are collected from a population. The correlation between the genetic information at each variant, referred to as the genotype, and the phenotypic information are assessed to identify the set of variants associated with the trait of interest. GWAS now are routinely performed on tens of thousands of individuals and millions of genetic variants.

One of the major challenges in GWAS is multiple hypothesis testing. Because each GWAS involves computing up to millions of statistical tests, the p-value threshold for significance, referred to as the per-marker threshold, must be adjusted to control the overall false positive rate. The Bonferroni correction [Sid67] assumes independence among the association tests. However, there is a substantial degree of correlation between the association statistics due to a phenomenon called linkage disequilibrium [RCB01], which renders the Bonferroni correction too conservative [GBB10]. The permutation test [WY93], which samples the null distribution of statistics by repeatedly permuting the phenotypes and

computing the association statistics for each permutation, is considered to be the gold standard because it accurately accounts for the correlation structure of the genome at the expense of computational cost. Several strategies aimed at speeding up the computational cost of the permutation test have recently been developed [Lin05, SM05, CB07, HKE09].

Recently, the linear mixed model (LMM) [KYE08, KSS10, LLL11, ZS14, LTB15] has become the standard practice for performing GWAS. The LMM can address two important challenges in GWAS: population structure and insufficient power. Population structure refers to a complex relatedness structure among individuals, which can generate false positives or spurious associations when utilizing traditional association study techniques [KYE08, KSS10]. LMM approaches can avoid these false positives by explicitly modeling these genetic relationships [KYE08, KSS10, LLL11, ZS14, LTB15, JKF15, JSH14]. Moreover, even when there is no population structure, LMM can increase the statistical power of GWAS [LLH13, YZG14, LTB15]. Due to these desirable properties, LMM has become a widely used method in current GWAS [CHP13, HMI14, CGX14, HBM15, FSC13].

However, the current approaches for multiple hypothesis testing correction cannot be applied to LMM. Even the gold standard, the permutation test, is not applicable to LMM, because the underlying idea is that each permutation represents a sample from the null distribution. This is not the case in LMM, because the phenotypes have a covariance structure induced by the complex patterns of relatedness among the individuals. Unfortunately, to date no available approach can correct for multiple testing in LMM, because almost all known multiple testing correction approaches are based on the permutation test and only aim to increase the efficiency of the permutation test [Lin05, SM05, CB07, Bro08, HKE09]. By performing simulations, we demonstrated that multiple testing burden changes with heritability, and that the permutation test inaccurately corrects for the multiple testing when heritability is non-zero.

In this chapter, we first set up the gold standard approach for multiple testing correction in LMM. Our approach is a bootstrapping resampling approach that is the equivalent of permutation test for LMM. Specifically, our parametric bootstrapping approach samples randomized null phenotypes from the distribution fitted by LMM. This approach straightforwardly accounts for the effect of between-individual genetic relatedness on phenotypes. However, similar to the permutation test, this approach is computationally expensive due to the large number of resamplings, and is therefore only suitable to small datasets.

To address this issue, we developed a new approach called multiple testing in transformed space (MultiTrans), which can efficiently correct for multiple testing for LMM. To efficiently approximate the results of parametric bootstrapping, we employ a strategy that directly samples statistics instead of sampling phenotypes. Both sampling phenotypes in bootstrapping and sampling of statistics in our new approach involve sampling from a multivariate normal distribution (MVN). However, the sampling of statistics is much more efficient because the time complexity of the sampling procedure is independent of the number of individuals. To obtain the covariance matrix of the MVN for statistics, previous strategies [Lin05,SM05,CB07,HKE09] that directly use the genotype correlation structure as the covariance matrix cannot be applied, because such a relationship no longer holds under LMM. Therefore, we developed a new approach to overcome this challenge, which transforms genotype dosages into a space where phenotypic correlation between related individuals can be accounted. Finally, to reduce computational cost in GWAS where linkage disequilibrium is expected to be local, we apply the sliding window-based sampling approach [HKE09]. We applied our approach to the Hybrid Mouse Diversity Panel (HMDP) dataset [BFO10], a yeast dataset [SK08], and the HapMap dataset [GBH03]; the results demonstrated that our method can perform multiple hypothesis correction as accurately as parametric bootstrapping, while reducing the time required from months to

hours. Applying our approach to a number of different phenotypes in these real datasets also provided an intuition that the per-marker threshold depends on both the heritability of the trait and the genetic relatedness between individuals. We expect that our method will be widely used to obtain correct per-marker threshold in future studies utilizing LMM.

5.2 Results

5.2.1 Overview of the method

In multiple testing correction, our goal is to find the per-marker threshold that gives an overall false positive rate of α. Let us assume the following linear model:

$$Y = \mu \mathbf{1}_n + X_i \beta_i + \mathbf{e} \tag{5.1}$$

Here, n is the number of individuals, μ is the mean of the phenotypic values, $\mathbf{1}_n$ is a vector of n ones, Y is a vector of length n with the phenotypic values, X_i is a vector of length n with the genotypic values of ith marker, β_i is the coefficient of ith marker, and $\mathbf{e}$ is a vector of length n sampled from $\mathcal{N}(0, \sigma^2 \mathbf{I})$ accounting for the residual errors. Let S_i and S_j be the test statistics for ith and jth markers under the linear model, accordingly. Under the assumption of a linear model (Equation (5.1)), we can derive the equality between covariance of the two statistics, $\texttt{Cov}(S_i, S_j)$, and the correlation of the genotypes, r_{ij}, as follows:

$$\texttt{Cov}(S_i, S_j) \equiv \frac{X_i^T X_j}{\sqrt{X_i^T X_i}\sqrt{X_j^T X_j}} \equiv \texttt{Cor}(X_i, X_j) \equiv r_{ij} \tag{5.2}$$

The derivation of this equality is described in detail in Materials and Methods. This property has been reported in previous studies [HKE09, KLE11, HKK14].

Let m be the number of markers, and Σ be the $m \times m$ covariance matrix between the

statistics whose (i, j)th element is $\Sigma_{i,j} = \texttt{Cov}(S_i, S_j)$. According to the central limit theorem [Was13], the statistics over multiple markers asymptotically follow a MVN. Thus, under the null hypothesis, when n is large, $S_i \sim \mathcal{N}(0, 1)$ and the vector of statistics $(S_1, ..., S_m)$ asymptotically follows a MVN with mean 0 and variance Σ. Figure 5.1 (a) shows a probability density function of bivariate normal distribution at two markers under the null hypothesis. The area outside the meshed rectangle region shows the critical region under the null hypothesis in which, if a p-value falls within this region, the null hypothesis is rejected. Figure 5.1 (b) shows the image when we project the MVN in Figure 5.1 (a) into the xy space. Let u be pointwise p-value that is shown as each point in the MVN. Four corners of the shaded rectangle are $(\Phi^{-1}(u/2), \Phi^{-1}(u/2)), (\Phi^{-1}(1 - u/2), \Phi^{-1}(u/2)), (\Phi^{-1}(u/2), \Phi^{-1}(1 - u/2))$ and $(\Phi^{-1}(1 - u/2), \Phi^{-1}(1 - u/2))$, where Φ is the cumulative density function of the standard normal distribution. Let p_α be the outside-rectangle probability in Figure 5.1 (b). Then, given an overall significance level α, the per-marker threshold is approximated by searching for the pointwise p-value u whose p_α is α. Utilizing the equality in Equation (5.2), the covariance matrix of the MVN could be estimated as $\Sigma = \{r_{ij}\}$ under the linear model (Equation (5.1)). However, in

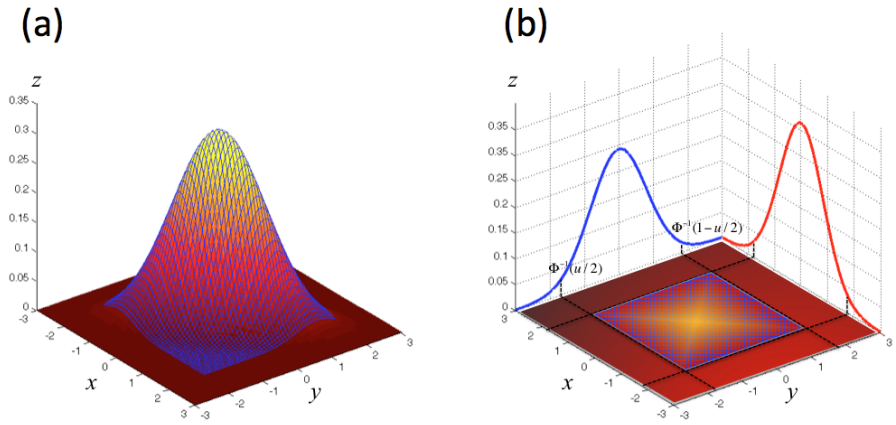


Figure 5.1. Probability density function of a bivariate MVN at two markers under the null hypothesis. (b) shows the image when we project the MVN (a) into the xy space.

LMM, the properties in Equation (5.2) are no longer valid. Let us assume the followning

LMM:

$$Y = \mu \mathbf{1}_n + X_i \beta_i^M + \mathbf{g} + \mathbf{e} \tag{5.3}$$

Here, β_i^M are the coefficients of ith marker under the LMM. LMM has an extra term $\mathbf{g}$ of the linear model (Equation (5.1)), which is a vector of length n sampled from $\mathcal{N}(0, \sigma_g^2 \mathbf{K})$ accounting for the effect of genetic relatedness, where $\mathbf{K}$ is a $n \times n$ kinship matrix that explains the genetic correlation between individuals. Under the LMM, $Y \sim \mathcal{N}(\mu \mathbf{1}_n + X_i \beta_i^M, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I})$ and the equality between covariance of statistics and correlation of genotypes in Equation (5.2) is no longer valid. Let S_i^M and S_j^M be the test statistics under the LMM and $\hat{V} = \hat{\sigma}_g^2 \mathbf{K}, + \hat{\sigma}_e^2 \mathbf{I}$ be the estimated covariance matrix by fitting the data into the LMM. Then, the covariance between the statistics in Equation (5.2) changes as follows:

$$\texttt{Cov}(S_i^M, S_j^M) = \frac{X_i^T \hat{V}^{-1} X_j}{\sqrt{X_i^T \hat{V}^{-1} X_i} \sqrt{X_j^T \hat{V}^{-1} X_j}} \tag{5.4}$$

$$= \texttt{Cor}(\hat{V}^{-1/2} X_i, \hat{V}^{-1/2} X_j) \equiv r_{ij}^M \tag{5.5}$$

That is, the covariance is equivalent to the correlation of the genotype data that is transformed by $\hat{V}^{-1/2}$ (which is why we call our method 'multiple-testing in transformed space', or MultiTrans). The details of the derivation are provided in the Materials and Methods. Note that the covariance of statistics of two markers that are in linkage disequilibrium with each other depends on $\hat{V}$, which in turn depends on the heritability (σ_g^2) of the trait. Thus, heritability affects the covariance of the statistics, which results in different per-marker thresholds. Utilizing the Equation (5.5), we can compute the $\Sigma^M = \{r_{ij}^M\}$ directly from genotypes and sample the test statistics from the MVN with Σ^M to approximate the true null distribution and find the correct per-marker threshold.

To efficiently sample the statistics from the MVN, we make a local linkage disequilib-

rium assumption that the statistics at distant markers are uncorrelated. Under this assumption, we adapt a sliding-window Monte-Carlo approach [HKE09] to accurately and efficiently estimate the per-marker thresholds.

5.2.2 Permutation is inaccurate in LMM

LMM has become one of the standard analysis methods for GWAS [KYE08,KSS10,LLL11, LLH13, ZS14, YZG14, LTB15] because it can explicitly model hidden factors, such as population structure, to avoid false positives, and can also increase the statistical power of the study. However, the permutation test, which has been widely considered to be the gold standard for multiple testing, is not applicable to LMM. The underlying assumption of the permutation test is that if we permute either the genotypes or phenotypes, we can generate the null distribution of our test statistics. However, under the LMM, permutation alters correlations between the individuals specific to LMM, and the correlation is no longer explained by the permuted genotypes or the phenotypes. Thus, applying LMMs to permuted data may result in spurious statistics. Alternatively, we can generate a null distribution for LMM by utilizing parametric bootstrapping, a resampling method that samples null phenotypes from MVN based on LMM and uses them to generate the null distribution (see Materials and Methods for the details of the parametric bootstrapping). A similar approach was used in a previous study of power calculation [KKW10], and it can be thought of as the gold standard approach for LMM.

To show that the permutation cannot approximate the true null distribution for LMM, whereas parametric bootstrapping can do so accurately, we evaluated p-values estimated from the permutation test and those estimated from the parametric bootstrapping for LMM under the null hypothesis. Because the HMDP dataset is known to contain a significant amount of population structure [FE12], we used 100 genotypes and a pheno-

type of low-density lipoprotein (LDL) estimates from this dataset. For the permutation test, we first permuted the phenotype $10,000$ times. Next, we estimated a p-value for each genotype-phenotype pair by fitting the data to the LMM (Equation (5.3)) using a kinship matrix, $\mathbf{K}$, estimated from the whole genome of the HMDP dataset. For parametric bootstrapping, we first fitted the data to the LMM and estimated its parameters, $\hat{\sigma}_g^2 = 0.702$ and $\hat{\sigma}_e^2 = 0.298$. Using these parameters, we sampled $10,000$ null phenotypes from MVN with the covariance matrix, $\hat{V} = \hat{\sigma}_g^2\mathbf{K} + \hat{\sigma}_e^2\mathbf{I}$. Then, we estimated a p-value for each genotype-phenotype pair by fitting the data to the LMM using a kinship matrix, $\mathbf{K}$, estimated from the whole genome of the HMDP dataset. Theoretically, under the null hypothesis, the p-values estimated from each marker and null phenotypes should be uniformly distributed. Thus, for each marker, if we rank $10,000$ p-values estimated from null phenotypes in ascending order, then the 5th-percentile p-value, which is 500th smallest one, should be close to 0.05. Figure 5.2 shows the 5th-percentile p-values estimated from the permutation test (red stars) and the parametric bootstrapping (blue dots). Each point represents the 5th-percentile p-value of a marker. As shown in the figure, the parametric bootstrapping gives accurate p-values very close to 0.05, which demonstrates that parametric bootstrapping can accurately approximate the null distribution for LMM. On the other hand, the permutation test yielded inflated p-values, which demonstrating that the distribution generated from the permutation cannot be used to approximate the true null distribution for LMM.

5.2.3 MultiTrans accurately approximates covariance between test statistics

As shown in the previous section, the parametric bootstrapping closely approximates the true null distribution for LMM, and can thus be used as the gold standard for multi-

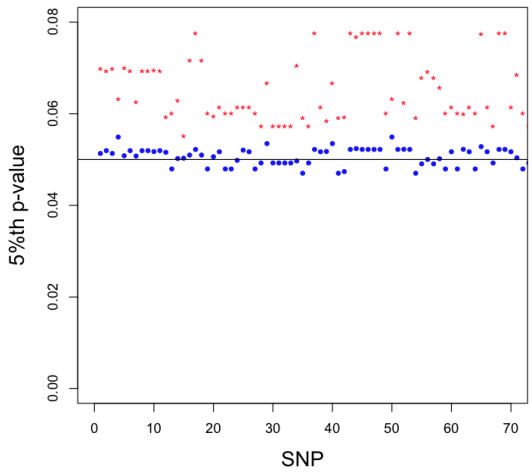


Figure 5.2. 5th-percentile p-values estimated from the permutation test and parametric bootstrapping for LMM under the null hypothesis. One hundred markers and LDL estimates from the HMDP dataset were used. The x-axis shows the markers and the y-axis shows the 5th-percentile p-value. The gray horizontal line shows a p-value of 0.05, each red star shows the 5th-percentile p-value of a marker estimated from the permutation test, and each blue dot shows the 5th-percentile p-value of a marker estimated from parametric bootstrapping.

ple testing in LMM. MultiTrans is rooted on in the idea of parametric bootstrapping. However, to efficiently approximate the results of parametric bootstrapping, MultiTrans samples statistics directly from MVN with a covariance matrix estimated from transformed genotypes. In this section, we show how accurately MultiTrans approximates the covariance matrix of test statistics using the transformation strategy (Equation (5.5)), by testing the difference between the empirical estimate of covariance of test statistics, $\texttt{Cov}(S_i^M, S_j^M)$, and the correlation of transformed genotypes, $\texttt{Cor}(\hat{V}^{-1/2}X_i, \hat{V}^{-1/2}X_j)$, utilizing simulated datasets.

We generated three sets of genotypes, with 100 markers each from the HMDP dataset, a yeast dataset, and the HapMap dataset. 10^5 phenotypes simulated for four different cases each with heritability, 0, 0.2, 0.5, and 0.8, and $\frac{\hat{\beta}}{\hat{\sigma}}\sqrt{N}$ was used as the test statistic. We compared the correlation of the genotypes and covariance of the test statistics be-

fore and after applying the transformation strategy. The term 'heritability' is defined as $\sigma_g^2/(\sigma_g^2+\sigma_e^2)$, which represents the fraction of variance explained by population structure, as in Yang et al [YBM10]. Figure 5.3 shows the histogram of the differences between the covariance of test statistics and the correlation of genotypes, estimated from a simulated dataset of HMDP. Gray bars represent the differences between the covariance of test statistics and the correlation of untranformed genotypes, r_{ij}. Black bars represent the differences between the covariance of test statistics and the correlation of genotypes transformed by the squared root of $\hat{V}^{-1/2}$, r_{ij}^M. As shown in Figure 5.3, the difference is centered at zero when we use tranformed genotypes, regardless of heritability. However, if we do not use transformation, the difference deviates widely from zero as the heritability increases, indicating that the naive genotype correlation cannot effectively approximate the covariance of statistics well. Figure 5.4 shows the scatter plot of the covariance of test statistics (x-axis) and the correlation of genotypes (y-axis). Red and Black dots represent cases in which we did or did not use genotype transformation, respectively. When heritability is zero (Figure 5.3 (a) and Figure 5.4 (a)), the equality in Equation (5.2) holds as expected. However, as the heritability increases (Figure 5.3 (b)-(d) and Figure 5.4 (b)-(d)), the discrepancy between the covariance of statistics and the correlation of genotypes increases. After applying our genotype transformation and using Equation (5.5) to approximate the covariance of statistics, the differences are calibrated back to zero. We applied the same strategy to simulated datasets from the yeast data (Figure 5.5 and Figure 5.6) and HapMap data (Figure 5.7 and Figure 5.8), and obtained consistent results across the three species.

5.2.4 MultiTrans accurately corrects for multiple testing

We examined the accuracy of our method, MultiTrans, for multiple testing in LMM. We compared MultiTrans with three different methods: Bonferroni correction; SLIDE

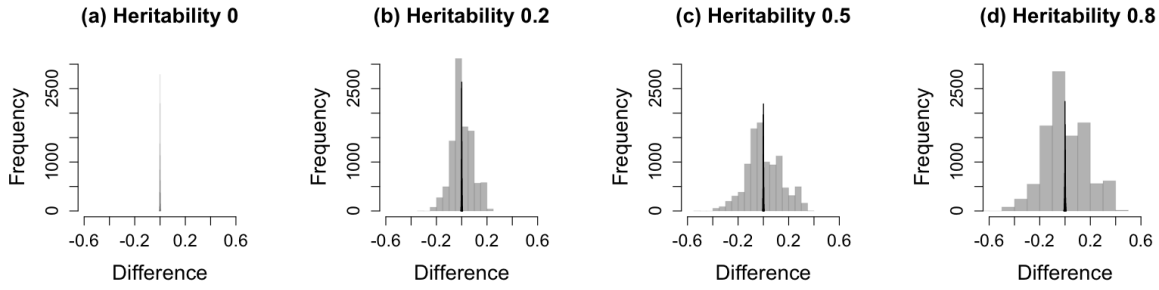


Figure 5.3. Histograms showing the differences between the covariance of statistics and the correlation of genotypes estimated from a simulated HMDP dataset. Heritability: (a) 0; (b) 0.2; (c) 0.5; and (d) 0.8. The x-axis represents the difference between the covariance of statistics and the correlation of genotypes, and the y-axis represents the frequencies. Gray bars represent the differences before applying genotype tranformation, and black bars represent the differences after applying genotype transformation.

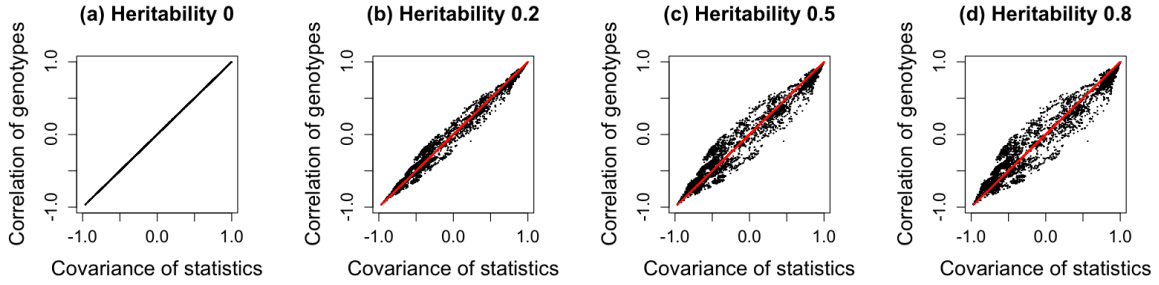


Figure 5.4. Scatter plots showing the covariance of statistics and the correlation of genotypes estimated from a simulated HMDP dataset. Heritability: (a) 0; (b) 0.2; (c) 0.5; and (d) 0.8. The x-axis represents the covariance of statistics, and the y-axis represents the corresponding correlation of genotypes. Red and black dots represent cases in which we did or did not use genotype transformation, respectively.

[HKE09], which is one of the MVN-based multiple testing correction method; and the standard parametric bootstrapping approach. Due to the computational cost of parametric bootstrapping, we applied each method only to chromosome 1 of the HMDP dataset. Table 5.1 shows the per-marker thresholds of different methods at the 5% significance levelWe simulated four different situations, each with heritability 0, 0.2, 0.5, and 0.8. Across the range of heritabilities, MultiTrans yielded very accurate per-marker thresholds very close to those of parametric bootstrapping. On the other hand, the Bonferroni correction gave very stringent thresholds. Previous studies showed that SLIDE closely approximates

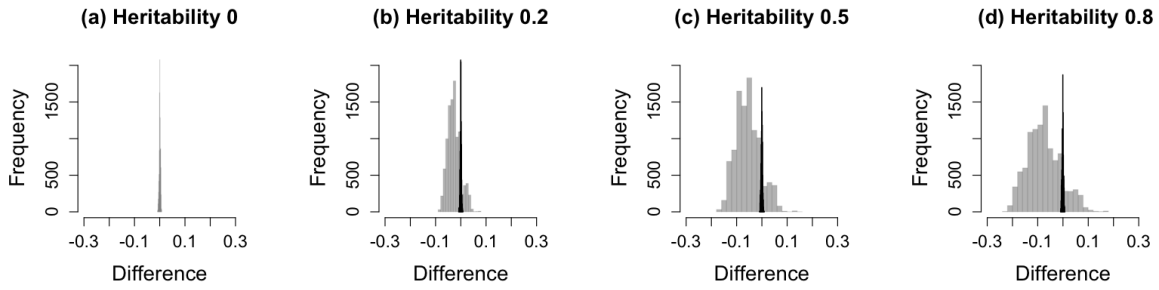


Figure 5.5. Histograms showing the differences between the covariance of statistics and the correlation of genotypes estimated from a simulated yeast dataset. Heritability: (a) 0; (b) 0.2; (c) 0.5; and (d) 0.8. The x-axis represents the difference between the covariance of statistics and the correlation of genotypes, and the y-axis represents the frequencies. Gray bars represent the differences before applying genotype tranformation, and black bars represent the differences after applying genotype transformation.

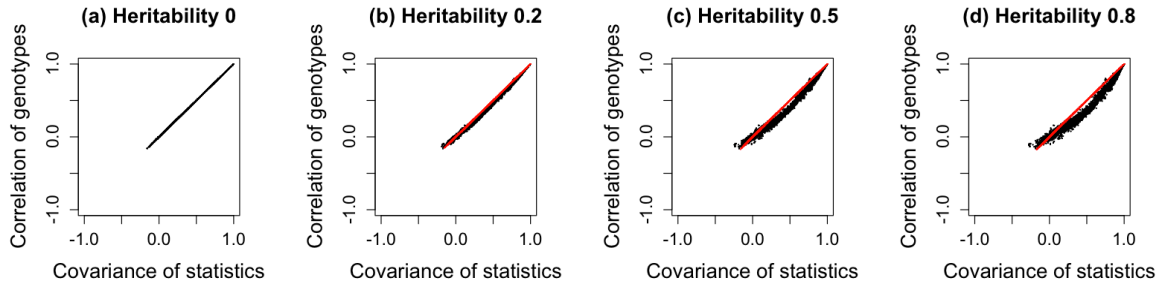


Figure 5.6. Scatter plots showing the covariance of statistics and the correlation of genotypes estimated from a simulated yeast dataset. Heritability: (a) 0; (b) 0.2; (c) 0.5; and (d) 0.8. The x-axis represents the covariance of statistics, and the y-axis represents the corresponding correlation of genotypes. Red and black dots represent cases in which we did or did not use genotype transformation, respectively.

the permutation test and gives accurate per-marker thresholds for the standard linear model [HKE09]. When the simulated heritability is zero, LMM is equivalent to the standard linear model. Thus, it is not surprising that SLIDE gives a per-marker threshold of 6.59E-05, very close to the threshold obtained from parametric bootstrapping, 6.71E-05. However, SLIDE performed worse as the heritability increased. This is expected based on the results in the previous section showing that the discrepancy between the covariance of statistics and the correlation of genotypes increases as the heritability increases if we do not account for phenotype correlations specific to LMM.

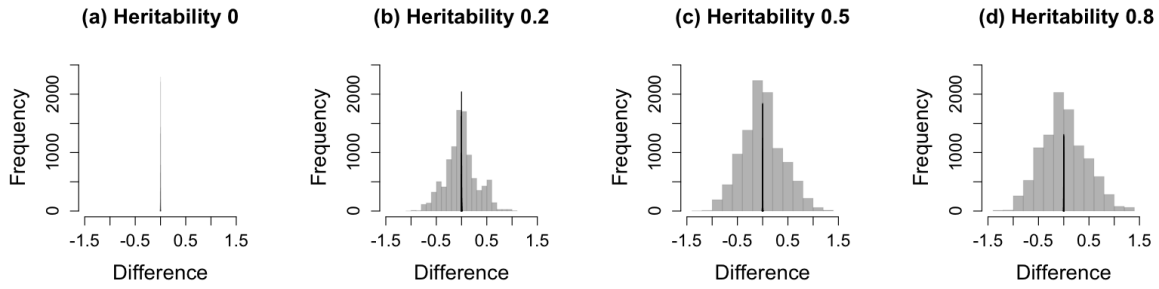


Figure 5.7. Histograms showing the differences between the covariance of statistics and the correlation of genotypes estimated from a simulated HapMap dataset. Heritability: (a) 0; (b) 0.2; (c) 0.5; and (d) 0.8. The x-axis represents the difference between the covariance of statistics and the correlation of genotypes, and the y-axis represents the frequencies. Gray bars represent the differences before applying genotype tranformation, and black bars represent the differences after applying genotype transformation.

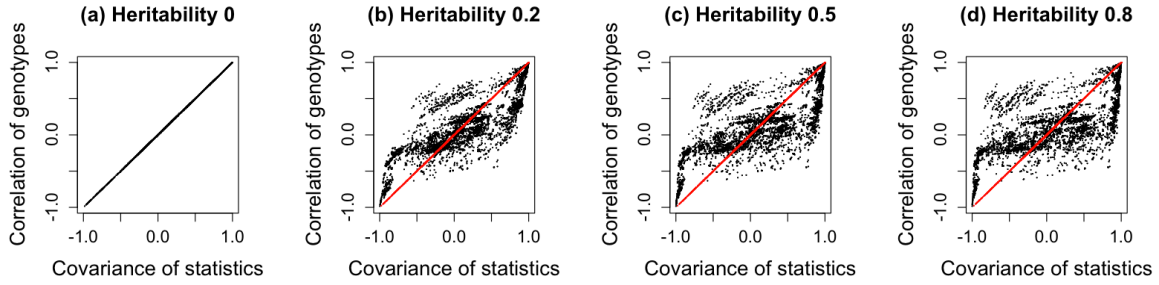


Figure 5.8. Scatter plots showing the covariance of statistics and the correlation of genotypes estimated from a simulated HapMap dataset. Heritability: (a) 0; (b) 0.2; (c) 0.5; and (d) 0.8. The x-axis represents the covariance of statistics, and the y-axis represents the corresponding correlation of genotypes. Red and black dots represent cases in which we did or did not use genotype transformation, respectively.

Heritability	Bonferroni	SLIDE	**MultiTrans**	**Bootstrapping**
0	5.19E-06	6.59E-05	**6.59E-05**	**6.71E-05**
0.2	5.19E-06	6.59E-05	**5.17E-05**	**5.29E-05**
0.5	5.19E-06	6.59E-05	**4.71E-05**	**4.85E-05**
0.8	5.19E-06	6.59E-05	**4.54E-05**	**4.48E-05**

Table 5.1. Per-marker thresholds at the 5% significance level for different simulated heritabilities of 0, 0.2, 0.5, and 0.8, applied to chromosome 1 of the HMDP dataset.

5.2.5 Per-marker threshold depends on both heritability and genetic relatedness

We applied MultiTrans to various datasets from different species and with different heritabilities to see how heritability affects the per-marker thresholds, as well as how per-marker thresholds changes in a dataset-specific manner. Due to the computational cost of parametric bootstrapping, in the previous section (Table 5.1) we tested each method only on chromosome 1, which contains 9629 markers. Taking advantage of the efficiency of MultiTrans, in this experiment we were able to apply MultiTrans to the whole genome in large datasets.

Figure (5.9) shows the per-marker thresholds of the whole genome of the HMDP dataset estimated from MultiTrans for four simulated situations, each with heritability 0, 0.2, 0.5, and 0.8, over a range of significance levels from 0.1% to 10%. The red, blue, green, and orange solid lines show the per-marker thresholds of MultiTrans, and they demonstrate how heritability affected the per-marker thresholds for the HMDP dataset; as the heritability increased the per-marker thresholds decreased. However, this was not reflected in the previous methods, the Bonferroni correction (blue solid line in Figure 5.9) and SLIDE (black dotted line in Figure 5.9), whose per-marker thresholds did not change as the heritability changed.

In addition, we applied MultiTrans to the whole genome of yeast and HapMap datasets. Table 5.2 shows the per-marker thresholds at a significance level of 5%, estimated from MultiTrans for the HMDP, yeast, and HapMap datasets. For each dataset, four different heritabilities (0, 0.2, 0.5, and 0.8) were simulated. For all datasets, the per-marker threshold decreased as the heritability increased. However, the amount that heritability affected the per-marker thresholds differed across the datasets. As heritability changed, the HMDP

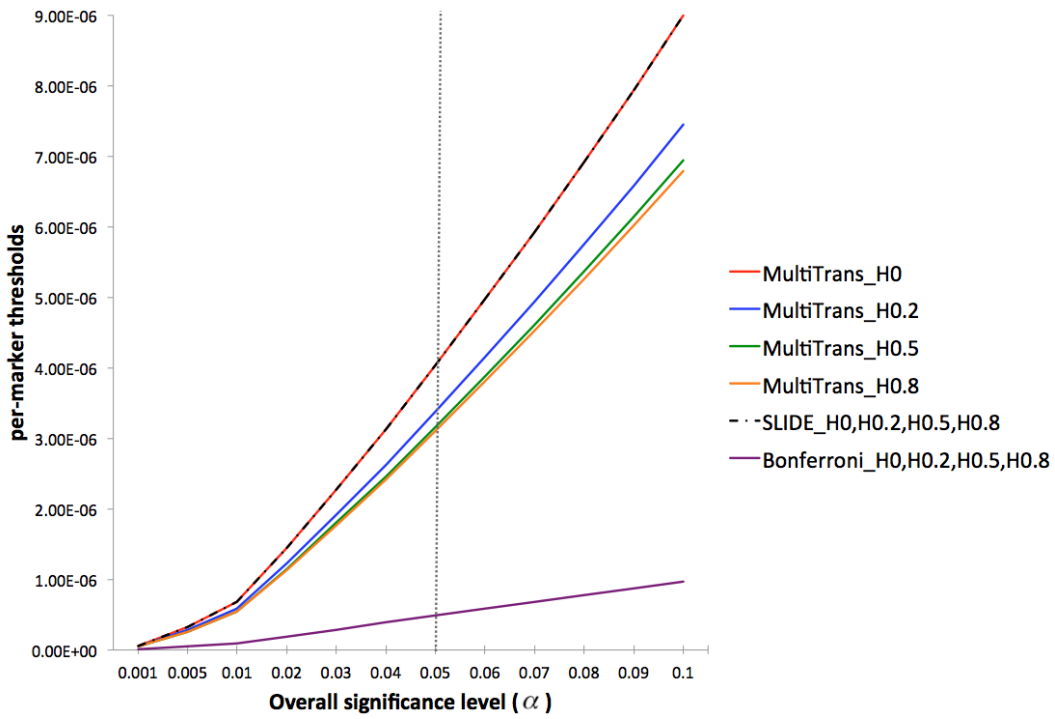


Figure 5.9. Per-marker thresholds for different heritabilities applied to the whole genome of the HMDP dataset. The x-axis represents the overall significance level, α, from 0.1% to 10%. The y-axis represents the corresponding per-marker thresholds. The gray vertical line shows the significance level, 5%. The red, blue, green, and orange solid lines show the result of MultiTrans when heritability is 0, 0.2, 0.5, and 0.8. The purple solid line shows the results of Bonferroni correction for all four heritabilities. The black dotted line shows the result of SLIDE for all four heritabilities.

and yeast datasets exhibited larger differences in their per-marker thresholds than the HapMap dataset. The reason that different datasets show different changes in per-marker threshold given the same changes in heritability is that not only the heritability but also the amount of genetic relatedness in genotypes may affect the per-marker thresholds. For example, if individuals are less related or unrelated in a study, even for a trait that is highly heritable, the correlation of genotypes, r_{ij} (Equation (5.2)), and the correlation of transformed genotypes, r_{ij}^M (Equation (5.5)), may be similar. This is because their kinship matrix $\mathbf{K}$ may be similar to the identity matrix $\mathbf{I}$, and $\hat{V} = \hat{\sigma}_g^2\mathbf{K} + \hat{\sigma}_e^2\mathbf{I} \approx (\hat{\sigma}_g^2 + \hat{\sigma}_e^2)\mathbf{I}$l therefore, the transformation with $\hat{V}^{1/2}$ may not significantly change the correlation between

the genotypes. In this case, the influence of heritability ($\hat{\sigma}_g^2$) on the per-marker thresholds may be very small. Figure 5.10 shows heatmaps of genetic relatedness reflected in kinship matrices for the HMDP, yeast, and HapMap datasets. The color of each pixel represents the strength of the relatedness, with yellow indicating strong correlation between individuals and red indicating no relatedness. Compared to the HDMP and yeast datasets, the HapMap dataset shows smaller relatedness between the individuals. In addition, we show the histogram of off-diagonal values of kinship matrices for the HMDP, yeast, and HapMap datasets (Figure 5.11). The figure shows that the individuals in HapMap are less related to each other than those in the HMDP and yeast datasets.

Datasets / Heritability	HMDP	Yeast	HapMap
0	4.03E-06	5.09E-05	7.29E-08
0.2	3.38E-06	4.65E-05	7.08E-08
0.5	3.16E-06	4.24E-05	7.07E-08
0.8	3.10E-06	3.87E-05	7.06E-08

Table 5.2. Per-marker thresholds at a 5% significance level estimated from MultiTrans for different simulated heritabilities of 0, 0.2, 0.5, and 0.8, applied to the whole genome of HMDP, yeast, and HapMap datasets.

5.2.6 MultiTrans applied to the real traits

Because MultiTrans is efficient and accurate, we were able to apply MultiTrans to a large number of real phenotypes in the HMDP, yeast, and HapMap datasets. As described above, these datasets have different genetic relatedness, and the phenotypes in each dataset have different heritabilities; therefore, each phenotype will have a unique per-marker threshold. Table 5.3 confirms that multiple phenotypes in the three datasets have different per-marker thresholds.

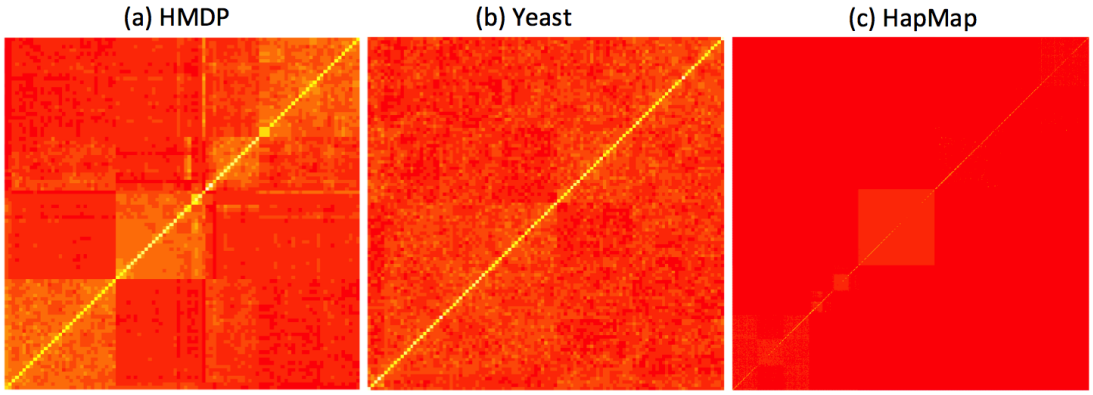


Figure 5.10. Heatmaps of genetic relatedness reflected in a kinship matrix for different datasets. (a) HMDP, (b) yeast, and (c) HapMap. Individuals are ordered from left to right on the x-axis, and from bottom to top on the y-axis. Each pixel of the heatmap shows the strength of the correlation between the individuals, with yellow indicating strong correlation and red indicating no correlation.

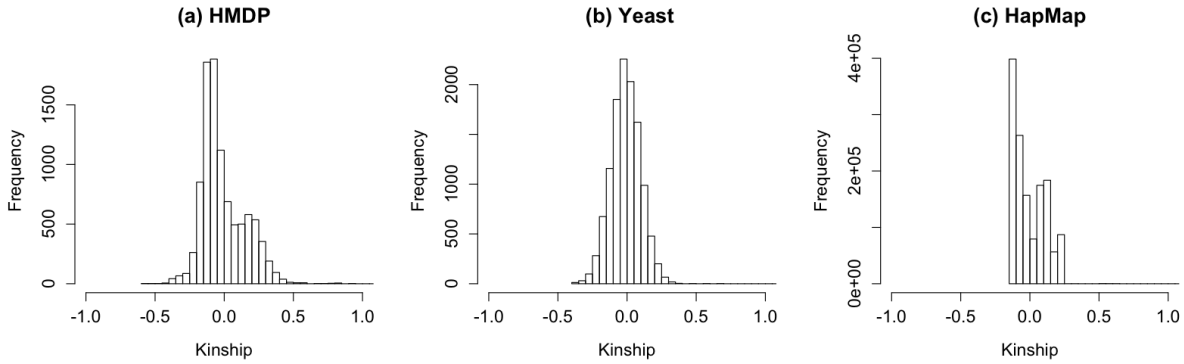


Figure 5.11. Histograms of off-diagonal values of kinship matrix (a) HMDP (b) Yeast (c) HapMap.

5.2.7 Efficiency of MultiTrans

To demonstrate the efficiency of MultiTrans, we compared the running time of MultiTrans and parametric bootstrapping, which can accurately correct p-values for multiple testing in LMM. Both MultiTrans and the parametric bootstrapping must calculate the inverse square root of the covariance matrix $\hat{V}^{-1/2}$ once. However, parametric bootstrapping needs to sample null phenotypes from MVN multiple times and estimate statistics for each

HMDP		
Phenotype	Heritability	MultiTrans;
Thioglycolate treated	0.036	3.91E-06
free fluid	0.653	3.13E-06
low-density lipoprotein	0.706	3.12E-06
glucose	0.735	3.11E-06
mesenteric fat pad (percentage)	0.712	3.10E-06
lean mass	0.865	3.09E-06
fat mass	0.884	3.08E-06
Yeast		
ProbeID	Heritability	MultiTrans
YMR073C	0.010	5.06E-05
YMR242C	0.111	4.82E-05
YLR447C	0.214	4.63E-05
YDR186C	0.310	4.48E-05
YHL012W	0.409	4.34E-05
YOL144W	0.503	4.23E-05
YFL018C	0.615	4.09E-05
YCR107W	0.700	3.99E-05
YMR312W	0.819	3.85E-05
YNL046W	0.911	3.73E-05
HapMap		
ProbeID	Heritability	MultiTrans
ILMN 1756694	0.013	7.28E-08
ILMN 1851657	0.156	7.13E-08
ILMN 1803219	0.225	7.08E-08
ILMN 1741165	0.401	7.07E-08
ILMN 1704746	0.728	7.06E-08

Table 5.3. Per-marker thresholds for various real phenotypes of HMDP, Yeast, and HapMap datasets estimated from MultiTrans.

of them, which takes a lot of time [YZG14,LTB15]. To compare the running time of MultiTrans and parametric bootstrapping, we estimated the running time of both methods utilizing four different datasets; HMDP [BFO10], HapMap [GBH03], 1000Genomes [AAA10], and NFBC (Northern Finland Birth Cohorts) [SSH09], which contains 99, 1184, 2504, and 5326 individuals, respectively. MultiTrans assumes local linkage disequilibrium; that the

statistics outside a range of a window are independent to each other, and apply a sliding-window approach (see Materials and Methods for the details of sliding window approach). The running time of MultiTrans depends on the size of the window, so we applied two different window sizes, 100 and 1000. Figure 5.12 shows the running times of MultiTrans and parametric bootstrapping for different numbers of individuals for $100,000$ markers. For both MultiTrans and the parametric bootstrapping, $10,000$ numbers of samplings were performed, and the running times were extrapolated from one chromosome. When the number of individuals was 5326, the parametric bootstrapping took about 5 months, which is impractical, whereas MultiTrans took only 2.57 hours or 3.71 hours using a window size of 100 or 1000, respectively. Even for 99 individuals, parametric bootstrapping took more than 22 days, whereas MultiTrans took only 13.35 minutes or 1.45 hours using a window size of 100 or 1000, respectively. The result shows that even for a small study, MultiTrans is 2421 times faster or 376 times faster than the parametric bootstrapping using a window size of 100 or 1000, respectively. The discrepancy between the running times of MultiTrans and parametric bootstrapping will increase not only as the number of individuals increases, but also as the numbers of samplings or markers increases (data not shown).

5.3 Discussion

Multiple testing correction is a very well-studied problem in the context of GWAS [Gen92, WY93, GB02, SM05, Lin05, CB07, HKE09], with the most widely utilized approach being the permutation test. In most modern GWAS, LMM is applied to account for the effect of population structure or increase statistical power. Unfortunately, in these studies, the permutation test is not only impractical due to the computational cost [CB07], but the assumptions required for permutation testing are not satisfied under LMM and may lead to spurious associations.

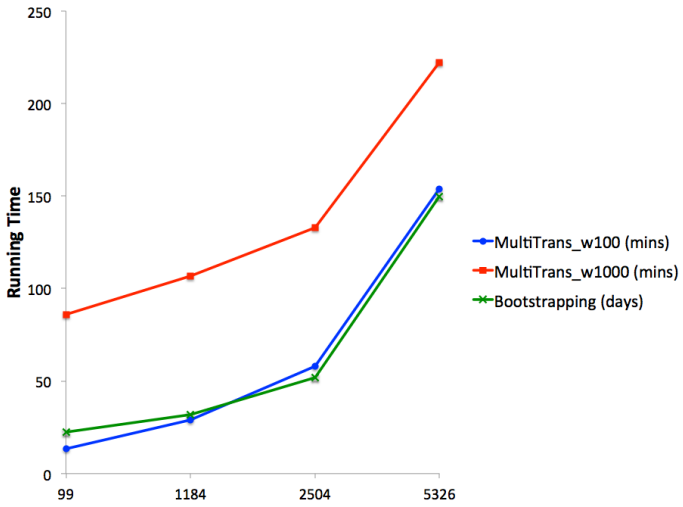


Figure 5.12. Comparison of running time of MultiTrans and the parametric bootstrapping. The running times evaluated for $100,000$ markers and $10,000$ samplingw. The x-axis shows the number of individuals, and the y-axis shows the running time. The blue and red lines show the running times of MultiTrans using window sizes of 100 and 1000, respectively, in minutes. The green line shows the running time of parametric bootstrapping in days.

Here, we showned that the heritability of a trait affects the significance threshold,as well how to perform multiple testing correction in the context of LMM association studies. Our proposed method, MultiTrans, accurately corrects for multiple hypothesis testing and is also efficient, making it applicable to large GWAS. In addition, we demonstrated the accuracy and efficiency of MultiTrans utilizing mouse, yeast, and human datasets.

Theoretically, parametric bootstrapping can be applied to LMM for multiple hypothesis testing. However, this approach is computationally very expensive. Instead of sampling phenotypes, MultiTrans samples statistics directly from a MVN whose covariance matrix is estimated from transformed genotypes and applies a sliding-window Monte Carlo approach to speed up the sampling procedure. Comparing the running time of MultiTrans and parametric bootstrapping, which can accurately correct the p-values for multiple test-

ing in LMM, we showed that the parametric bootstrapping approach is impractical even for a small study, whereas MultiTrans can dramatically reduce the running time.

Our results show that the heritability changes the covariance of statistics and per-marker thresholds. In addition, we made the novel observation that the per-marker threshold tends to decrease as the heritability increases for the HMDP, yeast, and HapMap datasets. We also provided an intuition regarding how genetic relatedness in datasets affects the per-marker threshold. To our knowledge, our study is the first study to explain the relationship between heritability, genetic relatedness, and the per-marker threshold.

The ideas behind our approach extend multivariate normal approaches for modeling the joint distribution of GWAS statistics to scenarios in which mixed models are utilized to compute the association statistics. In this chapter, we demonstrated how this extension can be used to compute the significance threshold for multiple testing correction; however, this framework can be utilized for other applications of MVNs as well. For example, similar extensions can be applied to fine mapping methods [HKK14, KYL14, HKWss], GWAS statistic imputation [LBR13, PZS14], joint testing [ZPG10], follow-up SNP selection [KLE11], etc.

5.4 Material and Methods

5.4.1 Previous multiple testing correction methods for non-LMM

5.4.1.1 Permutation test

The permutation test gives a simple way to compute the null sampling distribution for a test statistic by repeatedly permuting either genotypes or phenotypes and computing the association statistic for each permutation. The permutation test can be thought of as a re-sampling approach that samples individuals from a uniform distribution without replacement. The permutation test accurately accounts for the correlation structure of the genome, and therefore, has been used as the gold standard for GWAS. However, it is computationally expensive, and its running time is linearly dependent on the number of individuals.

5.4.1.2 Methods using multivariate normal approximation

Several previous studies proposed alternative approaches to permutation because the permutation test is computationally expensive especially when the number of individuals is large. The idea underlying these approaches is sampling of test statistics directly from MVN, taking advantage of the the fact that the statistics over multiple markers asymptotically follows a MVN [SM05, Lin05].

Below, we show how to obtain the covariance matrix of the MVN. Let m be the number of markers, S_i be a statistic for the ith marker, and $\Sigma = \{\texttt{Cov}(S_i, S_j)\}$ be the $m \times m$ covariance matrix between the statistics. Assuming the following linear model, we can

derive the covariance matrix for the MVN.

$$Y = \mu \mathbf{1}_n + X_i \beta_i + \mathbf{e}$$

Here, n is the number of individuals, μ is a mean of the phenotypic values, $\mathbf{1}_n$ is a vector of n ones, Y is a vector of length n with the phenotypic values, X_i is a vector of length n with the genotypic values of ith marker, β_i is their coefficients, and $\mathbf{e}$ is a vector of length n sampled from $\mathcal{N}(0, \sigma_e^2 \mathbf{I})$ accounting for the residual errors. Here, we assume that Y and X_i are normalized as mean 0 and variance 1. Then, the phenotype follows a MVN with a mean and variance as follows:

$$Y \sim \mathcal{N}(\mu \mathbf{1}_n + X_i \beta_i, \sigma_e^2 \mathbf{I})$$

The ordinary least-squares solution of β for the ith and jth marker are as follows:

$$\begin{aligned}
\hat{\beta}_i =& (X_i^T X_i)^{-1} X_i^T Y \sim \mathcal{N}\left(\beta_i, \frac{\sigma_e^2}{X_i^T X_i}\right) \\
\hat{\beta}_j =& (X_j^T X_j)^{-1} X_j^T Y \sim \mathcal{N}\left(\beta_j, \frac{\sigma_e^2}{X_j^T X_j}\right)
\end{aligned}$$

The statistics of the two markers are computed as follows:

$$\begin{aligned}
S_i =& \frac{\hat{\beta}_i}{\hat{\sigma}_e} \sqrt{X_i^T X_i} \sim \mathcal{N}\left(\beta_i \frac{\sqrt{X_i^T X_i}}{\sigma_e}, 1\right) \\
S_j =& \frac{\hat{\beta}_j}{\hat{\sigma}_e} \sqrt{X_j^T X_j} \sim \mathcal{N}\left(\beta_j \frac{\sqrt{X_j^T X_j}}{\sigma_e}, 1\right)
\end{aligned}$$

Here, the estimated values for μ, $\mathbf{e}$, and σ for the ith marker are as follows: $\hat{\mu} = \frac{\mathbf{1}_n^T X_i}{X_i^T X_i}$, $\hat{\mathbf{e}} = Y - \hat{\mu}\mathbf{1}_n - X\hat{\beta}$ and $\hat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}^T\hat{\mathbf{e}}}{n-2}}$. Then, we can prove that the covariance of the two statistics, $\texttt{Cov}(S_i, S_j)$, is equal to the correlation between the genotypes, r_{ij}, as follows [HKE09, HKK14, HKWss]:

$$\begin{aligned}
\texttt{Cov}(S_i, S_j) =& \texttt{Cov}\left(\frac{\hat{\beta}_i}{\sigma_e}\sqrt{X_i^T X_i}, \frac{\hat{\beta}_j}{\sigma_e}\sqrt{X_j^T X_j}\right) \\
=& \frac{1}{\sigma_e^2}\texttt{Cov}\left(\frac{X_i^T Y}{\sqrt{X_i^T X_i}}, \frac{X_j^T Y}{\sqrt{X_j^T X_j}}\right) \\
=& \frac{X_i^T X_j}{\sqrt{X_i^T X_i}\sqrt{X_j^T X_j}} \\
=& \texttt{Cor}(X_i, X_j) = r_{ij}
\end{aligned} \tag{5.6}$$

Previous studies showed that this relationship between genotype correlation and MVN covariance holds for binary traits as well, using different ways of derivations [SM05,HKE09].

Using the properties of Equation (5.6), we can sample the statistics directly from the MVN with mean 0 and variance $\Sigma = \{r_{ij}\}$ instead of permuting phenotypes. In fact, in this sampling, phenotype information is not needed. Specifically, under the null hypothesis, by the multivariate central limit theorem [Was13], if the number of individuals, n, is large, $S_i \sim \mathcal{N}(0, 1)$ and the vector of statistics $(S_1, ..., S_m)$ asymptotically follows a MVN with mean 0 and variance Σ. Given a pointwise p-value u, let $R(u)$ be the m-dimensional rectangle with corners $\Phi^{-1}(u/2)\mathbf{1}_m$ and $\Phi^{-1}(1-u/2)\mathbf{1}_m$, where Φ is the cumulative density function of the standard normal distribution and $\mathbf{1}_m$ is the vector of m ones. Then, the significance level p_α is approximated as the outside-rectangle probability as shown in

Figure 5.1,

$$p_\alpha = 1 - \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma|^{\frac{1}{2}}} \int_{R_{(u)}} e^{-\frac{1}{2}X^T\Sigma^{-1}X_{dX}} \tag{5.7}$$

Thus, given an overall significance threshold α, the per-marker threshold can be approximated by searching for a pointwise p-value u whose significance level p_α is α.

5.4.2 Multiple testing correction methods for LMM

5.4.2.1 Parametric bootstrapping re-sampling approach

Because no available approach can correct for correcting for multiple testing in LMM, we first set up the gold standard approach, which is the equivalent of the permutation test for LMM. We emphasize that the traditional permutation test and its variations do not work for LMM. The idea underlying permutation testing is that each permutation is a sample from the null distribution, which is not the case in LMM, because the permutation alters the dependency of the phenotype on the relatedness structure. If we permute phenotypes, relatedness structure between the individuals and its effect on phenotype is ignored, which can lead to an inflation of p-values.

We propose a resampling-based multiple hypothesis testing approach for LMM, which utilizes the parametric bootstrapping strategy. Figure 5.13 (a) shows an overview of the parametric bootstrapping applied to multiple hypothesis testing described as follows. First, by fitting to LMM, we estimate parameters $\hat{\sigma_g}^2$ and $\hat{\sigma}_e^2$ to generate a covariance matrix of the data, $\hat{V} = \hat{\sigma}_g^2\mathbf{K} + \hat{\sigma}_e^2\mathbf{I}$. Second, we sample size-n vectors of null phenotypes from the distribution from MVN with the covariance matrix $\hat{V}$. Third, using each size-n vector of those null phenotypes, we compute null statistics $(S_1, S_2, \cdots, S_m)$. This parametric bootstrapping approach can be thought of as the permutation-equivalent for LMM. A similar approach was used in a previous study for power calculation [KKW10].

Unfortunately, this parametric bootstrapping approach is computationally very expensive.

5.4.2.2 MultiTrans

5.4.2.2.1 MVN approximation for LMM As described in the previous section, the parametric bootstrapping strategy is impractical due to its high computational cost. To make the procedure efficient, we propose a new approach, MultiTrans. MultiTrans alternatively samples statistics directly from MVN without needing to generate any null phenotypes. Figure 5.13 (b) shows an overview of the MultiTrans. Once we obtain the null samples, we can obtain the per-marker threshold using Equation (5.7). However, the challenge is to charaterize the covariance of MVN for LMM.

5.4.2.2.2 Covariance of MVN in LMM For LMM, Equation (5.6) is no longer valid. That is, we cannot use the genotype correlation matrix as the covariance matrix of MVN for LMM. To derive the covariance matrix, we assume a LMM instead of the linear model as follows:

$$Y = \mu \mathbf{1}_n + X_i \beta_i^M + \mathbf{g} + \mathbf{e}$$

, where μ is a mean of the phenotypic values, $\mathbf{1}_n$ is a vector of n ones, Y is a vector of length n with the phenotypic values, X_i is a vector of length n with the genotypic values of ith marker, β_i^M is their coefficients under the LMM, $\mathbf{g}$ is a vector of length n sampled from $\mathcal{N}(0, \sigma_g^2 \mathbf{K})$ accounting for population structure effects where $\mathbf{K}$ is a $n \times n$ matrix that explains the correlation between the individuals induced by population structure, and $\mathbf{e}$ is a vector of length n sampled from $\mathcal{N}(0, \sigma^2 \mathbf{I})$ accounting for the residual errors. Under this model, the phenotype follows a MVN with a mean and variance as follows:

$$Y \sim \mathcal{N}(\mu \mathbf{1}_n + X_i \beta_i^M, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I})$$

Given the observed data, it is straightforward to fit LMM and estimate parameters σ_g^2 and σ_e^2 using standard strategies, which define the covariance matrix of phenotypes, $\texttt{Cov}(Y) = \hat{V} = \hat{\sigma}_g^2\mathbf{K}, +\hat{\sigma}_e^2\mathbf{I}$. Now we utilize the fact that after obtaining $\hat{V}$, the remaining regression procedure is equivalent to performing ordinary least-squares in the transformed space,

$$\hat{V}^{-1/2}Y \sim \mathcal{N}(\hat{V}^{-1/2}\mu\mathbf{1}_n + \hat{V}^{-1/2}X_i\beta_i^M, \mathbf{I})$$

, where both genotypes and phenotypes are transformed by a factor $\hat{V}^{-1/2}$. Assuming that $\hat{V}^{-1/2}X_i$ and $\hat{V}^{-1/2}Y$ are normalized as mean 0 and variance 1 (without loss of generality), the ordinary least-squares solution of β_i^M for ith marker and jth marker are as follows:

$$\begin{aligned}
\hat{\beta}_i^M =& (X_i^T\hat{V}^{-1}X_i)^{-1}X_i^T\hat{V}^{-1}Y \sim \mathcal{N}\left(\beta_i^M, (X_i^T\hat{V}^{-1}X_i)^{-1}\right) \\
\hat{\beta}_j^M =& (X_j^T\hat{V}^{-1}X_j)^{-1}X_j^T\hat{V}^{-1}Y \sim \mathcal{N}\left(\beta_j^M, (X_j^T\hat{V}^{-1}X_j)^{-1}\right)
\end{aligned}$$

The statistics are computed as follows:

$$\begin{aligned}
S_i =& \hat{\beta}_i^M\sqrt{X_i^T\hat{V}^{-1}X_i} \sim \mathcal{N}\left(\beta_i^M\sqrt{X_i^T\hat{V}^{-1}X_i}, 1\right) \\
S_j =& \hat{\beta}_j^M\sqrt{X_j^T\hat{V}^{-1}X_j} \sim \mathcal{N}\left(\beta_i^M\sqrt{X_j^T\hat{V}^{-1}X_j}, 1\right)
\end{aligned}$$

Accordingly, the correlation between the statistics changes from Equation (5.6) to the following and the correlation between the statistics are equal to the correlation between the marker transformed by the inverse square root of $\hat{V}$,

$$\begin{aligned}
\texttt{Cov}(S_i^M, S_j^M) =& \texttt{Cov}\left(\frac{X_i^T\hat{V}^{-1}Y}{\sqrt{X_i^T\hat{V}^{-1}X_i}}, \frac{X_j^T\hat{V}^{-1}Y}{\sqrt{X_j^T\hat{V}^{-1}X_j}}\right) \\
=& \frac{X_i^T\hat{V}^{-1/2}(\hat{V}^{-1/2})^TX_j}{\sqrt{X_i^T(\hat{V}^{-1/2})^T\hat{V}^{-1/2}X_i}\sqrt{X_j^T(\hat{V}^{-1/2})^T\hat{V}^{-1/2}X_j}} \\
=& \texttt{Cor}(\hat{V}^{-1/2}X_i, \hat{V}^{-1/2}X_j) = r_{ij}^M
\end{aligned}$$

Utilizing the covariance matrix estimated from transformed genotypes, we can generate a large number of samples, $(S_1, S_2, \cdots, S_m)$, to approximate MVN and correct p-values by integrating over the outside of the rectangle, as in Equation (5.7).

5.4.2.2.3 Sliding-window approach If m is large, the standard sampling approach that samples $(S_1, S_2, \cdots, S_m)$ from MVN using Cholesky decomposition [HMR96] is computationally very expensive. However, under the local linkage disequilibrium assumption, the statistics at distant markers are uncorrelated and we can split the region into small blocks to dramatically decrease computational cost. In addition, we can perform a sliding-window approach as follows to incorporate the inter-block correlations to accurately estimate the p-values [HKE09]. Let $f(S_1, S_2, \cdots, S_m)$ be the joint probability density function of the statistics. Under the local linkage disequilibrium assumption, the statistics at distant markers are uncorrelated. Thus given a window size w, we can assume that S_i is conditionally independent of $S_1, S_2, \cdots, S_{i-w-1}$ given $S_{i-w}, S_{i-w+1}, \cdots, S_{i-1}$. Utilizing the chain rule,

$$f(S_1, S_2, \cdots, S_m) = f(S_1)f(S_2|S_1)f(S_3|S_1, S_2)\cdots f(S_m|S_{m-w}, \cdots, S_{m-1})$$

Thus, we can sample S_i given $S_{i-w}, S_{i-w+1}, \cdots, S_{i-1}$, based on the conditional distribution $f(S_i|S_{i-w}, \cdots, S_{i-1})$ and efficiently generate a large number of samples.

5.4.3 HMDP dataset

We evaluated our approach using a HMDP (high resolution association mapping) mouse dataset [BFO10] which contains 102,987 SNPs in 99 individuals. SNPs with a minor allele frequency less than 5 % and missing more than 10 % are filtered. To test the difference between covariance of test statistics and correlation between the genotypes, we generated a simulated dataset by extracting 100 SNPs from chromosome 1. Seven phenotypes with

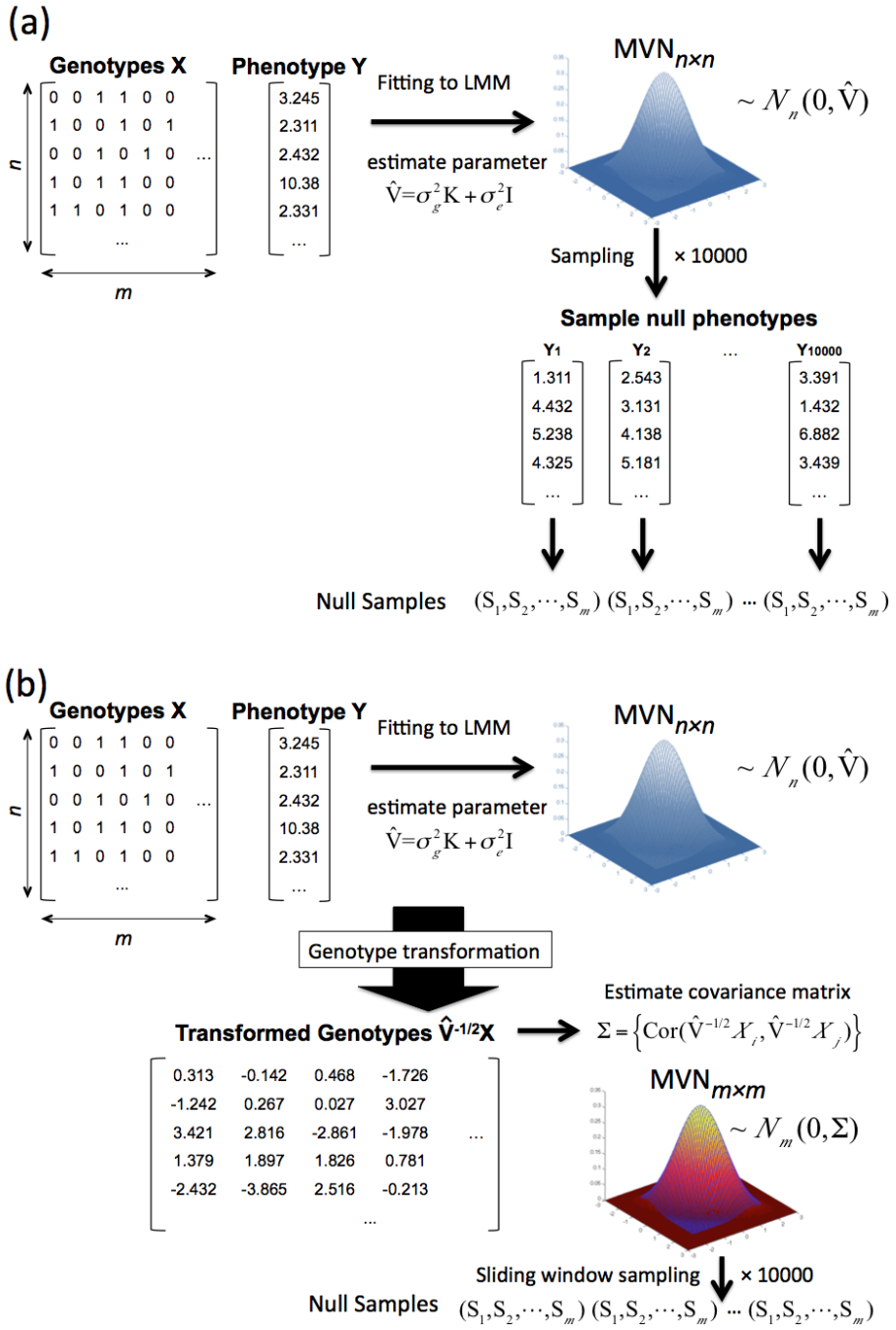


Figure 5.13. Overview of the re-sampling procedures of parametric bootstrapping (a) and MultiTrans (b). 10^4 sampling applied for both parametric bootstrapping and MultiTrans.

different heritabilities which were estimated from the HMDP dataset [BFO10] were used for section 2.6.

5.4.4 Yeast dataset

We evaluated our approach utilizing a yeast dataset [SK08] that contains 2,956 SNPs in 109 segregants. To test the difference between covariance of test statistics and correlation between the genotypes, we generated a simulated dataset by extracting 100 consecutive SNPs from chromosome 4. Ten gene expressions with different heritabilities, which were estimated from the yeast dataset [SK08], were used for section 2.6.

5.4.5 HapMap dataset

We evaluated our approach utilizing a HapMap Phase 3 dataset [GBH03] which contains 1,070,114 SNPs in 1,184 individuals. SNPs with a minor allele frequency less than 5 % and missing more than 10 % are filtered. To test the difference between the covariance of test statistics and correlation between the genotypes, we generated a simulated dataset by extracting 100 consecutive SNPs from chromosome 22. Five gene expressions with different heritabilities, which were estimated from the HapMap dataset [GBH03], were used for section 2.6.

5.4.6 Implementation

For the results of MultiTrans in section 2.3 Table 5.1 and section 2.5 Figure 5.9, a window size of 1000 was used, and 10^7 number of samplings were performed. For the results of the parametric bootstrapping in section 2.3 Table 5.1, 10^5 samplings were performed. To evaluate our method for various ranges of heritabilities, we applied our method for four different heritabilities, 0, 0.2, 0.5, and 0.8. To estimate p-values and the variance

components (σ_g^2 and σ_e^2) for LMM, one of the LMM-solver, pylmm [FE15] were used. In practice, however, other LMM-based methods such as EMMA [KYE08], EMMAX [KSS10], FaST-LMM [LLL11], etc, could be also used.

CHAPTER 6

Conclusion

Getting the benefit of high-throughput technologies that made it possible to cost effectively obtain DNA sequence information from individuals, GWAS have discovered a large number of genetic variants involved in disease and traits in past decade. Despite the tremendously successful, it has been reported that there exist various hidden confounding factors such as population structure [KCP02, FRP04, MCP04, COL05, HYH05, RZL05, VP05, BSK06, SSV06, FG06] or technical artifacts [FE12, LS07, LKS10] that complicates the GWAS analysis. Applying the standard polygenic model to the data with confounding effects may cause an inflation of the values of the association statistics leading to false positive identification [Esk15]. Many statistical approaches have been proposed for addressing this problem and most recently linear mixed model has emerged and successfully increased both accuracy and statistical power in many GWAS [KSS10, LLK12, LLH13, YZG14, LTB15].

The first contribution of my work is in eQTL studies. The last few years have seen the development of large efforts for understanding how the GWAS variants contribute to disease through eQTL studies. Especially, trans regulatory hotspots are of most interest where hundreds or even thousands of genes are trans regulated by a small number of genomic loci. However, as shown in studies of recombinant inbred (RI) mice [KYE08], many of these regulatory hotspots replicate poorly. Utilizing the linear mixed model, I explicitly modeled the confounding effects and developed an efficient approach that iden-

tifies regulatory hotspots while correcting for confounding effects in eQTL studies.

The second contribution of my work is in multiple hypothesis testing. As the size of GWAS data grows, in many cases analyzing multiple phenotypes simultaneously is preferable than analysis each phenotype one at a time. Despite the fact that confounding effects may cause serious problems in multiple phenotype analysis, none of the previous multiple phenotype approach were aware of this problem. I have shown that even a small bias induced by confounding effect could cause serious problem in multivariate analysis and developed a multiple phenotype analysis that can be applied to linear mixed model to correct for the confounding effects.

My last contribution is in multiple hypothesis testing correction for GWAS. Multiple hypothesis testing correction is an essential step in current GWAS which test tens of thousands of variants. Unfortunately, in the case when confounding is present, the confounding causes a violation of the basic assumption necessary for the standard approaches, such as permutation tests or false discovery rates, which is that the individuals in the sample are i.i.d. Shared confounding factors induces complex dependencies among the phenotype patterns of individuals and complicates multiple testing. I have shown that confounding effects affect the significance threshold and developed an efficient multiple hypothesis correction method for linear mixed model that can accommodate those confounding factors and increase statistical power.

References

AAA10. Gonalo R. Abecasis, David Altshuler, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Richard A. Gibbs, Matt E. Hurles, and Gil A. McVean. "A map of human genome variation from population-scale sequencing." *Nature*, **467**(7319):1061–73, 10 2010.

ABB00. O. Alter, P. O. Brown, and D. Botstein. "Singular value decomposition for genome-wide expression data processing and modeling." *Proc Natl Acad Sci U S A*, **97**(18):10101–6, 8 2000.

ADL08. David Altshuler, Mark J. Daly, and Eric S. Lander. "Genetic mapping in human disease." *Science*, **322**(5903):881–8, 11 2008.

AVF11. David L. Aylor, William Valdar, Wendy Foulds-Mathes, Ryan J. Buus, Ricardo A. Verdugo, Ralph S. Baric, Martin T. Ferris, Jeff A. Frelinger, Mark Heise, Matt B. Frieman, Lisa E. Gralinski, Timothy A. Bell, John D. Didion, Kunjie Hua, Derrick L. Nehrenberg, Christine L. Powell, Jill Steigerwalt, Yuying Xie, Samir N. P. Kelada, Francis S. Collins, Ivana V. Yang, David A. Schwartz, Lisa A. Branstetter, Elissa J. Chesler, Darla R. Miller, Jason Spence, Eric Yi Liu, Leonard McMillan, Abhishek Sarkar, Jeremy Wang, Wei Wang, Qi Zhang, Karl W. Broman, Ron Korstanje, Caroline Durrant, Richard Mott, Fuad A. Iraqi, Daniel Pomp, David Threadgill, Fernando Pardo-Manuel de Villena, and Gary A. Churchill. "Genetic analysis of complex traits in the emerging Collaborative Cross." *Genome Res*, **21**(8):1213–22, 8 2011.

BC57. J. Roger Bray and John T. Curtis. "An ordination of the upland forest communities of southern Wisconsin." *Ecological monographs*, **27**(4):325–349, 1957.

BFO10. Brian J. Bennett, Charles R. Farber, Luz Orozco, Hyun Min Kang, Anatole Ghazalpour, Nathan Siemers, Michael Neubauer, Isaac Neuhaus, Roumyana Yordanova, Bo Guan, Amy Truong, Wen-pin P. Yang, Aiqing He, Paul Kayne, Peter Gargalovic, Todd Kirchgessner, Calvin Pan, Lawrence W. Castellani, Emrah Kostem, Nicholas Furlotte, Thomas A. Drake, Eleazar Eskin, and Aldons J. Lusis. "A high-resolution association mapping panel for the dissection of complex traits in mice." *Genome Res*, **20**(2):281–90, 2 2010.

BGM07. Molly A. Bogue, Stephen C. Grubb, Terry P. Maddatu, and Carol J. Bult. "Mouse Phenome Database (MPD)." *Nucleic Acids Res*, **35**(Database issue):D643–9, 1 2007.

BK05. Rachel B. Brem and Leonid Kruglyak. "The landscape of genetic complexity across 5,700 gene expression traits in yeast." *Proc Natl Acad Sci U S A*, **102**(5):1572–7, 2 2005.

BMH07. William S. Branham, Cathy D. Melvin, Tao Han, Varsha G. Desai, Carrie L. Moland, Adam T. Scully, and James C. Fuscoe. "Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements." *BMC Biotechnol*, **7**:8, 2007.

Bro08. Brian L. Browning. "PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies." *BMC Bioinformatics*, **9**:309, 2008.

BSF13. Nicholas A. Bokulich, Sathish Subramanian, Jeremiah J. Faith, Dirk Gevers, Jeffrey I. Gordon, Rob Knight, David A. Mills, and J. Gregory Caporaso. "Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing." *Nat Methods*, **10**(1):57–9, 1 2013.

BSK06. Mario Berger, Hans H. Stassen, Karola Khler, Vera Krane, Detlev Mnks, Christoph Wanner, Katrin Hoffmann, Michael M. Hoffmann, Michael Zimmer, Heike Bickebller, and Tom H. Lindner. "Hidden population substructures in an apparently homogeneous population bias association studies." *Eur J Hum Genet*, **14**(2):236–44, 2 2006.

BWD05. Leonid Bystrykh, Ellen Weersing, Bert Dontje, Sue Sutton, Mathew T. Pletcher, Tim Wiltshire, Andrew I. Su, Edo Vellenga, Jintao Wang, Kenneth F. Manly, Lu Lu, Elissa J. Chesler, Rudi Alberts, Ritsert C. Jansen, Robert W. Williams, Michael P. Cooke, and Gerald de Haan. "Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'." *Nat Genet*, **37**(3):225–32, 3 2005.

BYC02. Rachel B. Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. "Genetic dissection of transcriptional regulation in budding yeast." *Science*, **296**(5568):752–5, 4 2002.

CB07. Karen N. Conneely and Michael Boehnke. "So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests." *Am J Hum Genet*, **81**(6):1158–68, 12 2007.

CGX14. Wei Chen, Yanqiang Gao, Weibo Xie, Liang Gong, Kai Lu, Wensheng Wang, Yang Li, Xianqing Liu, Hongyan Zhang, Huaxia Dong, Wan Zhang, Lejing Zhang, Sibin Yu, Gongwei Wang, Xingming Lian, and Jie Luo. "Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism." *Nat Genet*, **46**(7):714–21, 7 2014.

CHP13. Adrian Cortes, Johanna Hadler, Jenny P. Pointon, Philip C. Robinson, Tugce Karaderi, Paul Leo, Katie Cremin, Karena Pryce, Jessica Harris, Seunghun Lee, Kyung Bin Joo, Seung-Cheol C. Shim, Michael Weisman, Michael Ward,

Xiaodong Zhou, Henri-Jean J. Garchon, Gilles Chiocchia, Johannes Nossent, Benedicte A. Lie, ystein Frre, Jaakko Tuomilehto, Kari Laiho, Lei Jiang, Yu Liu, Xin Wu, Linda A. Bradbury, Dirk Elewaut, Ruben Burgos-Vargas, Simon Stebbings, Louise Appleton, Claire Farrah, Jonathan Lau, Tony J. Kenna, Nigil Haroon, Manuel A. Ferreira, Jian Yang, Juan Mulero, Jose Luis Fernandez-Sueiro, Miguel A. Gonzalez-Gay, Carlos Lopez-Larrea, Panos Deloukas, Peter Donnelly, Paul Bowness, Karl Gafney, Hill Gaston, Dafna D. Gladman, Proton Rahman, Walter P. Maksymowych, Huji Xu, J. Bart A. Crusius, Irene E. van der Horst-Bruinsma, Chung-Tei T. Chou, Raphael Valle-Oate, Consuelo Romero-Snchez, Inger Myrnes Hansen, Fernando M. Pimentel-Santos, Robert D. Inman, Vibeke Videm, Javier Martin, Maxime Breban, John D. Reveille, David M. Evans, Tae-Hwan H. Kim, Bryan Paul Wordsworth, and Matthew A. Brown. "Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci." *Nat Genet*, **45**(7):730–8, 7 2013.

Chu02. Gary A. Churchill. "Fundamentals of experimental design for cDNA microarrays." *Nat Genet*, **32 Suppl**:490–5, 12 2002.

CLE05. Alessandra C. Cervino, Guoya Li, Steve Edwards, Jun Zhu, Cathy Laurie, George Tokiwa, Pek Yee Lum, Susanna Wang, Lawrence W. Castellani, Lawrence W. Castellini, Aldons J. Lusis, Sonia Carlson, Alan B. Sachs, and Eric E. Schadt. "Integrating QTL and high-density SNP analyses in mice to identify Insig2 as a susceptibility gene for plasma cholesterol levels." *Genomics*, **86**(5):505–17, 11 2005.

CLS05. Elissa J. Chesler, Lu Lu, Siming Shou, Yanhua Qu, Jing Gu, Jintao Wang, Hui Chen Hsu, John D. Mountz, Nicole E. Baldwin, Michael A. Langston, David W. Threadgill, Kenneth F. Manly, and Robert W. Williams. "Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function." *Nat Genet*, **37**(3):233–42, 3 2005.

COL05. Catarina D. Campbell, Elizabeth L. Ogburn, Kathryn L. Lunetta, Helen N. Lyon, Matthew L. Freedman, Leif C. Groop, David Altshuler, Kristin G. Ardlie, and Joel N. Hirschhorn. "Demonstrating stratification in a European American population." *Nat Genet*, **37**(8):868–72, 8 2005.

CSE05. Vivian G. Cheung, Richard S. Spielman, Kathryn G. Ewens, Teresa M. Weber, Michael Morley, and Joshua T. Burdick. "Mapping determinants of human gene expression by regional and genome-wide association." *Nature*, **437**(7063):1365–9, 10 2005.

DR99. B. Devlin and K. Roeder. "Genomic control for association studies." *Biometrics*, **55**(4):997–1004, 12 1999.

DRW01. B. Devlin, K. Roeder, and L. Wasserman. "Genomic control, a new approach to genetic-based association studies." *Theor Popul Biol*, **60**(3):155–66, 11 2001.

DYK13. Lea K. Davis, Dongmei Yu, Clare L. Keenan, Eric R. Gamazon, Anuar I. Konkashbaev, Eske M. Derks, Benjamin M. Neale, Jian Yang, S. Hong Lee, Patrick Evans, Cathy L. Barr, Laura Bellodi, Fortu Benarroch, Gabriel Bedoya Berrio, Oscar J. Bienvenu, Michael H. Bloch, Rianne M. Blom, Ruth D. Bruun, Cathy L. Budman, Beatriz Camarena, Desmond Campbell, Carolina Cappi, Julio C. Cardona Silgado, Danielle C. Cath, Maria C. Cavallini, Denise A. Chavira, Sylvain Chouinard, David V. Conti, Edwin H. Cook, Vladimir Coric, Bernadette A. Cullen, Dieter Deforce, Richard Delorme, Yves Dion, Christopher K. Edlund, Karin Egberts, Peter Falkai, Thomas V. Fernandez, Patience J. Gallagher, Helena Garrido, Daniel Geller, Simon L. Girard, Hans J. Grabe, Marco A. Grados, Benjamin D. Greenberg, Varda Gross-Tsur, Stephen Haddad, Gary A. Heiman, Sian M. J. Hemmings, Ana G. Hounie, Cornelia Illmann, Joseph Jankovic, Michael A. Jenike, James L. Kennedy, Robert A. King, Barbara Kremeyer, Roger Kurlan, Nuria Lanzagorta, Marion Leboyer, James F. Leckman, Leonhard Lennertz, Chunyu Liu, Christine Lochner, Thomas L. Lowe, Fabio Macciardi, James T. McCracken, Lauren M. McGrath, Sandra C. Mesa Restrepo, Rainald Moessner, Jubel Morgan, Heike Muller, Dennis L. Murphy, Allan L. Naarden, William Cornejo Ochoa, Roel A. Ophoff, Lisa Osiecki, Andrew J. Pakstis, Michele T. Pato, Carlos N. Pato, John Piacentini, Christopher Pittenger, Yehuda Pollak, Scott L. Rauch, Tobias J. Renner, Victor I. Reus, Margaret A. Richter, Mark A. Riddle, Mary M. Robertson, Roxana Romero, Maria C. Rosàrio, David Rosenberg, Guy A. Rouleau, Stephan Ruhrmann, Andres Ruiz-Linares, Aline S. Sampaio, Jack Samuels, Paul Sandor, Brooke Sheppard, Harvey S. Singer, Jan H. Smit, Dan J. Stein, E. Strengman, Jay A. Tischfield, Ana V. Valencia Duarte, Homero Vallada, Filip Van Nieuwerburgh, Jeremy Veenstra-Vanderweele, Susanne Walitza, Ying Wang, Jens R. Wendland, Herman G. M. Westenberg, Yin Yao Shugart, Euripedes C. Miguel, William McMahon, Michael Wagner, Humberto Nicolini, Danielle Posthuma, Gregory L. Hanna, Peter Heutink, Damiaan Denys, Paul D. Arnold, Ben A. Oostra, Gerald Nestadt, Nelson B. Freimer, David L. Pauls, Naomi R. Wray, S. Evelyn Stewart, Carol A. Mathews, James A. Knowles, Nancy J. Cox, and Jeremiah M. Scharf. "Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture." *PLoS Genet*, **9**(10):e1003864, 10 2013.

Esk15. Eleazar Eskin. "Discovering genes involved in disease and the mystery of missing heritability." *Communications of the ACM*, **58**(10):80–87, 2015.

ETZ08. Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S. Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G. Bragi Walters, Steinunn Gunnarsdottir, Magali Mouy, Valgerdur Steinthorsdottir, Gudrun H. Eiriksdottir, Gyda Bjornsdottir, Inga Reynisdottir, Daniel Gudbjartsson, Anna Helgadottir, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Unnur Styrkarsdottir, Solveig Gretarsdottir, Kristinn P. Magnusson, Hreinn Stefansson, Ragnheidur Fossdal, Kristleifur Kristjansson, Hjortur G. Gislason, Tryggvi Stefansson, Bjorn G. Leifsson, Unnur Thorsteinsdottir, John R. Lamb, Jeffrey R. Gulcher, Marc L. Reitman, Augustine Kong, Eric E. Schadt, and Kari Stefansson. "Genetics of gene expression and its effect on disease." *Nature*, **452**(7186):423–8, 3 2008.

FBO11. Charles R. Farber, Brian J. Bennett, Luz Orozco, Wei Zou, Ana Lira, Emrah Kostem, Hyun Min Kang, Nicholas Furlotte, Ani Berberyan, Anatole Ghazalpour, Jaijam Suwanwela, Thomas A. Drake, Eleazar Eskin, Q. Tian Wang, Steven L. Teitelbaum, and Aldons J. Lusis. "Mouse genome-wide association and systems genetics identify Asxl2 as a regulator of bone mineral density and osteoclastogenesis." *PLoS Genet*, **7**(4):e1002038, 4 2011.

FCD03. Thomas L. Fare, Ernest M. Coffey, Hongyue Dai, Yudong D. He, Deborah A. Kessler, Kristopher A. Kilian, John E. Koch, Eric LeProust, Matthew J. Marton, Michael R. Meyer, Roland B. Stoughton, George Y. Tokiwa, and Yanqun Wang. "Effects of atmospheric ozone on microarray data quality." *Anal Chem*, **75**(17):4672–5, 9 2003.

FE12. Jonathan Flint and Eleazar Eskin. "Genome-wide association studies in mice." *Nat Rev Genet*, **13**(11):807–17, 11 2012.

FE15. Nicholas A. Furlotte and Eleazar Eskin. "Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model." *Genetics*, **200**(1):59–68, 5 2015.

FG06. Matthieu Foll and Oscar Gaggiotti. "Identifying the environmental factors that determine the genetic structure of populations." *Genetics*, **174**(2):875–91, 10 2006.

FRP04. Matthew L. Freedman, David Reich, Kathryn L. Penney, Gavin J. McDonald, Andre A. Mignault, Nick Patterson, Stacey B. Gabriel, Eric J. Topol, Jordan W. Smoller, Carlos N. Pato, Michele T. Pato, Tracey L. Petryshen, Laurence N. Kolonel, Eric S. Lander, Pamela Sklar, Brian Henderson, Joel N. Hirschhorn, and David Altshuler. "Assessing the impact of population stratification on genetic association studies." *Nat Genet*, **36**(4):388–93, 4 2004.

FRS07. Eric J. Foss, Dragan Radulovic, Scott A. Shaffer, Douglas M. Ruderfer, Antonio Bedalov, David R. Goodlett, and Leonid Kruglyak. "Genetic basis of proteome variation in yeast." *Nat Genet*, **39**(11):1369–75, 11 2007.

FSC13. Michaela Fakiola, Amy Strange, Heather J. Cordell, E. Nancy Miller, Matti Pirinen, Zhan Su, Anshuman Mishra, Sanjana Mehrotra, Gloria R. Monteiro, Gavin Band, Cline Bellenguez, Serge Dronov, Sarah Edkins, Colin Freeman, Eleni Giannoulatou, Emma Gray, Sarah E. Hunt, Henio G. Lacerda, Cordelia Langford, Richard Pearson, Nbia N. Pontes, Madhukar Rai, Shri P. Singh, Linda Smith, Olivia Sousa, Damjan Vukcevic, Elvira Bramon, Matthew A. Brown, Juan P. Casas, Aiden Corvin, Audrey Duncanson, Janusz Jankowski, Hugh S. Markus, Christopher G. Mathew, Colin N. A. Palmer, Robert Plomin, Anna Rautanen, Stephen J. Sawcer, Richard C. Trembath, Ananth C. Viswanathan, Nicholas W. Wood, Mary E. Wilson, Panos Deloukas, Leena Peltonen, Frank Christiansen, Campbell Witt, Selma M. B. Jeronimo, Shyam Sundar, Chris C. A. Spencer, Jenefer M. Blackwell, and Peter Donnelly. "Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis." *Nat Genet*, **45**(2):208–13, 2 2013.

FSL12. Nicoló Fusi, Oliver Stegle, and Neil D. Lawrence. "Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies." *PLoS Comput Biol*, **8**(1):e1002330, 1 2012.

GB02. Alan Genz and Frank Bretz. "Comparison of methods for the computation of multivariate t probabilities." *Journal of Computational and Graphical Statistics*, **11**(4):950–971, 2002.

GBB10. Xiaoyi Gao, Lewis C. Becker, Diane M. Becker, Joshua D. Starmer, and Michael A. Province. "Avoiding the high Bonferroni penalty in genome-wide association studies." *Genetic epidemiology*, **34**(1):100–105, 2010.

GBH03. Richard A. Gibbs, John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan-Yang . Y. Ch'ang, Wei Huang, Bin Liu, and Yan Shen. "The international HapMap project." *Nature*, **426**(6968):789–796, 2003.

Gen92. Alan Genz. "Numerical computation of multivariate normal probabilities." *Journal of computational and graphical statistics*, **1**(2):141–149, 1992.

GHC10. Eric R. Gamazon, R. Stephanie Huang, Nancy J. Cox, and M. Eileen Dolan. "Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci." *Proc Natl Acad Sci U S A*, **107**(20):9287–92, 5 2010.

GMB09. Stephen C. Grubb, Terry P. Maddatu, Carol J. Bult, and Molly A. Bogue. "Mouse phenome database." *Nucleic Acids Res*, **37**(Database issue):D720–30, 1 2009.

Gow66. John C. Gower. "Some distance properties of latent root and vector methods used in multivariate analysis." *Biometrika*, **53**(3-4):325–338, 1966.

GRG99. S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, and R. Aebersold. "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." *Nat Biotechnol*, **17**(10):994–9, 10 1999.

HAA07. John M. Hancock, Niels C. Adams, Vassilis Aidinis, Andrew Blake, Molly Bogue, Steve D. M. Brown, Elissa J. Chesler, Duncan Davidson, Christopher Duran, Janan T. Eppig, Valérie Gailus-Durner, Hilary Gates, Georgios V. Gkoutos, Simon Greenaway, Martin Hrabé de Angelis, George Kollias, Sophie Leblanc, Kirsty Lee, Christoph Lengger, Holger Maier, Ann-Marie M. Mallon, Hiroshi Masuya, David G. Melvin, Werner Müller, Helen Parkinson, Glenn Proctor, Eli Reuveni, Paul Schofield, Aadya Shukla, Cynthia Smith, Tetsuro Toyoda, Laurent Vasseur, Shigeharu Wakana, Alison Walling, Jacqui White, Joe Wood, and Michalis Zouberakis. "Mouse Phenotype Database Integration Consortium: integration [corrected] of mouse phenome data resources." *Mamm Genome*, **18**(3):157–63, 3 2007.

HBM15. Jrg Hagmann, Claude Becker, Jonas Mller, Oliver Stegle, Rhonda C. Meyer, George Wang, Korbinian Schneeberger, Joffrey Fitz, Thomas Altmann, Joy Bergelson, Karsten Borgwardt, and Detlef Weigel. "Century-scale methylome stability in a recently diverged Arabidopsis thaliana lineage." *PLoS Genet*, **11**(1):e1004920, 1 2015.

HE11. Buhm Han and Eleazar Eskin. "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies." *Am J Hum Genet*, **88**(5):586–98, 5 2011.

HE12. Buhm Han and Eleazar Eskin. "Interpreting meta-analyses of genome-wide association studies." *PLoS Genet*, **8**(3):e1002555, 3 2012.

HGB07. Hakon Hakonarson, Struan F. A. Grant, Jonathan P. Bradfield, Luc Marchand, Cecilia E. Kim, Joseph T. Glessner, Rosemarie Grabs, Tracy Casalunovo, Shayne P. Taback, Edward C. Frackelton, Margaret L. Lawson, Luke J. Robinson, Robert Skraban, Yang Lu, Rosetta M. Chiavacci, Charles A. Stanley, Susan E. Kirsch, Eric F. Rappaport, Jordan S. Orange, Dimitri S. Monos, Marcella Devoto, Hui-Qi Q. Qu, and Constantin Polychronakos. "A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene." *Nature*, **448**(7153):591–4, 8 2007.

HKE09. Buhm Han, Hyun Min Kang, and Eleazar Eskin. "Rapid and accurate multiple testing correction and power estimation for millions of correlated markers." *PLoS Genet*, **5**(4):e1000456, 4 2009.

HKK14. Farhad Hormozdiari, Emrah Kostem, Eun Yong Kang, Bogdan Pasaniuc, and Eleazar Eskin. "Identifying causal variants at loci with multiple signals of association." *Genetics*, **198**(2):497–508, 10 2014.

HKWss. Farhad Hormozdiari, Yang Kichaev, Gleb, Bogdan Wen-Yun, Pasaniuc, and Eleazar Eskin. "Identification of causal genes for complex traits." *Bioinformatics*, In press.

HMI14. Wen Huang, Andreas Massouras, Yutaka Inoue, Jason Peiffer, Miquel Rmia, Aaron M. Tarone, Lavanya Turlapati, Thomas Zichner, Dianhui Zhu, Richard F. Lyman, Michael M. Magwire, Kerstin Blankenburg, Mary Anna Carbone, Kyle Chang, Lisa L. Ellis, Sonia Fernandez, Yi Han, Gareth Highnam, Carl E. Hjelmen, John R. Jack, Mehwish Javaid, Joy Jayaseelan, Divya Kalra, Sandy Lee, Lora Lewis, Mala Munidasa, Fiona Ongeri, Shohba Patel, Lora Perales, Agapito Perez, LingLing Pu, Stephanie M. Rollmann, Robert Ruth, Nehad Saada, Crystal Warner, Aneisa Williams, Yuan-Qing Q. Wu, Akihiko Yamamoto, Yiqing Zhang, Yiming Zhu, Robert R. H. Anholt, Jan O. Korbel, David Mittelman, Donna M. Muzny, Richard A. Gibbs, Antonio Barbadilla, J. Spencer Johnston, Eric A. Stone, Stephen Richards, Bart Deplancke, and Trudy F. C. Mackay. "Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines." *Genome Res*, **24**(7):1193–208, 7 2014.

HMR96. Vassilis Hajivassiliou, Daniel McFadden, and Paul Ruud. "Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results." *Journal of econometrics*, **72**(1):85–134, 1996.

HWW05. Sonja Hillebrandt, Hermann E. Wasmuth, Ralf Weiskirchen, Claus Hellerbrand, Hildegard Keppeler, Alexa Werth, Ramin Schirin-Sokhan, Gabriele Wilkens, Andreas Geier, Johann Lorenzen, Jörg Köhl, Axel M. Gressner, Siegfried Matern, and Frank Lammert. "Complement factor 5 is a quantitative trait gene that modifies liver fibrogenesis in mice and humans." *Nat Genet*, **37**(8):835–43, 8 2005.

HYH05. Agnar Helgason, Brynds Yngvadttir, Birgir Hrafnkelsson, Jeffrey Gulcher, and Kri Stefnsson. "An Icelandic example of the impact of population structure on association studies." *Nat Genet*, **37**(1):90–5, 1 2005.

JKF15. Jong Wha J. Joo, Eun Yong Kang, Nick Furlotte, Brian Parks, Aldons J. Lusis, and Eleazar Eskin. "Efficient and Accurate Multiple-Phenotypes Regression

Method for High Dimensional Data Considering Population Structure." In *Research in Computational Molecular Biology*, pp. 136–153. Springer, 2015.

JSH14. Jong Wha J. Joo, Jae Hoon Sul, Buhm Han, Chun Ye, and Eleazar Eskin. "Effectively identifying regulatory hotspots while capturing expression heterogeneity in gene expression studies." *Genome Biol*, **15**(4):r61, 2014.

KAT13. Anna Kttgen, Eva Albrecht, Alexander Teumer, Veronique Vitart, Jan Krumsiek, Claudia Hundertmark, Giorgio Pistis, Daniela Ruggiero, Conall M. O'Seaghdha, Toomas Haller, Qiong Yang, Toshiko Tanaka, Andrew D. Johnson, Zoltn Kutalik, Albert V. Smith, Julia Shi, Maksim Struchalin, Rita P. S. Middelberg, Morris J. Brown, Angelo L. Gaffo, Nicola Pirastu, Guo Li, Caroline Hayward, Tatijana Zemunik, Jennifer Huffman, Loic Yengo, Jing Hua Zhao, Ayse Demirkan, Mary F. Feitosa, Xuan Liu, Giovanni Malerba, Lorna M. Lopez, Pim van der Harst, Xinzhong Li, Marcus E. Kleber, Andrew A. Hicks, Ilja M. Nolte, Asa Johansson, Federico Murgia, Sarah H. Wild, Stephan J. L. Bakker, John F. Peden, Abbas Dehghan, Maristella Steri, Albert Tenesa, Vasiliki Lagou, Perttu Salo, Massimo Mangino, Lynda M. Rose, Terho Lehtimki, Owen M. Woodward, Yukinori Okada, Adrienne Tin, Christian Mller, Christopher Oldmeadow, Margus Putku, Darina Czamara, Peter Kraft, Laura Frogheri, Gian Andri Thun, Anne Grotevendt, Gauti Kjartan Gislason, Tamara B. Harris, Lenore J. Launer, Patrick McArdle, Alan R. Shuldiner, Eric Boerwinkle, Josef Coresh, Helena Schmidt, Michael Schallert, Nicholas G. Martin, Grant W. Montgomery, Michiaki Kubo, Yusuke Nakamura, Toshihiro Tanaka, Patricia B. Munroe, Nilesh J. Samani, David R. Jacobs, Kiang Liu, Pio D'Adamo, Sheila Ulivi, Jerome I. Rotter, Bruce M. Psaty, Peter Vollenweider, Gerard Waeber, Susan Campbell, Olivier Devuyst, Pau Navarro, Ivana Kolcic, Nicholas Hastie, Beverley Balkau, Philippe Froguel, Tnu Esko, Andres Salumets, Kay Tee Khaw, Claudia Langenberg, Nicholas J. Wareham, Aaron Isaacs, Aldi Kraja, Qunyuan Zhang, Philipp S. Wild, Rodney J. Scott, Elizabeth G. Holliday, Elin Org, Margus Viigimaa, Stefania Bandinelli, Jeffrey E. Metter, Antonio Lupo, Elisabetta Trabetti, Rossella Sorice, Angela Dring, Eva Lattka, Konstantin Strauch, Fabian Theis, Melanie Waldenberger, H-Erich E. Wichmann, Gail Davies, Alan J. Gow, Marcel Bruinenberg, Ronald P. Stolk, Jaspal S. Kooner, Weihua Zhang, Bernhard R. Winkelmann, Bernhard O. Boehm, Susanne Lucae, Brenda W. Penninx, Johannes H. Smit, Gary Curhan, Poorva Mudgal, Robert M. Plenge, Laura Portas, Ivana Persico, Mirna Kirin, James F. Wilson, Irene Mateo Leach, Wiek H. van Gilst, Anuj Goel, Halit Ongen, Albert Hofman, Fernando Rivadeneira, Andre G. Uitterlinden, Medea Imboden, Arnold von Eckardstein, Francesco Cucca, Ramaiah Nagaraja, Maria Grazia Piras, Matthias Nauck, Claudia Schurmann, Kathrin Budde, Florian Ernst, Susan M. Farrington, Evropi Theodoratou,

Inga Prokopenko, Michael Stumvoll, Antti Jula, Markus Perola, Veikko Salomaa, So-Youn Y. Shin, Tim D. Spector, Cinzia Sala, Paul M. Ridker, Mika Khnen, Jorma Viikari, Christian Hengstenberg, Christopher P. Nelson, James F. Meschia, Michael A. Nalls, Pankaj Sharma, Andrew B. Singleton, Naoyuki Kamatani, Tanja Zeller, Michel Burnier, John Attia, Maris Laan, Norman Klopp, Hans L. Hillege, Stefan Kloiber, Hyon Choi, Mario Pirastu, Silvia Tore, Nicole M. Probst-Hensch, Henry Vlzke, Vilmundur Gudnason, Afshin Parsa, Reinhold Schmidt, John B. Whitfield, Myriam Fornage, Paolo Gasparini, David S. Siscovick, Ozren Polaek, Harry Campbell, Igor Rudan, Nabila Bouatia-Naji, Andres Metspalu, Ruth J. F. Loos, Cornelia M. van Duijn, Ingrid B. Borecki, Luigi Ferrucci, Giovanni Gambaro, Ian J. Deary, Bruce H. R. Wolffenbuttel, John C. Chambers, Winfried Mrz, Peter P. Pramstaller, Harold Snieder, Ulf Gyllensten, Alan F. Wright, Gerjan Navis, Hugh Watkins, Jacqueline C. M. Witteman, Serena Sanna, Sabine Schipf, Malcolm G. Dunlop, Anke Tnjes, Samuli Ripatti, Nicole Soranzo, Daniela Toniolo, Daniel I. Chasman, Olli Raitakari, W. H. Linda Kao, Marina Ciullo, Caroline S. Fox, Mark Caulfield, Murielle Bochud, and Christian Gieger. "Genome-wide association analyses identify 18 new loci associated with serum urate concentrations." *Nat Genet*, **45**(2):145–54, 2 2013.

KCP02. Rick A. Kittles, Weidong Chen, Ramesh K. Panguluri, Chiledum Ahaghotu, Aaron Jackson, Clement A. Adebamowo, Robin Griffin, Tyisha Williams, Flora Ukoli, Lucile Adams-Campbell, John Kwagyan, William Isaacs, Vincent Freeman, and Georgia M. Dunston. "CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification?" *Hum Genet*, **110**(6):553–60, 6 2002.

KFT07. Joost J. B. Keurentjes, Jingyuan Fu, Inez R. Terpstra, Juan M. Garcia, Guido van den Ackerveken, L. Basten Snoek, Anton J. M. Peeters, Dick Vreugdenhil, Maarten Koornneef, and Ritsert C. Jansen. "Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci." *Proc Natl Acad Sci U S A*, **104**(5):1708–13, 1 2007.

KKW10. Andrew Kirby, Hyun Min Kang, Claire M. Wade, Chris Cotsapas, Emrah Kostem, Buhm Han, Nick Furlotte, Eun Yong Kang, Manuel Rivas, Molly A. Bogue, Kelly A. Frazer, Frank M. Johnson, Erica J. Beilharz, David R. Cox, Eleazar Eskin, and Mark J. Daly. "Fine mapping in 94 inbred mouse strains using a high-density haplotype resource." *Genetics*, **185**(3):1081–95, 7 2010.

KLE11. Emrah Kostem, Jose A. Lozano, and Eleazar Eskin. "Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms." *Genetics*, **188**(2):449–60, 6 2011.

KSS10. Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Y. Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. "Variance component model to account for sample structure in genome-wide association studies." *Nat Genet*, **42**(4):348–54, 4 2010.

KTN13. Fredrik H. Karlsson, Valentina Tremaroli, Intawat Nookaew, Gran Bergstrm, Carl Johan Behre, Bjrn Fagerberg, Jens Nielsen, and Fredrik Bckhed. "Gut metagenome in European women with normal, impaired and diabetic glucose control." *Nature*, **498**(7452):99–103, 6 2013.

KVS12. Arthur Korte, Bjarni J. Vilhjlmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. "A mixed-model approach for genome-wide association studies of correlated traits in structured populations." *Nat Genet*, **44**(9):1066–71, 9 2012.

KYE08. Hyun Min Kang, Chun Ye, and Eleazar Eskin. "Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots." *Genetics*, **180**(4):1909–25, 12 2008.

KYL14. Gleb Kichaev, Wen-Yun Y. Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, and Bogdan Pasaniuc. "Integrating functional data to prioritize causal variants in statistical fine-mapping studies." *PLoS Genet*, **10**(10):e1004722, 10 2014.

KZW08. Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. "Efficient control of population structure in model organism association mapping." *Genetics*, **178**(3):1709–23, 3 2008.

LBR13. Donghyung Lee, T. Bernard Bigdeli, Brien P. Riley, Ayman H. Fanous, and Silviu-Alin A. Bacanu. "DIST: direct imputation of summary statistics for unmeasured SNPs." *Bioinformatics*, **29**(22):2925–7, 11 2013.

LBT05. Ruth E. Ley, Fredrik Bckhed, Peter Turnbaugh, Catherine A. Lozupone, Robin D. Knight, and Jeffrey I. Gordon. "Obesity alters gut microbial ecology." *Proc Natl Acad Sci U S A*, **102**(31):11070–5, 8 2005.

LDB96. D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. "Expression monitoring by hybridization to high-density oligonucleotide arrays." *Nat Biotechnol*, **14**(13):1675–80, 12 1996.

Lin05. D. Y. Lin. "An efficient Monte Carlo approach to assessing statistical significance in genomic studies." *Bioinformatics*, **21**(6):781–7, 3 2005.

LKS10. Jennifer Listgarten, Carl Kadie, Eric E. Schadt, and David Heckerman. "Correction for hidden confounders in the genetic analysis of gene expression." *Proc Natl Acad Sci U S A*, **107**(38):16465–70, 9 2010.

LLH13. Jennifer Listgarten, Christoph Lippert, and David Heckerman. "Nat Genet.", 5 2013.

LLK12. Jennifer Listgarten, Christoph Lippert, Carl M. Kadie, Robert I. Davidson, Eleazar Eskin, and David Heckerman. "Nat Methods.", 6 2012.

LLL11. Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. "FaST linear mixed models for genome-wide association studies." *Nat Methods*, **8**(10):833–5, 2011.

LS07. Jeffrey T. Leek and John D. Storey. "Capturing heterogeneity in gene expression studies by surrogate variable analysis." *PLoS Genet*, **3**(9):1724–35, 9 2007.

LTB15. Po-Ru R. Loh, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjálmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman, Paul M. Ridker, Benjamin M. Neale, Bonnie Berger, Nick Patterson, and Alkes L. Price. "Efficient Bayesian mixed-model analysis increases association power in large cohorts." *Nat Genet*, **47**(3):284–90, 3 2015.

LVB13. Yi Lu, Veronique Vitart, Kathryn P. Burdon, Chiea Chuen Khor, Yelena Bykhovskaya, Alireza Mirshahi, Alex W. Hewitt, Demelza Koehn, Pirro G. Hysi, Wishal D. Ramdas, Tanja Zeller, Eranga N. Vithana, Belinda K. Cornes, Wan-Ting T. Tay, E. Shyong Tai, Ching-Yu Y. Cheng, Jianjun Liu, Jia-Nee N. Foo, Seang Mei Saw, Gudmar Thorleifsson, Kari Stefansson, David P. Dimasi, Richard A. Mills, Jenny Mountain, Wei Ang, Ren Hoehn, Virginie J. M. Verhoeven, Franz Grus, Roger Wolfs, Raphale Castagne, Karl J. Lackner, Henrit Springelkamp, Jian Yang, Fridbert Jonasson, Dexter Y. L. Leung, Li J. Chen, Clement C. Y. Tham, Igor Rudan, Zoran Vatavuk, Caroline Hayward, Jane Gibson, Angela J. Cree, Alex MacLeod, Sarah Ennis, Ozren Polasek, Harry Campbell, James F. Wilson, Ananth C. Viswanathan, Brian Fleck, Xiaohui Li, David Siscovick, Kent D. Taylor, Jerome I. Rotter, Seyhan Yazar, Megan Ulmer, Jun Li, Brian L. Yaspan, Ayse B. Ozel, Julia E. Richards, Sayoko E. Moroi, Jonathan L. Haines, Jae H. Kang, Louis R. Pasquale, R. Rand Allingham, Allison Ashley-Koch, Paul Mitchell, Jie Jin Wang, Alan F. Wright, Craig Pennell, Timothy D. Spector, Terri L. Young, Caroline C. W. Klaver, Nicholas G. Martin, Grant W. Montgomery, Michael G. Anderson, Tin Aung, Colin E. Willoughby, Janey L. Wiggs, Chi P. Pang, Unnur Thorsteinsdottir, Andrew J. Lotery, Christopher J. Hammond, Cornelia M. van Duijn, Michael A. Hauser,

Yaron S. Rabinowitz, Norbert Pfeiffer, David A. Mackey, Jamie E. Craig, Stuart Macgregor, and Tien Y. Wong. "Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus." *Nat Genet*, **45**(2):155–63, 2 2013.

MA01. Brian H. McArdle and Marti J. Anderson. "Fitting multivariate models to community data: a comment on distance-based redundancy analysis." *Ecology*, **82**(1):290–297, 2001.

MAC08. Mark I. McCarthy, Gonalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." *Nat Rev Genet*, **9**(5):356–69, 5 2008.

MCP04. Jonathan Marchini, Lon R. Cardon, Michael S. Phillips, and Peter Donnelly. "The effects of human population structure on large genetic association studies." *Nat Genet*, **36**(5):512–7, 5 2004.

MLB09. Jacob J. Michaelson, Salvatore Loguercio, and Andreas Beyer. "Detection and interpretation of expression quantitative trait loci (eQTL)." *Methods*, **48**(3):265–76, 7 2009.

ND13. Alexandra C. Nica and Emmanouil T. Dermitzakis. "Expression quantitative trait loci: present and future." *Philos Trans R Soc Lond B Biol Sci*, **368**(1620):20120362, 2013.

NGZ10. Dan L. Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. "Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS." *PLoS Genet*, **6**(4):e1000888, 4 2010.

NLS07. Caroline M. Nievergelt, Ondrej Libiger, and Nicholas J. Schork. "Generalized analysis of molecular variance." *PLoS Genet*, **3**(4):e51, 4 2007.

OHP12. Paul F. O'Reilly, Clive J. Hoggart, Yotsawat Pomyen, Federico C. F. Calboli, Paul Elliott, Marjo-Riitta R. Jarvelin, and Lachlan J. M. Coin. "MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS." *PLoS One*, **7**(5):e34861, 2012.

PGJ11. Christopher C. Park, Greg D. Gale, Simone de Jong, Anatole Ghazalpour, Brian J. Bennett, Charles R. Farber, Peter Langfelder, Andy Lin, Arshad H. Khan, Eleazar Eskin, Steve Horvath, Aldons J. Lusis, Roel A. Ophoff, and Desmond J. Smith. "Gene networks associated with conditional fear in mice identified using a systems genetics approach." *BMC Syst Biol*, **5**:43, 2011.

PLW06. Jeremy L. Peirce, Hongqiang Li, Jintao Wang, Kenneth F. Manly, Robert J. Hitzemann, John K. Belknap, Glenn D. Rosen, Shirlean Goodwin, Thomas R. Sutter, Robert W. Williams, and Lu Lu. "How replicable are mRNA expression QTL?" *Mamm Genome*, **17**(6):643–56, 6 2006.

PNO13. Brian W. Parks, Elizabeth Nam, Elin Org, Emrah Kostem, Frode Norheim, Simon T. Hui, Calvin Pan, Mete Civelek, Christoph D. Rau, Brian J. Bennett, Margarete Mehrabian, Luke K. Ursell, Aiqing He, Lawrence W. Castellani, Bradley Zinker, Mark Kirby, Thomas A. Drake, Christian A. Drevon, Rob Knight, Peter Gargalovic, Todd Kirchgessner, Eleazar Eskin, and Aldons J. Lusis. "Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice." *Cell Metab*, **17**(1):141–52, 1 2013.

PRR07. Ethan O. Perlstein, Douglas M. Ruderfer, David C. Roberts, Stuart L. Schreiber, and Leonid Kruglyak. "Genetic basis of individual differences in the response to small-molecule drugs in yeast." *Nat Genet*, **39**(4):496–502, 4 2007.

PZS14. Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P. Strachan, Nick Patterson, and Alkes L. Price. "Fast and accurate imputation of summary statistics enhances evidence of functional enrichment." *Bioinformatics*, **30**(20):2906–14, 10 2014.

Qua01. J. Quackenbush. "Computational analysis of microarray data." *Nat Rev Genet*, **2**(6):418–27, 6 2001.

RCB01. D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. "Linkage disequilibrium in the human genome." *Nature*, **411**(6834):199–204, 5 2001.

ROC13. Stephan Ripke, Colm O'Dushlaine, Kimberly Chambert, Jennifer L. Moran, Anna K. Khler, Susanne Akterin, Sarah E. Bergen, Ann L. Collins, James J. Crowley, Menachem Fromer, Yunjung Kim, Sang Hong Lee, Patrik K. E. Magnusson, Nick Sanchez, Eli A. Stahl, Stephanie Williams, Naomi R. Wray, Kai Xia, Francesco Bettella, Anders D. Borglum, Brendan K. Bulik-Sullivan, Paul Cormican, Nick Craddock, Christiaan de Leeuw, Naser Durmishi, Michael Gill, Vera Golimbet, Marian L. Hamshere, Peter Holmans, David M. Hougaard, Kenneth S. Kendler, Kuang Lin, Derek W. Morris, Ole Mors, Preben B. Mortensen, Benjamin M. Neale, Francis A. O'Neill, Michael J. Owen, Milica Pejovic Milovancevic, Danielle Posthuma, John Powell, Alexander L. Richards, Brien P. Riley, Douglas Ruderfer, Dan Rujescu, Engilbert Sigurdsson, Teimuraz Silagadze, August B. Smit, Hreinn Stefansson, Stacy Steinberg,

Jaana Suvisaari, Sarah Tosato, Matthijs Verhage, James T. Walters, Douglas F. Levinson, Pablo V. Gejman, Kenneth S. Kendler, Claudine Laurent, Bryan J. Mowry, Michael C. O'Donovan, Michael J. Owen, Ann E. Pulver, Brien P. Riley, Sibylle G. Schwab, Dieter B. Wildenauer, Frank Dudbridge, Peter Holmans, Jianxin Shi, Margot Albus, Madeline Alexander, Dominique Campion, David Cohen, Dimitris Dikeos, Jubao Duan, Peter Eichhammer, Stephanie Godard, Mark Hansen, F. Bernard Lerer, Kung-Yee Y. Liang, Wolfgang Maier, Jacques Mallet, Deborah A. Nertney, Gerald Nestadt, Nadine Norton, Francis A. O'Neill, George N. Papadimitriou, Robert Ribble, Alan R. Sanders, Jeremy M. Silverman, Dermot Walsh, Nigel M. Williams, Brandon Wormley, Maria J. Arranz, Steven Bakker, Stephan Bender, Elvira Bramon, David Collier, Benedicto Crespo-Facorro, Jeremy Hall, Conrad Iyegbe, Assen Jablensky, Rene S. Kahn, Luba Kalaydjieva, Stephen Lawrie, Cathryn M. Lewis, Kuang Lin, Don H. Linszen, Ignacio Mata, Andrew McIntosh, Robin M. Murray, Roel A. Ophoff, John Powell, Dan Rujescu, Jim Van Os, Muriel Walshe, Matthias Weisbrod, Durk Wiersma, Peter Donnelly, Ines Barroso, Jenefer M. Blackwell, Elvira Bramon, Matthew A. Brown, Juan P. Casas, Aiden P. Corvin, Panos Deloukas, Audrey Duncanson, Janusz Jankowski, Hugh S. Markus, Christopher G. Mathew, Colin N. A. Palmer, Robert Plomin, Anna Rautanen, Stephen J. Sawcer, Richard C. Trembath, Ananth C. Viswanathan, Nicholas W. Wood, Chris C. A. Spencer, Gavin Band, Cline Bellenguez, Colin Freeman, Garrett Hellenthal, Eleni Giannoulatou, Matti Pirinen, Richard D. Pearson, Amy Strange, Zhan Su, Damjan Vukcevic, Peter Donnelly, Cordelia Langford, Sarah E. Hunt, Sarah Edkins, Rhian Gwilliam, Hannah Blackburn, Suzannah J. Bumpstead, Serge Dronov, Matthew Gillman, Emma Gray, Naomi Hammond, Alagurevathi Jayakumar, Owen T. McCann, Jennifer Liddle, Simon C. Potter, Radhi Ravindrarajah, Michelle Ricketts, Avazeh Tashakkori-Ghanbaria, Matthew J. Waller, Paul Weston, Sara Widaa, Pamela Whittaker, Ines Barroso, Panos Deloukas, Christopher G. Mathew, Jenefer M. Blackwell, Matthew A. Brown, Aiden P. Corvin, Mark I. McCarthy, Chris C. A. Spencer, Elvira Bramon, Aiden P. Corvin, Michael C. O'Donovan, Kari Stefansson, Edward Scolnick, Shaun Purcell, Steven A. McCarroll, Pamela Sklar, Christina M. Hultman, and Patrick F. Sullivan. "Genome-wide association analysis identifies 13 new risk loci for schizophrenia." *Nat Genet*, **45**(10):1150–9, 10 2013.

RZL05. Alexander P. Reiner, Elad Ziv, Denise L. Lind, Caroline M. Nievergelt, Nicholas J. Schork, Steven R. Cummings, Angie Phong, Esteban Gonzlez Burchard, Tamara B. Harris, Bruce M. Psaty, and Pui-Yan Y. Kwok. "Population structure, admixture, and aging-related phenotypes in African American adults: the Cardiovascular Health Study." *Am J Hum Genet*, **76**(3):463–77, 3

2005.

SAB12. Gulnara R. Svishcheva, Tatiana I. Axenovich, Nadezhda M. Belonogova, Cornelia M. van Duijn, and Yurii S. Aulchenko. "Rapid variance components-based method for whole-genome association analysis." *Nat Genet*, **44**(10):1166–70, 10 2012.

SBB07. Richard S. Spielman, Laurel A. Bastone, Joshua T. Burdick, Michael Morley, Warren J. Ewens, and Vivian G. Cheung. "Common genetic variants account for differences in gene expression among ethnic groups." *Nat Genet*, **39**(2):226–31, 2 2007.

Sid67. Zbyn Sidk. "Rectangular confidence regions for the means of multivariate normal distributions." *Journal of the American Statistical Association*, **62**(318):626–633, 1967.

SK08. Erin N. Smith and Leonid Kruglyak. "Gene-environment interaction in yeast gene expression." *PLoS Biol*, **6**(4):e83, 4 2008.

SM05. S. R. Seaman and B. Mller-Myhsok. "Rapid simulation of P values for product methods and multiple-testing adjustment in association studies." *Am J Hum Genet*, **76**(3):399–408, 3 2005.

SMD03. Eric E. Schadt, Stephanie A. Monks, Thomas A. Drake, Aldons J. Lusis, Nam Che, Veronica Colinayo, Thomas G. Ruff, Stephen B. Milligan, John R. Lamb, Guy Cavet, Peter S. Linsley, Mao Mao, Roland B. Stoughton, and Stephen H. Friend. "Genetics of gene expression surveyed in maize, mouse and man." *Nature*, **422**(6929):297–302, 3 2003.

SNF07. Barbara E. Stranger, Alexandra C. Nica, Matthew S. Forrest, Antigone Dimas, Christine P. Bird, Claude Beazley, Catherine E. Ingle, Mark Dunning, Paul Flicek, Daphne Koller, Stephen Montgomery, Simon Tavaré, Panos Deloukas, and Emmanouil T. Dermitzakis. "Population genomics of human gene expression." *Nat Genet*, **39**(10):1217–24, 10 2007.

SRR07. Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J. Hudson, Alexandre Montpetit, Alexey V. Pshezhetsky, Marc Prentki, Barry I. Posner, David J. Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. "A genome-wide association study identifies novel risk loci for type 2 diabetes." *Nature*, **445**(7130):881–5, 2 2007.

SSH09. Chiara Sabatti, Susan K. Service, Anna-Liisa L. Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G. Jones, Noah A. Zaitlen, Teppo Varilo, Marika Kaakinen, Ulla Sovio, Aimo Ruokonen, Jaana Laitinen, Eveliina Jakkula, Lachlan Coin, Clive Hoggart, Andrew Collins, Hannu Turunen, Stacey Gabriel, Paul Elliot, Mark I. McCarthy, Mark J. Daly, Marjo-Riitta R. Jrvelin, Nelson B. Freimer, and Leena Peltonen. "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population." *Nat Genet*, **41**(1):35–46, 1 2009.

SSV06. Michael F. Seldin, Russell Shigeta, Pablo Villoslada, Carlo Selmi, Jaakko Tuomilehto, Gabriel Silva, John W. Belmont, Lars Klareskog, and Peter K. Gregersen. "European population substructure: clustering of northern and southern populations." *PLoS Genet*, **2**(9):e143, 9 2006.

SVP12. Vincent Segura, Bjarni J. Vilhjlmsson, Alexander Platt, Arthur Korte, mit Seren, Quan Long, and Magnus Nordborg. "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations." *Nat Genet*, **44**(7):825–30, 7 2012.

VP05. Benjamin F. Voight and Jonathan K. Pritchard. "Confounding from cryptic relatedness in case-control association studies." *PLoS Genet*, **1**(3):e32, 9 2005.

Was13. Larry Wasserman. *All of statistics : a concise course in statistical inference*. Springer Science & Business Media, illustrated edition, 2013.

WKH04. Xiaosong Wang, Ron Korstanje, David Higgins, and Beverly Paigen. "Haplotype analysis in multiple crosses to identify a QTL gene." *Genome Res*, **14**(9):1767–72, 9 2004.

WS06. Jennifer Wessel and Nicholas J. Schork. "Generalized genomic distance-based regression methodology for multilocus association analysis." *Am J Hum Genet*, **79**(5):792–806, 11 2006.

WY93. Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for P-Value AdjustmentA Wiley-Interscience publicationVolume 279 of Wiley Series in Probability and StatisticsWiley series in probability and mathematical statistics: Applied probability and statistics, ISSN 0271-6356*. John Wiley & Sons, illustrated edition, 1993.

YBM10. Jian Yang, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, Andrew C. Heath, Nicholas G. Martin, Grant W. Montgomery, Michael E. Goddard, and Peter M. Visscher. "Common SNPs explain a large proportion of the heritability for human height." *Nat Genet*, **42**(7):565–9, 7 2010.

YBW03. 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." *Nature genetics*, **35**(1):57–64, 2003.

YPB06. Jianming Yu, Gael Pressoir, William H. Briggs, Irie Vroh Bi, Masanori Yamasaki, John F. Doebley, Michael D. McMullen, Brandon S. Gaut, Dahlia M. Nielsen, James B. Holland, Stephen Kresovich, and Edward S. Buckler. "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." *Nat Genet*, **38**(2):203–8, 2 2006.

YZG14. Jian Yang, Noah A. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. "Advantages and pitfalls in the application of mixed-model association methods." *Nat Genet*, **46**(2):100–6, 2 2014.

ZAK07. Keyan Zhao, María José Aranzana, Sung Kim, Clare Lister, Chikako Shindo, Chunlao Tang, Christopher Toomajian, Honggang Zheng, Caroline Dean, Paul Marjoram, and Magnus Nordborg. "An Arabidopsis example of association mapping in structured samples." *PLoS Genet*, **3**(1):e4, 1 2007.

ZKT12. Weidong Zhang, Ron Korstanje, Jill Thaisz, Frank Staedtler, Nicole Harttman, Lingfei Xu, Minjie Feng, Liane Yanas, Hyuna Yang, William Valdar, Gary A. Churchill, and Keith Dipetrillo. "Genome-wide association mapping of quantitative traits in outbred mice." *G3 (Bethesda)*, **2**(2):167–74, 2 2012.

ZPG10. Noah Zaitlen, Bogdan Paaniuc, Tom Gur, Elad Ziv, and Eran Halperin. "Leveraging genetic variability across populations for the identification of causal variants." *Am J Hum Genet*, **86**(1):23–33, 1 2010.

ZS12a. Matthew A. Zapala and Nicholas J. Schork. "Statistical properties of multivariate distance matrix regression for high-dimensional data analysis." *Front Genet*, **3**:190, 2012.

ZS12b. Xiang Zhou and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies." *Nat Genet*, **44**(7):821–4, 7 2012.

ZS14. Xiang Zhou and Matthew Stephens. "Efficient multivariate linear mixed model algorithms for genome-wide association studies." *Nat Methods*, **11**(4):407–9, 4 2014.

ZWL07. Eleftheria Zeggini, Michael N. Weedon, Cecilia M. Lindgren, Timothy M. Frayling, Katherine S. Elliott, Hana Lango, Nicholas J. Timpson, John R. B. Perry, Nigel W. Rayner, Rachel M. Freathy, Jeffrey C. Barrett, Beverley Shields, Andrew P. Morris, Sian Ellard, Christopher J. Groves, Lorna W.

Harries, Jonathan L. Marchini, Katharine R. Owen, Beatrice Knight, Lon R. Cardon, Mark Walker, Graham A. Hitman, Andrew D. Morris, Alex S. F. Doney, Mark I. McCarthy, and Andrew T. Hattersley. "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes." *Science*, **316**(5829):1336–41, 6 2007.