UNIVERSITY OF CALIFORNIA

Los Angeles

Design of efficient and statistically powerful approaches for human genetics

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Jae Hoon Sul

2013

ABSTRACT OF THE DISSERTATION

Design of efficient and statistically powerful approaches for human genetics

by

Jae Hoon Sul

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2013

Professor Eleazar Eskin, Chair

The advent of genotyping and sequencing technologies has enabled human genetics to discover numerous genetic variants associated with many diseases and traits over the past decades. One of the most effective approaches to detect those variants has been genome-wide association studies (GWASs) that scan all variants found in genomes. GWASs collect people with a disease (called cases) and people without a disease (called controls) and compare allele frequencies between cases and controls to identify genetic variants associated the disease. This simple yet effective approach has been widely utilized by many studies, and more than 1,600 GWASs have been published during the last decade.

An underlying assumption of GWAS is that cases and controls are sampled from the same population. If they are not, then a phenomenon called population structure may cause spurious associations. Correcting for population structure in GWASs has been a very important problem in human genetics, and several methods have been proposed. However, those methods fail to correct for complex structure or are computationally too challenging for current GWAS datasets. I will introduce a new statistical approach that correctly removes effects of population structure and reduces the computational time from years to hours.

Recently, sequencing technologies that enable a detection of rare variants have received considerable attention and been utilized by many GWASs. In these studies, rare variants in a gene are often grouped together to test the aggregated effect of rare variants on disease susceptibility. However, there are many different approaches to combine information of multiple rare variants, and it is unknown which approach is optimal in detecting associations of rare variants. I will introduce two novel approaches to better identify a group of rare variants involved in a disease. I will show using simulations that our approaches outperform previous methods, and using real sequencing data, I will show that our methods can identify an association reported by a previous study.

Finally, I will introduce a statistical approach to identify expression quantitative trait loci (eQTL) or genetic variants that are associated with gene expression in multiple tissues. Recent technological developments and cost decreases have enabled eQTL studies to collect expression data in multiple tissues, but most studies focus on finding eQTLs in each tissue separately. I will introduce a statistical approach that combines results from multiple tissues to better identify eQTLs. I will show by using simulations and multiple tissue data from mouse that our approach detects many eQTLs undetected by traditional eQTL methods.

The dissertation of Jae Hoon Sul is approved.

Wei Wang

Stott Parker

Nelson Freimer

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2013

To my family: Hyunjung, Hannah, Kyoungja, Jinseob, Kyuwhi, and Suja. Their love and support have been invaluable to me

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to thank my wife, Hyunjung, and daughter, Hannah, for their support and love. They have been so supportive and patient while I was in this degree, and without them, I would have never finished this dissertation. I also would like to give thanks to my parents and parents-in-law who supported me from Korea.

I would like thank my adviser Prof. Eleazar Eskin. He taught me statistics, genetics, writing and everything that I needed to do my research throughout the degree. I also learnt how to do research from him, and he gave me invaluable advice on my career. I also thank my committee members, Prof. Wei Wang, Prof. Stott Parker, and Prof. Nelson Freimer who provided feedback and suggestion on this dissertation. In particular, I would like to thank Prof. Nelson Freimer for his mentorship and for providing me opportunities to work and gain experience on real GWAS datasets.

Lastly, I would like to thank my current and past colleagues in our lab. I learnt a lot from them, and especially, I would like to thank Buhm Han and Hyun Min Kang for their mentorship and for their full support on my research.

VITA

1994-1998 Attended Northwest School in Seattle, Washington

1998-2002 B.Sc., Computer Science, University of Washington, Seattle, Washington

2002-2003 M.Eng., Computer Science, Cornell University, Ithaca, New York

2003-2007 Junior Research Engineer, Digital Media Research Lab in LG Electronics, Seoul, Korea

2008-2013 Doctoral Student and Research Assistant, University of California, Los Angeles, California

2010-2013 Teaching Assistant, University of California Los Angeles, Los Angeles, California

PUBLICATIONS

Cornelius A. Rietveld, Sarah E. Medland, Jaime Derringer, Jian Yang, Tnu Esko, Nicolas W. Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, Arpana Agrawal, Eva Albrecht, Behrooz Z. Alizadeh, Najaf Amin, John Barnard, Sebastian E. Baumeister, Kelly S. Benke, Lawrence F. Bielak, Jeffrey A. Boatman, Patricia A. Boyle, Gail Davies, Christiaan de Leeuw, Niina Eklund, Daniel S. Evans, Rudolf Ferhmann, Krista Fischer, Christian Gieger, Hkon K. Gjessing, Sara HŁgg, Jennifer R. Harris, Caroline Hayward, Christina Holzapfel, Carla A. Ibrahim-Verbaas, Erik Ingelsson, Bo Jacobsson, Peter K. Joshi, Astanand Jugessur, Marika Kaakinen, Stavroula

Kanoni, Juha Karjalainen, Ivana Kolcic, Kati Kristiansson, Zoltn Kutalik, Jari Lahti, Sang H. Lee, Peng Lin, Penelope A. Lind, Yongmei Liu, Kurt Lohman, Marisa Loitfelder, George McMahon, Pedro Marques Vidal, Osorio Meirelles, Lili Milani, Ronny Myhre, Marja-Liisa Nuotio, Christopher J. Oldmeadow, Katja E. Petrovic, Wouter J. Peyrot, Ozren Polasek, Lydia Quaye, Eva Reinmaa, John P. Rice, Thais S. Rizzi, Helena Schmidt, Reinhold Schmidt, Albert V. Smith, Jennifer A. Smith, Toshiko Tanaka, Antonio Terracciano, Matthijs J.H.M. van der Loos, Veronique Vitart, Henry Vlzke, Jrgen Wellmann, Lei Yu, Wei Zhao, Jri Allik, John R. Attia, Stefania Bandinelli, Franois Bastardot, Jonathan Beauchamp, David A. Bennett, Klaus Berger, Laura J. Bierut, Dorret I. Boomsma, Ute Bltmann, Harry Campbell, Christopher F. Chabris, Lynn Cherkas, Mina K. Chung, Francesco Cucca, Mariza de Andrade, Philip L. De Jager, Jan-Emmanuel De Neve, Ian J. Deary, George V. Dedoussis, Panos Deloukas, Maria Dimitriou, Gudny Eiriksdottir, Martin F. Elderson, Johan G. Eriksson, David M. Evans, Jessica D. Faul, Luigi Ferrucci, Melissa E. Garcia, Henrik Grnberg, Vilmundur Gudnason, Per Hall, Juliette M. Harris, Tamara B. Harris, Nicholas D. Hastie, Andrew C. Heath, Dena G. Hernandez, Wolfgang Hoffmann, Adriaan Hofman, Rolf Holle, Elizabeth G. Holliday, Jouke-Jan Hottenga, William G. Iacono, Thomas Illig, Marjo-Riitta JŁrvelin, Mika KŁhnen, Jaakko Kaprio, Robert M. Kirkpatrick, Matthew Kowgier, Antti Latvala, Lenore J. Launer, Debbie A. Lawlor, Terho LehtimŁki, Jingmei Li, Paul Lichtenstein, Peter Lichtner, David C. Liewald, Pamela A. Madden, Patrik K.E. Magnusson, Tomi E. MŁkinen, Marco Masala, Matt McGue, Andres Metspalu, Andreas Mielck, Michael B. Miller, Grant W. Montgomery, Sutapa Mukherjee, Dale R. Nyholt, Ben A. Oostra, Lyle J. Palmer, Aarno Palotie, Brenda Penninx, Markus Perola, Patricia A. Peyser, Martin Preisig, Katri RŁikknen, Olli T. Raitakari, Anu Realo, Susan M. Ring, Samuli Ripatti, Fernando Rivadeneira, Igor Rudan, Aldo Rustichini, Veikko Salomaa, Antti-Pekka Sarin, David Schlessinger, Rodney J. Scott, Harold Snieder, Beate St Pourcain, John M. Starr, **Jae Hoon Sul**, Ida Surakka, Rauli

Svento, Alexander Teumer, The LifeLines Cohort Study, Henning Tiemeier, Frank J.A. van Rooij, David R. Van Wagoner, Erkki Vartiainen, Jorma Viikari, Peter Vollenweider, Judith M. Vonk, Grard Waeber, David R. Weir, H.-Erich Wichmann, Elisabeth Widen, Gonneke Willemsen, James F. Wilson, Alan F. Wright, Dalton Conley, George Davey-Smith, Lude Franke, Patrick J.F. Groenen, Albert Hofman, Magnus Johannesson, Sharon L.R. Kardia, Robert F. Krueger, David Laibson, Nicholas G. Martin, Michelle N. Meyer, Danielle Posthuma, A. Roy Thurik, , Nicholas J. Timpson, Andr G. Uitterlinden, Cornelia M. van Duijn, Peter M. Visscher, Daniel J. Benjamin, David Cesarini, Philipp D. Koellinger, “GWAS of 126,559 individuals identifies genetic variants associated with educational attainment.” *Science*. 340, 1467-1471, 2013.

Oron Navon*, **Jae Hoon Sul***, Buhm Han, Lucia Conde, Paige Bracci, Jacques Riby, Christine Skibola, Eleazar Eskin, Eran Halperin, “Rare Variants Association Testing with Sequencing Errors and Pooling.” *Genetics*. 194, 769-779, 2013.

Jae Hoon Sul*, Buhm Han*, Chun Ye*, Ted Choi, and Eleazar Eskin, “Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches.” *PLoS Genetics*. 9, e1003491, 2013.

Jae Hoon Sul, and Eleazar Eskin, “Mixed models can correct for population structure for genomic regions under selection.” *Nature Reviews Genetics*. 14, 300-300, 2013.

Jurjen J. Luykx, Steven C. Bakker, Eef Lentjes, Marjolein Neeleman, Eric Strengman, Laura Mentink, Joseph DeYoung, Simone de Jong, **Jae Hoon Sul**, Eleazar Eskin, Kristel van Eijk, Jessica van Setten, Jacobine E. Buizer-Voskamp, Rita M. Cantor, Ake Lu, Marjolein van Amerongen, Eric P.A. van Dongen, Peter Keijzers, T Kappen, P Borgdorff, P Bruins, E M. Derks, Ren S. Kahn, Roel A. Ophoff, “Genome-Wide Association Study of Monoamine Metabolite Levels in Human Cerebrospinal Fluid.” *Molecular Psychiatry*. 2013.

Michael B. Miller, Saonli Basu, Julie Cunningham, Eleazar Eskin, Steven M. Malone, William S. Oetting, Nicholas Schork, **Jae Hoon Sul**, William G. Iacono and Matt McGue, “The Minnesota Center for Twin and Family Research Genome-Wide Association Study.” *Twin Research and Human Genetics*. 15, 767-774, 2012.

Jae Hoon Sul, Buhm Han, Dan He, and Eleazar Eskin, “An Optimal Weighted Aggregated Association Test for Identification of Rare Variants Involved in Common Diseases.” *Genetics*. 188, 181-8, 2011

Jae Hoon Sul, Buhm Han, and Eleazar Eskin, “Increasing Power of Groupwise Association Test with Likelihood Ratio Test.” *Journal of Computational Biology*. 18, 1611-1624, 2011.

Jae Hoon Sul, Buhm Han, and Eleazar Eskin, “Increasing Power of Groupwise Association Test with Likelihood Ratio Test.” *In Proceedings ofthe Fifteenth Annual Conference on Research in Computational Biology (RECOMB-2011)*. Vancouver, Canada: March 28-31, 2011

Hyun Min Kang*, **Jae Hoon Sul***, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin. “Variance component model to account for sample structure in genome-wide association studies.” *Nature Genetics*. 42, 348354, 2010.

Jason L. Stein, Xue Hua, Jonathan H. Morra, Suh Lee, Derrek P. Hibar, April J. Ho, Alex D. Leow, Arthur W. Toga, **Jae Hoon Sul**, Hyun Min Kang, Eleazar Eskin, Andrew J. Saykin, Li Shen, Tatiana Foroud, Nathan Pankratz, Matthew J. Huentelman, David W. Craig, Jill D. Gerber, April N. Allen, Jason J. Corneveaux, Dietrich A. Stephan, Jennifer Webster, Bryan M. DeChairo, Steven G. Potkin, Clifford R. Jack Jr., Michael W. Weiner, Paul M. Thompson, and the Alzheimers Disease Neuroimaging Initiative. "Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimers disease." *Neuroimage*. 51, 542-554, 2010.

Paul Ginsparg, Paul Houle, Thorsten Joachims, and **Jae-Hoon Sul**. "Mapping Subsets of Scholarly Information." *Proceedings of the National Academy of Sciences of the USA*, 10.1073, 101, 5236- 5240, 2004.

* These authors contributed equally to the work.

CHAPTER 1

Introduction

Humans have three billion letters or nucleotides of DNA sequences, and a majority of them (99.9%) are identical for all humans. It is the 0.1% of DNA sequences that make us all different, and they are referred as genetic variants. One of the primary goals in genetics studies is to identify DNA sequences or genetic variants that cause a difference in traits or diseases. For example, individuals who carry a mutation at a certain DNA position may be more susceptible to a certain disease than individuals without the mutation. Similarly, certain genetic variants may influence the height or weight of individuals. Identifying genetic variants related to or "associated" with traits or diseases is very important to uncover the roles of genetics in diseases, which is fundamental in understanding diseases and searching for their treatments.

One approach to detect genetic variants that cause a disease is the linkage analysis [LS94]. It collects DNA information of a family that consists of both unaffected and affected individuals. It then tries to detect the segments of DNA that segregate with a disease within a family. This was the predominant approach when a technology to obtain DNA information of individuals was limited; the linkage analysis used hundreds of markers such as microsatellite markers that track the number of short repeats in genomes. The linkage analysis has been successful in identifying genetic variants or genes that cause Mendelian disorders such as Huntington's disease [Wal07] and Cystic Fibrosis [KRB89]. Mendelian disorders are diseases in which there is usually a single gene that causes them. The linkage analysis, however, was shown to be unsuccessful in detecting genetic basis of the complex diseases, opposite to Mendelian diseases, that involve many genes or genetic variants such as autism and type I and II diabetes.

To identify genetic mechanisms of complex traits or diseases, association studies have been proposed. Association studies collect individuals with a disease called cases and individuals without a disease called controls and compare their DNA information. Specifically, an allele is one of the forms of genes or genetic variants, and association studies determine whether certain alleles are overrepresented in cases than in controls. The advent of microarray genotyping technologies enabled genetic studies to collect information on numerous single nucleotide polymorphisms (SNPs) usually in the order 500K to one million in a genome-wide level. Studies that perform associations studies on genetic variants collected genome-wide are called Genome-wide association studies (GWASs). GWASs have been very successful in detecting novel genetic variants associated with many complex disease and traits [Con07, SSH09, AHK00, BKK94], and more than 1,600 GWASs have been published over the past decade.

Current GWASs, however, suffer several drawbacks. First, they are susceptible to population stratification. GWASs assume that cases and controls are sampled from the same population. If there are not, a phenomenon called population stratification or structure may cause GWASs to detect spurious associations [PSR00]. For example, GWASs may collect individuals from two populations. If cases consist of more individuals from one population than those from the other population and controls have more individuals from the other population, then any genetic variants that differ between the two populations and does not cause a disease will be spuriously associated with the disease. Correcting for population structure in GWASs has been a very important problem in human genetics, and several statistical approaches have been proposed [DR99b, PPP06].

Another drawback of current GWASs is that they can only detect associations of common SNPs because the microarray genotyping technologies are designed to detect polymorphisms in genomes that are common in populations. Results of GWASs have indicated that common SNPs have small effects in complex traits, and they do not explain all causes related to diseases [MCC09a]. Recently, it has been suggested that rare variants that occur rarely among individuals may cause diseases [KPS07, CKP04, FWW04], and researchers have been interested in the role of rare variants in diseases. To detect rare variants, sequencing technologies that discover almost all DNA sequences of one's genome are used. While it cost $3 billion and took 13 years to sequence the first human genome [Con04], the advent of next generation sequencing technologies reduced the cost and time to sequence significantly; currently, it costs about $1,000 and takes a week to sequence a genome. This has allowed many GWASs to adopt sequencing technologies and to detect rare variants. Many methods that try to detect associations of rare variants have also been proposed recently [MB09, LL08, PKB10].

The underlying requirements of statistical methods that attempt to solve the problems in GWASs are that first they need to be efficient. GWASs generate an enormous amount of data; it collects DNA information of millions of genetic variants from thousands or tens of thousands of individuals. Current sequencing technologies can easily generate terabytes of information, and statistical methods or algorithms must be efficient enough to handle this large data. Another requirement is that methods need to statistically powerful. This means that they need to be able to detect associations of genetic variants effectively. Genetics data are often expensive to obtain; it is expensive to collect DNA of many individuals, to genotype or sequence their DNA, and to obtain their traits information or disease status. Hence, we want to utilize the available data as much as possible to find associations of genetic variants by designing statistically powerful approaches.

My thesis work focuses on developing efficient and statistically powerful methods to solve aforementioned and other problems in current human genetics. In what follows, I will briefly explain a background of a problem and a method that I developed.

Chapter 2: Variance component model to correct for population structure

Population structure as mentioned above may cause spurious associations in GWASs. To correct for population structure, several statistical methods such as genomic control [DR99b] and principal component analysis [PPP06] have been utilized. However, they fail to correct for complex population structure or hidden relatedness in which individuals in GWASs are related. A variance component model or mixed model [KZW08] was proposed and shown to effectively correct for the complex population structure in model organisms such as inbred mouse. It computes a pairwise relatedness between individuals and incorporates this relatedness to correct for population structure or hidden relatedness. However, its complicated algorithm is not efficient enough for large human GWAS datasets. In Chapter 2, I propose a variance component approach that reduces the computational time for analyzing large GWAS datasets from years to hours. I use the insight that genetic variants have small effects in humans, which allowed me to simplify the original variance component model. I show by using two human GWAS datasets that this method outperforms both principal component analysis and genomic control in correcting for population structure.

Chapter 3: Aggregated association test for rare variants

In traditional GWASs where common SNPs are collected, each common SNP is tested individually to detect an association between a disease and the SNP. Statistical power of this test is proportional to the minor allele frequency of SNPs; the more common the SNP is, the higher power we achieve. This means that we have very low power to detect associations of rare variants. In other words, it is extremely difficult to identify a single rare variant involved in a disease. To increase the statistical power of rare variants, a groupwise association test has been proposed that groups rare variants in a gene and discovers the aggregated effect of rare variants on disease susceptibility. The idea behind this approach is that if a certain gene is involved in a disease, many rare variants within the gene will disrupt the function of the gene and are associated with the disease. In Chapter 3, I present a method that groups rare variants and computes a weighted sum of differences between case and control mutation counts. I show by using simulated data that this approach is more powerful than previous methods. In addition, by using real mutation screening data of the susceptibility gene for ataxia telangiectasia, I show that this approach can identify an association reported by a previous study that used a different statistical approach.

Chapter 4: Likelihood ratio test to increase power of groupwise association test

As discussed in Chapter 3, we take a weighted sum of statistics among multiple rare variants to determine whether a gene is involved in a disease. However, not every rare variant has effects in a disease, and if we include non-causal rare variants into our weighted sum of statistics, we lose power. This means that we want to only include causal variants into our statistical framework, but it is not known which variants are causal. To overcome this, previous methods including the method discussed in Chapter 3 used the prior information that specifies how likely each variant is deleterious. Another source of information that can be used to determine causal variants is the observed data itself since case individuals are likely to carry more casual variants than control individuals. In Chapter 4, I introduce a likelihood ratio test (LRT) for rare variants that use both data and prior information to infer which variants are causal and uses this finding to determine whether a group of rare variants is involved in a disease. I show by simulations that this method outperforms previous methods. I also develop an efficient permutation test and decomposition of likelihood ratio to increase the computational speed of our method. With this optimization, we can efficiently compute a statistic for LRT and perform a permutation test at a genome-wide level.

Chapter 5: Combining mixed model and meta-analysis to detect eQTLs from multiple tissues

Expression Quantitative Trait Loci (eQTL) studies attempt to identify associations between genetic variants and gene expression. Until recently, eQTL studies collected gene expression data from a specific tissue and performed association studies. Recent advancements in gene expression technologies have allowed studies to collect gene expression from multiple tissues. One advantage of multiple tissue datasets is that studies can combine results from different tissues to identify eQTLs more accurately than examining each tissue separately. The idea of aggregating results of multiple tissues is closely related to the idea of meta-analysis which aggregates results of multiple studies. One challenge in applying meta-analysis to multiple tissues dataset is that studies usually collect tissues from the same individuals, which violates the assumption of meta-analysis that studies are independent. Another challenge is that eQTLs may have effects in only a single tissue, in all tissues, or in a subset of tissues. This heterogeneity in terms of effects presents a key challenge to detect eQTLs. In Chapter 5, I develop a statistical framework that combines mixed model and meta-analysis to overcome these two challenges. I show by using simulations and multiple tissue data from mouse that this approach detects many eQTLs that traditional eQTL methods do not detect.

CHAPTER 2

Variance component model to correct for population structure

2.1 Background

GWASs may utilize either case-control cohorts to test for associations with diseases or population cohorts to identify associations with quantitative traits. In both cases, it is assumed that the cohorts consist of unrelated individuals that share the same population background, although this may not hold in practice for cohorts used in many current GWASs. The presence of related individuals within a study sample results in sample structure, a term that encompasses population stratification and hidden relatedness. Population stratification refers to the inclusion of individuals from different populations within the same study sample. Hidden relatedness refers to the presence of unknown genetic relationships between individuals within the study sample [VP05, WAH06]. The effects of sample structure present in cohorts used for genetic association studies have been well documented and identified as a cause for some spurious associations [NAM01, HYH05].

Although limiting study samples entirely to unrelated individuals may be difficult or impossible, genotype data provides valuable information on the sample structure that can inform genetic association analysis. For example, the STRUCTURE software [PSR00] uses genotype data to partition the sample into subpopulations within which there is no sample structure and subsequently carries out association tests within the identified subpopulations. To eliminate the effects of hidden relatedness, one can estimate the proportion of genes identical by descent (IBD) between any pair of individuals in the sample and exclude from the analysis those individuals that appear closely related [VP05, Con07]. Population stratification and hidden relatedness, however, constitute just two extreme manifestations of sample structure, and methods are needed to correct for other forms of sample structure. In the genomic control approach [DR99a, BDR02], which has been widely adopted, the distribution of test statistics from the single-marker analysis is used to estimate the inflation factor, λ, with which the test statistics are subsequently rescaled, constraining the risk of false positives. The EIGENSTRAT software [PPP06, PPR06] uses principal components analysis (PCA) to detect and describe sample structure and has been widely used in GWASs. Some principal components may represent broad differences across individuals within a given data set, effectively capturing a few major axes of population structure, but it is unclear how to interpret the rest of the principal components as surrogates of sample structure [NS08, NJB08]. Currently, association studies typically use a combination of these strategies, first identifying close relatives to remove them from analysis, then correcting for broad sample structure using principal components or spatial information and finally correcting for the residual inflation with genomic control [Con07, SSH09, CGK09].

If we knew the complete genealogy of the population, we could, in principle, apply a variance component method to model the effects of the genetic relationships on the phenotypes; this approach would be similar in spirit to the classical polygenic model [Fis18] directly applied to association mapping [OAM01]. The variance component would capture the complex mixture of both population stratification and hidden relatedness that directly results from the genealogy and would correct for these relationships during the mapping. Although the exact genetic relationships between individuals in the samples are unknown, we could take advantage of the high-density genotype information to empirically estimate the level of relatedness between reportedly unrelated individuals.

In this chapter, we report an approach for correcting for sample structure within GWASs, based on a linear mixed model (also sometimes referred to as a mixed linear model) with an empirically estimated relatedness matrix to model the correlation between phenotypes of sample subjects. Similar variance component approaches have been used successfully in animal models [YPB06, ZAK07, KZW08]. However, applying even an efficient implementation of a variance component approach, such as EMMA (ref. [KZW08]), is computationally intractable for data sets consisting of thousands of individuals, owing to the heavy computational burden in the estimation of variance parameters. Capitalizing on the characteristics of complex traits in humans, we make a few simplifying assumptions that allow us to markedly increase the speed of computations, making our approach readily applicable to GWASs with tens of thousands of individuals assayed at hundreds of thousands of SNPs. For most genetic association studies in humans, because the effect of any given locus on the trait is very small [MCC09b], we need to estimate the variance parameters only once for each data set, and we can globally apply them to each marker. Our computational improvements reduce the running time for the analysis of a typical GWAS data set using a variance component model from years to hours. The advantage of the variance component approach is that the empirical relatedness matrix encodes a wide range of sample structures, including both hidden relatedness and population stratification. Principal componentbased methods, in contrast, by estimating major axes of the pairwise genetic similarity matrix, capture some, but not all, of the sample structure, as we show in this chapter.

We evaluate our method using two human GWAS data sets, from the 1966 Northern Finland Birth Cohort (NFBC66) [SSH09, Ran69] and the Wellcome Trust Case Control Consortium (WTCCC) [Con07]. The NFBC66 is based on a founder population, which is expected to minimize genetic heterogeneity, increasing the chances of mapping genes underlying traits of interest [VP04]. This is an ideal sample to evaluate our method because a detailed study [JRV08] of this data set has revealed the presence of substantial population structure that could influence the results of genetic association studies. In addition, we apply our method to the case-control studies for seven common complex diseases conducted by the WTCCC [Con07]. In both data sets, our method consistently outperforms both genomic control and principal component analysis. We term our method EMMA eXpedited (EMMAX) because it builds on the previous approach EMMA (ref. [KZW08]) and markedly reduces the computational cost.

2.2 Methods

2.2.1 Variance component model

We consider here the simplest form of Fisher [Fis18]'s polygenic model. Let $Z_{i,j}$ be the contribution of factor j to person i, then we assume that the phenotype y_i can be modeled as

$$y_i = \sum_{j=1}^{J} Z_{i,j} + \epsilon_i, \qquad \mathrm{E}(\epsilon_i) = 0, \qquad \mathrm{Cov}(\epsilon_{i1}, \epsilon_{i2}) = 0 \text{ if } i_1 \neq i_2 \tag{2.1}$$

with ϵ_i being a random variable representing environmental effects on the phenotype. In equation (2.1) and throughout the chapter, we include only variables accounting for the genetic factors, and all genetic factors contribute additively. This is purely a convenient assumption to simplify notation, and nongenetic factors can be modeled as additional regressors with a straight-forward extension. Epistatic loci can be incorporated by including additional interaction terms in equation (2.1) to model a diverse set of possible types of interactions

Let the vector $Y = \{y_i, \ldots, y_n\}$ contain the phenotypes of the individuals computed from a pedigree. Assuming that the environmental components are uncorrelated, the variance covariance structure of Y depends on the number of genes shared among subjects. In absence of dominance effects, we have

$$\text{Var}(Y) = 2\sigma_a^2 \Phi + \sigma_e^2 I, \tag{2.2}$$

where Φ is the matrix of kinship coefficients between each pair of individuals in the pedigree, and I is an identity matrix [FM96]. Let σ_a^2 represent the parameter for additive genetic variance and σ_e^2 represent the parameter for random environmental variance. An analysis of variance with random effects leads to the estimates of σ_a^2 and σ_e^2, and in turn to the evaluation of heritability $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ [FM96].

In linkage studies, this decomposition of variance is carried one step further. By tracking the transmission of marker genes in the vicinity of locus k, one can calculate the conditional kinship coefficients (Φ_k, probabilities that two genes sampled from two individuals at locus k are IBD), and decompose the variance Var(Y) to emphasize the contribution of the k-th locus

$$\text{Var}(Y) = 2\sigma_{ak}^2 \Phi_k + 2\Phi \sigma_a^2 + \sigma_e^2 I.$$

To investigate the contribution of locus k to the phenotype, one tests the null hypothesis that $\sigma_{ak}^2 = 0$. The values of the variance parameters are estimated with maximum likelihood procedures [Lan02].

In association studies, using a much denser set of genotypes, we aim to associate the phenotypes directly to the alleles at marker loci; in other words, our goal is to estimate fixed effects. Assuming additive effects only, equation (2.1) can be translated to the following regression framework:

$$y_i = \beta_0 + \sum_{k=1}^{M} \beta_k X_{ik} + \epsilon_i, \tag{2.3}$$

with Var$(\epsilon) = \sigma_e^2 I$, and X_k being the individuals' minor allele counts at locus $k \in \{1, 2, L, M\}$ (for simplicity, we assume all markers are biallelic). Our goal is to identify which elements in the $M \times 1$ vector β are different from 0.

While model (2.3) is fundamentally a multivariate one, association studies are typically carried out by testing the hypothesis $H_0 : \beta_k = 0$, for each of the M loci, one locus at a time, on the basis of model

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_{i\bar{k}} \tag{2.4}$$

where β_k is the effect size of marker k, and the error term $\eta_{i\bar{k}} = \sum_{s \neq k} \beta_s X_{is} + \epsilon_i$. With respect to model (2.3), model (2.4) is mis-specified if $\eta_{i\bar{k}}$ are assumed to be independently and identically distributed (i.i.d.): relevant regressors are omitted, or, in other words, we ignore the polygenic background of the trait.

The appropriate statistical methods to estimate β_k in (2.4) depends on the nature of the sample. If the n individuals are related with a known degree of relatedness, the variance covariance of $\eta_{i\bar{k}}$ in model (2.4) can be represented approximately as in (2.2). That is, the effect of the genotype at locus k can be modeled as a main effect, whereas the relationships among all individuals are taken into account by means of variance components of random polygenic effects [OAM01]. This model is sometimes referred to as an instance of a "mixed effect" model [YPB06].

If the n individuals are unrelated and there is no dependence across the genotypes, so that the $\eta_{i\bar{k}}$ values are i.i.d., a simple linear regression would make appropriate inference. However, these conditions are not easily met. First, because of linkage disequilibrium, X_k values corresponding to markers with close genomic position are correlated. Moreover, neither the homogeneity of population background nor the level of relatedness are easily controlled in the sampling stage. If the n individuals in the sample belong to distinct populations or are (albeit distantly) related, one can expect a substantial correlation between the rows and columns of X. This translates to bias in the estimate of β_k from equation (2.4), and the distribution of $\hat{\beta}_k$, a best unbiased linear estimation of β_k, is different from what is assumed in standard linear regression (that is, the $\eta_{i\bar{k}}$ values in (2.4) are not i.i.d.).

Using dense, genome-wide genotype data, it has become possible to estimate the degree of relationship or kinship matrix between independently ascertained subjects [LR99, EDB00, TH00] in the absence of genealogical information. With an estimated kinship matrix one can, in principle, use variance component techniques in linear mixed models (as in Ober et al. [OAM01]) to analyze population samples. If many SNPs are involved in a trait and the contribution of each SNP to the total trait variance is almost negligible, as appears to be the case for human quantitative traits [20,56], the variance components for $\eta_{i\bar{k}}$ in (2.4) can be approximated to $\eta_i = \sum_{s=1}^{M} \beta_k X_{ik} + \epsilon_i$ and may not need to be estimated separately for each SNP. Instead, one might estimate the values σ_a^2 and σ_e^2 from a variance decomposition model as in equation (2.2), keep them fixed, and then estimate the parameter β_k in equation (2.4) using a GLS procedure.

2.2.2 Application to quantitative traits

We used the following procedure to analyze human population samples in association studies for quantitative traits. Let n be the sample size, p the total number of genotyped SNPs and Y the vector of observed phenotypes.

1. Use the genotype data to calculate the $n \times n$ matrix $\hat{S}$ pairwise genetic relatedness between individuals, such as identity by state (IBS) or Balding-Nichols matrix, and normalize $\hat{S}$ to have sample variance 1 using a Gower's centered matrix

[MA01].

$$\hat{S}_N = \frac{(n-1)\hat{S}}{\text{Tr}(P\hat{S}P)} \tag{2.5}$$

where $P = I - \mathbf{1}\mathbf{1}'/n$ and $\mathbf{1}$ is vector of ones. It should be noted that $\hat{S}$ can be substituted to various other pairwise relatedness matrices estimated from the genotypes [PPP06, Rit09, MS00, Mil03], as long as the matrix is positive-semidefinite.

2. Use a variance component model to estimate the restricted maximum likelihood parameters (or alternatively, maximum likelihood parameter) of σ_a and σ_e in:

$$\text{Var}(Y) \;\; = \;\; \sigma_a^2 \hat{S}_N + \sigma_e^2 I \tag{2.6}$$

Test the hypothesis $H_0 : \sigma_a^2 = 0$. If the null hypothesis is rejected, proceed to step 3; otherwise, use ordinary least squares to estimate the coefficients of each of the SNPs genotyped.

3. For each marker, use GLS F-test [KK04], or alternatively a score test, to estimate the effects β_k and test the hypothesis $\beta_k \neq 0$ in the following model.

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_i \qquad \text{Var}(\eta) = V \propto \hat{\sigma}_a^2 \hat{S}_N + \hat{\sigma}_e^2 I. \tag{2.7}$$

The above model can be easily extended to have additional confounding variables by substituting β_0 for a multi-column matrix containing the confounding variables, such as sex and age. Note that these additional confounding variables should be included in the procedure of restricted maximum likelihood estimation of the variance component parameters. Multi-locus models can be incorporated by including additional interaction terms [MDC05, EMM06]. For the variance component estimation procedure in step 2, we use EMMA [KZW08]. We term our method EMMAX (EMMA eXpedited) because it markedly reduces the computational cost compared to the original EMMA by avoiding the repetitive variance component estimation procedure for each single marker. We investigated the effects of this simplification on the data sets we analyzed.

2.2.3 Application to case control data sets

Although EMMAX was developed with quantitative traits in mind, it can also be adapted to the analysis of case-control data sets. As the case-control phenotypes do not follow a normal distribution, applying a generalized linear mixed model using logit or probit link function is preferable to a linear mixed model. However, the computational cost of a generalized linear model with a correlated variance component is much higher, and currently available algorithms cannot handle thousands of individuals simultaneously [McC03].

When the hypothesis of additive model appears reasonable, the Armitage trend test [Arm55] can be used to test for the presence of a genetic effect. (See, for example, Devlin and Roeder [DR99a] and note the equivalence of an Armitage test to a score test in logistic regression for $H_0 : \beta = 0$ [AW90]). The Armitage test can be described as testing the significance of the slope coefficient in a linear regression of a 0-1 variable representing case/control status on the additively coded genotypes. Armitage [Arm55] suggested using a χ_1^2 test that is slightly different from the square of a standard t test in linear regression. The statistic proposed by Armitage is $\chi_0^2 = \beta/var(\beta)$, but instead of estimating the variance of the error terms using the residuals from the regression, we estimate it using the variance of the response variable. Therefore, χ_0^2 is equal to the square of the correlation between the response and the genotype variables, multiplied by the number of samples.

Despite this suggestion, Armitage indicated that the standard t statistic may be preferable, especially to construct confidence intervals. Therefore, it seems that one can carry out tests in the spirit of Armitage simply using a standard linear regression framework with a 0-1 quantitative response variable representing the case control status. Adopting this approach, we were immediately able to translate the problem to the methodology suggested for quantitative traits.

2.2.4 Genotype and phenotype data

We analyzed two data sets: one that contains measurements on quantitative traits (NFBC66), and one for binary disease traits (WTCCC). Genotype data were available for 5,546 Finnish subjects from NFBC66 (ref [SSH09]), all with genotyping completeness $>$95$\%$. We excluded subjects from further analysis because they had withdrawn consent (15), had discrepancy between reported sex and sex determined from the X chromosome (14), were sample duplications (2), were too related to another subject (77), had more than 5% missing genotypes (1), or had no phenotype data (111), leaving 5,326 subjects for analysis. For the relatedness criterion, we identified all pairs of subjects with probability of IBD $> 20\%$, and included one subject from each such pair in further analyses. In most cases, the subject with the most nonmissing phenotype data was chosen for analysis. If the two subjects had an equal amount of missing phenotype data, the subject with the most nonmissing genotype data was used.

Using these 5,326 subjects, we examined the 368,177 SNP markers for Hardy Weinberg Equilibrium (exact test), genotyping completeness, and minor allele frequency. Markers were excluded for more than two discordant genotype calls between different methods (4,711), Hardy-Weinberg equilibrium $P < 10^{-4}$ (5,260), genotyping completeness $< 95\%$ (2,535) and minor allele frequency $< 1\%$ (27,002), leaving 331,475 markers for analysis (some SNP markers failed quality checks on more than one criterion). We adjusted the nine phenotypes used in the original data for sex, pregnancy status and use of oral contraceptive, as described [SSH09], and adjusted height or sex only.

The NFBC66 database contains information on the birth locations of subjects and their parents, which can be used to derive ancestry information. Ref [SSH09] describe how six distinct linguistic and geographical groups can be identified in the northern provinces of Finland. Given the patterns of internal migrations and their variation over time, we can assign individuals in NFBC66 to one one these groups when both parents were born in a municipality within the same group. Approximately 50% of the sample can be assigned this way, and these individuals are used to compare the results of population stratification analysis based on genotypes.

We also obtained the genotypes of the WTCCC subjects collected for a GWAS of seven common diseases [Con07]. We applied the same quality-control criteria as suggested in the original paper. We also excluded the SNPs that the original studies excluded in their analysis. We considered a total of 404,862 SNPs after the quality control across 2,938 shared controls and 13,241 case individuals across seven diseases.

Additionally, it appears that using the simple identity-by-state (IBS) between individuals, rather than the more laboriously constructed kinship coefficients, may be sufficient, and in some cases more appropriate, to model the dependency in the sample. We investigate this assumption further in the Method, Table 2.6 and Figures 2.8b and 2.11.

2.2.5 Estimation of relatedness from high-density markers

Unlike a traditional variance component model which uses IBD (identity by descent) coefficients estimated from the pedigree [OAM01], our proposed method empirically estimate the genetic relatedness between the individuals from high-density markers. In model organism studies, Yu et al. [YPB06] estimated kinship coefficients from multi-locus genotypes using method-of-moment estimators [LSN95, Rit09], and Zhao et al. and Kang et al. [ZAK07, KZW08] demonstrated that using a haplotype-based IBS matrix or a simple IBS matrix more robustly corrects for the population structure resulting in a lower inflation factor than using the estimated IBD matrix from structured model organism samples. Several other methods [BSS08, CWW09, BN95] have been proposed to estimate IBD kinship coefficients or sample structure from multi-locus genotypes including the maximum-likelihood method implemented in PLINK software [PNT07, Mil03] and the PREST software [MS00]

The effectiveness of the empirically estimated pairwise relatedness in correcting for sample structure has not been comprehensively examined in a large-scale human association mapping studies, where the sample structure is much less heterogeneous than those among the strains of model organisms. For this reason, we compared three different empirical estimates of pairwise genetic relatedness from the NFBC66 samples. First is a simple IBS coefficient, and the second is a maximum-likelihood estimates (MLE) of IBD kinship coefficient [Mil03] implemented in the PLINK [PNT07] software. The third is the Balding-Nichols (BN) kinship coefficient [BN95].

The pairwise plots across these three methods suggest that the relatedness estimates computed by these methods are highly correlated with each other (Figure 2.11). The MLE-based IBD estimates [Mil03] shows a correlation of $r = 0.62$ with IBS coefficient, and $r = 0.48$ with BN coefficient. The MLE-based methods estimates 37% of the pairwise kinship coefficients to be positive, and those individual pairs show strong correlation of $r = 0.68$ between BN and IBS coefficients. Among the 63% of individual pairs where the MLE-based kinship coefficient are zero, a strong correlation of $r = 0.54$ is observed between the IBS and BN coefficients, suggesting that the unrelated individual pairs may still have different degrees of distant relatedness.

We applied either the simple IBS or the BN matrix as the surrogate of sample structure when applying EMMAX, and results with IBS matrix is reported unless specified or compared between the two methods. The MLE-based method does not guarantee that the estimated kinship matrix is positive semidefinite (all eigenvalues are non-negative), making it difficult to use in a variance component model. The EMMAX p-values across the two methods provide a very high concordance to each other (Table 2.6 and Figure 2.8B).

2.2.6 Methods for estimating marker specific inflation factors

Assuming that model 2.4 is true with $V = \text{Var}(\eta)$ and marker k has no effect on the phenotype, we define the inflation factor for marker k as the ratio between the expectation of the F statistics calculated from OLS for a model that includes k, to the expectation of the F statistics for the same model calculated from GLS. In fact, we do not compute this ratio explicitly, but simply provide an approximation. If one considers that as $n \longrightarrow \infty$, the expectation of the GLS F statistics under arbitrary V, as long as V is non singular, converges to 1; hence we simply need an approximation for the numerator of the ratio.

Specifically, let us assume, to simplify notation, that Y and X_k are centered to have zero sample mean so that $\hat{\beta}_0 = 0$ holds. In such a case, $V = \text{Var}(\eta)$ has to be centered to $V_C = PVP$ where $P = I - \mathbf{1}\mathbf{1}'/n$. In addition, for convenience purposes, we standardize X_k to satisfy $X_k^T X_k = n - 1$, where n is the number of individuals. Then the F-test statistic based on OLS [RD02] becomes

$$\begin{aligned} F_{OLS} &= \frac{((X_k'X_k)^{-1}X_k'Y)^2(X_k'X_k)(n-2)}{Y'(I - X_k(X_k'X_k)^{-1}X_k')Y} && (2.8) \\ &= \frac{(X_k'Y)^2(n-2)}{nY'Y - (X_k'Y)^2}. && (2.9) \end{aligned}$$

If $V = \sigma^2 I$, then F_{OLS} follows a F-distribution with (1, $n-2$) degree of freedom.

Then if n is large, F_{OLS} asymptotically converges to chi-square distribution with 1 degree of freedom. While the distribution of F_{OLS} is difficult to calculate when V has off-diagonal elements, the expected values of numerator and denominator in F_{OLS} are relatively easy to compute. The expectation of denominator becomes $n\text{Tr}(V_C) - X_k'V_CX_k$, and the expectation of numerator becomes $(n-2)X_k'V_CX_k$.

We can then take as operational definition of the marker specific inflation factor ζ_k at marker k,

$$\begin{aligned}\zeta_k &= \frac{(n-2)X_k'V_CX_k}{(n-1)\text{Tr}(V_C) - (X_k'V_CX_k)} &(2.10)\\ &\approx \frac{X_k'V_CX_k}{\text{Tr}(V_C)} &(2.11)\end{aligned}$$

Note that when $V = \sigma^2 I$, then $\zeta_k = 1$ holds regardless of the values of X_k. Let $\hat{S}_C = P\hat{S}_NP$. When we take for V the specific form assumed in (7), we can further simplify the expression above:

$$\begin{aligned}\zeta_k &= \frac{(n-2)X_k'(\sigma_a^2\hat{S}_C + \sigma_e^2 P)X_k}{(n-1)\text{Tr}(\sigma_a^2\hat{S}_C + \sigma_e^2 P) - (X_k'(\sigma_a^2\hat{S}_C + \sigma_e^2 P)X_k)}\\ &= \frac{\sigma_a^2(n-1)X_k'\hat{S}_CX_k + \sigma_e^2(n-1)(n-2)}{\sigma_a^2\left[(n-1)^2 - X_k'\hat{S}_CX_k\right] + \sigma_e^2(n-1)(n-2)}\\ &\approx \frac{\sigma_a^2X_k'\hat{S}_CX_k/(n-1) + \sigma_e^2}{\sigma_a^2 + \sigma_e^2}\\ &= h_a^2X_k'\hat{S}_CX_k/(n-1) + (1-h_a^2) &(2.12)\end{aligned}$$

where $h_a^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$ is the pseudo-heritability.

We are now in the position to discuss the meaning and implication of the marker specific inflation factors we defined. The introduced marker-specific inflation factors essentially estimate the effects of the mis-specification of variance component by using OLS in the place of GLS. From expression (2.12) it is clear that the amount of inflation at any given marker depends on the level of correlation between the marker genotypes and the GLS variance-covariance matrix. This validates the common intuition that cryptic population structure may affect tests differently at different markers and it illustrates the reasons of such variability. Expression (2.12) also clarifies how the same level of sample structure will affect differently the association tests for different phenotypes. The inflation will be stronger the higher is the ratio of σ_a^2 to σ_e^2, while for a trait that does not follow the polygenic model $\sigma_a^2 = 0$, no amount of population structure will have any impact on the association tests. Finally, it is useful to recall that the inflation factors ζ_k, while marker specific, are calculated independently of the observed association between marker and phenotype, being based on expectations of test statistics under the null model.

More generally, if multiple confounding variables need to be accounted for in addition to the intercept under the null model, Equation (2.9) can be rewritten in a general form of F statistic to get the expectation of numerator and denominator. Such a procedure is asymptotically equivalent to centering an arbitrary variance component V to $V_C = (I - G(G'G)^{-1}G)V(I - G(G'G)^{-1}G)$, given a non-singular matrix of confounding variables G that includes the intercept. In this case, the SNP vector X_k also needs to be regressed out with respected to G, and $(n-2)$ in Equation (2.9) needs to be replaced with $(n-q-1)$, where q is the number of columns in G.

This method can also be extended for estimating the effect of mis-specified variance component or errors in the variance component estimation. Before running GLS, let $\hat{V} = \hat{\sigma_a}^2 \hat{S}_N + \hat{\sigma_e}^2 I$ be the estimated variance-covariance matrix when V is the true one. Assuming that Y and X_k are centered, the F test statistics for GLS is

$$\begin{aligned} F_{GLS} &= \frac{((X_k'\hat{V}_C^{-1}X_k)^{-1}X_k'\hat{V}_C^{-1}Y)^2(X_k'\hat{V}_C^{-1}X_k)(n-2)}{Y'(\hat{V}_C^{-1} - \hat{V}_C^{-1}X_k(X_k'\hat{V}_C^{-1}X_k)^{-1}X_k'\hat{V}_C^{-1})Y} && (2.13) \\ &= \frac{(X_k'\hat{V}_C^{-1}Y)^2(n-2)}{(X_k'\hat{V}_C^{-1}X_k)Y'\hat{V}_C^{-1}Y - (X_k'\hat{V}_C^{-1}Y)^2} && (2.14) \end{aligned}$$

where $\hat{V}_C$ represents the centered matrix of $\hat{V}$. The ratio between expected numerator and denominator provides the inflation factor with mis-specified variance component.

$$\begin{aligned}
\zeta_k &= \frac{X_k' \hat{V}_C^{-1} V_C \hat{V}_C^{-1} X_k (n-2)}{(X_k \hat{V}_C^{-1} X_k) \mathrm{Tr}(\hat{V}_C^{-1} \hat{V}_C) - X_k' \hat{V}_C^{-1} V_C \hat{V}_C^{-1} X_k} && (2.15) \\
&\approx \frac{(n-1) X_k' \hat{V}_C^{-1} V_C \hat{V}_C^{-1} X_k}{(X_k \hat{V}_C^{-1} X_k) \mathrm{Tr}(\hat{V}_C^{-1} V_C)} && (2.16)
\end{aligned}$$

2.2.7 Accounting for large effect sizes at some SNPs

The accuracy of EMMAX relies on the assumption that the effect of each SNP on the phenotype is negligible for the purpose of estimating σ_a^2 and σ_e^2 in model 2.7. This is a reasonable assumption for most of current human GWAS, because a majority of genome-wide significant signals reported so far explain only a small fraction of phenotypic variance [MBC08]. For example, in a genome-wide study with 5,000 individuals, a genome-wide significance p-value of 7.2×10^{-8} corresponds to 0.58% of phenotypic variance explained. 10^{-10} corresponds to 0.84%, and 10^{-15} to 1.3%. A cumulative effect of several significant SNPs are still relatively small compared to the total genetic effects for most complex traits [MCC09b, LCP08, MBC08, Bog09].

However, a number of phenotypes do not comply with the “negligible effect” assumption. There are many Mendelian traits where a single locus explains the total phenotypic variance almost completely. Among complex traits, several autoimmune diseases including Rheumatoid arthritis and Type I diabetes are largely explained by HLA alleles with relative risks 4 or greater [BMS06, Cla09], with extremely significant with p-values smaller than 10^{-50} or 10^{-100}, explaining 50% or even larger variance of these traits [Con07]. In such cases, where a number of SNPs explains a considerable portion of the phenotypic variance, the negligible effect assumption is ungrounded, and the strategy described so far impractical, because the variance parameter estimation can be substantially biased due to the large effect SNPs.

In fact, it is possible to use EMMAX even in this context, provided that one conditions on the effects of the strongly associated SNPs. Specifically, one can condition on the effects of the implicated SNPs by modeling them as fixed effects when estimating σ_a^2 and σ_e^2 in model 2.7. It is crucial, then, to decide on the effect of which SNPs one should condition upon. If we know *a priori* the identity of associated loci with strong effect, such as the MHC region in the above example, the choice will be obvious. Otherwise, we may condition on the effects of SNPs with highly significant p-values. It is important to use a very stringent significance threshold to avoid loss of power. In our analysis, we conditioned on the SNPs explaining more than 1% of phenotypic variance. In RA and T1D, 58 and 135 significant SNPs in MHC and PTPN2 region are conditioned on. Note that this conditioning procedure is really recommended only if (1) there are a few genomic loci largely explaining the phenotypic variance, and (2) significant over-dispersion or under-dispersion of test statistics is observed after applying EMMAX. It should be noted that it is also possible to account for the large effect SNPs in a more sophisticated way using regularization-based methods such as ridge regression or LASSO [Was04], instead of a simple threshold-based conditioning.

2.2.8 URL

The EMMAX software is available at http://genetics.cs.ucla.edu/emmax.

2.3 Results

2.3.1 Revisiting principal component analysis in the NFBC66

To more closely examine the extent of sample structure within the NFBC66, we used PCA of the genotype covariance matrix [PPP06] and multidimensional scaling analysis (MDS) of the identity-by-state (IBS) matrix from NFBC66 samples. The first two coordinates identified by MDS are known to correlate well with geographical location of the linguistic groups [SSH09]. The first two principal components in the current sample correlate well with latitude and longitude of parental birthplaces for the subset of individuals with known ancestry (Figure 2.1). Indeed, we noted that PCA of genotypes and classical MDS of the IBS matrix lead to very similar results. There is a correlation coefficient of 0.9993 between the first components from PCA and MDS and a correlation coefficient of 0.9978 between the second components. The first five principal components separate to varying degrees the linguistic and geographic subgroups comprising Northern Finland (Figure 2.6), consistent with the previous analysis using MDS [SSH09]. Despite the clear correlation between geographical regions of origin and the first two principal components, clustering analyses of the IBS matrix using PLINK software or hierarchical clustering in R did not identify separate subgroups.

2.3.2 Association analysis

Performing a simple uncorrected association test for each of the nine phenotypes originally examined in ref. [SSH09], we made the following estimates of the genomic control parameters λ: body mass index, 1.031, C-reactive protein (CRP), 1.007, diastolic blood pressure, 1.031, glucose, 1.045, high density lipoprotein (HDL), 1.052, insulin plasma levels, 1.029, low density lipoprotein, 1.098, systolic blood pressure, 1.066, triglyceride, 1.023. These values are all higher than the ones obtained previously with a smaller sample size [SSH09], and substantially higher than what one would expect in a sample with no structure. In addition, the height phenotype, which has not been analyzed in the previous study [SSH09], has a λ value of 1.187. For reference, note that a conservative estimate of the 95% confidence interval of the inflation factor is between 0.992 and 1.008, assuming independence between the markers.

As hidden relatedness is a possible cause of inflated genomic control parameters, we reanalyzed the data after excluding a larger number of possibly related subjects (a genome-wide IBD estimates $> 10\%$ was used as a cutoff using PLINK software, excluding additional 611 individuals). This resulted in a slight reduction of λ for some phenotypes (Table 2.1).

As suggested in ref. [PPP06], we explored the effect of including a variable number of principal components in the association tests. Although including two or five principal components has a considerable effect on the λ values, further augmenting the number of principal components does not substantially decrease the genomic control parameter (Figure 2.2). It is often suggested that only principal components having predictive power for the phenotype should be included in the regression [NS08]. We identified principal components for each phenotype that have a t-test p-value $<$0.005 as predictors; the results of their inclusion in the association tests are reported in Figure 2.2.

2.3.3 Correcting for sample structure

We analyzed the ten NFBC66 phenotypes with EMMAX using a three-step procedure (see Methods). First, we computed a pairwise relatedness matrix from high-density markers, which we used to represent the sample structure. Second, we estimated the contribution of the sample structure to the phenotype using a variance component model, resulting in an estimated covariance matrix of phenotypes that models the effect of genetic relatedness on the phenotypes. Third, we applied a generalized least square (GLS) F-test [KK04], or alternatively, a score test [CA07], at each marker to detect associations accounting for the sample structure using the covariance matrix.

The second step also provides us with the fraction of phenotypic variance explained by the empirically estimated relatedness matrix. We call this fraction pseudoheritability because it resembles the heritability estimated from a pedigree [LW98] although it is not directly interchangeable with heritability of the trait because the estimated pairwise relatedness does not correspond exactly to the kinship coefficients. Nonetheless, the pseudoheritability estimates are concordant with the previous heritability estimates from a large family based study of Kosrae and Sardinia populations [LMP09, PCS06] Different methods for estimating the pairwise relatedness provide slightly different, but highly correlated estimates of pseudoheritability across the ten traits. (Table 2.4).

Using the estimated covariance matrix, we proceeded with the GLS F-test to test the effect of each marker on the phenotype and then apply genomic control to quantify the amount of residual inflation. The genomic control λ parameters we obtained with EMMAX are much lower than those obtained using either standard association methods or regression analysis including 100 principal components (Table 2.1). Figure 2.3 and Figure 2.7 illustrate the results using quantile-quantile plots of the P value distributions from these three tests. Only one of the ten phenotypes showed λ values with the 95% confidence interval of 0.992 - 1.008 with uncorrected or principal component analysis, while all of them fell in the confidence interval with EMMAX.

Unlike genomic control, the EMMAX model alters the ranking of SNPs by their statistics. This is especially important as many GWAS follow-up and multistage design studies take the approach of genotyping all SNPs exceeding some predefined threshold [EPD07, TJK09, ATG09]. We examined the extent to which the adoption of the EM-

MAX model changes the SNP rankings in comparison to the uncorrected and principal component analyses. We took the top k markers from the results of EMMAX, the uncorrected method, and regression including 100 principal components (as implemented in EIGENSOFT software), for k between 10 and 5,000. For each of these sets, we calculated the number of SNPs shared between the lists and the fraction of these shared SNPs relative to the number of unique SNPs in each pair of list. Although many of the top SNPs reported by each method overlap, a considerable number of highly ranked SNPs differ between the methods (Figure 2.4 and Table 2.5). In general, EMMAX results become similar to uncorrected analysis when the inflation of test statistics is small, but they become more similar to the PCA as the inflation increases. Notably, the PCA consistently shows larger departures from the uncorrected analysis than EMMAX does across all ten phenotypes. For example, when the overdispersion of test statistics was negligible, such as in the CRP phenotype, only 66% of the top 2,000 hits were concordant between the principal component and the uncorrected analysis, whereas 89% were concordant between EMMAX and the uncorrected analysis.

EMMAX prevents the overdispersion of test statistics using a statistical model that explicitly takes into account sample structure, rather than correcting the overdispersed test statistics caused by not taking into account genetic relatedness in the statistical model. Consequently, EMMAX can also prevent the overcorrection that would remove true positive associations. We identified 15 genome-wide significant loci with at least one of the uncorrected, 100 principal components-corrected, or EMMAX analyses after genomic control at the suggested P-value threshold [DG08] of 7.2×10^{-8} across the ten phenotypes (Table 2.2). In 13 out of the 15 loci, EMMAX P values become smaller than the uncorrected analysis. The two-sided binomial P value of the observed asymmetry is 9.8×10^{-4} if two methods have the same statistical power. With the 100 principal component-corrected analysis, 10 out of the 15 loci show smaller P value than the uncorrected analysis (binomial p-value of 0.12). While 12 out of the 15 loci are found by all methods to be genome-wide significant at $p < 7.2 \times 10^{-8}$, two known loci [KMG08], APOB (with triglyceride)) and HNF4A (with HDL), pass the threshold only with EMMAX. In contrast, the locus NR1H3 (with HDL), which is genome-wide significant only with uncorrected analysis, turns out to be the only locus whose association has not yet been replicated by an independent study among the 15 loci.

Because EMMAX estimates the variance parameters under the null hypothesis, one may suspect that the it is underpowered compared to the full mixed model, which estimates the variance parameters under the alternative hypothesis. This is comparable to the difference between the score statistic and the efficient score statistic [Hin79, WT98, CA07]. As most genetic variants associated to date with human complex traits are estimated to explain only a small fraction of phenotypic variance [MCC09b], the difference between the two approaches will be negligible in most cases. To assess the seriousness of this concern, we ran the original EMMA, which uses a full mixed effect model, on the 15 peak SNPs, and compared the resulting P values to those estimated with EMMAX using GLS. Overall, as expected, the P values from the full mixed effect model tend to be smaller than the P values from the GLS model, but the magnitude of the difference was very small (Figure 2.8A). However, the running times for EMMA were substantially longer. Because the original EMMA re-estimates the variance parameters at each marker, given the size of the NFBC data set, it took more than 10 min of CPU time per marker on an Intel Xeon 3GHz processor, even with an efficient C implementation of EMMA. A simple extrapolation suggests that it would take more than 6 years of CPU time to analyze a single GWAS data set using EMMA, taking a full mixed model approach. The total computational time using EMMAX for this data was 6.6 hours in a single CPU, and the procedure could be easily parallelized to speed it up further.

2.3.4 Application to Wellcome Trust Case Control Consortium data

We also applied our method to the WTCCC data set consisting of case-control studies for seven common diseases [Con07]. To analyze case-control phenotypes, we applied a linear model to the binary phenotypes, in the spirit of Armitage's test (see Methods). We performed association testing over the seven disease phenotypes using EMMAX, EIGENSTRAT, and uncorrected analysis. The values we observed for inflation factors λ were very similar to those in the original study, in which the test statistics were uncorrected: bipolar disease, bipolar disease, 1.11; coronary artery disease, 1.06; Crohn's disease, 1.10; hypertension, 1.06; rheumatoid arthritis, 1.03; type 1 diabetes, 1.04; and type 2 diabetes, 1.07. Consistent with our observations over the NFBC66 data, correcting for 100 principal components only partially reduced the inflation factors (Table 2.3 and Figure 2.7). When EMMAX was applied, the estimated inflation factors were below the upper bound of the confidence interval, suggesting that none of the phenotypes show significant inflation of test statistics.

However, we noticed that two of the phenotypes, rheumatoid arthritis and type 1 diabetes, show significant deflation of test statistics beyond the 95% confidence interval ($\lambda = 0.965$ for rheumatoid arthritis , $\lambda = 0.946$ for type 1 diabetes). This is not unexpected, considering that a substantial fraction of the phenotypic variance in these autoimmune diseases is explained by the HLA loci, leading to inaccurate estimation of variance parameters under the null hypothesis when the HLA effect is not accounted fort. In fact, the set of genome-wide significant SNPs ($P < 7.2 \times 10^{-8}$; ref. [DG08]) in this region account for 47% and 60% of the phenotypic variance of rheumatoid arthritis and type 1 diabetes, respectively [Con07]. We re-estimated the variance parameters by conditioning on the 57 and 134 SNPs within the extended human MHC region [BMS06] that explain more than 1% of phenotypic variance of rheumatoid arthritis and type 1 diabetes, respectively (as described in the Methods section). As a result, the genomic control λ increased to 0.989 for rheumatoid arthritis and 0.991 for type 1 diabetes. We performed this conditioning procedure only for estimating variance parameters and not in the SNP association test so that the P values would be consistent with the unconditioned analysis. Conditioning on the SNPs with such a strong effect may further improve the power to identify novel loci. A more sophisticated conditional analysis - for example, one including haplotype effects or epistatic interactions into covariates - may also better account for the strong effects in the autoimmune diseases [NHW07].

2.3.5 Marker specific inflation factors

Under certain conditions, one can expect the variance of the test statistics to be inflated by a constant across the genome [DR99a, BDR02], A formal model of hidden relatedness based on the coalescent theory [VP05] also suggests a constant inflation across the genome when the sample structure is entirely due to hidden relatedness [DR99a]. However, for a more complex genealogical relationship among individuals, it is not clear how the inflation of test statistics will behave.

Using the same variance component framework, we developed a method to estimate the marker-specific inflation of test statistics using the correlation between each marker and the empirically estimated kinship matrix (described in the Methods section). These estimates are concordant with the genome-wide genomic control inflation factor on average but showed substantial differences across the SNPs (Figure 2.5A). In the height phenotype, for example, the estimated marker-specific inflation factors have a mean of 1.107, s. d. of 0.090, and median value of 1.093. In light of this, we explored the relationship between marker-specific inflation factors and the overdispersion of test statistics with the uncorrected analysis. The distribution of height association P values for SNPs with inflation factor < 1.05 shows a less marked departure from uniform distribution than dies the distribution for SNPs with inflation factor > 1.20 (Figure 2.5B and 2.5C). Considering that SNPs with a higher inflation factor were identified without consideration of their possible association with the phenotype, it is reasonable to conclude that this excess of small p-values reflects overdispersion of test statistics.

These results underscore how correcting the test statistics using a single inflation factor may be inappropriate, possibly reducing power and not sufficiently controlling for false positives. To further demonstrate this point, we ran a simple simulation using the variance component model on which EMMAX is based. Although simulating data under this model puts our method at an advantage, and the approach is therefore less suited for comparison to other models, it does demonstrate that under some circumstances uniformly deflating P values may be inappropriate. We randomly simulated 100 sets of phenotypes solely from the sample structure with no SNP effects and examined the quantile-quantile plots across different methods. Although the inflation for most of the SNPs is corrected by genomic control as expected, we observed substantial fluctuations of the test statistics at the tail of the distribution (Figures 2.9A and 2.9B). More than 25% of the phenotypes showed inflation or deflation beyond the 95% confidence interval. This is because the SNPs with higher per-marker inflation are not sufficiently corrected by the constant genomic control inflation factor. In contrast, EMMAX results in P values close to the expected distribution (Figure 2.9C).

The finding that marker-specific inflation factors vary substantially across the genome has notable implications for the meta-analyses and multistage analyses. Such studies typically combine the test statistics after correcting for potential inflation using genomic control [ZSS08, TJK09, ATG09]. The disadvantages of using the same global correction rather than a marker specific one can become more serious when this step is done repeatedly. To better understand these effects in the context of meta-analysis, we first compared the marker-specific inflation factors between the two WTCCC control groups, collected from essentially the same population, We observed a very strong correlation ($r = 0.95$) (Figure 2.10A). We further compared the inflation factors across different populations and different genotyping platforms using the NFBC66 samples and WTCCC control samples. We observed a strong correlation of $r = 0.70$ (Figure 2.10B), suggesting that the marker specific inflation factors can be correlated across the multiple data sets used in meta-analysis or multistage analysis owing to the shared genetic history. If this is the case, the standard approach that corrects with genomic control before merging the P values from different studies may lead to further inaccuracies: tests at some markers would be excessively, or not sufficiently, deflated multiple times, resulting in an accumulation of errors.

2.4 Discussion

We report here the development of the EMMAX program, taking an expedited mixed linear model approach to correct for sample structure within human GWASs. We demonstrate its effectiveness with the analysis of two human GWAS data sets, including quantitative as well as disease traits. The proposed approach differs substantially from genomic control in that it accounts for inflation owing to population structure in a marker-specific manner, resulting in a modified ranking of association results. Accounting for marker-specific effects can reduce both false positives and false negatives. We discuss this issue in more detail in the Methods section.

There are several other methods that take into account pedigree-based or empirically estimated kinship matrices into the statistical test [TM07, GLB09, CWW09, RS09]. One of the key differences between these methods and the mixed model methods, including EMMAX, is that the mixed model methods have a procedure of estimating the contribution of the kinship matrix to the phenotypes, whereas the other methods do not. Estimating the phenotypic variance contributed by the sample structure enabled us to avoid undercorrection or overcorrection of the sample structure in the NFBC66 and WTCCC data sets.

The effective application of our method depends on an appropriate estimate of the variance parameters. The IBS or Balding-Nichols matrix [BN95] appears to be better than IBD estimates at capturing the long-distance relationships that result in variations at the population level. However, when the structure of the sample at hand is better described in terms of fairly recent hidden relatedness, methods based on the estimation of IBD may have an advantage. In principle, our approach is also suitable for association testing in a data set including individuals from a heterogeneous population with admixed background. In such cases, it is important to consider SNP ascertainment bias in estimating the degree of relatedness between individuals. Because many SNP probes in genotyping arrays are selected from European populations, the marker-based pairwise distance between two individuals may appear to be larger between unrelated European samples than between unrelated individuals from other populations. To resolve the resulting ascertainment bias, each SNP may be differently weighted when the IBS similarity matrix is computed. A general framework has ben presented [KZW08] for computing the similarity matrix with a different weight for each marker. Different weighting schemes can also be used to account for heterogeneous distribution of effect size from each marker or each genomic region.

Besides the choice of the kinship matrix, the estimation of variance parameters is also a crucial part of the EMMAX approach. In our analysis of the NFBC data, we show that estimating these parameters under the null hypothesis does not lead to appreciable bias in the association P values. The example of rheumatoid arthritis and type 1 diabetes in the WTCCC dataset, in contrast, reveals the difficulties encountered by EMMAX when there are SNPs explaining a large fraction of phenotypic variance. In such cases, we show that estimating variance parameters conditionally on the SNPs with stronger effects alleviates the problems.

Finally, whereas the analysis presented here relies on decomposing the variance in two terms, a genetic relatedness component and a component representing residual effects, future studies may need to account for additional variance components to more precisely model the heterogeneous phenotypic variance. In expression quantitative trait loci mapping, for example, one may want to add additional variance components to account for technical bias [KYE08]. When multiple variance components are involved, one would need to make use of algorithms such as PROC MIXED implemented in SAS, as EMMA is developed for two variance components only; this would increase the running time of the first step of our procedure. However, because the same variance components estimated from the null hypothesis would be used across the genome-wide markers, the the overall computational time should still be acceptable.

Reference to published article

Hyun Min Kang*, **Jae Hoon Sul***, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin. "Variance component model to account for sample structure in genome-wide association studies." *Nature Genetics*. 42, 348354, 2010.

Genomic control inflation factor				
Phenotypes	**Uncorrected**	**IBD**< 0.1	**ES100**	**EMMAX**
CRP	1.007	1.007	1.019	0.993
TG	1.023	1.010	1.019	1.002
INS	1.029	1.022	1.013	1.005
DBP	1.031	1.019	1.028	1.007
BMI	1.031	1.024	1.016	0.995
GLU	1.045	1.033	1.030	1.008
HDL	1.052	1.056	1.036	1.004
SBP	1.066	1.056	1.021	1.006
LDL	1.098	1.089	1.040	1.002
HEIGHT	1.187	1.151	1.074	1.003

Table 2.1: Comparison of genomic control inflation factors obtained with different models; ES100, EIGENSOFT correcting for 100 principal components; IBD < 0.1, uncorrected analysis after excluding 611 individuals whose PLINKs IBD estimates with another individual is greater than 0.1; phenotype abbreviations are CRP, C-reactive protein; TG, triglyceride; INS, insulin plasma levels; DBP, diastolic blood pressure; BMI, body mass index; GLU, glucose; HDL, high-density lipoprotein; SBP, systolic blood pressure; LDL, low density lipoprotein.

Trait	rsID	Chr	Base Position[a]	Closest Gene(s)	Uncorrected+GC	ES100+GC	EMMAX+GC
HDL	rs3764261	16	55,550,825	CETP	7.0×10^{-31}	3.8×10^{-31}	$\mathbf{3.7 \times 10^{-32}}$
CRP	rs2794520	1	15,7945,440	CRP	4.8×10^{-23}	3.6×10^{-23}	$\mathbf{3.0 \times 10^{-23}}$
LDL	rs646776	1	109,620,053	CELSR2	5.4×10^{-14}	7.7×10^{-15}	$\mathbf{3.8 \times 10^{-15}}$
CRP	rs2650000	12	119,873,345	LEF1	2.1×10^{-12}	7.0×10^{-12}	$\mathbf{1.9 \times 10^{-12}}$
HDL	rs1532085	15	56,470,658	LIPC	$\mathbf{4.3 \times 10^{-12}}$	7.9×10^{-11}	1.0×10^{-11}
GLU	rs560887	2	169,471,394	G6PC2	1.1×10^{-11}	4.1×10^{-12}	$\mathbf{3.1 \times 10^{-12}}$
LDL	rs693	2	21,085,700	APOB	9.6×10^{-11}	$\mathbf{1.5 \times 10^{-11}}$	2.8×10^{-11}
TG	rs1260326	2	27,584,444	GCKR	1.9×10^{-10}	$\mathbf{5.9 \times 10^{-11}}$	1.8×10^{-10}
HDL	rs255049	16	66,570,972	LCAT	3.9×10^{-9}	$\mathbf{1.2 \times 10^{-9}}$	1.4×10^{-8}
LDL	rs11668477	19	11,056,030	LDLR	1.4×10^{-8}	3.2×10^{-8}	$\mathbf{4.1 \times 10^{-9}}$
GLU	rs2971671	7	44,177,862	GCK	1.8×10^{-8}	$\mathbf{1.7 \times 10^{-9}}$	1.6×10^{-8}
HDL	rs7120118	11	47,242,866	NR1H3[b]	$\mathbf{4.8 \times 10^{-8}}$	$\mathit{6.6 \times 10^{-5}}$	$\mathit{1.1 \times 10^{-6}}$
TG	rs10096633	8	19,875,201	LPL	2.0×10^{-8}	$\mathbf{1.1 \times 10^{-8}}$	1.9×10^{-8}
TG	rs673548	2	21,091,049	APOB	$\mathit{8.0 \times 10^{-8}}$	$\mathit{1.2 \times 10^{-7}}$	$\mathbf{6.4 \times 10^{-8}}$
HDL	rs1800961	20	42,475,778	HNF4A	$\mathit{1.5 \times 10^{-7}}$	$\mathit{9.5 \times 10^{-8}}$	$\mathbf{1.8 \times 10^{-8}}$

Table 2.2: Fifteen peak associated SNPs with genome-wide significance; these SNPs had P values below the suggested [DG08] genome-wide significance threshold of 7.2×10^8 in the uncorrected the 100 principal componentscorrected (ES100) or the EMMAX analysis after genomic control (+GC). Traits are HDL, high-density lipoprotein; CRP, C-reactive protein; LDL, low density lipoprotein; GLU, glucose; TG, triglyceride. rsID, reference SNP ID assigned by dbSNP; Chr, chromosome; boldface indicates the strongest P values across the three methods; italics indicate P values that did not surpass the significance threshold. [a] Positions are based on National Center for Biotechnology Information build 36.1. [b]NR1H3 is the locus whose association with HDL that has not yet been replicated by other independent studies.

Phenotypes	**Uncorrected**	**ES100**	**EMMAX**
BD	1.105	1.071	0.998
CAD	1.063	1.048	1.006
CD	1.098	1.055	1.000
HT	1.055	1.051	0.997
RA	1.028	1.031	0.965 (0.989[a])
T1D	1.043	1.028	0.946 (0.991[a])
T2D	1.065	1.042	0.996

Table 2.3: Comparison of genomic control inflation factor obtained with different models in seven WTCCC phenotypes. ES100, EIGENSOFT correcting for 100 principal components; BD, bipolar disorder; CAD, coronary artery disease; CD, Crohns disease; HT, hypertension; RA, rheumatoid arthritis; T1D, type 1 diabetes; T2D, type 2 diabetes. [a]The variance component parameters (σ_a^2 and σ_e^2) are estimated by conditioning on the large-sized SNP effects explaining 1% or more phenotypic variance.

Phenotype	IBS matrix		BN matrix		Reference h^2	
	p-value ($\sigma_a^2 = 0$)	$\hat{h}^2_{IBS}$	p-value ($\sigma_a^2 = 0$)	$\hat{h}^2_{BN}$	**Kosrae** h^2	**Sardinia** h^2
CRP	1.7×10^{-2}	0.134	2.3×10^{-2}	0.116	0.245	0.296
TG	2.3×10^{-4}	0.178	2.4×10^{-3}	0.152	0.274	0.322
INS	8.3×10^{-4}	0.205	3.1×10^{-3}	0.152	N/A	0.260
DBP	4.7×10^{-4}	0.199	5.6×10^{-4}	0.167	0.289	0.186
BMI	3.9×10^{-6}	0.279	1.9×10^{-6}	0.242	0.473	0.426
GLU	4.2×10^{-5}	0.229	2.4×10^{-5}	0.197	0.188	0.362
HDL	5.5×10^{-11}	0.384	1.0×10^{-11}	0.324	0.391	0.486
SBP	2.7×10^{-8}	0.283	2.0×10^{-8}	0.233	0.243	0.253
LDL	1.4×10^{-17}	0.452	1.2×10^{-18}	0.384	0.414	0.425
HEIGHT	2.8×10^{-45}	0.738	2.5×10^{-48}	0.625	0.790	0.798

Table 2.4: P-values for test of the null hypothesis $\sigma_a^2 = 0$ for all traits; pseudo-heritability estimates $h_a^2 = \sigma_a^2/(\sigma_a^2+\sigma_e^2)$, and heritability estimates from Kosrae population [LMP09] and Sardinia population [PCS06]. A simple IBS matrix and Balding-Nichols (BN) matrix is used as estimates of relatedness.

Phenotype	**Uncorr. vs EMMAX**	**Uncorr. vs ES100**	**ES100 vs EMMAX**	**Uncorr.** λ
CRP	0.891 (0.94)	0.635 (0.78)	0.660 (0.79)	1.007
TG	0.856 (0.92)	0.569 (0.72)	0.612 (0.76)	1.023
INS	0.826 (0.90)	0.535 (0.70)	0.603 (0.75)	1.029
DBP	0.843 (0.91)	0.607 (0.75)	0.646 (0.78)	1.031
BMI	0.790 (0.88)	0.544 (0.70)	0.607 (0.75)	1.031
GLU	0.775 (0.87)	0.528 (0.69)	0.604 (0.75)	1.045
HDL	0.693 (0.82)	0.500 (0.66)	0.576 (0.73)	1.052
SBP	0.684 (0.81)	0.481 (0.65)	0.597 (0.75)	1.066
LDL	0.624 (0.77)	0.474 (0.64)	0.587 (0.74)	1.098
HEIGHT	0.453 (0.62)	0.386 (0.55)	0.497 (0.66)	1.187

Table 2.5: Comparison of top 2,000 hits obtained with uncorrected analysis, EIGENSOFT with 100 PCs (ES100), and EMMAX. The numbers in second to fourth column represents the proportion of shared SNPs between each pair of analysis, when selecting top 2,000 SNPs in each analysis. The values in parentheses are Cohen's kappa coefficients as a measure of the agreement between two tests. For clarity we have ordered the phenotypes with reference to their genomic control parameters and reported these as well in the last column.

Phenotypes	**Uncorrected**	**EMMAX-IBS**	**EMMAX-BN**	**Concordance**
CRP	1.007	0.993	0.992	0.969 (0.98)
TG	1.023	1.002	1.000	0.969 (0.98)
INS	1.029	1.005	1.005	0.951 (0.97)
DBP	1.031	1.007	1.005	0.955 (0.98)
BMI	1.031	0.995	0.992	0.942 (0.97)
GLU	1.045	1.008	1.004	0.946 (0.97)
HDL	1.052	1.004	1.000	0.919 (0.96)
SBP	1.066	1.006	1.001	0.940 (0.97)
LDL	1.098	1.002	0.999	0.915 (0.96)
HEIGHT	1.187	1.003	0.994	0.838 (0.91)

Table 2.6: Comparison of genomic control inflation factors obtained with uncorrected analysis and EMMAX with IBS matrix and Balding-Nichols (BN) matrix. The "Concordance" column represents the proportion of shared SNP between top 2000 associations between EMMAX-IBS and EMMAX-BN method. The values in the parentheses are kappa statistic

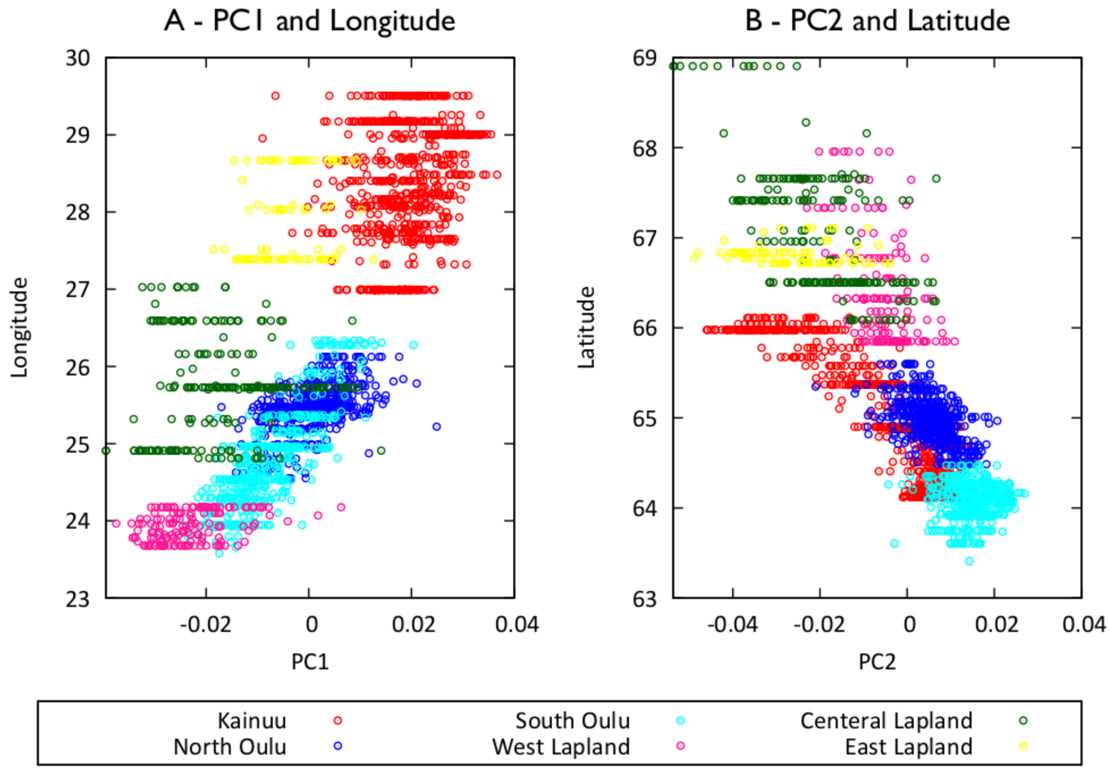


Figure 2.1: Scatter plots of the first two principal components against latitude and longitude. Only individuals of known ancestry are included in the plot. Latitude and longitude are defined as the average latitude and longitude of the parents' birthplaces. Colors indicate linguistic or geographic subgroups.

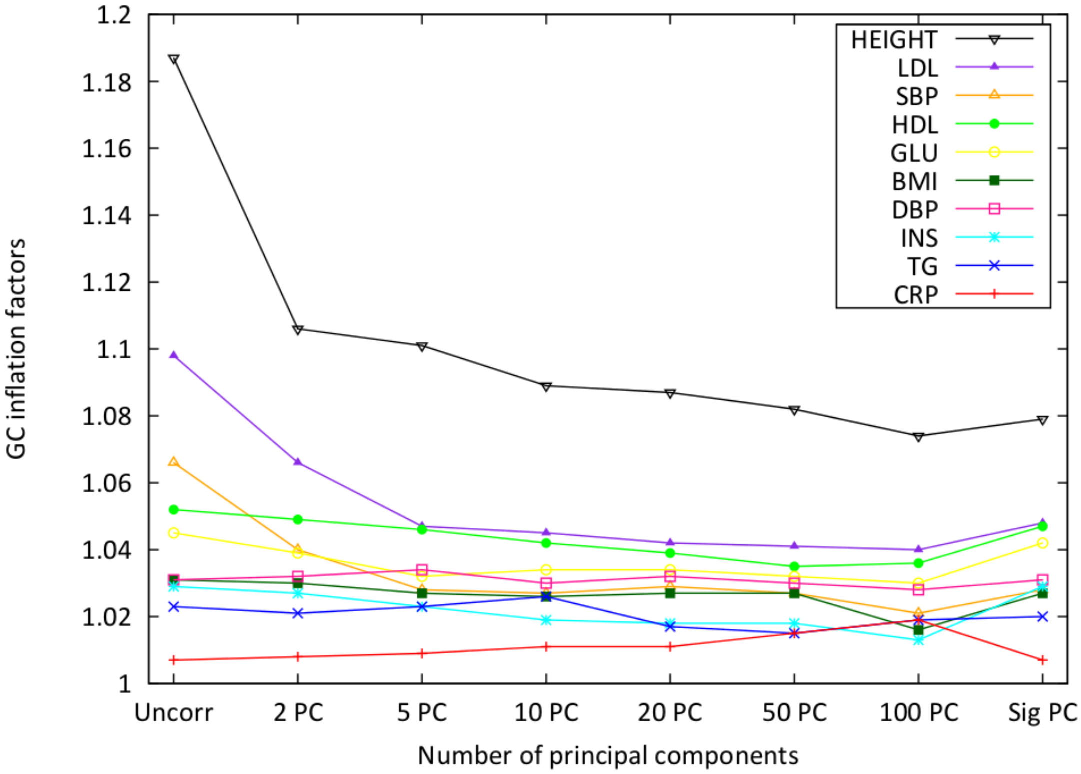


Figure 2.2: The genomic control parameters for ten traits change with the number of principal components used for adjustment. Sig PC, significant principal components, includes the principal components (PC) that have a t-test P value < 0.005 as predictors for each of the phenotypes. LDL, low density lipoprotein; SBP, systolic blood pressure; HDL, high-density lipoprotein; GLU, glucose; BMI, body mass index; DBP, diastolic blood pressure; INS, insulin plasma levels; TG, triglyceride; CRP, C-reactive protein.

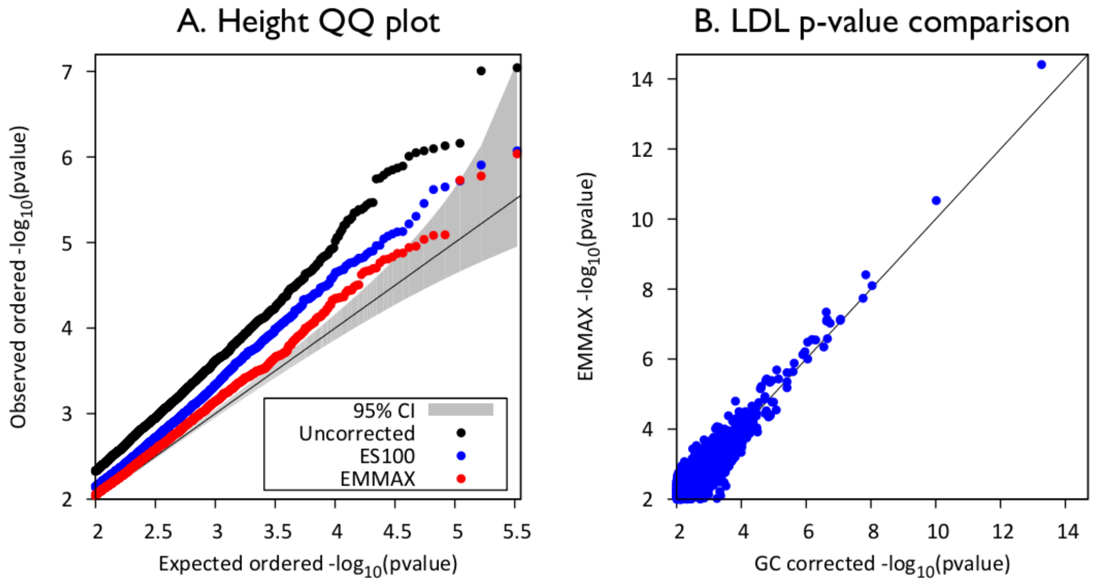


Figure 2.3: Comparison of P value distributions across different methods with NFBC66 data. (a) Quantile-quantile plot of the height phenotype, which shows the largest inflation of test statistics, before application of genomic control. The shadowed region represents a conservative 95% confidence interval (CI) computed from the beta distribution assuming independence markers. ES100 indicates EIGENSOFT correcting for 100 principal components. (b) Comparison of LDL association P values between uncorrected and EMMAX analysis after application of genomic control in a logarithmic scale.

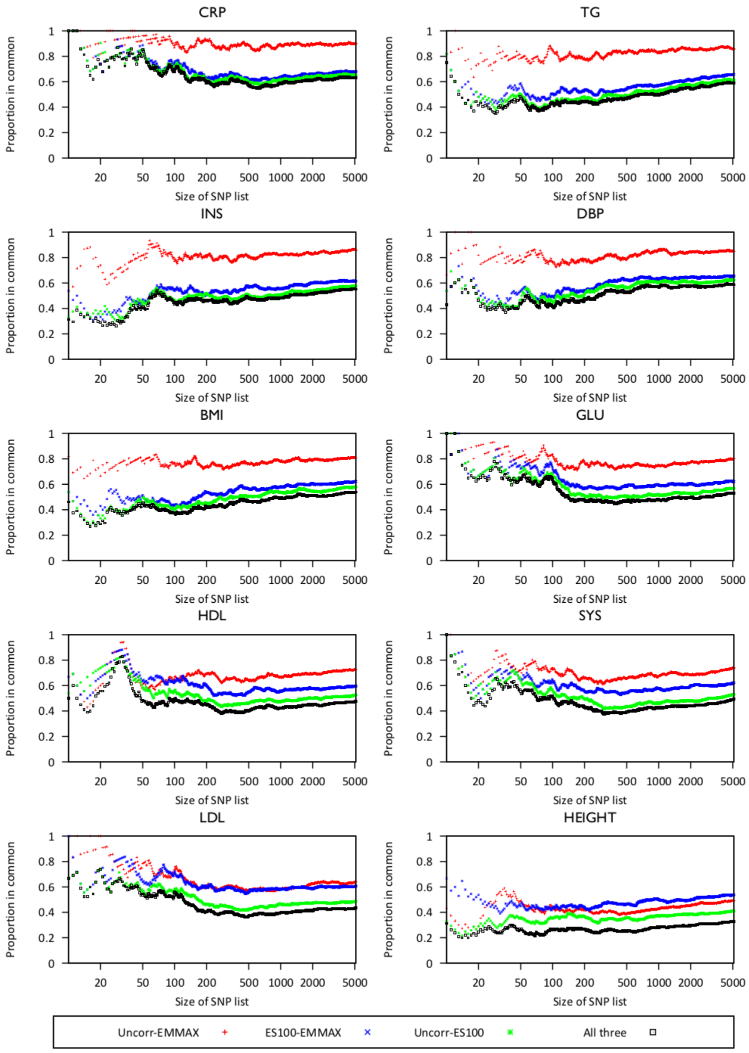


Figure 2.4: Rank concordance comparison of strongly associated SNPs between different methods. The ten NFBC66 phenotypes (abbreviated as in Figure 2.2) are ordered by their genomic control inflation factors. Rank concordance is presented as CAT plots [IWS05]. The proportion of SNPs shared between sets of the top k SNPs for different methods are shown for $10 \leq k \leq 5000$. Pairs of sets being compared are indicated in key at bottom; for example, Uncorr-EMMAX, comparison of uncorrected set and EMMAX set. ES100 indicates EIGENSOFT correcting for 100 principal components.

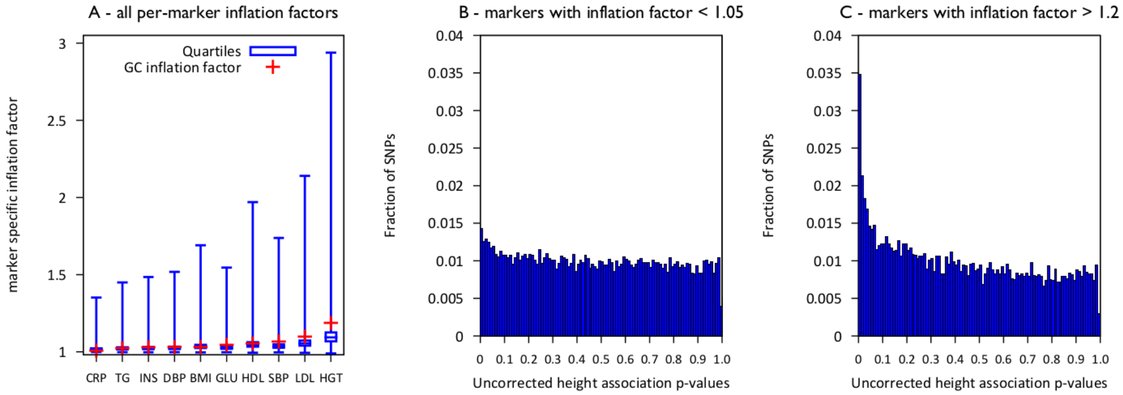


Figure 2.5: Distribution of the marker-specific inflation factors from NFBC66 data sets. (a) Box plots of the marker-specific inflation factors across ten phenotypes, in addition to the genomic control inflation factor for each phenotype. Abbreviations are as in Figure 2.2. (b,c) Distributions of P values of the height phenotype association when the estimated per-marker inflation factors are less than 1.05 (35,988 SNPs; b) and when they are greater than 1.2 (15,874 SNPs; c).

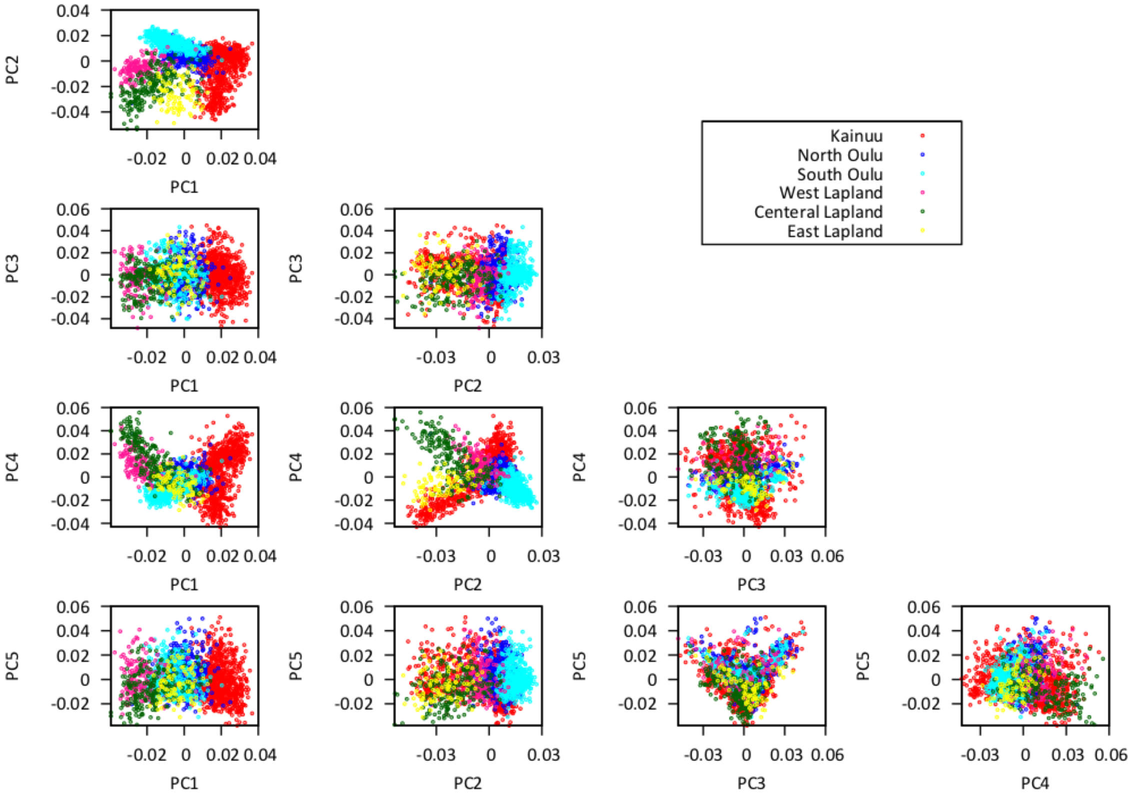


Figure 2.6: Scatter plots of the first 5 principal components for individuals of known ancestry. The different linguistic/geographic subgroups are color-coded.

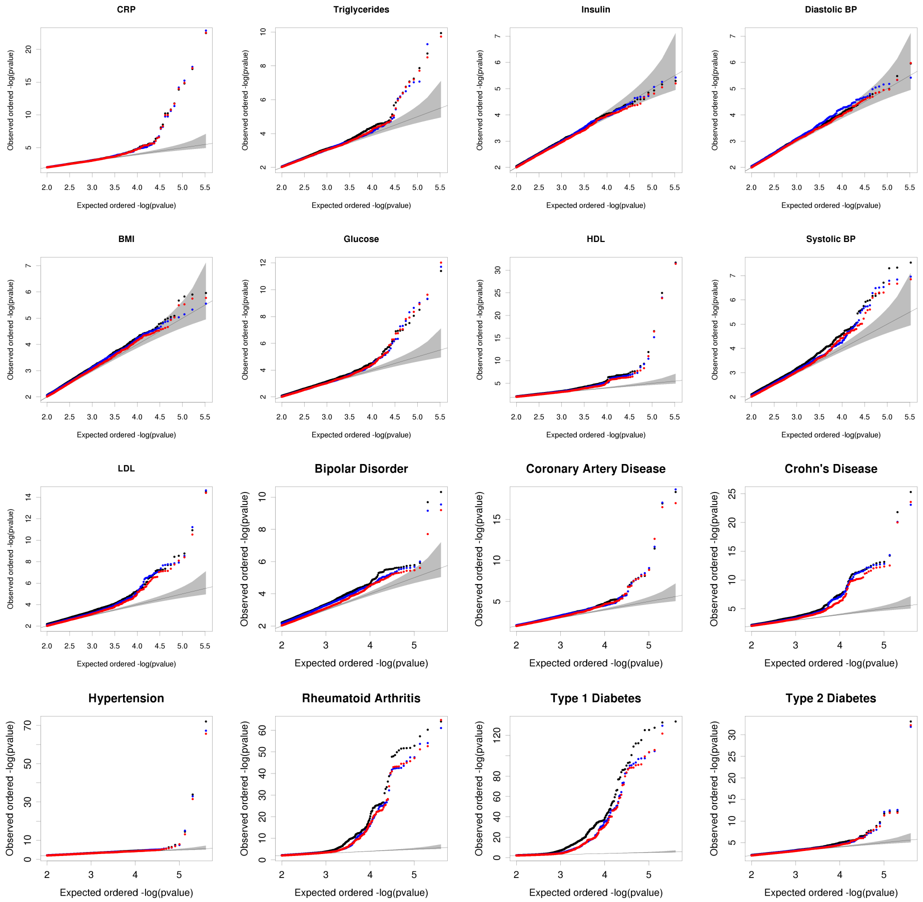
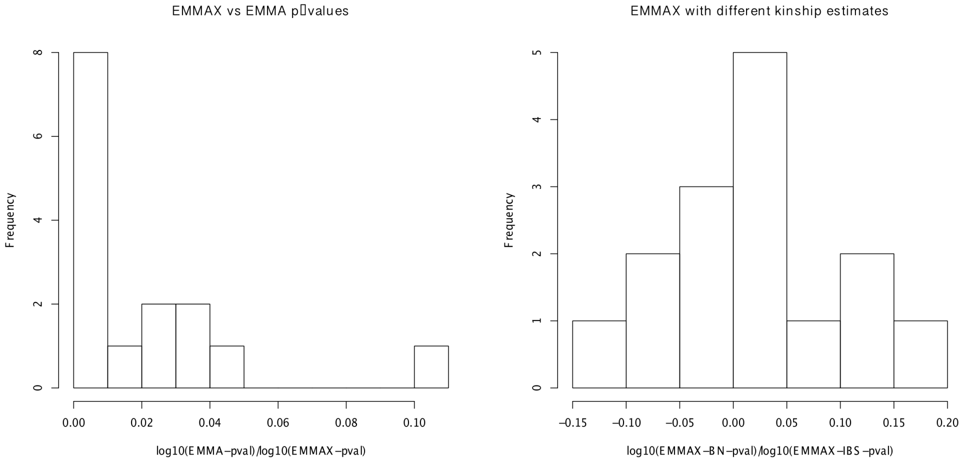


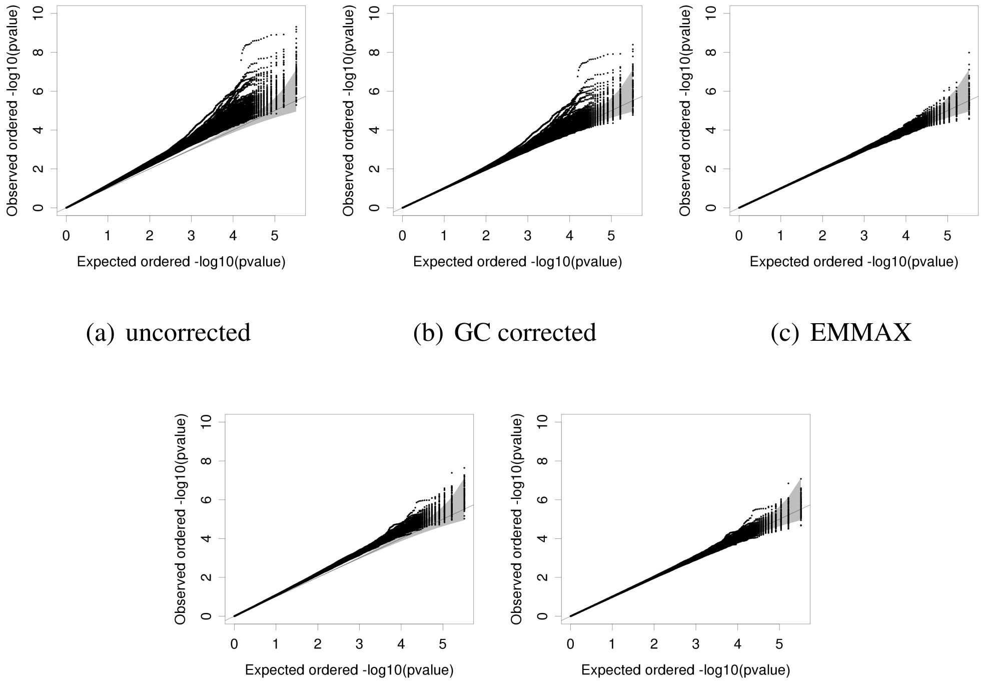
Figure 2.7: QQ-plots on the log10 scale of the association p-values obtained for nine traits according to three different models for 9 NFBC66 metabolic trais and 7 WTCCC disease phenotypes. In black, results from the unadjusted analysis; in blue results from the analysis conducted using 100 PC, and in red results from EMMAX.



(a) EMMAX vs EMMA

(b) EMMAX-IBS vs EMMAX-BN

Figure 2.8: Comparison of p-values obtained running EMMAX using IBS matrix with the corresponding value obtained using (a) the original EMMA and (b) EMMAX with Balding-Nichols (BN) matrix for the SNPs whose p-value under EMMAX was smaller than 7.2×10^{-8}.



(a) uncorrected

(b) GC corrected

(c) EMMAX

(d) ES100

(e) ES100 + GC

Figure 2.9: QQ plots of 100 randomly generated phenotypes under the variance component model using a (a) uncorrected analysis, (b) genomic control adjustment, (c) EMMAX, (d) EIGENSOFT with 100 PCs, and (e) genomic control adjustment after applying EIGENSOFT with 100 PCs.

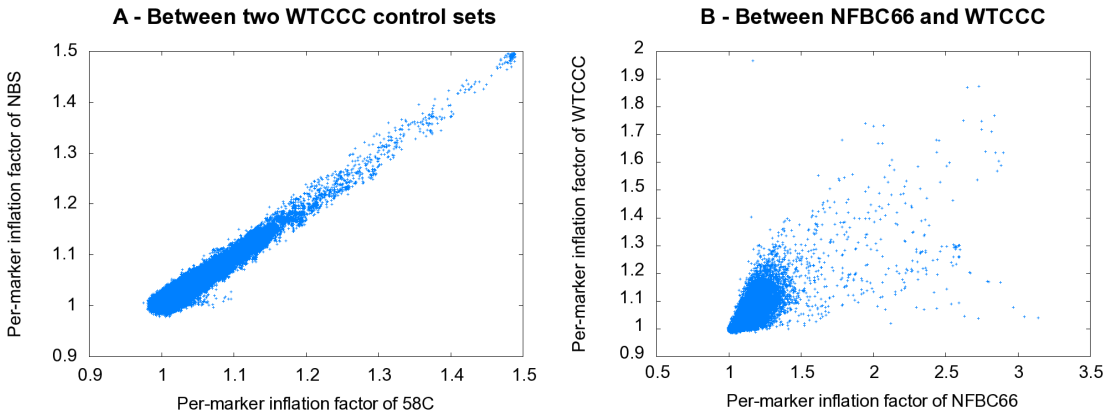


Figure 2.10: Concordance of per-marker inflation factor (A) between two different control sets (58C and NBS) in WTCCC data set, and (B) between NFBC66 samples and WTCCC control samples using the 50,298 overlapping markers.

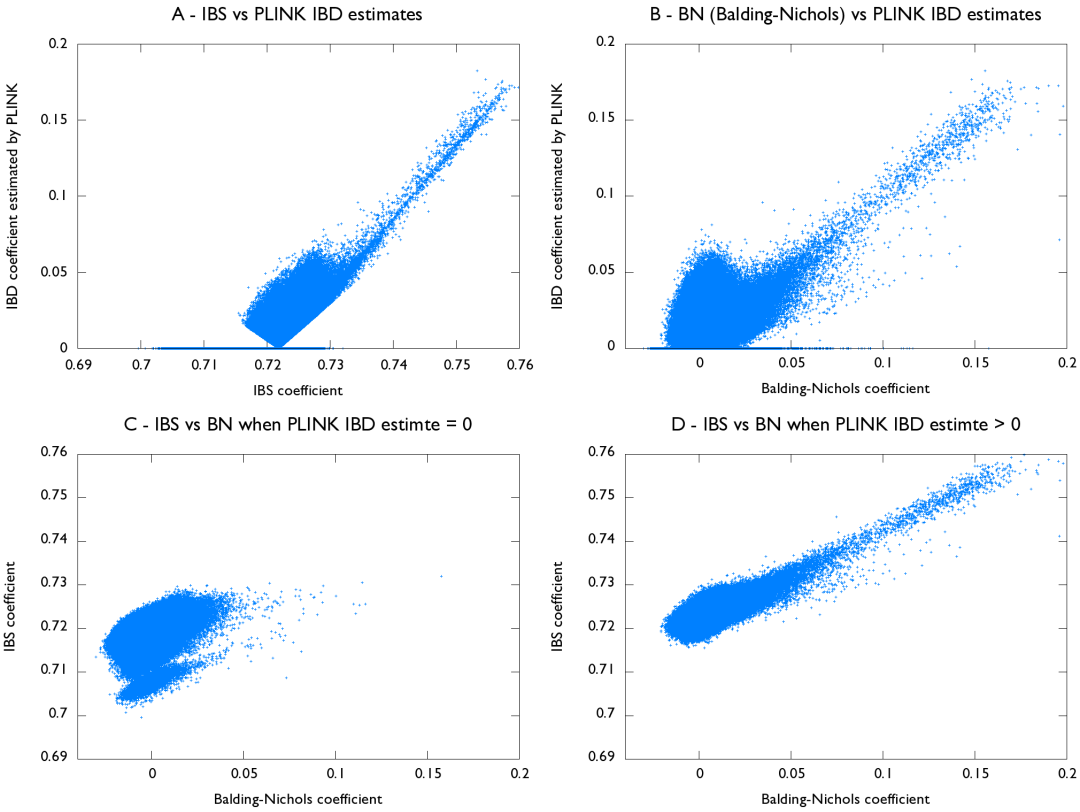


Figure 2.11: Comparisons (A) between the IBS coefficients and IBD estimates computed by PLINK (B) between the Balding-Nichols (BN) coefficients and IBD estimates from PLINK, (C) between IBS and BN coefficients when IBD estimates are zero (D) IBS and BN coefficients when IBD estimates are positive.

CHAPTER 3

Aggregated association test for rare variants

3.1 Background

Over the past few years, genome-wide association studies (GWAS) have identified many disease-causing variants [CSS93, BKK94, AHK00]. Most of these studies are conducted by collecting common variants and perform a series of single marker tests where each variant is tested individually in order to discover associations. However, only a small portion of disease heritability is explained by common variants, and several recent studies consider rare variants that collectively affect diseases [GGS08, KPS07, CKP04, FWW04, JFO08, BB08, RPF07, BVE08, Con08, XRL08, WMM08]. Since each rare variant is present in only a small number of individuals, single marker tests have low power to identify these variants involved in disease. Hence, groupwise association tests that group rare variants in genes have received considerable attention as methods that increase the power of studies on rare variants, and a number of methods have been proposed such as the Cohort Allelic Sums Test (CAST) [MT07], the Combined Multivariate and Collapsing (CMC) method [LL08], a weighted-sum statistic [MB09] and recently a variable-threshold approach [PKB10].

A groupwise association test is more complex than a single SNP association because there are many different ways of combining information across multiple variants. How the information from different variants is combined affects the statistical power of the association test which also depends on the actual effect sizes of the variants on the disease phenotype. The challenge in developing groupwise association testing methods is that the underlying disease-risk model is not known.

In this chapter, we focus on a disease-risk model that is motivated by filling a blind spot in traditional GWAS. In this model, all variants including common variants make an equally small contribution to disease-risk, that is, rarer variants are assumed to have higher effect sizes than common variants. Since each variant contributes only a small amount to the total disease-risk, the single marker test is not likely to detect associations in this disease-risk model, and thus this model describes associations usually not found in traditional GWAS. This is the same model discussed in [MB09]. Under this model, a weighted-sum statistic by Madsen and Browning (MB) is shown to be more powerful than other grouping methods such as CAST and CMC [MB09].

We propose a new method for the groupwise association test called Rare variant Weighted Aggregate Statistic (RWAS). RWAS computes a weighted sum of differences between case and control mutation counts where weights are estimated from data to increase power of studies. The optimal weights that maximize the power can be derived when the effect sizes of variants are known. When the true effect sizes are not known, RWAS approximates the optimal weights under the assumption that each variant makes an equally contribution to population disease risk. Simulations show that RWAS outperforms MB and the approximated weights achieve nearly the same power as the optimal weights under this assumed disease model. We also show how prior information on whether or not a variant is likely to be involved in a disease can be incorporated into RWAS. We first show through simulations that prior information greatly influences the statistical power of studies. Then, by using the real mutation screening data of the susceptibility gene for ataxia telangiectasia along with information of how likely a variant is deleterious [TOB09], we demonstrate that prior information plays a key role in this association study and RWAS is able to successfully detect the association in real data. The software package implementing RWAS is publicly available at http://genetics.cs.ucla.edu/rarevariants.

3.2 Methods

3.2.1 Optimal Weighted Aggregate Statistic (OWAS)

We consider an association study in which multiple variants within a gene affect the trait. For each variant, a difference in mutation counts between case and control individuals is computed, and a weighted sum of differences is used as a statistic for the group. This is in fact equivalent to computing a weighted sum of z-scores of variants where the z-score of a variant is computed from an allele frequency difference between cases and controls [Esk08, HKS08, ZPG10].

First, we assume that there are M rare variants in a group given $N/2$ case and $N/2$ control individuals. Let p_i denote population minor allele frequency (MAF) of variant i, and let $\hat{p}_i^+$ and $\hat{p}_i^-$ denote the observed MAF of case and control individuals in the sample, respectively. Then, z-score of variant i (or the association statistic at variant i), denoted as z_i, is calculated as

$$z_i = \frac{\hat{p}_i^+ - \hat{p}_i^-}{\sqrt{2/N}\sqrt{\hat{p}_i^\pm(1-\hat{p}_i^\pm)}} \quad \text{(where } \hat{p}_i^\pm = (\hat{p}_i^+ + \hat{p}_i^-)/2) \tag{3.1}$$

z-score approximately follows a normal distribution with variance equal to 1 and with mean equal to $\lambda_i\sqrt{N}$ (called the non-centrality parameter or NCP)

$$\begin{aligned} z_i &\sim \mathcal{N}(\lambda_i\sqrt{N}, 1) \\ \lambda_i\sqrt{N} &= \frac{p_i^+ - p_i^-}{\sqrt{2p_i^\pm(1-p_i^\pm)}}\sqrt{N} \quad \text{(where } p_i^\pm = (p_i^+ + p_i^-)/2) \end{aligned} \tag{3.2}$$

where p_i^+ and p_i^- are the true MAF of case and control individuals, respectively. Denoting γ_i as relative risk of variant i, p_i^+ and p_i^- are

$$p_i^+ = \frac{\gamma_i p_i}{(\gamma_i - 1)p_i + 1} \tag{3.3}$$

$$p_i^- = p_i \quad \text{(assuming the disease prevalence is very small)} \tag{3.4}$$

Let w_i be a weight of variant i. Then, a weighted sum of z-scores (S) and its distribution are

$$S = \frac{\sum_{i=1}^{M} w_i z_i}{\sqrt{\sum_{i=1}^{M} w_i^2}} \sim \mathcal{N}\left(\frac{\sqrt{N}\sum_{i=1}^{M} w_i \lambda_i}{\sqrt{\sum_{i=1}^{M} w_i^2}}, 1\right) \tag{3.5}$$

The greatest power is achieved when the NCP is maximized, which is equivalent to maximizing $\frac{\sum w_i \lambda_i}{\sum w_i^2}$ term. Using the Cauchy-Schwartz inequality, the NCP is maximized when $w_i = \lambda_i$. Therefore, the optimal weight for variant i is λ_i, and we call the weighted association method based on the optimal weights, Optimal Weighted Aggregate Statistic (OWAS). OWAS is optimal under any disease-risk models, but determining optimal weights requires knowledge of relative risk and population MAF of variants according to the definitions of λ_i, p_i^+, and p_i^- (Equations 3.2, 4.6, 4.7). We can estimate the population MAF from observed MAF of case and control individuals (see section 3.2.5 for details), but obtaining or estimating relative risk is often not easy. We note that if the number of cases ($N^+/2$) and controls ($N^-/2$) are unequal, we replace $\sqrt{N}$ above with $\sqrt{\frac{2N^+N^-}{N^++N^-}}$ and replace $\hat{p}_i^\pm$ and $p_i^\pm$ in (Equations 3.1 and 3.2) with $\frac{N^+\hat{p}_i^+ + N^-\hat{p}_i^-}{N^++N^-}$ and $\frac{N^+p_i^+ + N^-p_i^-}{N^++N^-}$, respectively, and above results hold.

3.2.2 Rare variant Weighted Aggregate Statistic (RWAS)

Setting the weights for OWAS requires knowledge of the effect sizes which are unknown. To set the weights for our method without knowledge of the effect sizes, we assume a disease-risk model in which all variants have constant population attributable risk (PAR). In this model, each group of variants has a certain level of the group PAR, and each variant in the group has the same marginal PAR. Let ω denote the marginal PAR which is the group PAR divided by the number of causal variants in a group. Given ω and p_i of variant i, its relative risk, γ_i, is

$$\gamma_i = \frac{\omega}{(1-\omega)p_i} + 1 \tag{3.6}$$

Then, it follows from (Equations 4.6, 4.7),

$$\begin{aligned} p_i^+ &= \omega + p_i(1-\omega) \\ p_i^- &= p_i \end{aligned} \tag{3.7}$$

The optimal weights (Equation 3.2) can be written as

$$\lambda_i = \frac{\omega(1-p_i)}{\sqrt{2p_i^\pm(1-p_i^\pm)}} \approx \omega\sqrt{\frac{1-p_i}{p_i}} \quad (\text{assuming } p_i^\pm \approx p_i) \tag{3.8}$$

Since ω in (Equation 3.8) is fixed for all variants, we can ignore it and derive an analytically approximated form of the optimal weights as

$$w_i = \sqrt{\frac{1-p_i}{p_i}} \tag{3.9}$$

We call the weighted sum of z-scores whose weights are (Equation 3.9) Rare variant Weighted Aggregate Statistic (RWAS). The statistic of RWAS, S_{RWAS}, can be formulated as

$$S_{RWAS} = \frac{\sum w_i z_i}{\sqrt{\sum w_i^2}} \approx \frac{\sum \frac{\hat{p}_i^+ - \hat{p}_i^-}{\hat{p}_i^\pm}}{\sqrt{\frac{2}{N}}\sqrt{\sum \frac{1-\hat{p}_i^\pm}{\hat{p}_i^\pm}}} \sim \mathcal{N}\left(\frac{\sum \frac{p_i^+ - p_i^-}{p_i^\pm}}{\sqrt{\frac{2}{N}}\sqrt{\sum \frac{1-p_i^\pm}{p_i^\pm}}}, 1\right) \quad \big(\text{assuming } \hat{p}_i^\pm \approx p_i\big) \tag{3.10}$$

We compare S_{RWAS} to the standard normal distribution to obtain a p-value.

3.2.3 Approximation of MB to a sum of z-scores

Our methods (OWAS and RWAS) adopt a weighted sum of z-scores approach, and MB can also be approximated as a weighted sum of z-scores with weights equal to 1 (or unweighted sum of z-scores). MB computes a statistic, denoted as z_{MB}, as follows. It can be decomposed into a sum of z_{MB_i} over M variants where i corresponds to ith variant and M is the number of variants.

$$z_{MB} = \frac{x - \hat{\mu}}{\hat{\sigma}} = \sum_{i=1}^{M} z_{MB_i} = \sum_{i=1}^{M} \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i} \tag{3.11}$$

where for variant i, x_i is the sum of ranks or genetics scores of cases, $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the average and standard deviation of x_i in the null distribution, respectively. We use the sum of genetic scores of cases as x_i since its power is very similar to the power of the sum of ranks [MB09].

Then, x_i, $\hat{\mu}_i$ and $\hat{\sigma}_i$ can be approximated as (see section 3.2.6 for details)

$$x_i = \frac{\sqrt{N}}{2} \frac{\hat{p}_i^+}{\sqrt{\hat{p}_i^-(1 - \hat{p}_i^-)}}, \quad \hat{\mu}_i = \frac{\sqrt{N}}{2} \frac{p_i}{p_i(1 - p_i)}, \quad \hat{\sigma}_i = \sqrt{1/2} \tag{3.12}$$

and the standardized statistic at variant i, z_{MB_i}, can be derived as

$$\begin{aligned} z_{MB_i} = \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i} &\approx \frac{\frac{\sqrt{n}}{2}\left(\frac{\hat{p}_i^+}{\sqrt{\hat{p}_i^-(1-\hat{p}_i^-)}} - \frac{p_i}{\sqrt{p_i(1-p_i)}}\right)}{1/\sqrt{2}} \\ &\approx \sqrt{\frac{N}{2}} \frac{\hat{p}_i^+ - \hat{p}_i^-}{\sqrt{\hat{p}_i^-(1 - \hat{p}_i^-)}} \quad \text{(assuming } p_i \approx \hat{p}_i^-\text{)} \end{aligned}$$

Finally, a sum of z_{MB_i} over M variants is equivalent to the original statistic of MB.

$$z_{MB} = \frac{x - \hat{\mu}}{\hat{\sigma}} = \sum_{i=1}^{M} z_{MB_i} = \sum_{i=1}^{M} \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i} = \sum_{i=1}^{M} \sqrt{\frac{N}{2}} \frac{\hat{p}_i^+ - \hat{p}_i^-}{\sqrt{\hat{p}_i^-(1 - \hat{p}_i^-)}} \tag{3.13}$$

Note that (Equation 3.13) shows MB is an unweighted sum of z-scores. One difference between z_{MB_i} in (Equation 3.13) and z-score used in our methods (Equation 3.1) is the way they estimate the population MAF that appears in the denominator of z-score; MB estimates it only from control individuals, but we estimate it from all case and control individuals (see section 3.2.5 for details).

3.2.4 RWAS with prior information

RWAS can be directly extended to incorporate prior knowledge about the degree that each variant is believed to be causal. Note that the underlying truth is that each variant is either causal or not. Thus, let V^i be the variable indicating the "causal status" of variant i, such that $V^i = 1$ if variant i is causal and $V^i = 0$ if not. Let $V = \{V^1, ..., V^M\}$ denote the causal statuses of all M variants. V can have 2^M possible values. Let v_j be the jth value out of 2^M possible values. That is, $v_j = \{v_j^1, ..., v_j^M\}$ is an ordered set of 0 and 1 that represents a specific scenario of causal statuses.

Assume that we have prior knowledge that the probability of variant i being causal is c_i. Then, the probability of each scenario v_j can be computed as

$$P(v_j) = \prod_{i=1}^{M} c_i^{v_j^i} (1 - c_i)^{1-v_j^i} \ . \tag{3.14}$$

Then, the expected non-centrality parameter of the weighted sum of z-scores statistic is

$$\begin{aligned} E[\text{NCP}] &= \sum_{j=1}^{2^M} P(v_j) \sqrt{N} \frac{\sum_{i=1}^{M} w_i (v_j^i \lambda_i)}{\sqrt{\sum_{i=1}^{M} w_i^2}} && (3.15) \\ &= \frac{\sum_{i=1}^{M} c_i w_i \lambda_i}{\sqrt{\sum_{i=1}^{M} w_i^2}} \ . && (3.16) \end{aligned}$$

The Cauchy-Schwartz inequality shows that this quantity is maximized when $w_i = c_i \lambda_i$. Thus, the prior knowledge $\{c_i\}$ can be easily incorporated in RWAS by multiplying the prior probability into each weight.

3.2.5 Estimation of population MAF in OWAS

There can be several ways to estimate population MAF. For example, MB estimates it from control individuals [MB09]. We choose to estimate population MAF in the following way. Denoting p_i as population MAF of variant i, we first assume that its true overall sample frequency is equal to observed overall sample frequency.

$$p_i^+ + p_i^- = \hat{p}_i^+ + \hat{p}_i^- \tag{3.17}$$

p_i^+ and p_i^- are defined in terms of p_i (Equations 4.6, 4.7), and we can rewrite (Equation 3.17) as

$$\frac{\gamma_i p_i}{(\gamma_i - 1)p_i + 1} + p_i = 2\hat{p}_i^\pm \quad \left(\text{where } \hat{p}_i^\pm = \frac{\hat{p}_i^+ + \hat{p}_i^-}{2}\right) \tag{3.18}$$

We can compute p_i in terms of γ_i and $\hat{p}_i^\pm$ by finding the root of (Equation 3.18).

$$p_i \;=\; \frac{b + \sqrt{b^2 + 8(\gamma_i - 1)\hat{p}_i^\pm}}{2(\gamma_i - 1)} \quad \text{where } b = 2\hat{p}_i^\pm(\gamma_i - 1) - (\gamma_i + 1) \tag{3.19}$$

3.2.6 Approximation of x_i, $\hat{\mu}_i$ and $\hat{\sigma}_i$ of MB

In this section, we show that x_i, $\hat{\mu}_i$ and $\hat{\sigma}_i$ of MB can be approximated as (Equation 3.12). First, MB calculates a weight of variant i ($\hat{w}_i$) as

$$\hat{w}_i = \sqrt{N \cdot q_i(1 - q_i)} \quad \text{where } q_i = \frac{m_i^U + 1}{2n_i^U + 2} \tag{3.20}$$

N is the total number of case and control individuals, m_i^U is the number of mutations for variant i in control individuals, and n_i^U is the number of control individuals.

MB then calculates the genetic score (γ_j) of each individual j.

$$\gamma_j = \sum_{i=1}^{M} \frac{I_{ij}}{\hat{w}_i}$$

where M is the number of variants, and I_{ij} is the number of mutations observed in individual j at variant i. MB ranks all individuals (both cases and controls) by their genetic scores and calculates the sum of the ranks of cases as its test statistic (x).

$$x = \sum_{j \in \text{cases}} \text{rank}(\gamma_j)$$

Madsen and Browning reports that x can also be computed using the sum of genetic scores instead of the sum of ranks, and the two methods have very similar power. Hence, we will compute x as the sum of genetic scores.

$$x = \sum_{j \in \text{cases}} \gamma_j$$

First, we observe that the sum of genetic scores of cases is equivalent to the sum of observed MAF of each variant in cases divided by the weight of the variant. In other words, we sum the number of mutations per variant instead of the number of mutations per individual.

$$\sum_{j \in \text{cases}} \gamma_j = \sum_{i=1}^{M} \frac{N/2 \cdot \hat{p_i}^+}{\sqrt{N q_i (1 - q_i)}} \tag{3.21}$$

Assuming $q_i \approx \hat{p_i}^-$ since q_i is an estimate of MAF of variant i in controls, the statistic of variant i, x_i, in (Equation 3.21) is

$$x_i = \frac{\sqrt{N}}{2} \frac{\hat{p_i}^+}{\sqrt{\hat{p_i}^-(1 - \hat{p_i}^-)}} \tag{3.22}$$

Next, we derive the statistic of the null distribution denoted as x_i^* . First, $\hat{p_i}^+$ and $\hat{p_i}^-$ have the following distribution under the null distribution.

$$\hat{p_i}^+ \sim \mathcal{N}\left(p_i, \frac{p_i(1 - p_i)}{N/2}\right) \tag{3.23}$$

$$\hat{p_i}^- \sim \mathcal{N}\left(p_i, \frac{p_i(1 - p_i)}{N/2}\right) \tag{3.24}$$

By multiplying $\hat{p}_i^+$ in (Equation 3.23) by $\frac{\sqrt{N}}{2\sqrt{\hat{p_i}^-(1-\hat{p_i}^-)}}$ and assuming $\hat{p_i}^- \approx p_i$, we can derive x_i^* that is approximately equivalent to x_i in (Equation 3.22). x_i^* and its distribution are then

$$x_i^* = \frac{\sqrt{N}}{2} \frac{\hat{p_i}^+}{\sqrt{\hat{p_i}^-(1 - \hat{p_i}^-)}} \approx \frac{\sqrt{N}}{2} \frac{\hat{p_i}^+}{\sqrt{p_i(1 - p_i)}} \sim \mathcal{N}\left(\frac{\sqrt{N}}{2} \frac{p_i}{p_i(1 - p_i)}, \frac{1}{2}\right) \tag{3.25}$$

Thus, the mean ($\hat{\mu}_i$) of x_i^* is $\frac{\sqrt{N}}{2} \frac{p_i}{p_i(1-p_i)}$, the standard deviation ($\hat{\sigma}_i$) is $\sqrt{1/2}$, and x_i is (Equation 3.22)

3.2.7 Web Resources

The software package for RWAS is publicly available online at
http://genetics.cs.ucla.edu/rarevariants.

3.3 Results

3.3.1 Power comparison between RWAS and MB

We evaluate the power of our novel method, RWAS, in the constant PAR disease-risk model where all variants have the same PAR. This was the model used to estimate the power of MB, and MB was shown to be more powerful than other competing methods [MB09]. Throughout all experiments, we use the sum of genetic scores of case individuals as a statistic for MB, rather than using the sum of ranks of cases suggested by Madsen and Browning. One reason is that both sums yield similar results [MB09], and another reason is that the sum of genetic scores allows RWAS and MB to be compared in the same sum of z-scores framework (see Material and Methods for approximation of MB to a sum of z-scores method). The power of RWAS is also compared to the power of OWAS that is the optimal weighted sum of z-scores and from which the weights of RWAS are derived. OWAS uses the effect sizes of variants for its weights, and hence the power of OWAS can be thought of as the upper bound of power that can be achieved in the weighted sum of z-scores approach. In this experiment, OWAS knows the group PAR that generated datasets (see below), computes relative risk of each variant using (Equation 4.10), and estimates population MAF as described in Section 3.2.5.

We use the exactly same simulation parameters as in Madsen and Browning to estimate the power of methods. In the simulations, a total of 10,000 datasets are generated, each with 1000 case and 1000 control individuals having 100 variants. The power of a method is estimated as the number of significant datasets among the 10,000 datasets using a significance threshold of 2.5×10^{-6} based on the Bonferroni correction assuming 20,000 genes genomewide. Among 100 variants, 50 variants are disease-risk contributing variants (D-variants) and 50 variants are disease-risk neutral variants (N-variants). For each variant, we sample its minor allele frequency (MAF) in controls using Wright's formula [Wri31, Ewe04] with the same parameter values as in Madsen and Browning (see [MB09] for details). According to (Equation 4.10), relative risk of D-variants is calculated from MAF of variants in controls and the marginal PAR that is the group PAR divided by the number of D-variants while relative risk of N-variants is 1. MAF of variants in cases can then be calculated using relative risk and MAF of variants in controls according to (Equation 4.6). We independently sample mutations of each variant in case and control individuals according to its MAF in cases and controls, respectively.

The results of power simulations demonstrate that RWAS consistently outperforms MB when the group PAR varies from 1% to 5% (Figure 3.1). For example, at the group PAR of 3%, RWAS has 78% power while MB has 40% power. The power simulations also show that the power of RWAS is very close to the power of OWAS. Although OWAS has higher power than RWAS at all group PAR levels, the difference in power between the two methods is small; the power of RWAS is about 2-4% smaller than that of OWAS. Therefore, the analytical approximation of the optimal weights in RWAS reduces its power by only a small amount in this disease model, and it can achieve high power even if it is not given the true effect sizes of variants.

3.3.2 Type I error rates of RWAS and OWAS

To check whether type I error rates (false positive rates) of RWAS and OWAS are correctly controlled, we create 100 million datasets without any causal variant. Each dataset has 1000 cases and 1000 controls with 100 variants, and we measure type I error rates of RWAS and OWAS on the 100 million null datasets under three different significance thresholds; 0.05, 0.01, 2.5×10^{-6}. The reason why we use a very large number of datasets is because the significance threshold for power is very low (2.5×10^{-6}). The proportion of significant datasets is an estimate of the type I error rate for each method.

The type I error rates for RWAS are 0.0503, 0.0089, and 1.2×10^{-7}, and those for OWAS are 0.0502, 0.0091, 1.8×10^{-7} for the significance threshold of 0.05, 0.01, and 2.5×10^{-6}, respectively. This indicates that the type I error rates are correctly controlled for RWAS and OWAS when the significance thresholds are 0.05 and 0.01. When the significance threshold is 2.5×10^{-6}, RWAS and OWAS both have lower type I error rates than the expected rate.

3.3.3 Power of RWAS with the different numbers of variants

Since the number of variants in a gene may be more than 100, we evaluate effects of the number of variants in a gene on the power of groupwise tests. We create five different datasets with five different numbers of total variants; 100, 200, 300, 400, and 500. In all five datasets, the number of causal variants is 50, and the group PAR is 3%. The number of case and control individuals is the same as the previous experiment.

Figure 3.2 shows that as the number of total variants in a gene increases, the power of all methods decreases. For example, when a gene contains 100 variants, RWAS achieves 78% power while it has 6% power when there are 500 variants in a gene.

This is because there are more non-causal variants in a gene as the number of variant increases. A large number of non-causal variants reduce our ability to detect causal variants and power of the groupwise tests.

3.3.4 Power of RWAS with prior information

Prior information can reduce or remove influence of non-causal variants, and in this experiment, we observe how prior information influences the power of RWAS. The prior information we consider is the probability of a variant being causal to a disease, denoted as c_i. We generate datasets with pre-defined true c_i values, and we evaluate how the power of RWAS changes when different prior information is given to RWAS. We first generate datasets that contain 100 variants split into two groups, each with 50 variants. We set c_i of the first group to 0.8, and c_i of the second group to 0.2. Then, five different types of prior information are given to RWAS; 1) "correct c_i" that is equivalent to true c_i of datasets, 2) "uniform incorrect c_i" in which $c_i = 1$ for all variants, 3) "three fourths correct c_i" that corresponds to $\frac{3}{4}$ of true c_i of the first and second groups, 4) "half correct c_i" that matches a half of true c_i of the first and second groups, and 5) "very incorrect c_i" in which c_i of the first and second groups is 0.2 and 0.8, respectively, which is opposite to true c_i of datasets. The single marker test and MB are also tested to compare their power to RWAS.

We follow the same set of experimental framework as the previous experiment in this power simulation with two changes. The first change is that we have two different c_i values assigned to the two groups of variants as mentioned earlier. For each dataset, a variant is causal with the probability proportional to its c_i. Relative risk of a causal variant is given by (Equation 4.10) whereas a non-causal variant has relative risk of 1. The other change is that the same set of control MAF is assigned to the two groups; MAF of 50 variants in control individuals are sampled using the Wright's formula and assigned to each group. The reason is that we want to observe only the effect of prior information on the power of studies, but the power is also dependent on MAF of variants.

Results show that the power of RWAS with “correct c_i” is always the highest among different prior information applied to RWAS (Figure 3.3). By knowing the correct prior information, the power increases as much as 7%; at the group PAR of 3%, the power of RWAS with “correct c_i” is 84% while the power of RWAS with “uniform incorrect c_i” is 77%. However, if RWAS is given incorrect prior information, it may suffer power loss as the power of RWAS with “very incorrect c_i” is more than 70% lower than the power of RWAS with “correct c_i” at the group PAR of 3% and 4%. This shows that when prior information is not very accurate, RWAS may achieve higher power by assuming that every variant is causal. The results also indicate that as RWAS is given more correct prior information, its power increases; the power of RWAS with “three fourths correct c_i” is higher than the power of RWAS with “half correct c_i.” Results of the experiment demonstrate that prior information may considerably influence the power of studies and higher power can be achieved by knowing correct prior information.

3.3.5 RWAS with prior information on real mutation screening data

To evaluate RWAS and effects of prior information on real sequencing data, we use mutation screening data of the susceptibility gene for ataxia telangiectasia [TOB09]. This gene is called *ATM*, and it is also known as an intermediate-risk gene for breast cancer. Tavtigian *et al.* [TOB09] collected data from seven *ATM* mutation screening studies in breast cancer cases and controls as well as data from their own mutation screening, which resulted in collecting 2531 case and 2245 control individuals (called “bona fide case-control studies”). They further increased the number of cases and controls by adding 17 case-only or control-only mutation screening studies, but we focus on the bona fide case-control studies in our experiment because adding the case-only and control-only studies does not yield substantial changes in results [TOB09].

Tavtigian *et al.* discovered 170 rare missense variants in the *ATM* dataset, and used the missense analysis programs, Align-GVGD [TDY06] and SIFT [NH03], to find how likely each variant is deleterious. Align-GVGD categorizes variants into seven grades: C0 (most likely neutral), C15, C25, C35, C45, C55, and C65 (most likely deleterious). Since the absolute deleteriousness of grades is not reported, we arbitrarily assign c_i of 0.05, 0.2, 0.35, 0.5, 0.65, 0.8, and 0.95 to the 7 grades, respectively. SIFT yields scores for variants ranging from 1.00 (most likely neutral) to 0.00 (most likely deleterious) in steps of 0.01. There is a pre-defined threshold (0.05) in SIFT scores such that variants whose SIFT scores are ≤ 0.05 are considered deleterious while other variants are considered neutral. Hence, we assigned c_i of 1 to variants with SIFT scores ≤ 0.05 and c_i of 0 to other variants.

We first apply RWAS to the case-control studies without prior information, and RWAS yields a p-value of 0.3946. The p-value indicates no significant difference in mutation counts between cases and controls, and Tavtigian *et al.* also reported that they did not find a significant association by comparing frequency in cases versus controls or by using CMC [TOB09]. However, when RWAS is applied with prior information from Align-GVGD, it yields a p-value of 0.0078, which indicates a significant association between rare variants and the disease. The result is consistent with the results of [TOB09]; a significant p-value was obtained by performing a log-linear trend test with output of Align-GVGD. Therefore, this suggests that prior information may be useful in association studies and that RWAS can be applied to detect an association in real data.

Interestingly, RWAS reports a non-significant p-value of 0.0881 when using SIFT scores as prior information while [TOB09] found a significant association with SIFT scores. It may be because the binary classification of variants according to SIFT scores is not as informative as output of Align-GVGD in predicting how likely each variant is causal. In other words, variants that are considered deleterious (SIFT scores $\leq$ 0.05) may be deleterious to different degrees, but SIFT scores do not capture this. The relative degree of how deleterious a variant is important in RWAS because more deleterious variants receive higher weights. Hence, this experiment shows that methods to determine prior information of variants play a key role in the real data analysis, and different prior information may yield different results.

3.4 Discussion

In this chapter, we presented Rare variant Weighted Aggregate Statistic (RWAS) to detect associations with a group of rare variants. We first developed the Optimal Weighted Aggregate Statistic (OWAS) that maximizes the power of studies under the weighted sum of z-scores statistic, but we need to know the effect sizes of variants to use OWAS. We then developed RWAS by analytically approximating the optimal weights, and it can be applied without the knowledge of effect sizes. The simulations demonstrate that RWAS outperforms a weighted sum statistic by Madsen and Browning [MB09] in the same disease-risk model discussed in [MB09]. The simulations also show that the power of RWAS is very close to the power of OWAS, suggesting that RWAS achieves nearly optimal power in the disease-risk model we focused on.

We then extended RWAS to incorporate prior information of variants, and we considered the probability of a variant being causal to a disease as prior information in this chapter. To determine effects of prior information on association studies, we used both simulated data and real mutation screening data for the susceptibility gene for ataxia telangiectasia. The results of simulated data show that power can be increased by incorporating correct prior information, and this is confirmed in the real data since RWAS is able to detect an association in the real data with prior information while it is not able to do so without the information. Hence, this suggests that it would be advantageous to incorporate prior information into association studies and RWAS can be used to find associations in such association studies.

Many studies suggest that rare variants are not in linkage disequilibrium with each other [LL08, PC02, Pri01]. To compute the p-values, our statistic assumes that these variants are independent. However, in the case that the rare variants are linked, we can apply a permutation test to obtain p-values in order to apply the method.

Reference to published article

Jae Hoon Sul, Buhm Han, Dan He, and Eleazar Eskin, "An Optimal Weighted Aggregated Association Test for Identification of Rare Variants Involved in Common Diseases." *Genetics*. 188, 181-8, 2011

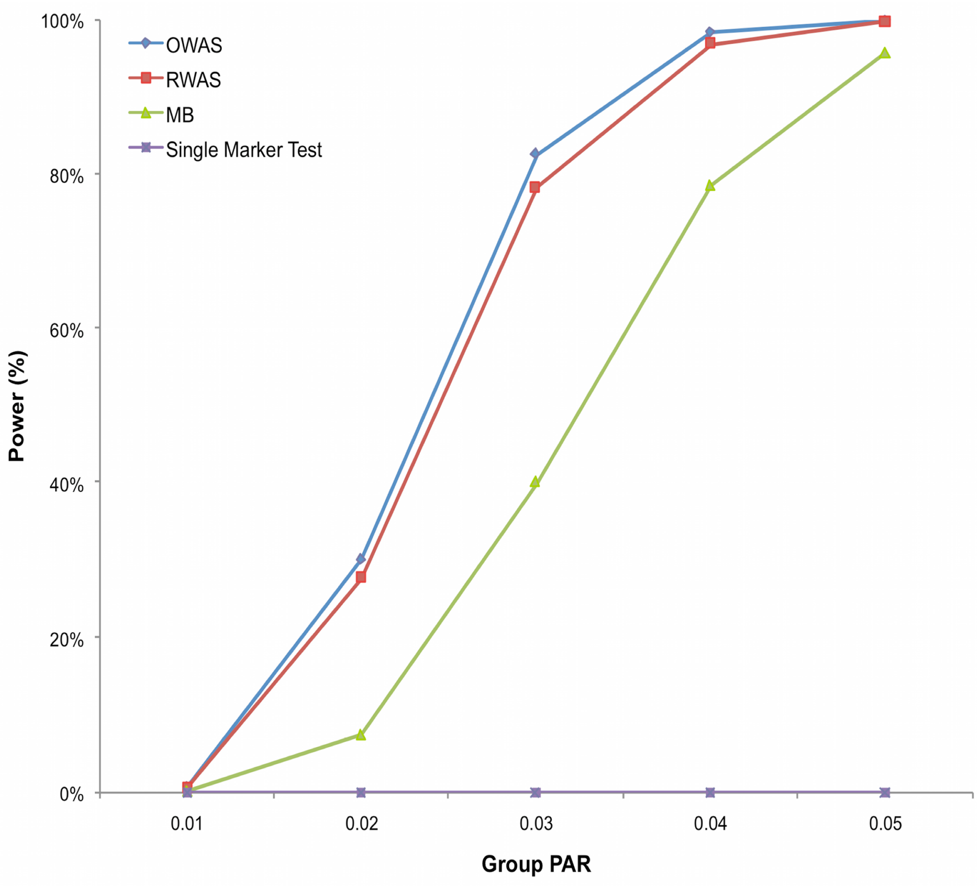


Figure 3.1: Power comparison in the constant PAR model. There are 5 group PAR levels (1%, 2%, 3%, 4%, and 5%), and for each group PAR, 10,000 datasets were generated. Each dataset contained 1000 case and 1000 control individuals having 100 variants (50 D-variants and 50 N-variants). Four different methods (RWAS, OWAS, MB, and the single marker test) were tested, and their power was estimated as the number of significant datasets among the 10,000 datasets using a significance threshold of 2.5×10^{-6}.

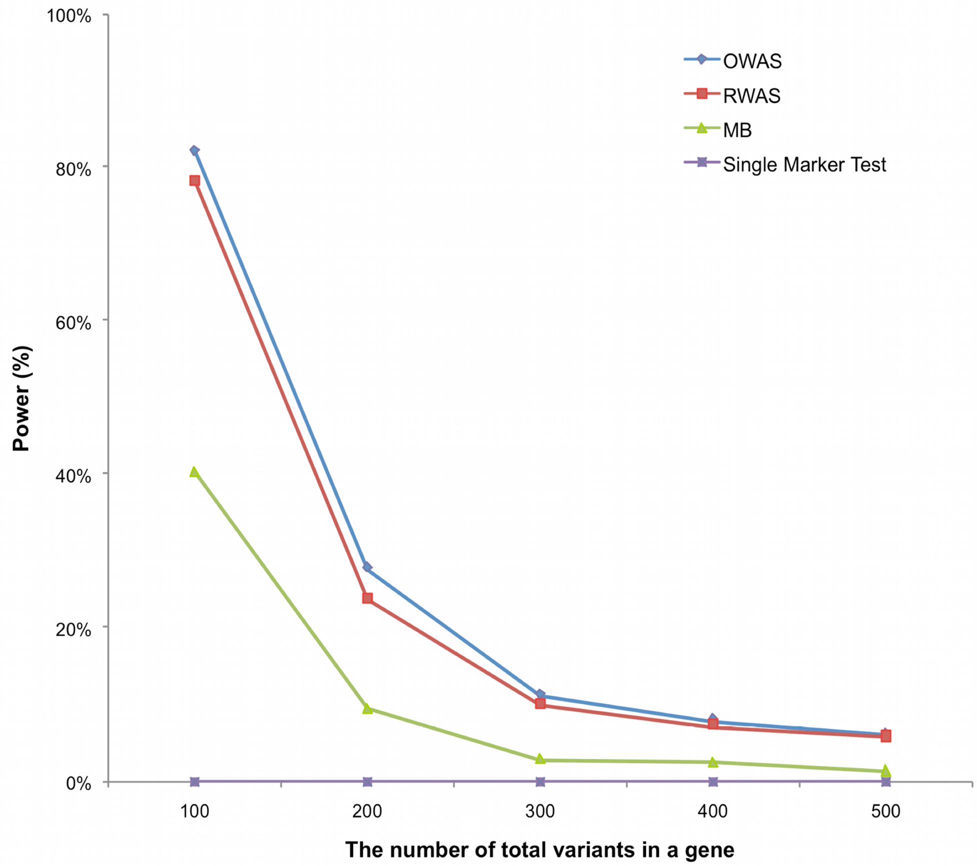


Figure 3.2: Power comparison with the different numbers of total variants in a gene. We simulated 5 different numbers of total variants; 100, 200, 300, 400, and 500. All simulations had 50 causal variants, the group PAR of 3%, and 1000 case and 1000 control individuals. We created 10,000 datasets for each of 5 different simulations. The plot shows the power of RWAS, OWAS, MB, and the single marker test.

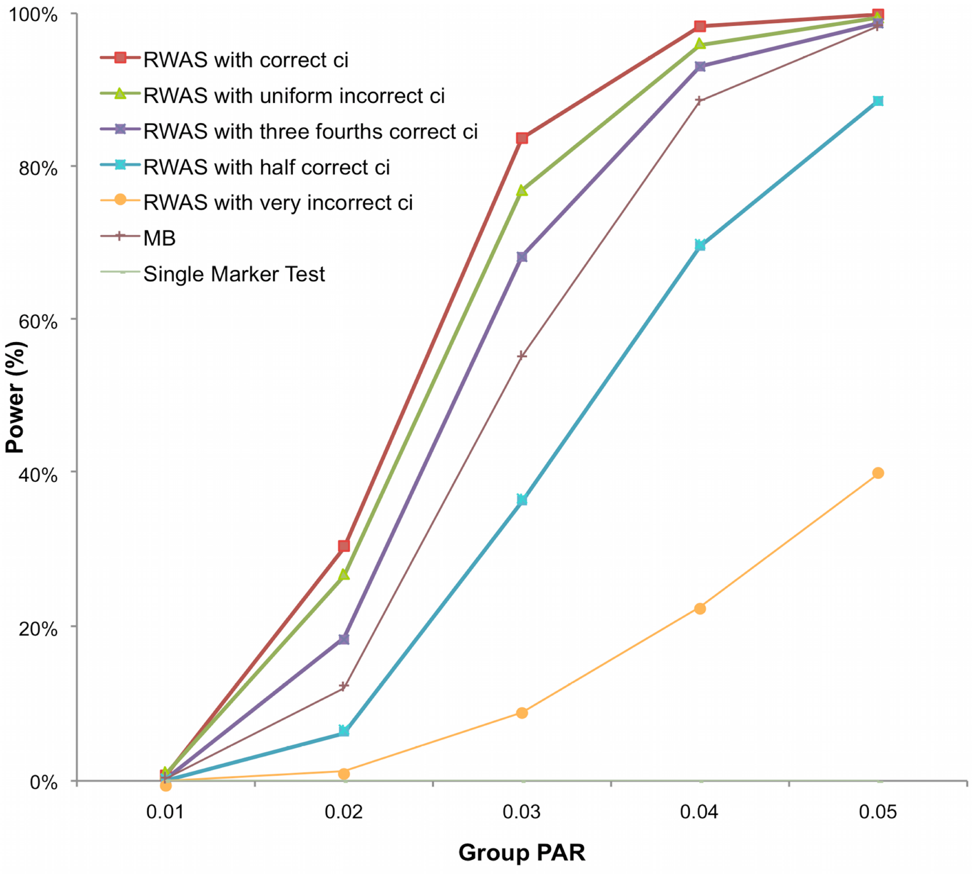


Figure 3.3: Power of RWAS with different prior information. For each group PAR, 10,000 datasets were generated, and each dataset contained 1000 case and 1000 control individuals having 100 variants with pre-defined true c_i values. c_i of 50 variants was 0.8, and c_i of the other 50 variants was 0.2. Five different types of prior information were given to RWAS; "correct c_i" (same c_i as true c_i of datasets), "uniform incorrect c_i" ($c_i = 1$ for all variants), "three fourths correct c_i" (equal to $\frac{3}{4}$ of true c_i), "half correct c_i" (equal to a half of true c_i), and "very incorrect c_i" (opposite c_i to true c_i of datasets). The single marker test, MB, and RWAS with the five different types of prior information were tested.

CHAPTER 4

Likelihood ratio test to increase power of groupwise association test

4.1 Background

Current genotyping technologies have enabled cost-effective genome-wide association studies (GWAS) on common variants. Although these studies have found numerous variants associated with complex diseases [CSS93, BKK94, AHK00], common variants explain only a small fraction of disease heritability. This has led studies to explore effects of rare variants, and recent studies report that multiple rare variants affect several complex diseases [GGS08, KPS07, CKP04, FWW04, JFO08, BB08, RPF07, BVE08, Con08, XRL08, WMM08]. However, the traditional statistical approach that tests each variant individually by comparing the frequency of the variant in individuals who have the disease (cases) with the frequency in individuals who do not have the disease (controls) yields low statistical power when applied to rare variants due to their low occurrences.

Identifying genes involved in diseases through multiple rare variants is an important challenge in genetics today. The main approach currently proposed is to group variants in genes and detect associations between a disease and these groups. The rationale behind this approach is that multiple rare variants may affect the function of a gene. By grouping variants, we may observe a larger difference in mutation counts between case and control individuals and hence, power of studies increases. Recently, several methods have been developed for the groupwise approach such as the Cohort Allelic Sums Test (CAST) [MT07], the Combined Multivariate and Collapsing (CMC) method [LL08], a weighted-sum statistic by Madsen and Browning (MB) [MB09], a variable-threshold approach (VT) [PKB10], and Rare variant Weighted Aggregate Statistic (RWAS) [SHH11].

In combining information from multiple rare variants, a groupwise association test faces two major challenges. The first is unknown effect sizes of variants on the disease phenotype. To address this challenge, MB and RWAS discuss a disease risk model in which rarer variants are assumed to have higher effect sizes than common variants [MB09, SHH11]. This model provides a simulation framework that would be appropriate for testing the groupwise tests on rare variants because it describes associations usually not found in traditional GWAS. RWAS is shown to outperform other grouping methods under this disease risk model [SHH11]. The second challenge is that only a subset of the rare variants in the gene will have an effect on the disease and which of these variants are causal is unknown. Including non-causal variants in a groupwise association test may reduce power because it decreases the relative contribution of the true causal variants to the statistic [SHH11]. RWAS and VT attempt to overcome this challenge by utilizing prior information of which variants are likely deleterious, and prior information can be obtained from bioinformatics tools such as Align-GVGD [TDY06] , SIFT [NH03] and PolyPhen-2 [ASP10]. By incorporating prior information into the methods, RWAS and VT reported that they achieved higher power [PKB10, SHH11].

These methods do not achieve the best performance even under the assumptions of their disease model, as we show below, and we improve on the previous methods by taking advantage of the following ideas. First, observational data can give us a clue to which variants are causal in data because casual variants occur more frequently in cases than in controls. Hence, a method that infers causal variants from data would outperform methods that do not, and previous methods fall into the latter category. In addition, previous methods such as RWAS, MB, and VT compute their statistics using a linear sum of mutation counts. In these methods, a variant having large discrepancy in mutation counts between cases and controls has the same effect on a statistic as the sum of two variants having small discrepancies with half the size of the large one. However, the large discrepancy should contribute more than the sum of small discrepancies because a variant that causes the large difference in mutation counts is more likely to be involved in a disease. To emphasize the large discrepancy, a nonlinear combination of mutation counts is necessary. Finally, the set of rare variants in the gene and their distribution among cases and controls can be used to estimate the effect sizes of the rare variants on the disease. This estimate can then be used to improve the statistical power of the method.

In this chapter, we present a novel method for the groupwise association test based on a likelihood ratio test (LRT). LRT computes and compares likelihoods of two models; the null model that asserts no causal variants in a group and the alternative model that asserts at least one causal variant. To compute likelihoods of the models, LRT assumes that some variants are causal and some are not (called "causal statuses of variants") and computes the likelihood of the data under each possible causal status. This allows LRT to compute likelihoods of the null and alternative models, and a statistic of LRT is a ratio between likelihoods of the two models.

LRT takes advantage of both prior information and data to compute likelihoods of underlying models, and hence it uses more information than previous methods to identify a true model that generated data. Simulations show that LRT is more powerful than previous methods such as RWAS and VT using the same set of prior information.

We also show by using real mutation screening data of the susceptibility gene for ataxia telangiectasia that LRT is able to detect an association previously reported by [TOB09] and [SHH11].

Another improvement of LRT is that it computes its statistic using a nonlinear combination of mutation counts as opposed to a linear sum of counts in the previous methods. Simulations show that this difference creates different decision boundaries (nonlinear versus linear decision boundaries) that determine whether a group of variants is associated with a disease. Moreover, we demonstrate that the nonlinear decision boundary allows LRT to detect more associations than the linear boundary.

Unfortunately, to compute the LRT statistic directly, we must consider a number of possible models exponential in the number of rare variants in the gene. In addition, we must perform this computation once for each permutation and we must perform millions of permutations to guarantee that we control false positives when trying to obtain genome-wide significance. We address these computational challenges by decomposing the computation of LRT and developing an efficient permutation test. Unlike the standard approach to compute the LRT statistic which requires exponential time complexity, we make a few assumptions and derive a method for computing the LRT statistic whose time complexity is linear. For the permutation test, we further decompose LRT and take advantage of the distribution of allele frequency. These techniques allow us to compute a statistic of each permutation efficiently, and hence we can perform a large number of permutations to obtain genome-wide significance. We provide the software package for LRT at http://genetics.cs.ucla.edu/rarevariants.

4.2 Methods

4.2.1 Likelihood Ratio Test

We consider likelihoods of two models under LRT; the likelihood of the null model (L_0) and the likelihood of the alternative model (L_1). The null model assumes that there is no variant causal to a disease while the alternative model assumes there is at least one causal variant. To compute the likelihood of each model, let D^+ and D^- denote a set of haplotypes in case and control individuals, respectively. We assume there are M variants in a group, and let V^i be the indicator variable for the "causal status" of variant i; $V^i = 1$ if variant i is causal, and $V^i = 0$ if not causal. Let $V = \{V^1, ..., V^M\}$ represent the causal statuses of M variants, and there exist 2^M possible values for V. Among them, let $v_j = \{v_j^1, ..., v_j^M\}$ be jth value, consisting of 0 and 1 that represent one specific scenario of causal statuses [SHH11]. We use c_i to denote the probability of variant i being causal to a disease. Then, assuming that the causal statuses are independent between variants, we can compute the prior probability of each scenario v_j as

$$P(v_j) = \prod_{i=1}^{M} c_i^{v_j^i}(1 - c_i)^{1-v_j^i} \ . \tag{4.1}$$

We define $L(D^+, D^-|v_j)$ as the likelihood of observing case and control haplotypes given jth scenario. Then, L_0 and L_1 can be defined as

$$\begin{aligned} L_0 &= L(D^+, D^-|v_0)P(v_0) && (4.2) \\ L_1 &= \sum_{j=1}^{2^M-1} L(D^+, D^-|v_j)P(v_j) && (4.3) \end{aligned}$$

where v_0 is a scenario where $v_0^i = 0$ for all variants; no causal variants. In section 4.2.6 we describe how we can compute $L(D^+, D^-|v_j)$. The computation is based on the no linkage disequilibrium (LD) assumption, which is reasonable on rare variants, because very low or no LD is expected between rare variants [LL08, PC02, Pri01].

The statistic of LRT is a ratio between L_1 and L_0, L_1/L_0, and we perform a permutation test to compute a p-value of the statistic.

4.2.2 Decomposition of LRT to increase computational efficiency

We decompose L_0 and L_1 in (Equations 4.2, 4.3) such that we compute likelihoods of variants instead of likelihoods of haplotypes to reduce the computational complexity. To compute L_1 in (Equation 4.3), we need to compute likelihoods of 2^M scenarios of causal statuses, which is computationally expensive if there are many rare variants in a group. To decompose likelihoods of haplotypes, we need to make one assumption, and it is low disease prevalence.

Assume there are $N/2$ case and $N/2$ control individuals. Let $H_k = \{H_k^1, H_k^2, \ldots, H_k^M\}$ denote kth haplotype, where $H_i^k \in \{0, 1\}$. $H_k^i = 1$ if ith variant in kth haplotype is mutated, and $H_k^i = 0$ if not. Let p_i denote population minor allele frequency (MAF) of variant i, and p_i^+ and p_i^- represent the true MAF of case and control individuals, respectively. We denote relative risk of variant i by γ_i. Then, L_0 and L_1 of (Equations 4.2, 4.3) can be decomposed into (see section 4.2.7 for the derivation)

$$L_0 = \prod_{i=1}^{M} \left\{ (1-c_i) \prod_{H_k \in D^+} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \prod_{H_k \in D^-} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \right\} \tag{4.4}$$

$$\begin{aligned} L_0 + L_1 = \prod_{i=1}^{M} \Bigg\{ & (1-c_i) \prod_{H_k \in D^+} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \prod_{H_k \in D^-} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \\ & + c_i \prod_{H_k \in D^+} {p_i^+}^{H_k^i} (1-p_i^+)^{1-H_k^i} \prod_{H_k \in D^-} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \Bigg\} \end{aligned} \tag{4.5}$$

where p_i^+ and p_i^- are

$$p_i^+ = \frac{\gamma_i p_i}{(\gamma_i - 1)p_i + 1} \tag{4.6}$$

$$p_i^- = p_i \quad \text{(assuming the disease prevalence is very small)} \tag{4.7}$$

We estimate the population MAF of a variant (p_i) using an observed overall sample frequency.

$$p_i = \frac{\hat{p}_i^+ + \hat{p}_i^-}{2}$$

where $\hat{p}_i^+$ and $\hat{p}_i^-$ represent observed case and control MAF, respectively.

This decomposition reduces the time complexity of computing L_1 from exponential to linear, substantially increasing the computational efficiency.

4.2.3 Efficient permutation test for LRT

We propose a permutation test that is substantially more efficient than a naive permutation test that permutes case and control statuses in each permutation. The naive permutation test is computationally expensive because every haplotype of case and control individuals needs to be examined in each permutation, and hence it requires more computation as the number of individuals increases. Moreover, to compute a p-value at a genome-wide level, more than 10 million permutations are necessary assuming a significance threshold of 2.5×10^{-6} (computed from the overall false positive rate of 0.05 and the Bonferroni correction with 20,000 genes genome-wide). It is often computationally impractical to perform this large number of permutations with the naive permutation test. Hence, we develop a permutation test that does not permute case and control statuses, and this makes the time complexity independent of the number of individuals and allows the permutation test to be capable of performing more than 10 million permutations.

First, we reformulate L_0 and L_1 (Equations 4.4, 4.5) such that they are composed of terms that do not change and terms that change per each permutation (see section

4.2.8 for the derivation).

$$\begin{aligned} L_0 &= \prod_{i=1}^{M} X_i && (4.8) \\ L_0 + L_1 &= \prod_{i=1}^{M} \left\{ X_i + K_i Y_i^{N\hat{p}_i^+} \right\} && (4.9) \end{aligned}$$

where

$$\begin{aligned} X_i &= (1-c_i)p_i^{2Np_i}(1-p_i)^{2N-2Np_i} \\ K_i &= c_i(1-p_i^+)^N(1-p_i)^N \left(\frac{p_i}{1-p_i}\right)^{2Np_i} \\ Y_i &= \left(\frac{p_i^+}{1-p_i^+} \cdot \frac{1-p_i}{p_i}\right) \end{aligned}$$

In (Equations 4.8 and 4.9), it is only a $\hat{p}_i^+$ term that changes when the dataset is permuted because p_i and p_i^+ are invariant per permutation, meaning X_i, K_i, and Y_i are constant. $N\hat{p}_i^+$ follows the hypergeometric distribution with the mean equal to Np_i and the variance equal to $\frac{N}{2}p_i(1-p_i)$ under permutations. Hence, we sample $N\hat{p}_i^+$ from the hypergeometric distribution, and since this sampling strategy does not permute and examine haplotypes of individuals, it is more efficient than the naive permutation test when studies have a large number of individuals.

To speed up sampling from the hypergeometric distribution, we pre-compute hypergeometric distributions of all rare variants (e.g. variants whose MAF are less than 10%) before performing the permutation test. Computing the hypergeometric distribution requires several factorial operations, which is computationally expensive. The pre-computation of distributions allows the permutation test to avoid having the expensive operations repeatedly per permutation, and the number of pre-computed distributions is limited due to the small range of MAF. For common variants, we sample $N\hat{p}_i^+$ from the normal distribution, which approximately follows the hypergeometric distribution when $\hat{p}_i^+$ is not close to 0 or 1.

We find that our permutation test is efficient enough to calculate a p-value of the LRT statistic at a genome-wide level. For example, using a dataset that contains 1000 cases and 1000 controls with 100 variants, 10 million permutations take about 10 CPU minutes using one core of a Quad-Core AMD 2.3 GHz Opteron Processor. Note that the time complexity of our method is $O(N + kMP)$ where N is the total number of individuals, M is the number of variants, P is the number of permutations, and k is the number of iterations in the local search algorithm discussed below (see "Estimating PAR of a group of variants using LRT" section for more details). We find that k is very small in permutations and $MP \gg N$ for a large number of permutations (e.g. 100 millions). Thus, the time complexity of our method becomes approximately $O(MP)$, and this shows that the amount of computation our method needs mostly depends on the number of variants and the number of permutations.

We note that our permutation test can also be applied to previous grouping methods such as RWAS [SHH11]. RWAS assumes that its statistic (a weighted sum of z-scores of variants) approximately follows the normal distribution, and the p-value is obtained accordingly. Since the permutation test does not make any assumptions on the distribution of a statistic, it may provide a more accurate estimate of a p-value and improve the power of previous methods.

4.2.4 Power Simulation Framework

The effect sizes and the causal statuses of variants are two major factors that influence the power of the groupwise association test. To simulate these two factors, we adopt the same simulation framework as one discussed in Sul *et al.* and Madsen and Browning [MB09, SHH11]. In this framework, population attributable risk (PAR) defines the effect sizes of variants, and we assign the predefined group PAR to a group of variants. The group PAR divided by the number of causal variants is the marginal PAR, denoted as ω, and every variant has the same ω.

The effect size of a variant also depends on its population MAF in this simulation framework. We assign each variant population MAF (p_i) sampled from Wright's formula [Wri31, Ewe04], and we use the same set of parameter values for the formula as discussed in Sul *et al.* and Madsen and Browning (see [MB09, SHH11] for details). Using ω and population MAF, we can compute relative risk of variant i (γ_i) as following.

$$\gamma_i = \frac{\omega}{(1-\omega)p_i} + 1 \tag{4.10}$$

(Equation 4.10) shows that rarer variants have the higher effect sizes. Given relative risk and population MAF of a variant, we compute the true case and control MAF of the variant according to (Equations 4.6 and 4.7). We then use the true case and control MAF to sample mutations in case and control individuals, respectively.

To simulate the causal status of a variant, we assign each variant the probability of being causal to a disease. Let c_i denote this probability for variant i, and in each dataset, a variant is causal with the probability c_i, and not causal with the probability $1 - c_i$. Relative risk of a causal variant is defined in (Equation 4.10) while that of non-causal variant is 1.

Given all parameters of variants, we generate 1,000 datasets, and each dataset has 1,000 case and 1,000 control individuals with 100 variants. Since we are interested in comparing power of the groupwise tests, we only include datasets that have at least two causal variants. The number of significant datasets among the 1,000 datasets is used as an estimate of power with the significance threshold of 2.5×10^{-6}.

4.2.5 Estimating PAR of a group of variants using LRT

We need a few model parameters to compute the LRT statistic, and we use data, prior information, and the LRT statistic itself to estimate the parameters. More specifically, we need to know relative risk of variant i, γ_i, to compute p_i^+ in (Equation 4.6). According to (Equation 4.10), γ_i depends on population MAF (p_i) and the marginal PAR (ω) which is the group PAR divided by the number of causal variants. We can estimate p_i from observational data, and we use prior information (c_i) of variants to compute the expected number of causal variants, which we use as an estimate of the number of causal variants.

To estimate the group PAR, we use the LRT statistic because we are likely to observe the greatest statistic when the statistic is computed using the group PAR that generated observational data. We apply a local search algorithm to find the value of PAR that maximizes the LRT statistic; we compute the statistic assuming a very small PAR value (0.1%), and iteratively compute statistics using incremental values of PAR (0.2%, 0.3%, etc.) until we observe a decrease in the LRT statistic. After we find the maximum LRT statistic, we perform the permutation test with the same local search algorithm to find the significance of the statistic.

4.2.6 Computation of $L(D^+, D^-|v_j)$ in LRT

We show how the likelihood of haplotypes under certain causal statuses of variants, $L(D^+, D^-|v_j)$, can be computed. Let H_k denote kth haplotype, and $H_k = \{H_k^1, H_k^2, \dots, H_k^M\}$. $H_k^i = 1$ if ith variant in kth haplotype is mutated, and $H_k^i = 0$ otherwise. Let p_i denote population minor allele frequency (MAF) of ith variant, and we can compute the probability of a haplotype H_k under the assumption of no linkage disequilibrium as

$$P(H_k) = \prod_{i=1}^{M} p_i^{H_k^i}(1-p_i)^{1-H_k^i} \tag{4.11}$$

Then, we define the likelihood of haplotypes as

$$L(D^+, D^-|v_j) = \prod_{H_k \in D^+} P(H_k|+, v_j) \prod_{H_k \in D^-} P(H_k|-, v_j) \tag{4.12}$$

where $+$ and $-$ denote case and control statuses. In order to compute $P(H_k|+/-, v_j)$, we first denote F as disease prevalence and $\gamma_{v_j}^{H_k}$ as the relative risk of kth haplotype under v_j. We define $\gamma_{v_j}^{H_k}$ as

$$\gamma_{v_j}^{H_k} = \prod_{i=1}^{M} \gamma_i^{v_j^i H_k^i}$$

Let H_0 denote the haplotype with no variants, and using Bayes' theorem and independence between H_k and v_j, and between disease status ($+$ and $-$) and v_j, we can define the $P(H_k| + /-, v_j)$ as

$$\begin{aligned} P(H_k|+, v_j) &= \frac{P(H_k, +|v_j)}{P(+)} = \frac{P(+|H_k, v_j)P(H_k)}{F} \\ &= \frac{\gamma_{v_j}^{H_k} P(+|H_0, v_j)P(H_k)}{F} && (4.13) \\ P(H_k|-, v_j) &= \frac{P(H_k, -|v_j)}{P(-)} = \frac{(1 - P(+|H_k, v_j))P(H_k)}{1 - F} && (4.14) \end{aligned}$$

$P(+|H_0, v_j)$, or the probability of having a disease given no variants in the haplotype under jth causal statuses, can be computed as

$$\sum_{k=0}^{2^M-1} \gamma_{v_j}^{H_k} P(+|H_0, v_j)P(H_k) = F$$

$$P(+|H_0, v_j) = \frac{F}{\sum_{k=0}^{2^M-1} \gamma_{v_j}^{H_k} P(H_k)}$$

4.2.7 Decomposition of likelihoods of haplotypes into likelihoods of variants in LRT

First, we consider two variants case. We have 4 possible causal statuses, denoted as $v_{00}, v_{01}, v_{10}, v_{11}$ and 4 possible haplotypes, denoted as $H_{00}, H_{01}, H_{10}, H_{11}$. Let p_1 and

p_2 denote population MAF of two variants and p_1^+ and p_2^+ are MAF of case individuals at two variants. The original LRT statistic based on (Equations 4.2 and 4.3) compute the following likelihoods

$$\begin{aligned} L_0 &= (1-c_1)(1-c_2)\prod_{H_k\in D^+}P(H_k|+,v_{00})\prod_{H_k\in D^-}P(H_k|-,v_{00}) \\ L_0+L_1 &= (1-c_1)(1-c_2)\prod_{H_k\in D^+}P(H_k|+,v_{00})\prod_{H_k\in D^-}P(H_k|-,v_{00}) \\ &+ (1-c_1)c_2\prod_{H_k\in D^+}P(H_k|+,v_{01})\prod_{H_k\in D^-}P(H_k|-,v_{01}) \\ &+ c_1(1-c_2)\prod_{H_k\in D^+}P(H_k|+,v_{10})\prod_{H_k\in D^-}P(H_k|-,v_{10}) \\ &+ c_1c_2\prod_{H_k\in D^+}P(H_k|+,v_{11})\prod_{H_k\in D^-}P(H_k|-,v_{11}) \end{aligned} \tag{4.15}$$

Our first assumption for decomposition is that F or disease prevalence is very small. Then, we can decompose $P(H_k|-,v_j)$ for all causal statuses j, as

$$P(H_k|-,v_j) = p_1^{H_k^1}(1-p_1)^{1-H_k^1}\times p_2^{H_k^2}(1-p_2)^{1-H_k^2} = P(H_k|+,v_{00}) \tag{4.16}$$

Then, we decompose $P(H_k|+,v_j)$ for different v_j, and first, let's consider v_{11} where two variants are both causal. We make another assumption here, which is the independence between rare variants; there is no linkage disequilibrium (LD) [LL08, PC02, Pri01]. If variants are independent, $P(H_{00}|+,v_{11})$ can be formulated as

$$\begin{aligned} P(H_{00}|+,v_{11}) &= \frac{P(H_{00})}{P(H_{00})+P(H_{10})\gamma_1+P(H_{01})\gamma_2+P(H_{11})\gamma_1\gamma_2} \\ &= \frac{(1-p_1)(1-p_2)}{(1-p_1)(1-p_2)+p_1(1-p_2)\gamma_1+(1-p_1)p_2\gamma_2+p_1p_2\gamma_1\gamma_2} \\ &= \frac{(1-p_1)\times(1-p_2)}{((1-p_1)+p_1\gamma 1)\times((1-p_2)+p_2\gamma_2)} \\ &= (1-p_1^+)(1-p_2^+) \end{aligned}$$

The last derivation comes from (Equation 4.6) where $p_i^+ = \frac{p_i\gamma_i}{(1-p_i)+p_i\gamma_i}$. Similarly, we can define the probabilities of other haplotypes (H_{01}, H_{10}, H_{11}) as

$$\begin{aligned}
P(H_{01}|+, v_{11}) &= (1 - p_1^+)p_2^+ \\
P(H_{10}|+, v_{11}) &= p_1^+(1 - p_2^+) \\
P(H_{11}|+, v_{11}) &= p_1^+ p_2^+
\end{aligned}$$

Combining these probabilities, we have the following decomposition of $P(H_k|+, v_{11})$.

$$P(H_k|+, v_{11}) = {p_1^+}^{H_k^1}(1 - p_1^+)^{1-H_k^1} \times {p_2^+}^{H_k^2}(1 - p_2^+)^{1-H_k^2} \quad (4.17)$$

Using the similar derivation, decomposition of $P(H_k|+, v_{01})$ and $P(H_k|+, v_{10})$ is

$$P(H_k|+, v_{01}) = {p_1}^{H_k^1}(1 - p_1)^{1-H_k^1} \times {p_2^+}^{H_k^2}(1 - p_2^+)^{1-H_k^2} \quad (4.18)$$

$$P(H_k|+, v_{10}) = {p_1^+}^{H_k^1}(1 - p_1^+)^{1-H_k^1} \times {p_2}^{H_k^2}(1 - p_2)^{1-H_k^2} \quad (4.19)$$

By the 4 decompositions (Equations 4.16, 4.17, 4.18, and 4.19), we can finally decompose the likelihoods of haplotypes (Equation 4.15) as

$$\begin{aligned}
L_0 =&(1-c_1)(1-c_2)\prod_{H_k\in D^+} p_1^{H_k^1}(1-p_1)^{1-H_k^1}p_2^{H_k^2}(1-p_2)^{1-H_k^2}\\
&\prod_{H_k\in D^-} p_1^{H_k^1}(1-p_1)^{1-H_k^1}p_2^{H_k^2}(1-p_2)^{1-H_k^2}\\
L_0+L_1 =&(1-c_1)(1-c_2)\prod_{H_k\in D^+} p_1^{H_k^1}(1-p_1)^{1-H_k^1}p_2^{H_k^2}(1-p_2)^{1-H_k^2}\\
&\prod_{H_k\in D^-} p_1^{H_k^1}(1-p_1)^{1-H_k^1}p_2^{H_k^2}(1-p_2)^{1-H_k^2}\\
&+(1-c_1)c_2\prod_{H_k\in D^+} {p_1}^{H_k^1}(1-p_1)^{1-H_k^1}{p_2^+}^{H_k^2}(1-p_2^+)^{1-H_k^2}\\
&\prod_{H_k\in D^-} p_1^{H_k^1}(1-p_1)^{1-H_k^1}p_2^{H_k^2}(1-p_2)^{1-H_k^2}\\
&+c_1(1-c_2)\prod_{H_k\in D^+} {p_1^+}^{H_k^1}(1-p_1^+)^{1-H_k^1}{p_2}^{H_k^2}(1-p_2)^{1-H_k^2}\\
&\prod_{H_k\in D^-} p_1^{H_k^1}(1-p_1)^{1-H_k^1}p_2^{H_k^2}(1-p_2)^{1-H_k^2}\\
&+c_1c_2\prod_{H_k\in D^+} {p_1^+}^{H_k^1}(1-p_1^+)^{1-H_k^1}{p_2^+}^{H_k^2}(1-p_2^+)^{1-H_k^2}\\
&\prod_{H_k\in D^-} p_1^{H_k^1}(1-p_1)^{1-H_k^1}p_2^{H_k^2}(1-p_2)^{1-H_k^2}
\end{aligned}$$

$$\begin{aligned}
L_0+L_1 = &\Bigg((1-c_1)\prod_{H_k\in D^+} p_1^{H_k^1}(1-p_1)^{1-H_k^1}\textstyle\prod_{H_k\in D^-} p_1^{H_k^1}(1-p_1)^{1-H_k^1}\\
&+c_1\prod_{H_k\in D^+} {p_1^+}^{H_k^1}(1-p_1^+)^{1-H_k^1}\textstyle\prod_{H_k\in D^-} p_1^{H_k^1}(1-p_1)^{1-H_k^1}\Bigg)\times\\
&\Bigg((1-c_2)\prod_{H_k\in D^+} p_2^{H_k^2}(1-p_2)^{1-H_k^2}\textstyle\prod_{H_k\in D^-} p_2^{H_k^2}(1-p_2)^{1-H_k^2}\\
&+c_2\prod_{H_k\in D^+} {p_2^+}^{H_k^2}(1-p_2^+)^{1-H_k^2}\textstyle\prod_{H_k\in D^-} p_2^{H_k^2}(1-p_2)^{1-H_k^2}\Bigg)
\end{aligned}$$

If we generalize the above equation to M variants, we have the likelihood of M variants as in (Equations 4.4 and 4.5)

4.2.8 Reformulation of L_0 and L_1 in LRT for an efficient permutation test

First, computation of L_0 (Equation 4.4) can be reformulated as

$$\begin{aligned}
L_0 &= \prod_{i=1}^{M} \left\{ (1-c_i) \prod_{H_k \in D^+} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \prod_{H_k \in D^-} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \right\} \\
&= \prod_{i=1}^{M} \left\{ (1-c_i) \prod_{H_k \in D^\pm} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \right\} \\
&= \prod_{i=1}^{M} \left\{ (1-c_i) p_i^{2Np_i} (1-p_i)^{2N-2Np_i} \right\} \triangleq \prod_{i=1}^{M} X_i
\end{aligned}$$

Similarly, we can reformulate L_1 as

$$\begin{aligned}
L_1 &= \prod_{i=1}^{M} \left\{ (1-c_i) \prod_{H_k \in D^+} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \prod_{H_k \in D^-} {p_i}^{H_k^i} (1-p_i)^{1-H_k^i} \right. \\
&+ \left. c_i \prod_{H_k \in D^+} {p_i^+}^{H_k^i} (1-p_i^+)^{1-H_k^i} \prod_{H_k \in D^-} {p_i^-}^{H_k^i} (1-p_i^-)^{1-H_k^i} \right\} \\
&= \prod_{i=1}^{M} \left\{ X_i + c_i {p_i^+}^{N\hat{p}_i^+} (1-p_i^+)^{N-N\hat{p}_i^+} {p_i^-}^{N\hat{p}_i^-} (1-p_i^-)^{N-N\hat{p}_i^-} \right\} \\
&= \prod_{i=1}^{M} \left\{ X_i + c_i (1-p_i^+)^N \left(\frac{p_i^+}{1-p_i^+} \right)^{N\hat{p}_i^+} (1-p_i^-)^N \left(\frac{p_i^-}{1-p_i^-} \right)^{N\hat{p}_i^-} \right\}
\end{aligned}$$

Using the fact that $N\hat{p}_i^+ + N\hat{p}_i^- = 2Np_i$ under permutations,

$$\begin{aligned}
L_1 &= \prod_{i=1}^{M} \left\{ X_i + c_i (1-p_i^+)^N (1-p_i^-)^N \left(\frac{p_i^+}{1-p_i^+} \right)^{N\hat{p}_i^+} \left(\frac{p_i^-}{1-p_i^-} \right)^{2Np_i - N\hat{p}_i^+} \right\} \\
&= \prod_{i=1}^{M} \left\{ X_i + c_i (1-p_i^+)^N (1-p_i^-)^N \left(\frac{p_i^-}{1-p_i^-} \right)^{2Np_i} \left(\frac{p_i^+}{1-p_i^+} \cdot \frac{1-p_i^-}{p_i^-} \right)^{N\hat{p}_i^+} \right\} \\
&\triangleq \prod_{i=1}^{M} \left\{ X_i + K_i Y_i^{N\hat{p}_i^+} \right\}
\end{aligned}$$

4.3 Results

4.3.1 Type I error rate of LRT

We examine the type I error rate of LRT by applying it to "null" datasets that contain no causal variants. We measure the type I error rates under three significance thresholds; 0.05, 0.01, and 2.5×10^{-6} (the significance threshold for the power simulation). A large number of null datasets are necessary to accurately estimate the type I error rate under the lowest significance threshold (2.5×10^{-6}). Thus, we create 10 million datasets, and each dataset contains 1000 case and 1000 control individuals with 100 variants. We estimate the type I error rate as the proportion of significant datasets among the 10 million datasets.

To efficiently measure the type I error rates of LRT, we use the following approach. We first test LRT on all 10 million datasets with 100,000 permutations. This small number of permutations makes it possible to test LRT on all null datasets and allows us to estimate the type I error rates under the 0.05 and 0.01 significance thresholds. As for the lowest significance threshold, we need to test LRT with a very large number of permutations (e.g. 100 million) to obtain a genome-wide level p-value. To reduce the amount of computation, we exclude datasets whose p-values cannot be lower than 2.5×10^{-6} with 100 million permutations. More specifically, to obtain a p-value less than 2.5×10^{-6}, the number of significant permutations (permutations whose LRT statistics are greater than the observed LRT statistic) must be less than 250 with 100 million permutations. We exclude datasets already having more than 250 significant permutations after the 100,000 permutations. We then apply the adaptive permutation test on the remaining datasets; we stop the permutation test when the number of significant permutations is greater than 250. The proportion of datasets whose permutation tests do not stop until 100 million permutations is the type I error rate under the

2.5×10^{-6} threshold.

We find that the type I error rates of LRT are 0.0500946, 0.0100042, and 2.6×10^{-6} for the significance thresholds of 0.05, 0.01, and 2.5×10^{-6}, respectively. This shows that the type I error rates are well controlled for LRT under the three different thresholds.

4.3.2 Power comparison between LRT and previous grouping methods

We compare power between LRT and previous methods using two simulations. We design these simulations to observe how LRT's implicit inference of which variants are causal affects the power compared to methods which do not make this kind of inference. In the first simulation, we generate datasets in which all variants have true $c_i = 0.1$. This means that only a subset of variants is causal, and causal statuses of variants vary per datasets. In the second simulation, all 100 variants in datasets are causal; true c_i of all variants is 1.

We test four different methods in this experiment; LRT, Optimal Weighted Aggregate Statistic (OWAS), MB, and VT. OWAS computes a difference in mutation counts between case and control individuals for each variant, or z-score of a variant, and assigns weights to z-scores according to the non-centrality parameters of z-scores [SHH11]. Sul *et al.* reported that OWAS achieves slightly higher power than RWAS [SHH11]. Thus, we test OWAS instead of their proposed method, RWAS, to compare power between a weighted sum of z-scores approach and the LRT approach. Since OWAS needs to know the effect sizes of variants, we give OWAS the true group PAR that generated data. OWAS divides the true group PAR by the expected number of causal variants to compute the marginal PAR (ω) and then computes relative risk of variants (Equation 4.10). We also apply our permutation test for LRT (see Material and Methods) to OWAS to estimate its p-value more accurately. To test VT, we use an

R package available online [PKB10]. LRT, OWAS, and VT are given prior information that is equivalent to true c_i of datasets, and we perform 10 million permutations to estimate p-values of their statistics.

Results of the two simulations show that LRT outperforms the previous groupwise tests in the first simulation, and it has almost the same power as OWAS in the second simulation. In the first simulation, LRT has higher power than other tests at all group PAR values (Figure 4.1A); at the group PAR of 5%, LRT achieves 94.5% power while OWAS and VT achieve 53.7% and 83.6% power, respectively. This shows that data may provide useful information about causal statuses of variants, and a method that takes advantage of data achieves higher power than those that do not. When prior information, however, can alone identify which variants are causal as in the second simulation, LRT and OWAS have almost the identical power (Figure 4.1B). This is because both methods know which variants are causal from prior information. Hence, this experiment demonstrates that LRT is generally a more powerful approach than the weighted sum of z-scores approach because it achieves higher power in studies where prior information cannot specify which variants are causal.

4.3.3 Comparison of decision boundaries between LRT and the weighted sum of z-scores

We show decision boundaries of LRT and the weighted sum of z-scores method to visualize the way each method combines information from multiple variants and to determine how decision boundaries affect power of studies. A decision boundary determines whether a group of variants is statistically associated with a disease; a statistic for the group is significant if it is above the boundary while it is not significant if it is below the boundary. Methods that combine information linearly has a linear decision boundary, and those methods include RWAS and MB that compute a statistic based on a linear sum of mutation counts or z-scores. LRT, on the other hand, has a nonlinear decision boundary since its statistic is computed using a nonlinear combination of mutation counts.

For this experiment, we perform two simulations similar to ones in the previous experiment with a fewer number of variants. In each simulation, we generate 10,000 datasets containing 500 case and 500 control individuals with only two variants. Population MAF of both variants is 1%, and the true case MAF is calculated assuming the group PAR of 2%. Both variants have true $c_i = 0.5$ in the first simulation while they have true $c_i = 1$ in the second simulation, which is similar to c_i of the two simulations in the previous experiment. We perform the single marker test to compute z-score of each variant in each dataset, meaning that we have two statistics per dataset. These statistics are represented in a two-dimensional graph where each dimension corresponds to z-score of each variant. We then test LRT and OWAS on each dataset to determine whether their statistics on a group of the two variants are significant using the significance threshold of 0.05.

Figure 4.2 shows results of the two simulations; Figures 4.2A and 4.2B are results of testing LRT and OWAS, respectively, on the first simulation ($c_i = 0.5$), and Figures 4.2C and 4.2D are results on the second simulation ($c_i = 1$). In each figure, a point represents one of the 10,000 datasets, and its x and y axes correspond to z-scores of the first and second variants, respectively. The red points are datasets whose LRT or OWAS statistics on a group of variants are significant, and the blue points are non-significant statistics.

We find that LRT achieves higher power by using the nonlinear decision boundary as there are more number of red points in Figure 4.2A (LRT) than Figure 4.2B (OWAS). A curved line separates significant and non-significant associations, which indicates the nonlinear decision boundary of LRT (Figure 4.2A). On the other hand, a straight line or a linear decision boundary segregates the statistics of OWAS (Figure 4.2B). The nonlinear decision boundary allows LRT to emphasize causal variants more strongly than non-casual variants while OWAS considers both casual and non-casual variants to be equally important.

When all variants in datasets are causal, however, they all should be emphasized equally, and hence a linear decision boundary would best detect associations. Hence, the decision boundary of LRT becomes linear in the second simulation (Figure 4.2C) since LRT knows every variant is causal and equally important. The decision boundary of OWAS is also linear in this simulation (Figure 4.2D), and this explains why LRT and OWAS have the same power in the second simulation of the previous experiment. This experiment shows that because the decision boundary of LRT can become both linear and nonlinear depending on the causal statuses of variants, LRT is more powerful than previous methods that have a fixed decision boundary.

4.3.4 LRT on real mutation screening data of ATM

We apply LRT to real mutation screening data of the susceptibility gene for ataxia telangiectasia [TOB09]. This gene, called *ATM*, is also an intermediate-risk susceptibility gene for breast cancer. Tavtigian *et al.* conducted mutation screening studies and collected data from 987 breast cancer cases and 1021 controls. Tavtigian *et al.* increased the number of cases and controls to 2531 and 2245, respectively, by collecting data from seven published *ATM* case-control mutation screening studies. This dataset is called "bona fide case-control studies," and 170 rare missense variants are present in this dataset. Sul *et al.* also analyzed the dataset with RWAS [SHH11].

To obtain prior information of variants in the dataset, Tavtigian *el al.* used two missense analysis programs, Align-GVGD [TDY06] and SIFT [NH03]. A difference between the two programs is that while SIFT classifies a variant as either deleterious

(SIFT scores ≤ 0.05) or neutral (SIFT scores > 0.05), Align-GVGD classifies a variant into seven grades from C0 (most likely neutral) to C65 (most likely deleterious). To convert the seven grades of Align-GVGD into c_i values, we arbitrarily assign c_i values from 0.05 to 0.95 in increments of 0.15 to the seven grades. As for converting SIFT scores into c_i values, we assign c_i value of 1 to variants whose SIFT scores are ≤ 0.05 and c_i of 0 to other variants. This is the same conversion used in [SHH11].

When LRT uses prior information from Align-GVGD, it yields a p-value of 0.0058, which indicates a significant association between the group of rare variants and the disease. This result is consistent with the findings of [TOB09] and [SHH11]; Tavtigian *et al.* and Sul *et al.* both obtained significant p-values when they used outputs of Align-GVGD as prior information. The result shows that we can apply LRT to real data to discover an association.

LRT yields a non-significant p-value of 0.39341 when it does not use prior information, and this is also consistent with results of [TOB09] and [SHH11]; Tavtigian *et al.* and Sul *et al.* reported non-significant p-values when they analyzed the data without prior information. When SIFT scores are used as prior information, LRT similarly reports a non-significant p-value of 0.08384, and Sul *et al.* also obtained a non-significant p-value [SHH11]. However, the analysis of Tavtigian *et al.* with SIFT scores showed a significant association [TOB09]. According to [SHH11], the reason for this difference may be that LRT and RWAS need to know the relative degree of how deleterious a variant is to better detect an association. However, it may be difficult to know this relative deleteriousness of variants with SIFT scores because variants are either deleterious or neutral. Thus, this experiment shows that more informative prior information such as the seven grades of Align-GVGD may yield better results with LRT.

4.4 Discussion

We developed a likelihood ratio test (LRT) to increase power of association studies on a group of rare variants. The power of statistical methods that group rare variants depends on which rare variants to group or to exclude from the group because including non-causal variants in the group decreases power [SHH11]. Although prior information of variants from bioinformatics tools provides information of how likely each variant is functional or deleterious, determining whether a variant is causal or not only from prior information is often infeasible. LRT takes advantage of data to identify causal variants, and when it is not possible to identify causal variants from prior information, we showed that LRT outperforms previous methods.

We then showed decision boundaries of LRT and one of previous grouping methods, Optimal Weighted Aggregate Statistic (OWAS). The two methods have the same linear decision boundary when datasets contain only causal variants, and thus they achieve the same power. When only a subset of them is causal, OWAS still has a linear decision boundary since its statistic is computed as a linear sum of differences in mutation counts. However, the decision boundary of LRT becomes nonlinear in this case because LRT places more emphasis on a variant that causes a large difference in mutation counts between cases and controls than a variant that causes a small difference. We showed by simulations that the nonlinear decision boundary detects more associations than the linear decision boundary. Hence, this suggests that LRT is a more powerful approach in finding associations with a group of rare variants because it is capable of changing its decision boundary depending on causal statuses of variants to better detect associations.

To evaluate LRT on real data, we used mutation screening data of the *ATM* gene [TOB09]. Tavtigian *et al.* and Sul *et al.* both found the significant association in the data [TOB09, SHH11], and we showed that LRT also detected the association using the output of Align-GVGD as prior information of variants. This shows that LRT can be applied to detect an association in real association studies.

One of the two assumptions that we made to efficiently compute the LRT statistic and its p-value is the independence between variants. Several studies suggest that there would be very low linkage disequilibrium between rare variants due to their low occurrences [LL08, PC02, Pri01]. However, if non-negligible LD is expected between variants, especially when common variants are in linkage disequilibrium in the group, we can change our permutation test as follows to take into account LD and to correctly control the false positive rate. Instead of separately sampling $N\hat{p}_i^+$ of each common variant from the normal distribution, we sample $N\hat{p}_i^+$ of all common variants from the multivariate normal distribution (MVN). This approach is similar to the approach of Han *et al* who used the MVN framework to correct for multiple testing on correlated markers [HKE09a]. The covariance matrix of the MVN we create consists of correlations (r) between common variants, and hence $N\hat{p}_i^+$ sampled from this MVN takes into account LD between variants. For rare variants, we use our proposed method that samples $N\hat{p}_i^+$ of each rare variant from the hypergeometric distribution because LD between rare variants is expected to be very low.

The other assumption of our method is the low disease prevalence, and this assumption does not influence the false positive rate of our method while it may affect the power. The false positive rate of LRT is controlled even though the disease we consider is highly prevalent because we perform the permutation test. Therefore, LRT can still be applied to association studies involving diseases with high prevalence while its power may not be as high as the power it achieves on diseases with low prevalence.

Reference to published article

Jae Hoon Sul, Buhm Han, and Eleazar Eskin, “Increasing Power of Groupwise Association Test with Likelihood Ratio Test.” *Journal of Computational Biology*. 18, 1611-1624, 2011.

Jae Hoon Sul, Buhm Han, and Eleazar Eskin, “Increasing Power of Groupwise Association Test with Likelihood Ratio Test.” *In Proceedings ofthe Fifteenth Annual Conference on Research in Computational Biology (RECOMB-2011).* Vancouver, Canada: March 28-31, 2011

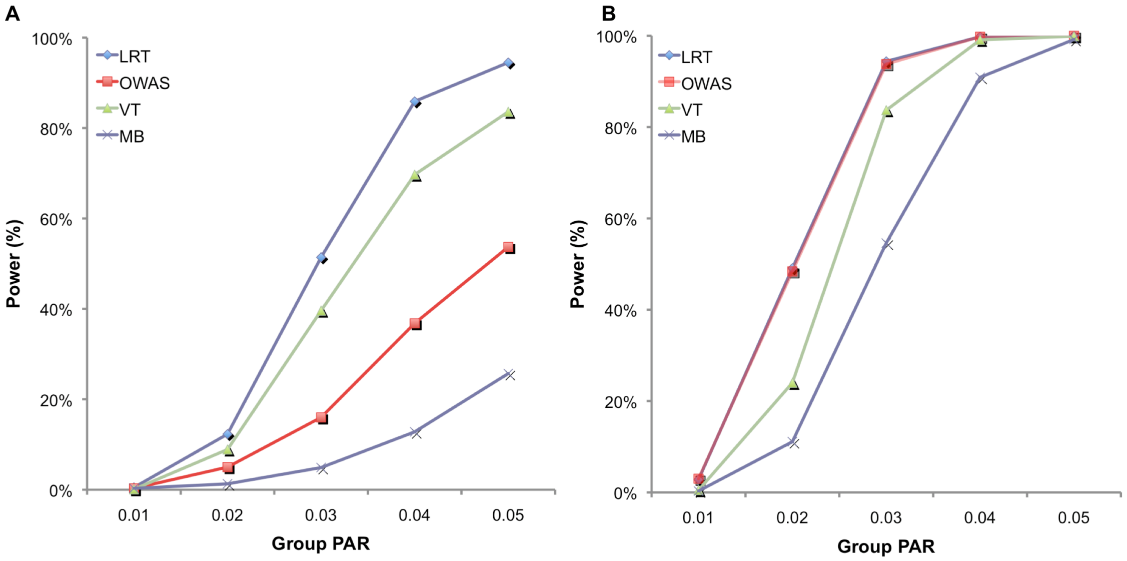


Figure 4.1: Power comparison among four different groupwise association tests on datasets where $c_i = 0.1$ for all variants (A) and $c_i = 1$ (B) over different group PAR values

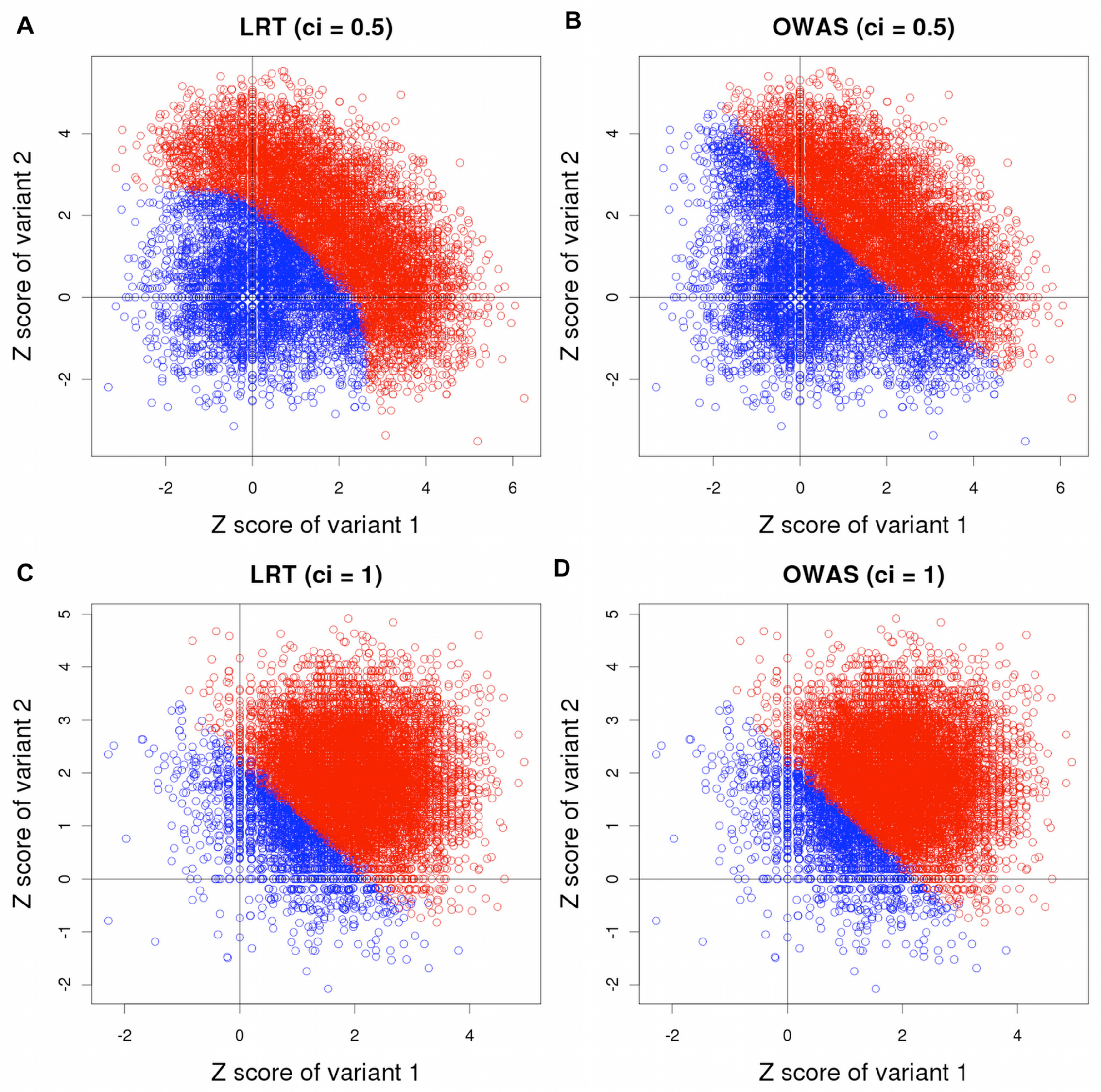


Figure 4.2: Plots showing decision boundaries of LRT and OWAS on datasets with two variants whose $c_i = 0.5$ (A and B) and $c_i = 1$ (C and D). X-axis and Y-axis correspond to z-scores of two variants, and red points are significant statistics according to LRT or OWAS while blue points are non-significant.

CHAPTER 5

Combining mixed model and meta-analysis to detect eQTLs from multiple tissues

5.1 Background

Advances in genotyping and gene expression technologies have enabled researchers to study associations between genetic variants and gene expression levels. These studies often treat expression levels as quantitative traits and apply statistical tests to identify genomic locations known as expression Quantitative Trait Loci (eQTLs) that segregate the traits. Genome-wide maps of eQTLs for several organisms including budding yeast [BYC02, BK05], Arabidopsis [KFT07], mouse [CLS05, BWD05] and human [CSE05, SNF07] have been successfully generated. Furthermore, recent technological developments and cost decreases in microarrays allow studies to collect expression data in more than one tissue in human [CSE05, ETZ08, SBB07] and mouse [CLS05, BWD05]. A collection of expression data from multiple tissues enables studies to explore the tissue-specific nature of eQTLs as well as their global effects on different types of tissues.

Multiple tissue datasets can potentially allow studies to more effectively identify eQTLs by combining information from multiple tissues. Due to a limited sample size, a standard single tissue eQTL method or "tissue-by-tissue" approach that examines each tissue individually may not detect an eQTL in any one tissue, or it may overestimate the proportion of tissue specific eQTLs [FWD12]. However, if a genetic variant is associated with the expression of a gene in more than one tissue, we can aggregate information from multiple tissues to increase statistical power. This idea is similar to the idea of meta-analysis in genome-wide association studies (GWAS) that combines results of several studies on the same phenotype. In our case, each tissue is considered as a separate "study" in the meta-analysis.

One key difficulty in combining results from multiple tissues is that it is not known in which tissues a genetic variant has an effect. For example, a variant may influence gene expression in all tissues, may have different effects on different tissues, or may have an effect in some tissues but may not have any effect in other tissues. This phenomenon, different effect sizes among tissues, is called heterogeneity. Meta-analysis methods have different assumptions on the distribution of effect sizes, and to better detect eQTLs, studies will perform best if they apply a meta-analysis method whose assumptions are consistent with the actual effect sizes of eQTLs in multiple tissues. For instance, if an eQTL has an effect in all tissues, studies would perform best if they utilize the fixed-effects model (FE) [BFJ08, COC54, MH59] that assumes no heterogeneity. On the other hand, to effectively detect an eQTL whose effects on gene expression differ across tissues, studies will perform best if they apply the random-effects model (RE) [DL86, IPE07a, IPE07b, EMI07, HE11] that considers heterogeneity.

Another challenge in applying meta-analysis to multi-tissue datasets is that studies often collect multiple tissues from the same individuals, which may cause the expression between tissues of the same individual to be correlated. This correlation may cause false positives for standard meta-analysis methods which assume a disjoint set of individuals in each study.

In this chapter, we present a novel approach called "Meta-Tissue" that identifies eQTLs from multiple tissues by utilizing meta-analysis. The critical advance of our methodology is that we extend meta-analysis to a mixed model framework. We apply the mixed model to account for the correlation of expression between tissues, and perform meta-analysis to combine results from multiple tissues. Since we do not know in advance the distribution of effect sizes for eQTLs among different tissues, we utilize both the FE and RE models to identify as many eQTLs as possible, and for RE, we use a recently developed random-effects model [HE11] that achieves higher statistical power than the traditional random-effects model. We first show by simulations that Meta-Tissue is more powerful than the tissue-by-tissue approach in detecting eQTLs when eQTLs have effects in multiple tissues, while controlling for the false positive rate correctly.

We then apply Meta-Tissue to a mouse expression dataset. This dataset is ideal for evaluating methods for discovering eQTLs for several reasons. The data are generated through a cross which limits the genetic diversity in the dataset, and all variants have similar frequencies which eliminate effects of allele frequency on power. In addition, the dataset contains gene expression from many different tissues and different numbers of individuals for the tissues, allowing us to compare results between different scenarios. We analyze four tissues from 50 samples per each tissue and ten tissues from 22 samples. We apply Meta-Tissue to both datasets and demonstrate that Meta-Tissue detects many eQTLs that are undetected by the tissue-by-tissue method.

In addition to accurately detecting eQTLs from multiple tissues, our method can also predict whether an eQTL affects or does not affect expression in a specific tissue. Predicting the existence or absence of an effect is a very difficult problem in meta-analysis, and it is known that making predictions based on p-values is not effective [HE12]. One of the reasons is that a non-significant p-value is not necessarily evidence of an absence of an effect since the study may be underpowered. Our method instead computes the posterior probability of the presence or absence of an effect for each study building on recent work in interpretation of meta-analysis [HE12]. Applying the framework to the four and ten tissue datasets, we identify more eQTLs that are predicted to have effects in all tissues compared to the p-value based approach, which are interesting potential candidates with possible global regulatory mechanisms. Meta-Tissue is publicly available at `http://genetics.cs.ucla.edu/metatissue/`.

5.2 Methods

5.2.1 Mouse Strains

F1N2 mice from a C57BL6/N x 129/OlaHsd cross were produced as follows. Male ES cell chimeric founders (E14 ES line [HHH87]) were crossed to C57BL6/N females (Harlan Laboratories). Male agouti offspring were backcrossed to C57BL6/N females, and F1N1 offspring were intercrossed to produce F1N2 animals, Figure 5.1. All animals were maintained in ventilated microisolator caging (Allentown), fed a standard lab chow diet (Harlan Teklad) and provided water ad libidem. F1N2 animals were group housed with littermates until 9 weeks of age. Mice selected for tissue harvest were singly housed for one additional week, to minimize socialization effects. Only males were used, to avoid estrus related effects on gene expression. While the production crosses segregated various gene targeted alleles, all mice selected for this study carried only wild type genomes and did not carry any engineered genomic alterations such as gene knockouts.

5.2.2 Gene Expression

Animals were sacrificed by cervical dislocation and immediately dissected. A set of thirty tissues were collected from each animal in a prescribed order, beginning with the pancreas. Each tissue was briefly rinsed in PBS and deposited in RNAlater (Ambion), held at room temperature to allow diffusion of RNAlater into the tissue, and then stored at -86C.

Tissue homogenization, total RNA isolation, cDNA production, in vitro transcription and fluorescent labeling were performed as per Affymetrix gene chip recommended protocols. The hybridization mixes were analyzed using Affymetrix U74Av2 expression microarrays, washed and scanned using Affymetrix instrumentation and protocols.

We consider the 10588 probes for which we have annotations. For each tissue type, we filter out array outliers which show an average correlation of < 0.98 with respect to all other arrays.

The mice were genotyped at 140 SNPs that are polymorphic between 129S1/SvImJ and C57BL/6J from the JAX SNP Genotyping Panel [PDC04]. We use 135 out of the 140 SNPs that are polymorphic in all tissues for our analysis.

5.2.3 Normalization and selection of individuals

In our analysis, we consider the gene expression levels of $G = 10588$ probes collected in 4 tissues (liver, spleen, cortex and heart) over $N = 50$ individuals. To be consistent with the different tissue datasets we analyze, we randomly chose 50 individuals from those datasets that have more than 50 individuals. We first used RMA to perform background adjustment on the raw expressions and then quantile normalization to normalize the adjusted expressions. For 10 tissues, we collect the same number of gene expression levels over $N = 22$ individuals.

5.2.4 Power simulation framework

Our power simulation assumes that we collect four tissues from 100 individuals, and considers four scenarios where an eQTL has an effect in (1) one tissue, (2) in two tissues, (3) in three tissues, and (4) in all four tissues. To generate the gene expression level of individuals that considers the repeated measurements from the same individuals, we first sample gene expression from the multivariate normal distribution:

$$\mathbf{y}^e = \mathbf{e} \tag{5.1}$$

where $\mathbf{y}^e$ is a vector of size 400 corresponding to gene expression of 100 individuals in 4 tissues, and $\mathbf{e} \sim \mathcal{N}(0, \sigma_v^2\mathbf{D} + \sigma_e^2\mathbf{I})$ where $\mathbf{D}$ is a 400 by 400 matrix representing correlation between individuals across the tissues. More specifically, $\mathbf{D}_{ij} = 1$ if i and j are the same individual between two tissues, and $\mathbf{D}_{ij} = 0$ otherwise. $\mathbf{I}$ is an identity matrix with size of 400. σ_v^2 and σ_e^2 are coefficients of the two variance components, and we use the real mouse dataset to obtain realistic values of the two coefficients. We estimate σ_v^2 and σ_e^2 for every pair between a gene expression and a SNP, and find that on average, $\sigma_v^2 = 0.0988$ and $\sigma_e^2 = 0.9039$. We use these values for our simulation.

After sampling $\mathbf{y}^e$, we add a SNP effect to $\mathbf{y}^e$ for tissues in which an effect exists using the following equation:

$$\mathbf{y}_t = \mathbf{x}\beta_t + \mathbf{y}_t^e$$

where $\mathbf{y}_t$ is gene expression of 100 individuals in tissue t ($t \in \{1, 2, 3, 4\}$), $\mathbf{y}_t^e$ is $\mathbf{y}^e$ on tissue t (size of 100), and $\mathbf{x}$ is SNP information of 100 individuals. $\beta_t = 0$ if an eQTL does not have an effect in tissue t, and $\beta_t > 0$ if an eQTL has an effect. Since the goal is to compare the relative power between methods, we vary the effect size (β_t) depending on the scenario to avoid too high or too low power. Specifically, we set $\beta_t = 1.5, 1.175, 1.0, 0.75$ for the scenarios (1), (2), (3), (4), respectively.

5.2.5 Linear model for Tissue-by-Tissue approach

We assume an additive linear model to represent the relationship between the expression of one gene and one SNP. We can write that relationship in the following way for an arbitrary gene g and SNP j at tissue t:

$$\mathbf{y}_t^g = 1\alpha_t + \mathbf{x}_j\beta_t + \mathbf{e}, \tag{5.2}$$

where $\mathbf{y}$ is a size N vector denoting gene expression levels of N individuals, $\mathbf{x}_j$ is a size N vector denoting SNP, 1 is a vector of ones, and $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$. To assess the significance of an association between a SNP and a gene, we perform a standard *F*-test for the null hypothesis $\beta_t = 0$ and also obtain an estimate of β_t using the `lm` function in `R`. In the tissue-by-tissue approach, if any single tissue turns out to be significant ($\beta_t \neq 0$), the pair of SNP and gene expression are reported as a significant eQTL. TBT can also find tissues in which an eQTL exists by examining which β_t is non-zero.

5.2.6 Meta-Tissue - Linear mixed model

We use a linear mixed model to take into account the fact that eQTL studies collect multiple tissues from the same individuals. This is called a "repeated measures design," and the mixed model is often used to model the correlation induced by the repeated measurements such as in longitudinal data. Let T be the number of tissues, and for simplicity, we assume there are N individuals for each tissue, but individuals collected in one tissue do not necessarily completely overlap with those in another tissue; it is possible that some individuals may provide all tissues while others may provide a subset. We also assume that we have SNP information for all individuals. We apply the following linear mixed model to assess the statistical significance between gene expression g and SNP j:

$$\mathbf{Y}^g = \mathbf{1}\boldsymbol{\alpha} + \mathbf{X}_j\boldsymbol{\beta} + \mathbf{u} + \mathbf{e}, \tag{5.3}$$

Here is a description of each variable in above equation. Let $NT = N \times T$.

– $\mathbf{Y}$ is an $NT \times 1$ matrix denoting expression levels of N individuals in T tissues. In other words, the first N rows are expression of N individuals in the first tissue, the next N are expression in the second tissue, and so on. Expression values of each tissue are normalized to $\mathcal{N}(0, 1)$.

– $\mathbf{1}$ is an $NT \times T$ matrix denoting the intercepts for T tissues. The first column of $\mathbf{1}$ denotes the intercept for the first tissue; the first N rows are ones, and the next $NT - N$ are zeros. In the second column that denotes the intercept for the second tissue, the first N rows are zeros, the next N rows are ones, and the next $NT - 2T$ rows are zeros.

– $\boldsymbol{\alpha}$ is a $T \times 1$ matrix denoting coefficients of intercepts.

– $\mathbf{X}_j$ is an $NT \times T$ matrix denoting SNP for T tissues. This is similar to the $\mathbf{1}$ matrix, and we replace ones in the $\mathbf{1}$ matrix with SNP information. For example, in the first column, the first N rows are SNP information of N individuals in the first tissue, and the next $NT - N$ rows are zeros.

– $\boldsymbol{\beta}$ is a $T \times 1$ matrix denoting coefficients of SNP effects in T tissues.

– $\mathbf{u}$ is the random effect of the mixed model due to the repeated measurements of individuals, and $\mathbf{u} \sim \mathcal{N}(0, \sigma_v^2\mathbf{D})$ where $\mathbf{D}$ is an $NT \times NT$ matrix representing how individuals are shared across the tissues (discussed in the Power simulation framework section). $\mathbf{e}$ represents random errors and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2\mathbf{I})$ where $\mathbf{I}$ is an identity matrix. To efficiently estimate the two variance components (σ_v^2 and σ_e^2), we use the efficient mixed-model association (EMMA) package [KZW08].

To estimate $\boldsymbol{\beta}$ and its covariance, we apply the generalized least squares. Let $\Sigma = \hat{\sigma}_v^2\mathbf{D} + \hat{\sigma}_e^2\mathbf{I}$. Then, the estimated $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_j'\Sigma^{-1}\mathbf{X}_j\right)^{-1}\mathbf{X}_j'\Sigma^{-1}\mathbf{Y}^m \tag{5.4}$$

5.2.7 Meta-Tissue - Meta-analysis

Given the estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_T)$, we combine information from multiple tissues by applying meta-analysis to $\hat{\boldsymbol{\beta}}$. If the effect of eQTL is the same for all tissues, applying fixed effects model (FE) meta-analysis will be a powerful approach. If the effects of eQTL differs by tissues, applying random effects model (RE) meta-analysis will be a powerful approach [HE11].

5.2.7.1 Fixed effects model

Fixed effects model (FE) is a meta-analysis method that assumes the effect size of a variant is fixed across datasets [COC54, MH59], and its statistic is computed based on the inverse-variance-weighted effect size [Fle93]. Let $B_1, \ldots, B_T$ and $V_1, \ldots, V_T$ be the estimates of effect-size and the standard error of B_i, respectively, in T tissues. Let μ be the unknown true effect size. The null hypothesis of FE is $\mu = 0$; in other words, effect size in all tissues is zero. A statistic of FE (S_{FE}) and its distribution under the null hypothesis are

$$S_{FE} = \frac{\sum_{i=1}^{T} V_i^{-1} B_i}{\sqrt{\sum_{i=1}^{T} V_i^{-1}}} \sim \mathcal{N}(0,1) \tag{5.5}$$

A p-value of S_{FE} is obtained from the standard normal distribution.

5.2.7.2 Random effects model

Our Meta-Tissue method leverages new random effects model (RE) [HE11] to detect eQTLs from multiple tissues while taking into account heterogeneity of effect sizes in different tissues. The assumption of the random effects model is that the effect size of a variant is different among datasets and follows a probability distribution with mean μ and variance τ^2. The null hypothesis of the random effects model is equivalent to that of the fixed effects model; that is, $\mu = 0$. The traditional random effects model, however, assumes a conservative null hypothesis model. The new random effects model corrects this conservative null hypothesis model and outperforms the traditional random effects model. More specifically, a statistic of RE (S_{RE}) is defined as

$$S_{RE} = \sum \log \left(\frac{V_i}{V_i + \hat{\tau}^2} \right) + \sum \frac{B_i^2}{V_i} - \sum \frac{\left(B_i - \hat{\mu}\right)^2}{V_i + \hat{\tau}^2} \tag{5.6}$$

where $\hat{\mu}$ and $\hat{\tau}^2$ are estimated mean and variance of the effect size, and the maximum likelihood estimates of the two parameters are calculated iteratively as following

$$\hat{\mu}_{(n+1)} = \frac{\sum \left(V_i + \hat{\tau}^2_{(n)}\right)^{-1} B_i}{\sum \left(V_i + \hat{\tau}^2_{(n)}\right)^{-1}} \qquad \hat{\tau}^2_{(n+1)} = \frac{\sum \frac{\left(B_i - \hat{\mu}_{(n+1)}\right)^2 - V_i}{\left(V_i + \hat{\tau}^2_{(n)}\right)^2}}{\sum \left(V_i + \hat{\tau}^2_{(n)}\right)^{-2}}$$

The initial value of $\hat{\tau}^2$ is estimated using approaches in the traditional random effects model [DL86, HT02, HT96]. We obtain a p-value of S_{RE} from p-value tables that are constructed from numerous null statistics.

5.2.7.3 Accounting for covariance of effect size estimates

Since we use linear mixed model to account for the fact that multi-tissue eQTL studies often collect multiple tissues from the same individuals, our estimates of effect size, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_T)$ in Equation (5.4) can become correlated. The covariance structure is estimated using the standard formula of the generalized least squares,

$$\text{var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}_j' \Sigma^{-1} \mathbf{X}_j\right)^{-1} \tag{5.7}$$

It is important that the meta-analysis methods account for this covariance structure of effect size estimates.

To take into account the covariance structure in meta-analysis, we use an extension [HSE13] of the Lin and Sullivan approach [LS09] . Given $\hat{\boldsymbol{\beta}}$ and their covariance $\Omega = \text{var}(\hat{\boldsymbol{\beta}})$, the optimal fixed effects model meta-analysis statistic is

$$S_{Lin} = \frac{\boldsymbol{e}^T \Omega^{-1} \hat{\boldsymbol{\beta}}}{\boldsymbol{e}^T \Omega^{-1} \boldsymbol{e}}$$

where $\boldsymbol{e}$ is the vector of ones ($\boldsymbol{e} = (1, ..., 1)$). The variance of the statistic is given

$$\text{var}(S_{Lin}) = \frac{1}{\boldsymbol{e}^T \Omega^{-1} \boldsymbol{e}}$$

Note that if $\hat{\boldsymbol{\beta}}$ is independent (Ω is a diagonal matrix), S_{Lin} and $\text{var}(S_{Lin})$ are equivalent to the inverse-variance weighted effect size estimate (the numerator of equation (5.5)) and its variance.

It can be shown that this approach is equivalent to building a new "un-correlated" variance of $\hat{\boldsymbol{\beta}}$,

$$\text{var}_{new}(\hat{\boldsymbol{\beta}}) = Diag(\Omega^{-1}\boldsymbol{e})^{-1}$$

and then giving $\hat{\boldsymbol{\beta}}$ and $\text{var}_{new}(\hat{\boldsymbol{\beta}})$ as input to the traditional meta-analysis approaches assuming independent estimates [HSE13]. This "un-correlating" idea allows us flexibility to use the correlated estimates in any meta-analysis framework requiring independent estimates. We use $\hat{\boldsymbol{\beta}}$ and its "un-correlated" variance for the fixed effects model (which gives equivalent results to the Lin and Sullivan approach [LS09]), random effects model, heterogeneity estimation (Q and I^2), and the m-value estimation [HE12].

5.2.7.4 Predicting effects of eQTLs in multiple tissues

To predict whether an eQTL has effects in a specific tissue, Meta-Tissue computes a statistic called the "m-value" proposed by Han and Eskin [HE12] that specifies the posterior probability that an effect exists in a tissue. First, we denote B as a vector of B_i; $B = \{B_1, B_2, \ldots, B_T\}$. Let R_i be a random variable whose value is 1 if dataset i has an effect and 0 otherwise. We also denote R as a vector of R_i, and since each R_i has two values, R has 2^T possible values. Let r_j be one of those 2^T values, and let $U = \{r_1, \ldots, r_{2^T}\}$ denote a vector of r_j. To estimate the m-value m_i, we need to compute the probability, $P(R_i = 1|B)$, which is the probability of dataset i having effects given the observed effect sizes. We can compute this probability using the Bayes' theorem

$$m_i = P(R_i = 1|B) = \frac{\sum_{r \in U_i} P(B|R = r)P(R = r)}{\sum_{r \in U} P(B|R = r)P(R = r)}$$

where U_i is a set of r_j in which ith value is 1. The equation shows that we need to compute $P(B|R = r)$ and $P(R = r)$ terms for every r to compute m_i. We can compute $P(R = r)$ as

$$P(R = r) = \frac{Beta(|r| + \gamma, T - |r| + \delta)}{Beta(\gamma, \delta)}$$

where $|r|$ denotes the number of 1's in r and $Beta$ denotes the beta function. γ and δ are set to one [HE12]. The probability of B given r, $P(B|R = r)$, is computed as

$$P(B|R = r) = \bar{D} \cdot N(\bar{B}; 0, \bar{V} + \sigma^2) \prod_{i \in q_0} N(B_i; 0, V_i)$$

where

$$\bar{B} = \frac{\sum_{i \in q_1} W_i B_i}{\sum_{i \in q_1} W_i} \quad \text{and} \quad \bar{V} = \frac{1}{\sum_{i \in q_1} W_i}$$

$N(B; a, b)$ denotes the probability density function of the normal distribution with mean equal to a and variance equal to b, and q_0 and q_1 denote the indices of 0 and 1 in r, respectively. $W_i = V_i^{-1}$ is the inverse variance, and $N(0, \sigma^2)$ is the prior for the effect size; $\sigma = 0.2$ when an effect is small while $\sigma = 0.4$ when an effect is large for binary traits [SB09, MHM07]. For quantitative traits, there is no general guidelines for the normally distributed priors, so we choose to use the default value $\sigma = 0.2$. $\bar{D}$ is a scaling factor defined as

$$\bar{D} = \frac{1}{(\sqrt{2\pi})^{T-1}} \sqrt{\frac{\prod_i W_i}{\sum_i W_i}} \cdot \text{exp} \left\{ -\frac{1}{2} \left(\sum_i W_i B_i^2 - \frac{(\sum_i W_i B_i)^2}{\sum_i W_i} \right) \right\}$$

More detailed derivations of $P(B|R = r)$ and $P(R = r)$ terms are discussed in Han and Eskin [HE12].

5.2.8 Practical issues in combining mixed model and meta-analysis

There are subtle issues in our framework combining mixed model and meta-analysis. First, the effect size estimates from linear model or mixed model are typically t-distributed, while most of meta-analysis methods assume normally distributed effect sizes. Second, our approach simultaneously considers all tissues using Equation (5.3), but the error model is slightly different from the tissue-by-tissue approach in Equation (5.2). In the tissue-by-tissue approach, the error $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$ is fit in each tissue separately, while in our new approach, the error is fit in all tissues together, which is often less powerful than the former. We correct for these subtle differences using simple heuristics (see following sections).

5.2.8.1 t-distributed effect size estimates

There are subtle issues in our framework combining mixed model and meta-analysis. First, the effect size estimates from linear model or mixed model are typically t-distributed, while most of meta-analysis methods assume normally distributed effect sizes. Let $\hat{\beta}$ and $\text{var}(\hat{\beta})$ be the effect size estimate and the variance estimate from a linear model. Assume that under the null, $\frac{\hat{\beta}}{\sqrt{\mathrm{var}(\hat{\beta})}}$ will approximately follow t-distribution with k degree of freedom. The p-value is calculated

$$p_t = 2\left(1 - \Phi_{t(k)}\left(\frac{|\hat{\beta}|}{\sqrt{\mathrm{var}(\hat{\beta})}}\right)\right)$$

where $\Phi_{t(k)}$ is the cummulative density function of the t-distribution with k degree of freedom. If we directly use $\hat{\beta}$ and $\mathrm{var}(\hat{\beta})$ in the meta-analysis approach assuming normally distributed effect size, false positive rate will increase. This issue is particularly important in model organisms where the sample size is moderate.

To correct for this, we use simple heuristic replacing $\sqrt{\mathrm{var}(\hat{\beta})}$ with

$$\frac{|\hat{\beta}|}{|\Phi^{-1}(p_t/2)|}$$

where Φ^{-1} is the inverse of the cummulative density function of the standard normal distribution. That is, we increase the variance of $\hat{\beta}$ according to the difference between the t-distribution and the normal distribution to prevent an excessive false positive rate in the meta-analysis.

5.2.8.2 Differences in error models

Another issue is that our approach simultaneously considers all tissues using Equation (3), but the error model is slightly different from the tissue-by-tissue approach in Equation (2). In the tissue-by-tissue approach, the error $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$ is fit in each tissue separately, while in our new approach, the error is fit in all tissues together. Certainly, the tissue-by-tissue model is more desirable because we cannot always expect that the true variance of error term (σ^2) to be the same across tissues. In other words, in our new framework, we are imposing an unrealistic assumption that the error variance is the same for all tissues, or constant error variance assumption (CEVA). We find that our approach is often less powerful when the truth deviates from CEVA. To compensate for the effect of this assumption, we apply the following idea. Before using our mixed model in Equation (3), we standardize the gene expressions in each tissue to follow $\mathcal{N}(0, 1)$. Note that this does not completely solve the problem because gene expression values include not only the error term but also the genetic effects.

To further correct for the effect of our assumption, we use the following heuristic. We first run tissue-by-tissue approach to obtain the effect size estimate $\hat{\beta}_{TBT}$ and its standard error STD_{TBT}. Second, we run our mixed model in Equation (3) assuming that $\sigma_v^2 = 0$. That is, we intentionally ignore the correlations of multiple tissue expressions from the same individuals. Under this simplified model, the estimate $\hat{\beta}_{COMB}$ turns out to be exactly the same as $\hat{\beta}_{TBT}$. Let STD_{COMB} be the standard error of $\hat{\beta}_{COMB}$ under this model. Although the effect size estimates are the same ($\hat{\beta}_{TBT} = \hat{\beta}_{COMB}$), their standard errors are different in two models because their error models are different. Therefore, the ratio between the two standard errors can be a measure of the effect of CEVA.

Finally, we run our standard mixed model by estimating σ_v^2 and σ_e^2 using the EMMA package. Let $\hat{\beta}_{MIX}$ and STD_{MIX} be the effect size estimate and its standard error under this model. Then we heuristically obtain a new standard error

$$STD_{NEW} = STD_{MIX} \cdot \frac{STD_{TBT}}{STD_{COMB}}$$

That is, we correct for the effect of CEVA using the ratio between STD_{TBT} and STD_{COMB}. What we use in the subsequent meta-analysis are $\hat{\beta}_{MIX}$ and STD_{NEW}. We find that this simple heuristic effectively corrects for the effect of CEVA in many cases.

5.3 Results

5.3.1 Meta-Tissue

The main idea of Meta-Tissue is that it combines the effect size estimates from multiple tissues using a “meta-analysis” approach. Meta-analysis techniques are widely applied to combine the results of GWAS studies. In our case, we consider each tissues as a “study.” This has the advantage of increasing the statistical power to detect eQTLs shared across tissues. There are several challenges corresponding to the inherent differences between combining GWAS studies and expression quantitative trait loci studies in multiple tissues. The first challenge is that we expect that there may be differences in effect sizes between tissues. For this reason, we utilize both the random-effects model which allows Meta-Tissue to detect eQTLs when heterogeneity is present, and the fixed-effects model when it is not. A second challenge is that in many multi-tissue eQTL study designs, multiple tissues are collected from the same individuals which induce correlation between measurements of expression levels in different tissues. However, meta-analysis methods assume that studies are independent and may be susceptible to false positives. To overcome this challenge, we utilize the linear mixed model to correct our effect size estimates before performing the meta-analysis.

We assume that multi-tissue eQTL studies collect expression values of G genes from N individuals in T tissues. However, those N individuals are not necessarily the same for all T tissues; some individuals may provide a subset of tissues. The studies also collect genotype information of M SNPs from the individuals. To determine eQTLs in a specific tissue, or pairs of SNP and gene that are significantly correlated, eQTL studies often use the following linear model.

$$\mathbf{y}_t^g = 1\alpha_t + \mathbf{x}_j\beta_t + \mathbf{e},$$

where $\mathbf{y}_t^g$ is gene expression g of individuals in tissue t, x_j is information on SNP j, and 1 is a vector of ones. β_t is the effect size of SNP j on gene g in tissue t, and if it is not zero, we claim the pair of SNP j and gene g as an eQTL. The Tissue-By-Tissue (TBT) approach computes β_t for every tissue ($t \in \{1 \ldots T\}$), and determines whether at least one β_t is not zero.

To increase the statistical power to detect eQTLs, Meta-Tissue utilizes meta-analysis that combines β_t from T tissues. A naive approach to apply meta-analysis to multi-tissue eQTL datasets is directly using β_t computed from the linear model for TBT. This approach, however, violates the main assumption of meta-analysis that β_t is independent for T tissues. Because multiple tissues are often collected from the same individuals, there exists correlation between gene expression values across different tissues, and this leads to correlated β_t.

To apply meta-analysis to correlated β_t, Meta-Tissue uses a linear mixed model to explicitly capture correlation between β_t:

$$\mathbf{Y}^g = \mathbf{1}\boldsymbol{\alpha} + \mathbf{X}_j\boldsymbol{\beta} + \mathbf{u} + \mathbf{e},$$

where $\mathbf{Y}^g$ and $\mathbf{X}_j$ contain gene expression and SNP information in all T tissues, and Figure 5.2 shows how they are encoded using a simple example. $\mathbf{u}$ is the random effect of a mixed model due to the fact that multiple tissues are collected from the same individuals. $\mathbf{u}$ follows the multivariate normal distribution whose covariance matrix ($\mathbf{D}$ matrix in Figure 5.2) represents sharing of individuals in multiple tissues. Meta-Tissue applies the generalized least squares to estimate $\boldsymbol{\beta}$ and its covariance or correlation between β_t. Meta-Tissue "un-correlates" β_t using the covariance it estimated and use the "un-correlated" β_t for meta-analysis (see the Materials and Methods section for more details).

There is a fundamental difference between Meta-Tissue and the TBT approach. The statistical test in Meta-Tissue tests whether or not a gene is involved in an eQTL in any of the tissues. In other words, the null hypothesis of Meta-Tissue assumes that no effect is present in any of the tissues for a specific gene. A rejection of this null hypothesis is effectively predicting the presence of an effect in at least one of the tissues. However, the tissue-by-tissue approach tests whether or not an eQTL is present in each tissue. Hence, the null hypothesis of TBT assumes that no effect is present in a specific tissue. This means that Meta-Tissue performs one test per gene and TBT performs one test per gene in each tissue. In our comparisons of Meta-Tissue and TBT, we adjust the significant thresholds so that the overall false positive rate of implicating any tissue of a gene in an eQTL is constant for both methods.

Once we identify a significant association using Meta-Tissue, this means that at least one of the tissues contains an eQTL. In order to identify which subset of the tissues contain an eQTL, we utilize a recently developed meta-analysis interpretation framework which computes an m-value statistic for each tissue [HE12]. The m-value estimates the posterior probability that an effect is present in a study included in a meta-analysis. Utilizing the m-values, we can predict tissues in which an effect is present.

5.3.2 Power comparison by simulation

We first simulate gene expression data to compare the power between the traditional Tissue-By-Tissue approach (TBT), Meta-Tissue FE, and Meta-Tissue RE. We create a dataset that has 100 individuals with one SNP and one gene expression level simulating one eQTL. We set the minor allele frequency to 30%. We simulate four tissues and consider four scenarios where a SNP has the same effect in (1) a single tissue, (2) in two tissues, (3) in three tissues, and (4) in all four tissues. The first three scenarios correspond to eQTLs with heterogeneity while eQTLs have no heterogeneity in the last scenario. We check I^2 statistics [HT02] of eQTLs that measure the magnitude of heterogeneity in each scenario and verify that eQTLs have high levels of heterogeneity in the first three scenarios, but very low levels in the last scenario (Figure 5.3). We assume that each individual provides four tissues, and hence this simulation corresponds to a repeated measures design. We use the mixed model discussed in the Materials and Methods section to generate the gene expression levels of individuals while taking into account the repeated measures design. We generate 1,000 datasets (each a potential eQTL) and the power is estimated as a proportion of eQTLs detected at a significance threshold of 5×10^{-8} for meta-analysis methods. We choose this threshold because the number of tests we perform in mouse datasets is on the order of one million (135 SNPs $\times$ 10,588 genes). The significance threshold adjusted for one million tests as in typical GWAS is 5×10^{-8}. For TBT, we apply a significance threshold of $1.25 \times 10^{-8} (5 \times 10^{-8}/4)$ such that the overall false positive rate of TBT is the same as that for Meta-Tissue as discussed in the previous section.

To apply the proposed methods to the simulations, we use the following approach. For TBT, we perform a standard *F*-test using a linear model to obtain a p-value for each pair of a SNP and a gene expression level in each tissue (see Materials and Methods). The tissue-by-tissue approach declares a SNP-gene expression pair as an eQTL if the p-value for the association statistic is below the threshold for any one of the tissues. For Meta-Tissue, we first perform generalized least squares (GLS) to correct for the fact that individuals are shared among tissues. Meta-Tissue then combines information from multiple tissues to obtain either fixed effect or random effect meta-analysis p-values as described in the Materials and Methods section. A SNP-expression pair is considered as an eQTL if its meta-analysis p-value is below the significance threshold. As a separate simulation, we verify that both of our implementations (Meta-Tissue FE and RE) control the false positive rates (see section 5.3.3). This simulation also shows that utilizing the mixed model is critical for controlling false positives when expression levels from multiple tissues are collected from the same individual.

Figure 5.4 shows that Meta-Tissue methods are more powerful than TBT when effects exist in multiple tissues; Meta-Tissue RE is the most powerful when an eQTL has effects in two or three tissues, and Meta-Tissue FE outperforms TBT and Meta-Tissue RE when the effects exist in all tissues. The TBT approach has higher power than Meta-Tissue methods when the effects exist in a single tissue. These results show that TBT is an ideal approach to detect an eQTL that is specific to a certain tissue while Meta-Tissue approaches are ideal for detecting an eQTL that has effects in more than one tissue. As the number of tissues with effects increases, the power of Meta-Tissue methods increases while that of TBT decreases. These results suggest an integrated approach in eQTL studies to apply TBT for detecting tissue-specific eQTLs and Meta-Tissue methods for detecting eQTLs shared between tissues.

5.3.3 False positive rates of meta analysis

To measure the false positive rate of our proposed method, we simulate a multiple tissue dataset where there is no eQTL; a SNP has no effect. We consider 10,000 gene expression levels and 100 SNPs simulating a million pairs of gene and SNP. The number of tissues is four, and we use SNP data from real four tissue dataset from mouse where we have 50 individuals per each tissue. In this dataset, 34% of individuals are shared between two tissues on average, as discussed in the Results section. To generate gene expression for individuals, we use Equation (1) where $\sigma_v = \sigma_e = 0.5$. $\mathbf{D}$ matrix is the same as $\mathbf{D}$ matrix used in the four tissues analysis. We use `mvtnorm` package in `R` to generate gene expression from the multivariate normal distribution.

Table 5.1 shows that Meta-Tissue that uses the linear mixed model to account for the correlation in gene expression between tissues has correct false positive rates. We use the significance threshold of 0.05, and Meta-Tissue FE and RE have false positive rates of 0.533 and 0.0417, respectively. We also measure the false positive rate when meta-analysis methods do not use the linear mixed model, but the linear model that assumes all tissues are independent. Table 5.1 shows that meta-analysis methods have inflation of false positives in this case; Meta-Tissue FE and RE have false positive rates of 0.11 and 0.097. This shows that meta-analysis methods need to consider that individuals are shared across the tissues when combining results from multiple tissues.

Methods	Linear mixed model	Linear model
Meta-Tissue FE	0.053323	0.10992
Meta-Tissue RE	0.041781	0.0972

Table 5.1: False positive rates of Meta-Tissue FE and RE using the linear mixed model and the linear model.

5.3.4 Simulation of heterogeneity in multiple tissues using mouse data

To verify the results of the previous power simulation in real multiple tissue data, we simulate heterogeneity using a liver tissue expression from mouse. This dataset contains 108 samples, 135 SNPs and 10,588 probe expression levels. We detect 389 eQTLs in this single tissue dataset using the standard linear model with a p-value threshold of 5×10^{-8}, which corresponds to the false discovery rate (FDR) of 0.017% level. We consider these detected associations as the gold standard for measuring accuracy of methods in this simulation. We then split the 108 samples into three groups of 36 samples to simulate three tissues, and this means that eQTLs have effects in all three tissues. In our simulations, we expect to find fewer eQTLs because each of our "tissues" only has 36 samples compared to the original 108 samples. We then consider three scenarios similar to scenarios in the previous power simulation; (1) eQTLs have effects only in the first tissue by permuting expression of the second and third tissues, (2) eQTLs have effects only in the first and second tissues by permuting expression of the third tissue, and (3) eQTLs have effects in all three tissues without any permutation. Permuting the expression of a specific tissue removes effects of eQTLs from the tissue, and hence allows simulation of heterogeneity. We apply Meta-Tissue FE, Meta-Tissue RE, and TBT to this multiple tissue dataset and measure how many eQTLs out of the original 389 eQTLs each method can recover using the same threshold ($5 \times 10^{-8}/3$ for TBT). Because the number of eQTLs methods recover can change depending on how we split the 108 samples, we perform ten iterations of the experiment where we divide individuals differently in each iteration, and average the results.

The result of this simulation shows that Meta-Tissue methods recover the most eQTLs when eQTLs have effects in more than one tissue (Figure 5.5). When effects exist in two out of three tissues, Meta-Tissue RE recovers the most eQTLs; it recovers 144 eQTLs out of the 389 eQTLs on average, and this is 27% and 133% more than the number of eQTLs Meta-Tissue FE and TBT recover, respectively. When eQTLs have effects in all tissues, Meta-Tissue FE recovers the most eQTLs, and when effects exist in a single tissue, TBT does. This result is consistent with the previous power simulation in which Meta-Tissue methods were more powerful than TBT when eQTLs have effects in multiple tissues.

5.3.5 Detecting eQTLs in multiple tissue mouse data

We apply Meta-Tissue to detect eQTLs in multiple tissues from mouse. Our data consists of two sets; one with four tissues (cortex, heart, liver, spleen), and the other with ten tissues (bone marrow, hippocampus, kidney, pancreas, stomach, white fat, and the four tissues). The four tissue dataset has 50 samples per each tissue while the ten tissue dataset has 22 samples per tissue. In both datasets, not all individuals provided all different types of tissues; on average, 34% of individuals are shared between two tissues in the four tissue dataset while 11% of individuals are shared in the ten tissues dataset. The number of SNPs (135 SNPs) and the number of probes (10,588) are the same as those of the liver tissue.

Figures 5.6A (four tissues) and 5.6B (ten tissues) show the number of eQTLs detected by Meta-Tissue RE, Meta-Tissue FE, and TBT using a threshold of 5×10^{-8} (5×10^{-8}/the number of tissues for TBT). The number substantially increases by using Meta-Tissue RE or FE, showing up to two fold and twelve fold increases compared to TBT in the four and ten tissue datasets, respectively. These results indicate that methods that combine results of multiple tissues outperform a method that uses results of each tissue separately as all meta-analysis methods detect more eQTLs than TBT. Moreover, these results suggest a possibility that there exist a considerable number of eQTLs with different effect sizes across tissues as Meta-Tissue RE consistently identifies more eQTLs than Meta-Tissue FE. In addition to the number of eQTLs (SNP-expression pairs), we also analyze the number of eSNPs (unique SNPs influencing gene expression) and eProbes (unique probes for gene expression). Similar to the results of the number of eQTLs, Meta-Tissue detects more eSNPs and eProbes than TBT (Figure 5.7).

Another important implication comes from comparing the two datasets. TBT finds substantially fewer number of eQTLs in the ten tissue dataset than in the four tissue dataset. This is possibly because the sample size of each tissue is decreased from 50 to 22. On the other hands, the meta-analytic methods find more eQTLs. One possible reason is that the total sample size is slightly increased from 200 to 220. Therefore, the results demonstrate that by using information from multiple tissues and leveraging meta-analysis methods, we may be able to detect eQTLs even if the sample size for each tissue is small.

In addition to the number of eQTLs that different methods detect, we also analyze the overlap of eQTLs using Venn diagrams (Figures 5.6C and 5.6D). The Venn diagrams show the number of eQTLs detected only by each of the three methods, by both TBT and each of Meta-Tissue methods, by both Meta-Tissue methods, and by all three methods. In the four tissue dataset, the three methods detect 493 unique eQTLs overall, and a majority of eQTLs (95.1% of total eQTL) are detected by either of Meta-Tissue methods. There are, however, 24 eQTLs (4.9% of total eQTLs) that only TBT detects, and they are likely to be tissue-specific eQTLs. In the ten tissue dataset, almost all eQTLs (99.3% of total eQTLs) are detected by Meta-Tissue RE or FE, and there are 4 eQTLs (0.7% of total eQTLs) detected only by TBT, which may be due to the low statistical power due to the limited number of samples.

Instead of the common genome-wide significance threshold (e.g. 5×10^{-8}) to identify eQTLs, an alternative approach is to use the false discovery rate (FDR) approach, and we use the QVALUE package in R [Sto02] to compute a q-value for each SNP-expression pair. We consider only *cis*-eQTLs for the FDR approach; we consider an eQTL as *cis* if a SNP is on the same chromosome as the probe for gene expression. While typical eQTL studies consider 1 Mb as a distance between a SNP and a probe for *cis*-eQTLs, we consider a much longer distance due to a small number of genotyped SNPs (135 SNPs). Figures 5.8A and 5.8B show the number of eQTLs detected by Meta-Tissue methods and TBT using FDR of 0.05 level in four and ten tissues, respectively, and Figures 5.8C and 5.8D are Venn digrams showing the overlap of eQTLs. The results using the FDR approach are consistent with those using the common genome-wide significance threshold; Meta-Tissue RE detects most eQTLs among the three methods, and a majority of eQTLs (86% and 93% of total eQTLs for four and ten tissues) are detected either by Meta-Tissue RE or FE.

5.3.6 Measuring heterogeneity in mouse data

The number of eQTLs detected only by TBT or by RE in Figures 5.6 and 5.8 indicates that there can be several eQTLs with different effect sizes in different tissues. To measure the magnitude of heterogeneity of eQTLs, we use the Cochran's Q statistic [DL86] and the I^2 statistic [HT02]. We make a plot whose x-axis is the I^2 statistic and whose y-axis is the log of p-value of Cochran's Q statistic, and a histogram showing the distribution of I^2 statistics. Figures 5.9, 5.10, and 5.11 show the heterogeneity of eQTLs detected by TBT, FE, and RE, respectively, in the four tissues of mouse data. These plots show that the eQTLs detected by RE show higher level of heterogeneity than the eQTLs detected by FE, as expected. Given the p-value threshold of $0.05/k$ where k is the number of eQTLs detected, 65, 17, and 53 eQTLs show statistically significant heterogeneity in TBT, Meta-Tissue FE, and Meta-Tissue RE, respectively, using the p-value of Cochran's Q statistic.

5.3.7 Predicting the presence of effects in multiple tissue data

Our Meta-Tissue approach not only detects more eQTLs from multiple tissues but also provides an interpretation framework that predicts whether an eQTL has effects in a specific tissue. Meta-Tissue computes a statistic called m-value [HE12], and it is the posterior probability that an effect exists in a specific tissue. If the m-value is greater than a threshold t, we predict that an effect exists, and if it is less than $1 - t$, we predict that an effect does not exist. Another approach to predict an effect is to use a p-value. In this approach, an effect exists if a p-value is less than a significance threshold and does not exist otherwise.

We first apply this prediction framework to the 3-way split liver tissue dataset that we previously generated. Recall that the liver tissue has 389 eQTLs, and we simulated three tissues from it and three scenarios in which we varied heterogeneity of eQTLs. For this simulation, we consider only the scenario where eQTLs have effects in the first two tissues out of three since this corresponds to heterogeneity in which the number of eQTLs that TBT and Meta-Tissue recover is relatively large. We measure how accurately Meta-Tissue and the p-value approach predict the presence and absence of effects of the 389 eQTLs in the three tissues. More specifically, Meta-Tissue makes a correct prediction if m-values are greater than 0.9 in the first two tissues and the m-value is less than 0.1 in the third tissue ($t = 0.9$). We consider an m-value prediction to be ambiguous if any of the three tissues has the m-value between 0.1 and 0.9. If the prediction is not either correct or ambiguous, it is considered as an incorrect prediction. For the p-value approach, p-values of the first two tissues need to be less than the significance threshold (5×10^{-8}/3) and p-value of the third tissue needs to be greater than the threshold for a correct prediction. Otherwise, the prediction is an incorrect prediction since the p-value approach does not have the notion of the ambiguous prediction. In the original 3-way split liver tissue experiment, we had ten simulations which differed in how the individuals were divided. Over the ten simulations, Meta-Tissue and TBT recovered 146 eQTLs out of total 389 eQTLs on average (Figure 5.5). Since we use m-values for the interpretation purpose (not for detecting eQTLs), we apply m-values to only those 146 eQTLs. We also predict effects of the 146 eQTLs using the p-value approach.

Meta-Tissue makes the correct prediction for 35% (51/146) of the eQTLs and predicts the ambiguous prediction for 56% (82/146). The p-value approach only makes the correct prediction for 11% (16/146) of the eQTLs. The number of correct predictions of Meta-Tissue is more than three times greater. In addition, given the advantage of the fact that Meta-Tissue can make ambiguous predictions, the number of incorrect predictions for Meta-Tissue (13/146) is ten times fewer than that for the p-value approach (130/146). The results demonstrate that by combining the meta-analysis method and the interpretation framework, we may predict effects of eQTLs more accurately than the approach utilizing p-values.

We then apply our interpretation framework to the four and ten multiple tissue datasets from mouse to predict effects of eQTLs that were discovered using Meta-Tissue and TBT (493 and 568 eQTLs in four and ten tissue datasets, respectively). We calculate the m-value for each eQTL per each tissue and make a prediction that the eQTL affects expression in that tissue if the m-value is greater than 0.9. We also compare our approach to the p-value approach as in the previous simulation using the same threshold (5×10^{-8}/the number of tissues).

First, we apply the two approaches to the four tissue dataset, and Table 5.2 lists the number of eQTLs predicted to have effects across various combinations of tissues (e.g. eQTLs affecting expression in heart/liver, heart/cortex, heart/liver/cortex). The results show that Meta-Tissue consistently categorizes more eQTLs having effects in multiple tissues than the p-value approach. Among those eQTLs, ones that influence expression levels in all tissues are particularly interesting because they may provide insights into the global regulatory mechanisms of eQTLs. Meta-Tissue predicts 283 such eQTLs while the p-value approach predicts 15 eQTLs. The small number of predictions in p-value approach is expected because even if the effect exists in all T tissues, given power p of tissue-by-tissue approach, we can predict the global effect only with probability p^T.

We next predict effects of eQTLs in the ten tissue dataset, and for this dataset, we would expect to detect a fewer number of eQTLs having effects across all tissues since it becomes less likely that all p-values or m-values pass the threshold as we try to detect effects in more tissues. Table 5.3 shows the number of eQTLs predicted to affect expression across different numbers of tissues considered (e.g. eQTLs having effects across any two tissues, any three tissues). Similar to the results of the four tissue dataset, Meta-Tissue predicts more eQTLs with effects in several tissues than the p-value approach. Unlike the four tissues, we detect a fewer number of eQTLs having effects in all ten tissues; 134 and zero such eQTLs by Meta-Tissue and the p-value approach, respectively. The results indicate the intrinsic difficulty in detecting eQTLs influencing expression across many different tissues.

Tissues	Meta-Tissue	p-values
Cortex/Heart	7	6
Cortex/Liver	1	2
Cortex/Spleen	4	2
Heart/Liver	7	3
Heart/Spleen	7	4
Liver/Spleen	10	2
Cortex/Heart/Liver	28	7
Cortex/Heart/Spleen	49	1
Cortex/Liver/Spleen	17	0
Heart/Liver/Spleen	24	2
All four tissues	283	15

Table 5.2: The number of eQTLs predicted to have effects by Meta-Tissue and the p-value approach across various combinations of the four tissues. Meta-Tissue uses m-value statistics to predict effects; if m-value is greater than 0.9, the effect exists. The p-value approach uses p-values to make predictions; the effect exists if p-value is less than the significance threshold (5×10^{-8}/the number of tissues).

	Meta-Tissue	p-values
2 tissues	12	10
3 tissues	7	0
4 tissues	20	4
5 tissues	33	0
6 tissues	36	1
7 tissues	88	0
8 tissues	99	0
9 tissues	124	0
10 tissues	134	0

Table 5.3: The number of eQTLs predicted to have effects by Meta-Tissue and the p–value approach across different numbers of tissues considered in the ten tissue dataset (eQTLs having effects across any two tissues, any three tissues, etc.).

5.4 Discussion

We presented a statistically powerful approach to detect eQTLs from multiple tissues. Our approach, Meta-Tissue, takes advantage of two meta-analysis methods that differ in their assumptions on effects of eQTLs in different tissues. The first method assumes that effects exist in all tissues with the same magnitude, and this assumption allows us to detect eQTLs shared across all tissues. The second method assumes that effect sizes of variants are different among studies. By assuming the heterogeneity, we may be able to accurately describe the nature of eQTLs whose patterns of genetic regulation differ across tissues. Meta-analysis methods, however, assume that studies are independent, and this assumption is unlikely to be true in multi-tissue dataset since studies collect multiple tissues from the same individuals. This may cause correlation in expression between tissues, and to correct for the correlation, we utilized a mixed model that enables the meta-analysis method to achieve correct false positive rates.

To measure the performance of Meta-Tissue, we first showed by simulations that our methods are generally more powerful than a naive approach that looks at results of each tissue individually. Next, by using data from mouse liver tissue, we simulated the heterogeneity in effect sizes across a subset of tissues as well as in all tissues. Meta-Tissue methods were shown to recover more original eQTLs from multiple tissues than the naive tissue-by-tissue approach when effects exist in multiple tissues. We then observed that Meta-Tissue detects many eQTLs that the naive approach does not detect in four and ten tissue datasets from mouse. However, we note that there are a few tissue-specific eQTLs that only the naive approach detects, and hence we recommend that eQTL studies also apply the naive approach in addition to Meta-Tissue.

In addition to detecting more eQTLs, Meta-Tissue can also accurately predict whether an effect exists in a specific tissue. Meta-Tissue calculates the posterior probability that an eQTL has an effect in a certain tissue, and we demonstrated that this probability is more effective in predicting the effect than a p-value is by using the same liver tissue simulation. We then predicted effects of eQTLs that we found in the four and ten tissue datasets and showed our method predicts more eQTLs having effects in multiple tissues than the p-value approach.

Our approach is fundamentally different from previous approaches that also attempt to detect eQTLs from multiple tissues, and to the best of our knowledge, Meta-Tissue is the first method to apply both a mixed model and meta-analysis methods to eQTL mapping. A traditional approach to detect associations from repeated measurements from same individuals such as multiple tissue data is MANOVA. However, MANOVA is not directly applicable to our multiple tissue data because not all samples provided all different types of tissues, and hence our data are not completely "repeated measurements." Meta-Tissue is more general than MANOVA since Meta-Tissue can be applied to both "repeated measures design" in which individuals are shared across all tissues and to a scenario in which only a subset of individuals are shared. Another advantage of our method is that Meta-Tissue can take into account population structure by adding an additional variance component term in our mixed model. This may be important to multiple tissue datasets in which individuals are sampled from different populations, which may cause inflation of false positives.

Meta-Tissue leverages the recently developed random effects model [HE11] that achieves higher power than the traditional random effects model [DL86, IPE07a, IPE07b, EMI07]. Han and Eskin showed that the traditional random effects model never achieves higher power than the fixed effects model due to its conservative null hypothesis. We apply the traditional RE to our power simulation (Figure 5.12), the heterogeneity experiment with the liver tissue (Figure 5.13), and the four and ten tissue datasets of mouse data (Figure 5.14), and we observe the same phenomenon; the traditional RE is always less powerful than FE and the recently developed RE.

There are a few other methods that attempt to detect eQTLs from the multiple tissue data such as Sparse Bayesian Multiple Regression and the GFlasso approach proposed by Petretto et al. [PBL10] and Kim et al. [KX09] However, a key difference between these methods and Meta-Tissue is that they attempt to detect multiple variants ("multi-locus") associated with multiple traits while our method focuses on an association of a single variant. Another difference and one main advantage of Meta-Tissue is that since it is a meta-analysis method, studies can combine results of many published eQTL analyses without actual data assuming that those analyses are independent; only results of an eQTL analysis such as effect size estimates are needed when the analyses are independent. Meta-Tissue has another advantage that it is simpler and more computationally efficient than other methods that involve computationally challenging algorithms such as Bayesian variable selection and regularized linear regression including Lasso. While we applied Meta-Tissue to the multi-tissue dataset with a small number of genotyped SNPs and samples (135 SNPs and about a total of 200 samples across tissues), our algorithm and software are efficient enough to be applied to larger eQTL studies where there are hundreds of individuals genotyped at hundreds of thousands SNPs.

Reference to published article

Jae Hoon Sul*, Buhm Han*, Chun Ye*, Ted Choi, and Eleazar Eskin, "Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches." *PLoS Genetics*. 9, e1003491, 2013.

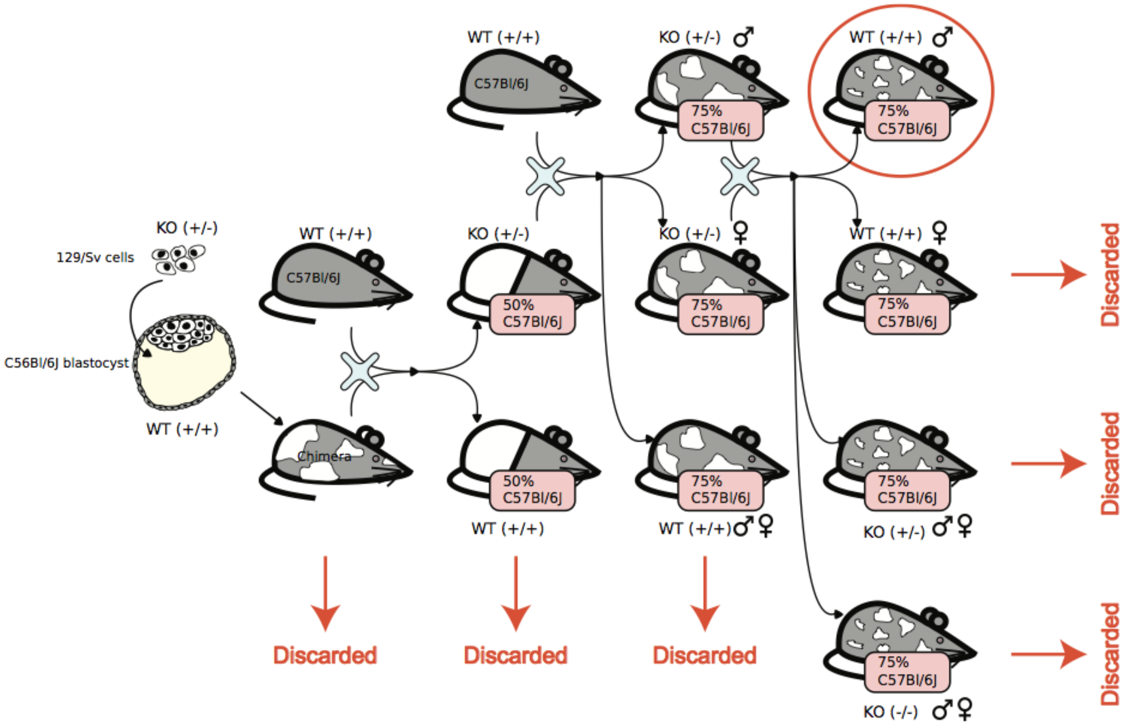
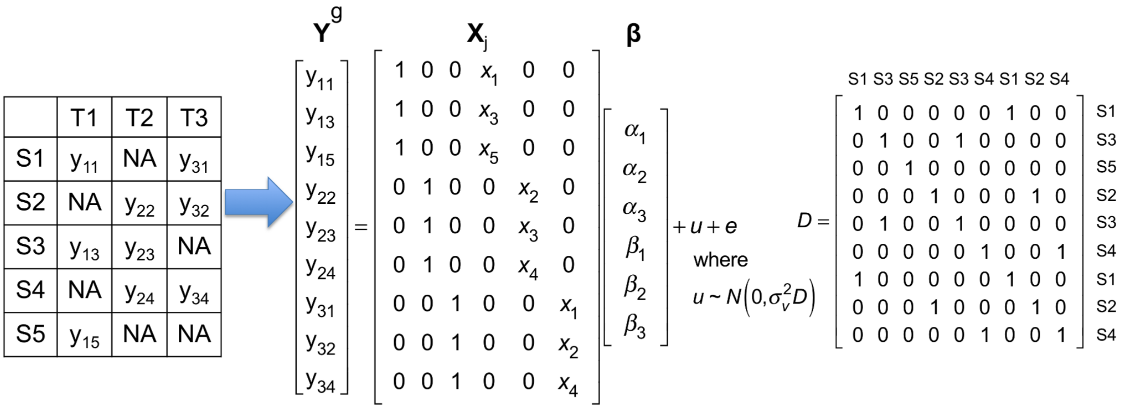


Figure 5.1: The mice were generated by creating a chimera with heterozygous 129/Sv cells in a C56Bl/6J blastocyst. The chimera was crossed with a wildtype C56Bl/6J to obtain heterozygous KOs and homozygous WTs. The heterozygous KOs were backcrossed to wildtype C56Bl/6J to obtain animals that are 75% C56Bl/6J. The male and female heterozygous KOs are intercrossed and only the resulting wildtype males are used in this study. The complicated structure of the cross is due to the fact that the knockouts were designed to be used subsequently for other studies.



	T1	T2	T3
S1	y_{11}	NA	y_{31}
S2	NA	y_{22}	y_{32}
S3	y_{13}	y_{23}	NA
S4	NA	y_{24}	y_{34}
S5	y_{15}	NA	NA

$$\overset{\mathbf{Y}^g}{\begin{bmatrix} y_{11} \\ y_{13} \\ y_{15} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{31} \\ y_{32} \\ y_{34} \end{bmatrix}} = \overset{\mathbf{X}_j}{\begin{bmatrix} 1 & 0 & 0 & x_1 & 0 & 0 \\ 1 & 0 & 0 & x_3 & 0 & 0 \\ 1 & 0 & 0 & x_5 & 0 & 0 \\ 0 & 1 & 0 & 0 & x_2 & 0 \\ 0 & 1 & 0 & 0 & x_3 & 0 \\ 0 & 1 & 0 & 0 & x_4 & 0 \\ 0 & 0 & 1 & 0 & 0 & x_1 \\ 0 & 0 & 1 & 0 & 0 & x_2 \\ 0 & 0 & 1 & 0 & 0 & x_4 \end{bmatrix}} \overset{\boldsymbol{\beta}}{\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}} + u + e \quad \text{where} \quad u \sim N\left(0, \sigma_v^2 D\right)$$

$$D = \begin{array}{c} \begin{array}{ccccccccc} \text{S1} & \text{S3} & \text{S5} & \text{S2} & \text{S3} & \text{S4} & \text{S1} & \text{S2} & \text{S4} \end{array} \\ \left[\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{c} \\ \text{S1} \\ \text{S3} \\ \text{S5} \\ \text{S2} \\ \text{S3} \\ \text{S4} \\ \text{S1} \\ \text{S2} \\ \text{S4} \end{array}$$

Figure 5.2: A simple example showing how gene expression and SNP in multi-tissue eQTL studies are encoded in the mixed model of Meta-Tissue. This example has five samples (S1, S2, S3, S4, and S5) in three tissues (T1, T2, and T3). The leftmost table shows which tissues are collected from each sample; y_{ij} means gene expression of jth sample in ith tissue, and NA means the tissue is not collected. In this example, each tissue has gene expression measured in three samples. $\mathbf{Y}^g$ is a vector containing expression of samples in all tissues; there are a total of 9 gene expression values. In the $\mathbf{X}_j$ matrix, x_i denotes genotype of ith sample. The $\boldsymbol{\beta}$ matrix contains three intercepts (α_t) and three β_t for the three tissues. $\mathbf{u}$ is the random effect of the mixed model, and $\mathbf{u} \sim \mathcal{N}(0, \sigma_v^2\mathbf{D})$. $\mathbf{D}$ is 9×9 matrix whose entry at ith row and jth column is 1 if the ith and jth entries of $\mathbf{Y}^g$ are collected from the same individual, and 0 otherwise.

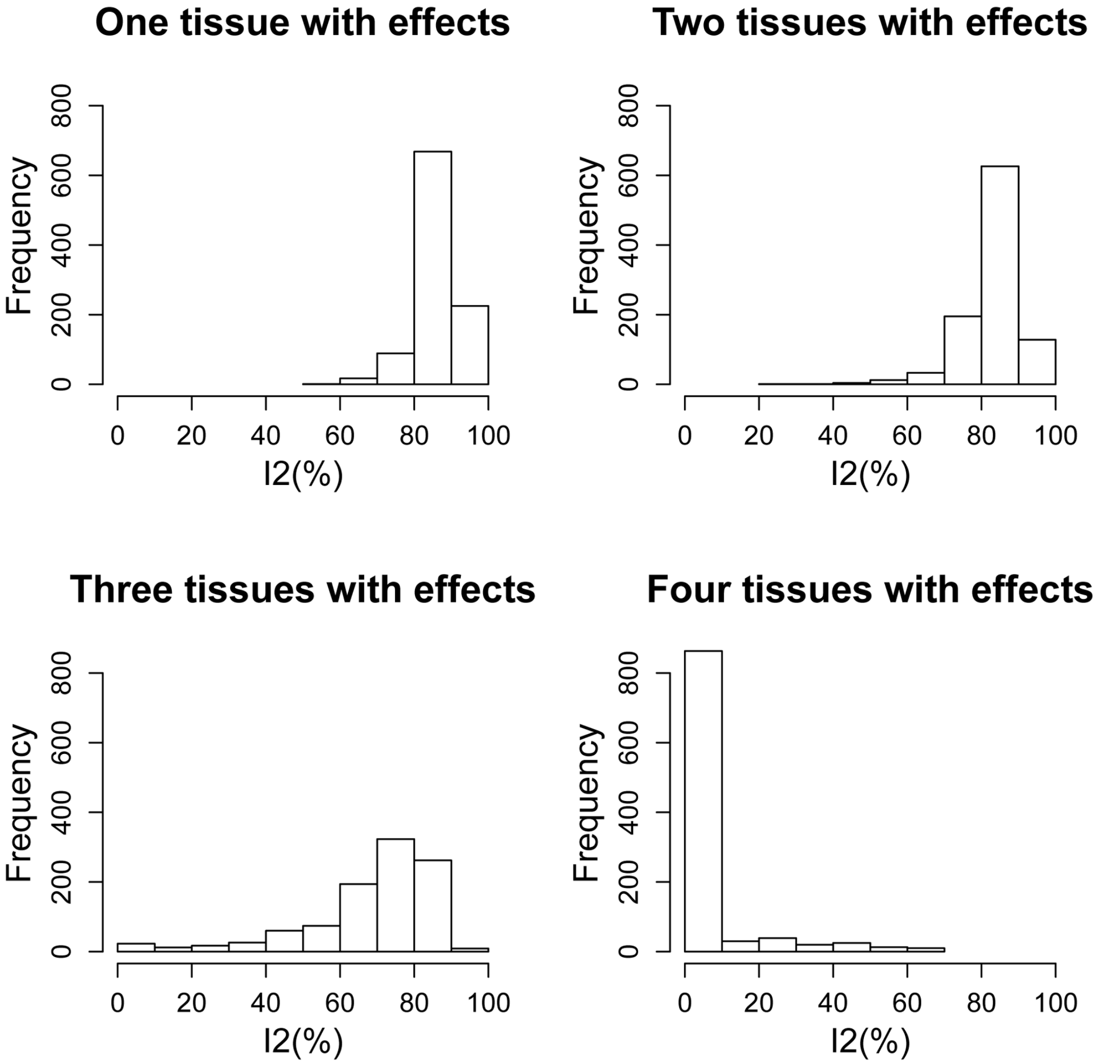


Figure 5.3: Histograms showing the distribution of I2 statistics in the power simulation. There are four scenarios in the power simulation where an eQTL has an effect 1) in one tissue, 2) in two tissues, 3) in three tissues, and 4) in all four tissues. There are 1,000 eQTLs in each scenario, and the histograms show the distribution of I2 statistics of the 1,000 eQTLs.

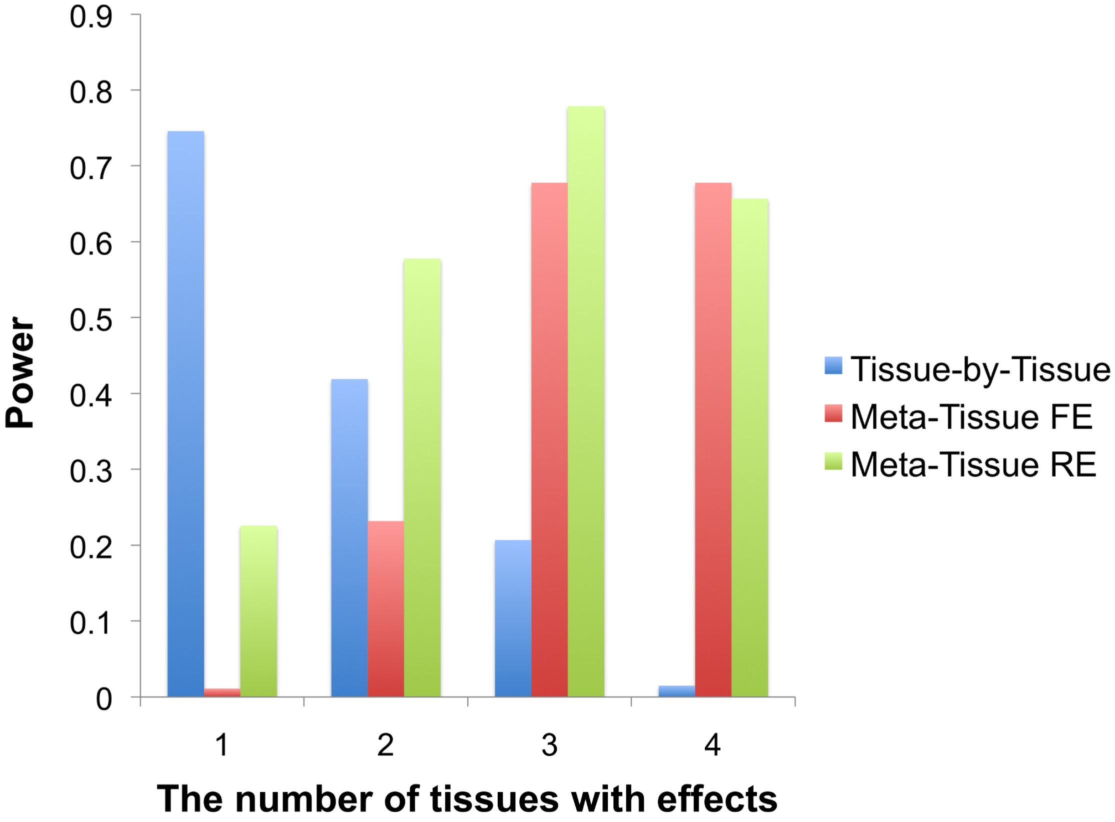


Figure 5.4: Power comparison between the tissue-by-tissue approach, Meta-Tissue fixed effects model (FE), and Meta-Tissue random effects model (RE) using simulated data. X-axis indicates the number of tissues having effects out of four tissues, and Y-axis is the power.

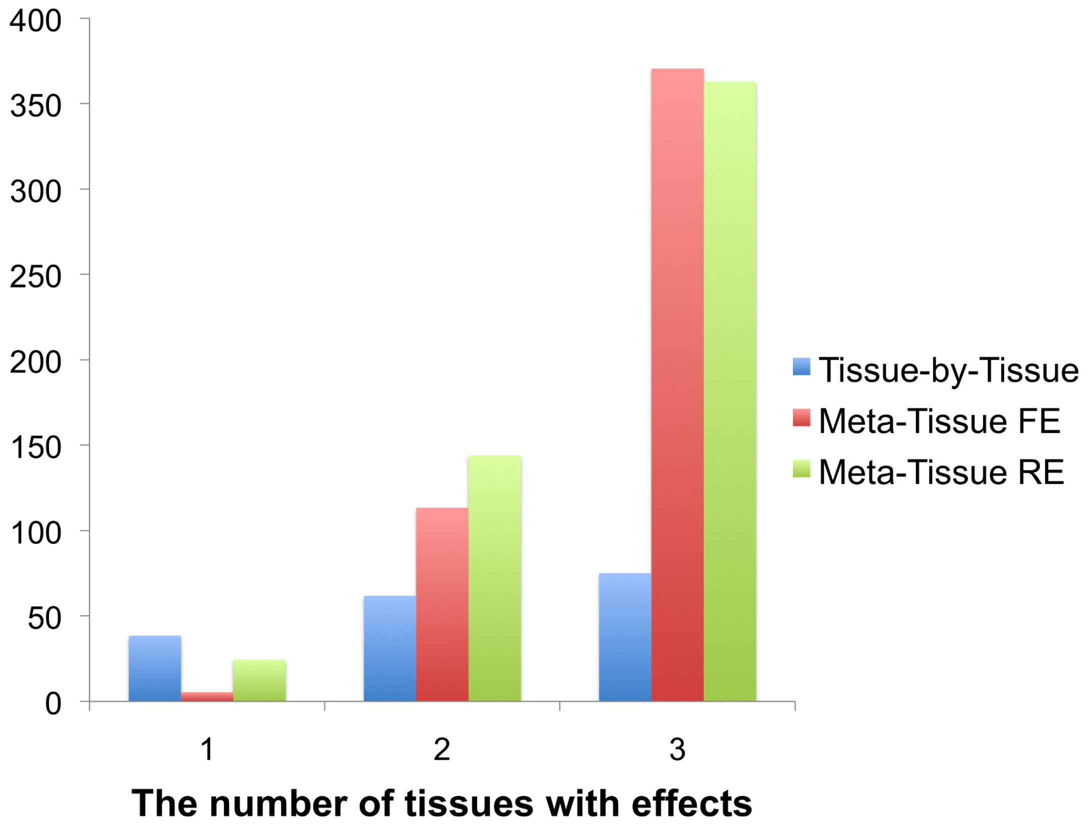


Figure 5.5: The average number of eQTLs that the tissue-by-tissue approach, Meta-Tissue FE, and Meta-Tissue RE recover from three tissues generated from the liver tissue. The liver tissue has 108 samples from which we simulate three tissues of 36 samples. X-axis indicates the number of tissues having effects out of three tissues. The original liver tissue has 389 eQTLs.

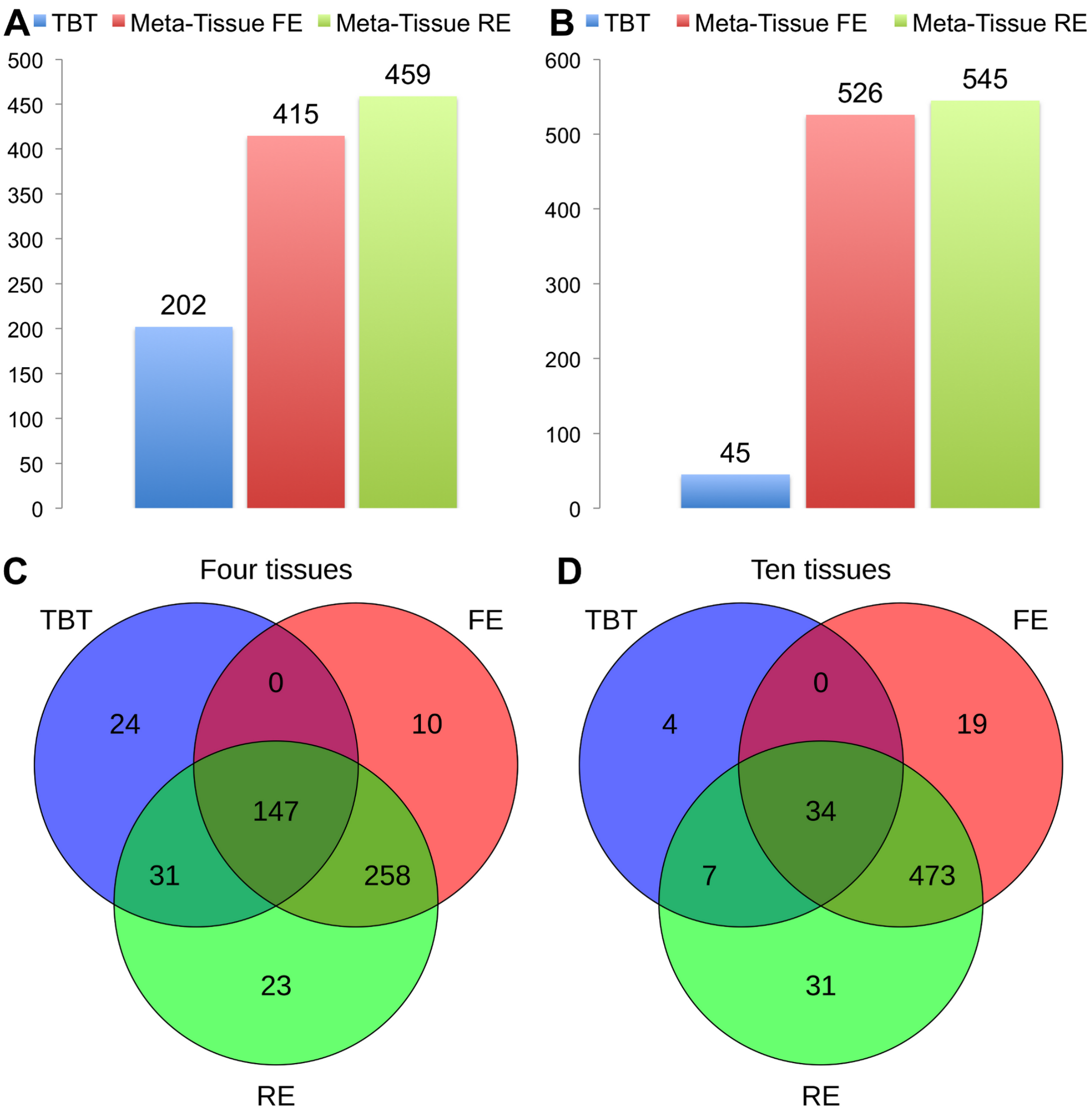


Figure 5.6: The number of eQTLs detected by the tissue-by-tissue approach (TBT), Meta-Tissue FE, and Meta-Tissue RE in A) four and B) ten tissues of mouse, and the overlap of eQTLs detected by the three methods in C) four and D) ten tissues. The datasets consist of the gene expression levels from 50 individuals (four tissues) and 22 individuals (ten tissues). We apply a p-value threshold of 5×10^{-8} for Meta-Tissue and a threshold of 5×10^{-8}/the number of tissues for tissue-by-tissue. The Venn diagrams (C and D) show the number of eQTLs detected by either TBT, FE, or RE, by TBT and either of FE and RE, by FE and RE, and by all three methods.

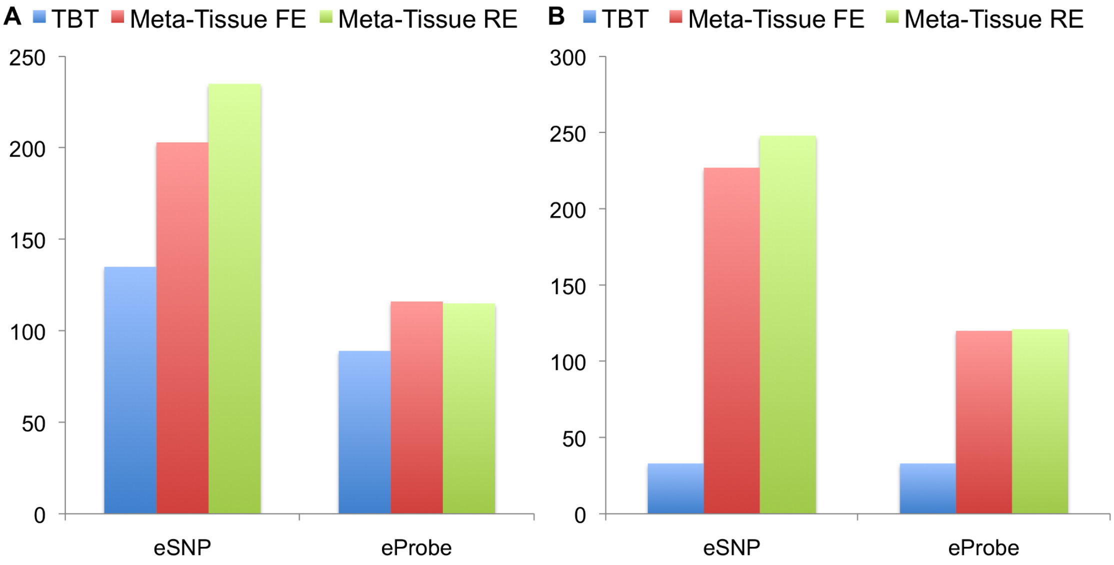


Figure 5.7: The number of eSNPs and eProbes detected by the tissue-by-tissue (TBT) approach, Meta-Tissue FE, and Meta-Tissue RE in A) four tissues and B) ten tissues of mouse. We apply a p-value threshold of 5×10^{-8} for Meta-Tissue and a threshold of 5×10^{-8}/the number of tissues for TBT.

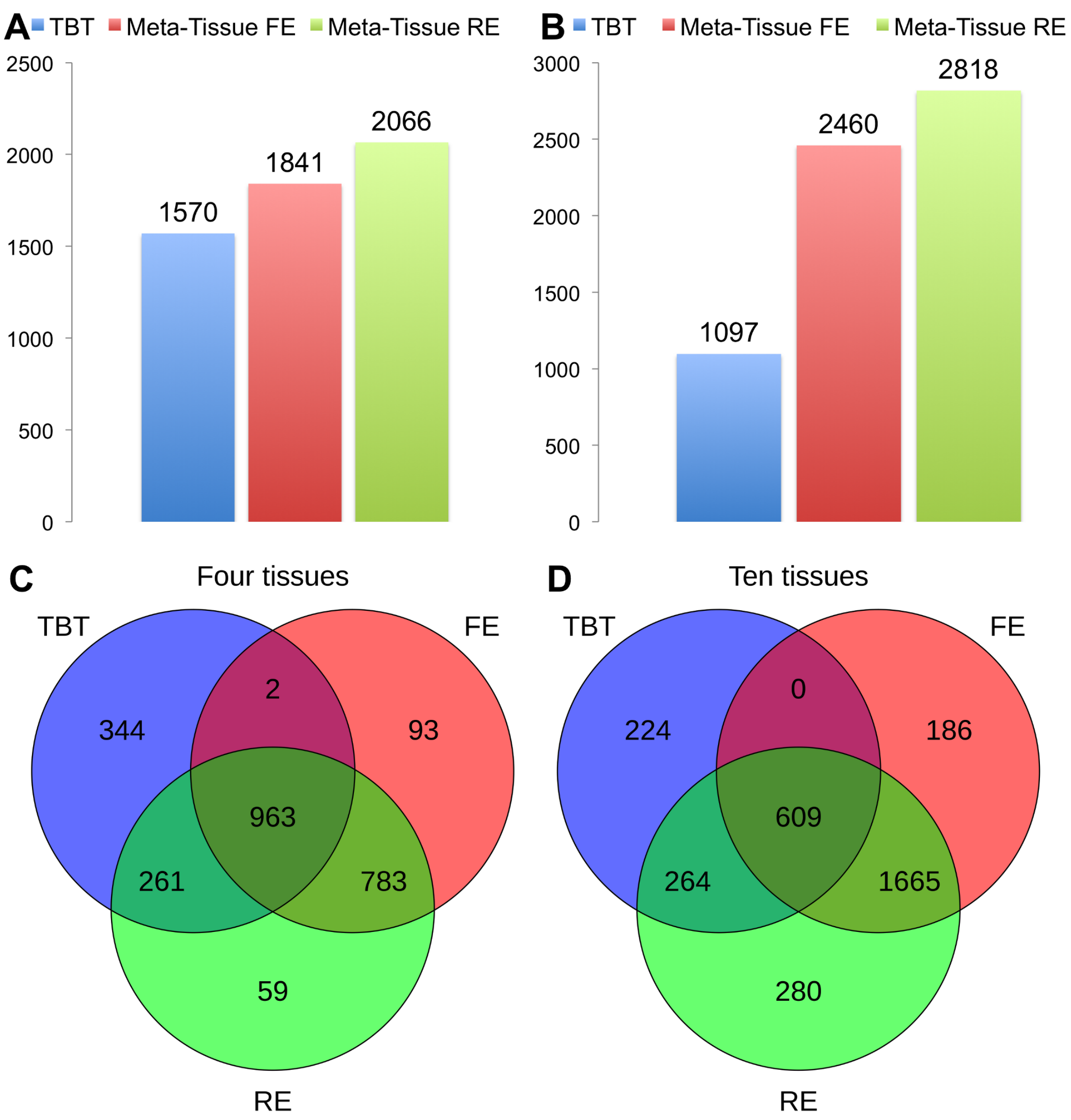


Figure 5.8: The number of eQTLs detected by the tissue-by-tissue approach (TBT), Meta-Tissue FE, and Meta-Tissue RE in A) four and B) ten tissues of mouse using FDR of 5%, and the overlap of eQTLs detected by the three methods in C) four and D) ten tissues. We consider only cis-eQTLs for the FDR approach, and a pair of SNP-probe for gene expression are considered cis if a SNP and a probe are on the same chromosome.

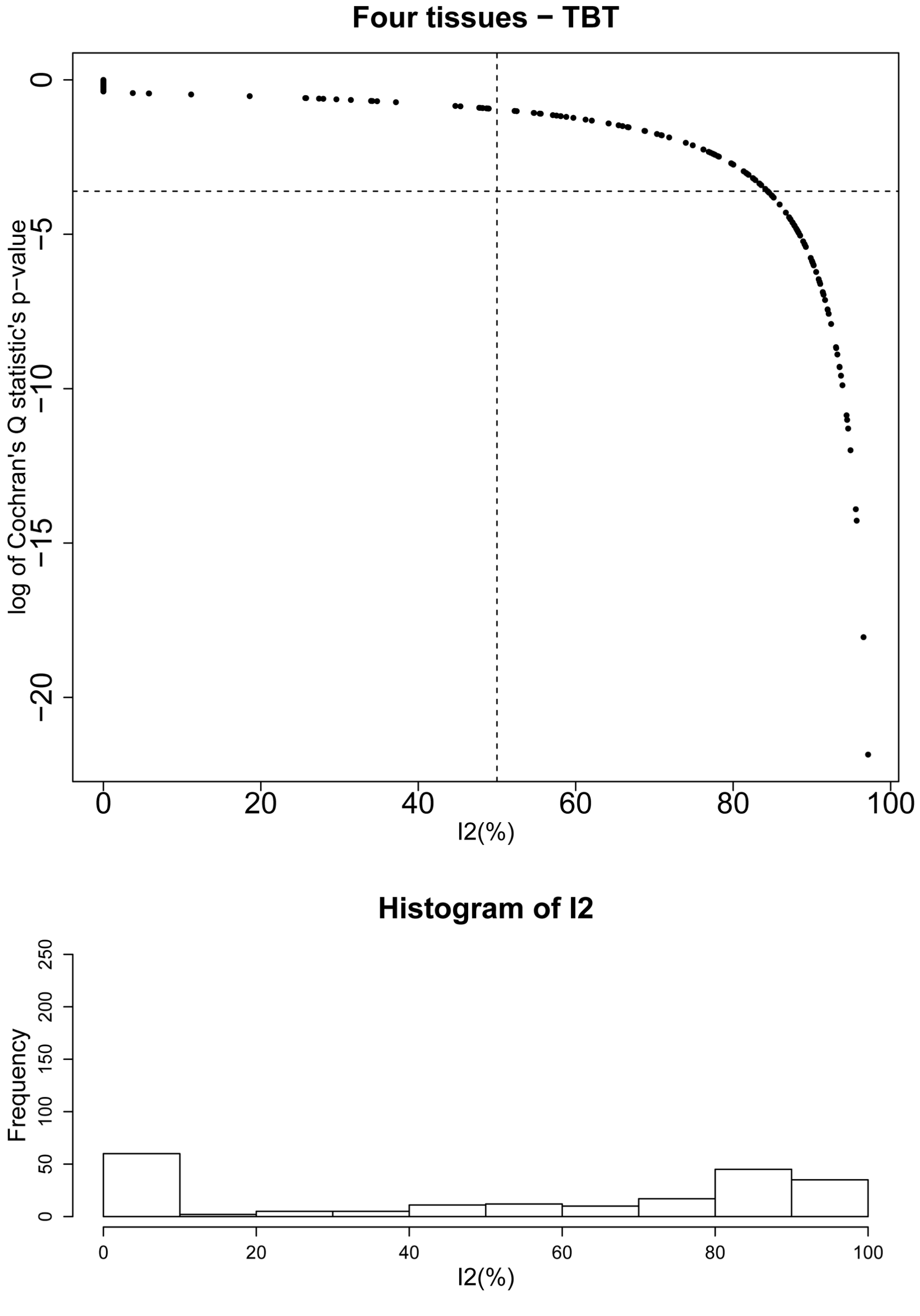


Figure 5.9: A plot showing heterogeneity of eQTLs detected by the tissue-by-tissue approach. X-axis of the top plot indicates I2 statistic and Y-axis indicates log of p–value of Cochrans Q statistic. The vertical dashed line is drawn at I2 = 50%, and the horizontal dash line is drawn at p-value = 0.05/the number of eQTLs detected. The bottom histogram shows the distribution of I2 statistic.

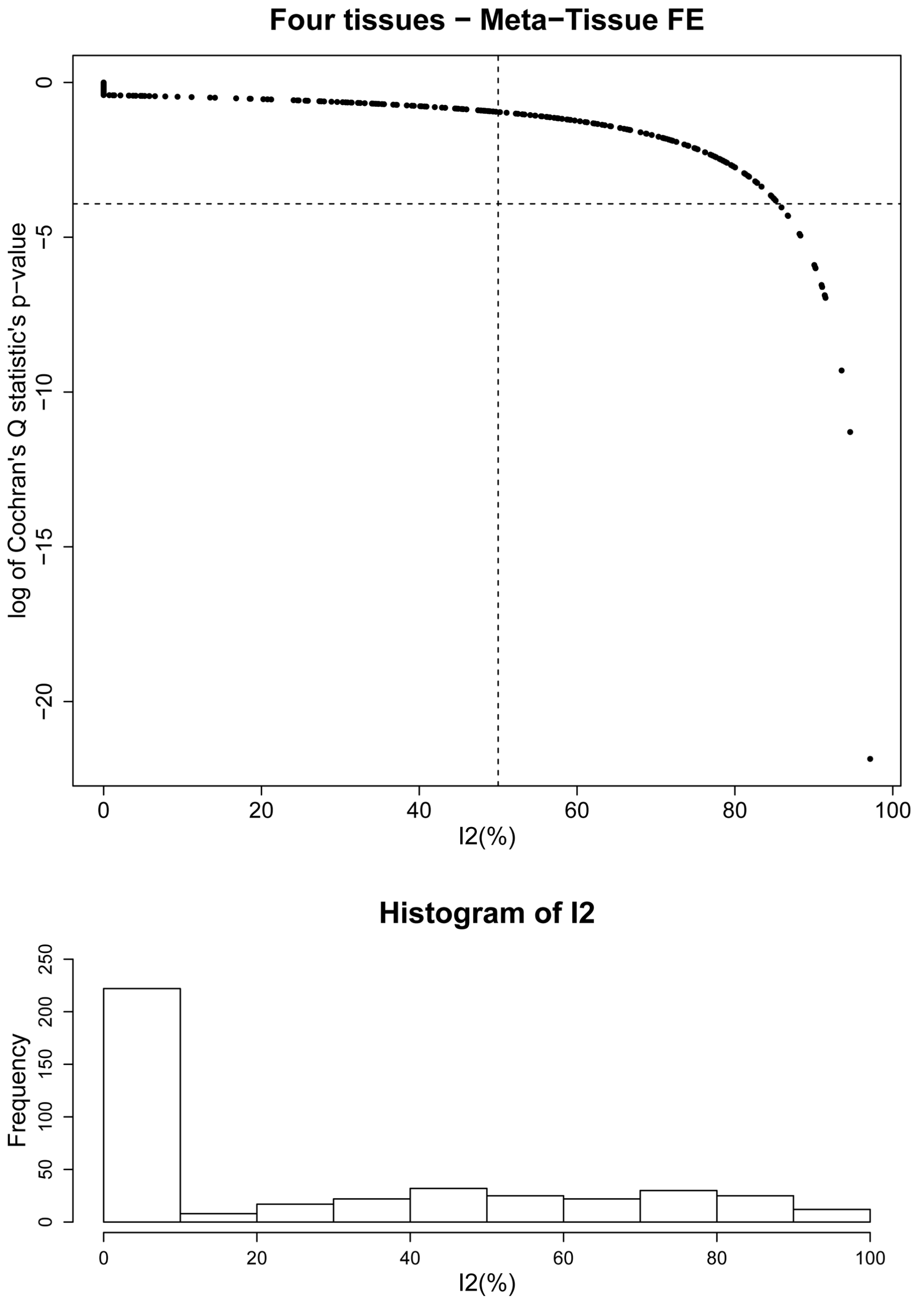


Figure 5.10: A plot showing heterogeneity of eQTLs detected by Meta-Tissue FE. X-axis of the top plot indicates I2 statistic and Y-axis indicates log of p-value of Cochrans Q statistic. The vertical dashed line is drawn at I2 = 50%, and the horizontal dash line is drawn at p-value = 0.05/the number of eQTLs detected. The bottom histogram shows the distribution of I2 statistic.

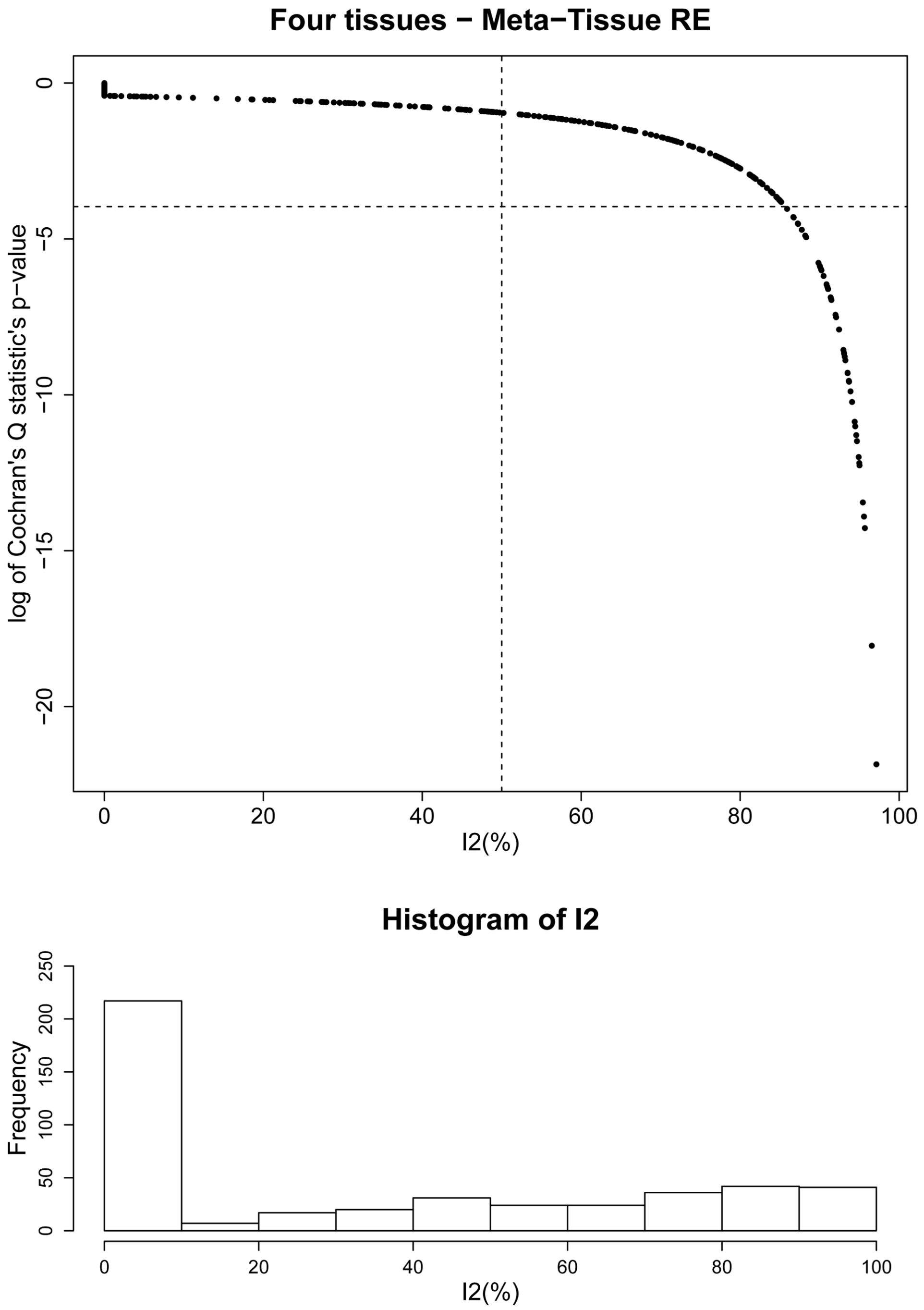


Figure 5.11: A plot showing heterogeneity of eQTLs detected by Meta-Tissue RE. X-axis of the top plot indicates I2 statistic and Y-axis indicates log of p-value of Cochrans Q statistic. The vertical dashed line is drawn at I2 = 50%, and the horizontal dash line is drawn at p-value = 0.05/the number of eQTLs detected. The bottom histogram shows the distribution of I2 statistic.

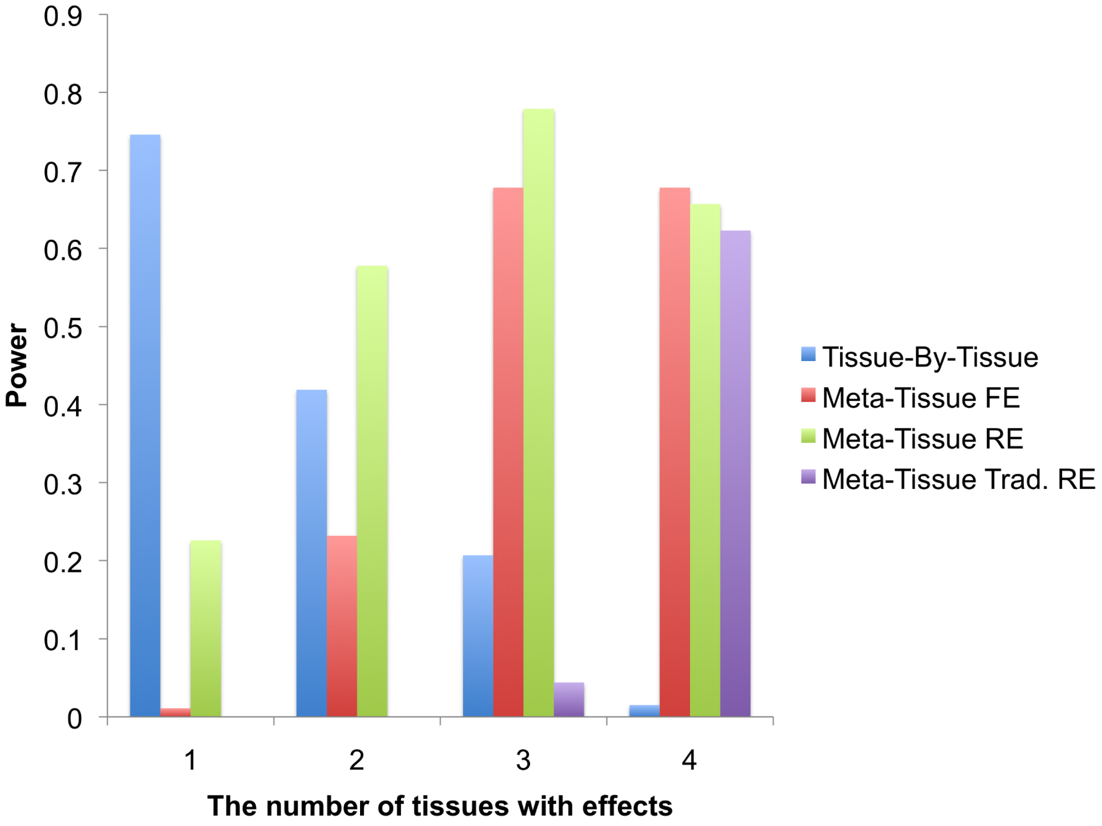


Figure 5.12: Power comparison between the tissue-by-tissue approach, Meta-Tissue fixed effects model (FE), Meta-Tissue random effects model (RE), and Meta-Tissue traditional random effects model using simulated data. X-axis indicates the number of tissues having effects out of four tissues, and Y-axis is the power.

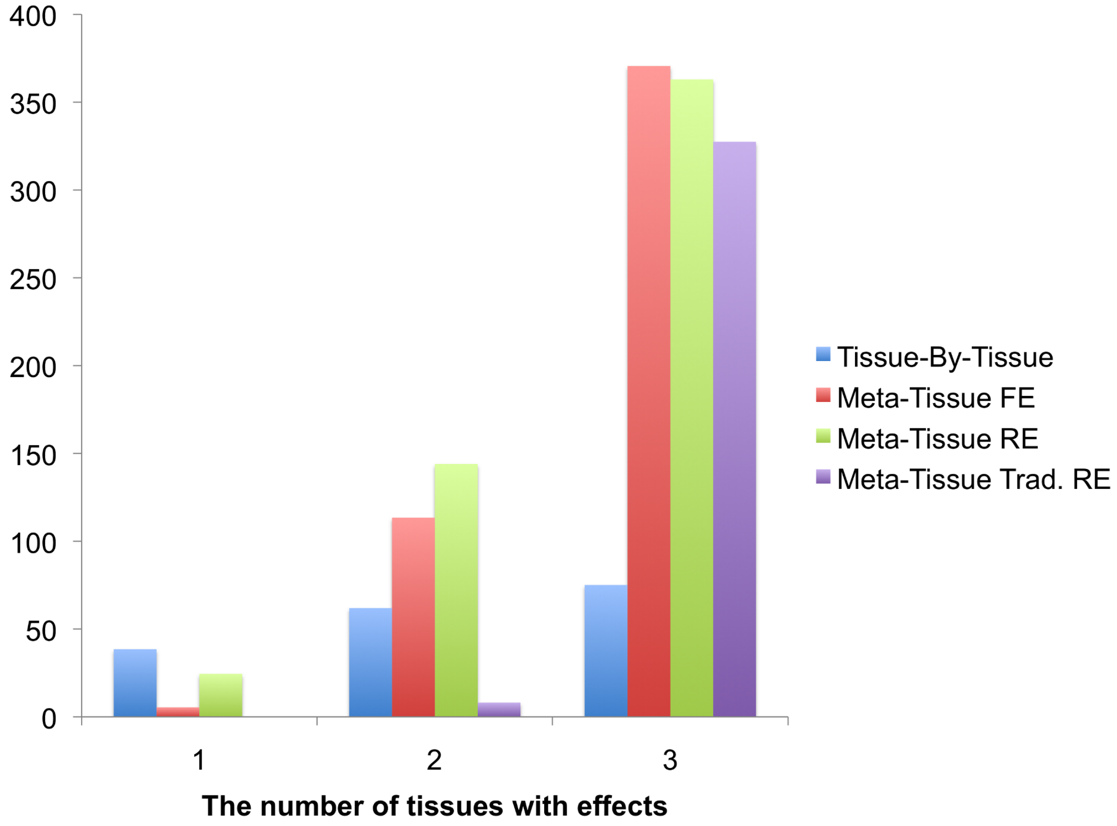


Figure 5.13: The average number of eQTLs that the tissue-by-tissue approach, Meta-Tissue FE, Meta-Tissue RE, and Meta-Tissue traditional RE recover from three tissues generated from the liver tissue. Effects of eQTLs exist in only two tissues. The original liver tissue has 389 eQTLs.

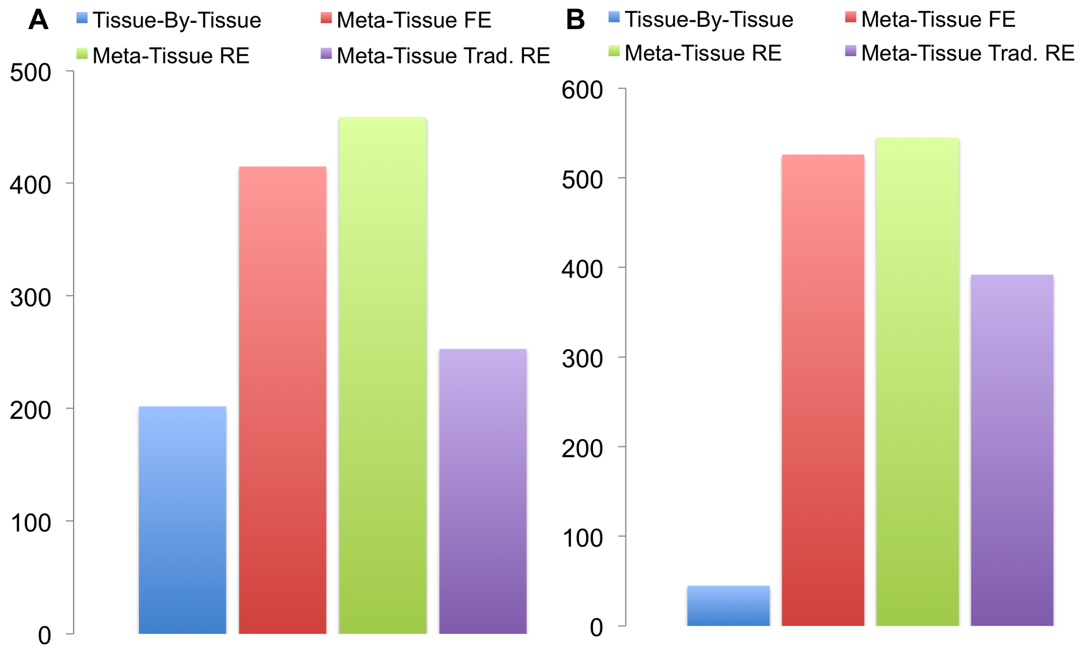


Figure 5.14: The number of eQTLs detected by the tissue-by-tissue approach, Meta-Tissue FE, Meta-Tissue RE, and Meta-Tissue traditional RE in A) four tissues and in B) ten tissues of mouse.

CHAPTER 6

Conclusions

Designing efficient and statistically powerful approaches has become very important in the field of genetics as the vast amount of data are generated to uncover genetic basis of complex traits and diseases. As the sequencing costs decrease at a rate that exceeds Moore's law, studies will soon be able to sequence thousands or tens of thousands individuals to identify genetic variants associated with diseases. This means that traditional approaches in genetics that were designed for hundreds of markers collected in less than a hundred individuals need to be re-developed so that they are efficient enough for ever-growing data. Also, current research in statistics and biostatistics has improved the accuracy or power of statistical approaches significantly, and these new techniques need to be fully utilized in the genetics studies.

The first problem I tackled in my thesis was to develop a statistical approach that efficiently corrects for population structure. Before this research, the linear mixed model had been mostly applied to datasets in which at most hundreds of individuals were collected, and it was not possible to apply it to human GWASs that usually collect thousands of individuals. My approach was one of the first work that enabled GWASs to adopt the mixed model. Several other mixed model approaches have been proposed after our method was published to either further improve the speed or the power of mixed model [LLK12, ZS12]. This shows that our work was a pioneer that opened many research opportunities for other researchers. There are several unanswered questions regarding the mixed model such as finding an optimal set of SNPs to create a kinship matrix and controlling population structure on a set of SNPs that have very different minor allele frequency between populations. Hence, I believe that there still exist several research opportunities that will improve the mixed model in GWASs.

Next, I worked on developing both efficient and powerful approaches to detect associations of rare variants. I proposed two different methods, RWAS and LRT, and while LRT is generally more powerful, RWAS is easier to understand and more efficient. Numerous statistical methods have been developed to identify a group of rare variants involved in a disease even after our papers were published. However, to the best of my knowledge, none of them attempts to identify causal variants explicitly and uses this information in the association. Finding causal variants and utilizing this information is critical in detecting associations of rare variants since including non-causal variants reduces the power of studies significantly as shown in Chapter 3. Therefore, I believe that LRT that attempts to detect causal variants from both data and prior information is still more powerful or comparable to current methods, and will be useful in finding the role of rare variants in diseases.

Lastly, I developed a statistical framework that combines the mixed model and meta-analysis to better identify eQTLs from multiple tissues. One of the main challenges in this project was to apply meta-analysis to a set of correlated studies or tissues. Most meta-analysis methods assume that studies are independent, and this is true for GWASs since it is highly unlikely that same individuals are collected in more than one GWAS. However, eQTL studies usually collect multiple tissues from the same individuals, which causes effects of a genetic variant in multiple tissues to be correlated and violates the independence assumption. To overcome this challenge, I utilized the mixed model to obtain correlation of effect sizes in multiple tissues and incorporated this correlation into meta-analysis. I showed that this approach correctly controls false positives and achieves higher power than a traditional eQTL method that examines each tissue individually. This approach is one of the first work that combines information from multiple tissues to better identify eQTLs, and I believe that many multiple tissue eQTL studies will utilize my approach.

In addition to problems that I focused on my thesis, there are many other problems in genetics that require efficient algorithms such as imputation [HFS12], sequence read mapping [LTP09], genotype discover/calling [MHB10], and multiple testing correction [HKE09b]. There are several efficient methods proposed for each of this problem, and active research is in progress. I believe that more efficient and powerful approaches will be necessary in future to utilize the tremendous amount of genetics data, and the work I have presented in this thesis will be useful for other researchers to develop such approaches.

References

[AHK00] David Altshuler, Joel N. Hirschhorn, Mia Klannemark, Cecilia M. Lindgren, Marie-Claude . C. Vohl, James Nemesh, Charles R. Lane, Stephen F. Schaffner, Stacey Bolk, and Carl Brewer. "The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes." *Nature genetics*, **26**(1):76–80, 2000.

[Arm55] P. Armitage. "Tests for linear trends in proportions and frequencies." *Biometrics*, **11**(3):375–386, 1955.

[ASP10] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. "A method and server for predicting damaging missense mutations." *Nature Methods*, **7**(4):248–249, 2010.

[ATG09] Shahana Ahmed, Gilles Thomas, Maya Ghoussaini, Catherine S. Healey, Manjeet K. Humphreys, Radka Platte, Jonathan Morrison, Melanie Maranian, Karen A. Pooley, Robert Luben, Diana Eccles, D. Gareth Evans, Olivia Fletcher, Nichola Johnson, Isabel Dos Santos Silva, Julian Peto, Michael R. Stratton, Nazneen Rahman, Kevin Jacobs, Ross Prentice, Garnet L. Anderson, Aleksandar Rajkovic, J. David Curb, Regina G. Ziegler, Christine D. Berg, Saundra S. Buys, Catherine A. McCarty, Heather Spencer Feigelson, Eugenia E. Calle, Michael J. Thun, W. Ryan Diver, Stig Bojesen, Brge G. Nordestgaard, Henrik Flyger, Thilo Drk, Peter Schrmann, Peter Hillemanns, Johann H. Karstens, Natalia V. Bogdanova, Natalia N. Antonenkova, Iosif V. Zalutsky, Marina Bermisheva, Sardana Fedorova, Elza Khusnutdinova, SEARCH, Daehee Kang, Keun-Young Y. Yoo, Dong Young Noh, Sei-Hyun H. Ahn, Peter Devilee, Christi J. van Asperen, R. A. E. M. Tollenaar, Caroline Seynaeve, Montserrat Garcia-Closas, Jolanta Lissowska, Louise Brinton, Beata Peplonska, Heli Nevanlinna, Tuomas Heikkinen, Kristiina Aittomki, Carl Blomqvist, John L. Hopper, Melissa C. Southey, Letitia Smith, Amanda B. Spurdle, Marjanka K. Schmidt, Annegien Broeks, Richard R. van Hien, Sten Cornelissen, Roger L. Milne, Gloria Ribas, Anna Gonzlez-Neira, Javier Benitez, Rita K. Schmutzler, Barbara Burwinkel, Claus R. Bartram, Alfons Meindl, Hiltrud Brauch, Christina Justenhoven, Ute Hamann, The GENICA Consortium, Jenny Chang-Claude, Rebecca Hein, Shan Wang-Gohrke, Annika Lindblom, Sara Margolin, Arto Mannermaa, Veli-Matti M. Kosma, Vesa Kataja, Janet E. Olson, Xianshu Wang, Zachary Fredericksen, Graham G. Giles, Gianluca Severi, Laura Baglietto, Dallas R. English, Susan E. Hankinson, David G. Cox, Peter Kraft, Lars J.

Vatten, Kristian Hveem, Merethe Kumle, Alice Sigurdson, Michele Doody, Parveen Bhatti, Bruce H. Alexander, Maartje J. Hooning, Ans M. W. van den Ouweland, Rogier A. Oldenburg, Mieke Schutte, Per Hall, Kamila Czene, Jianjun Liu, Yuqing Li, Angela Cox, Graeme Elliott, Ian Brock, Malcolm W. R. Reed, Chen-Yang Y. Shen, Jyh-Cherng C. Yu, Giu-Cheng C. Hsu, Shou-Tung T. Chen, Hoda Anton-Culver, Argyrios Ziogas, Irene L. Andrulis, Julia A. Knight, kConFab, Australian Ovarian Cancer Study Group, Jonathan Beesley, Ellen L. Goode, Fergus Couch, Georgia Chenevix-Trench, Robert N. Hoover, Bruce A. J. Ponder, David J. Hunter, Paul D. P. Pharoah, Alison M. Dunning, Stephen J. Chanock, and Douglas F. Easton. "Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2." *Nat Genet*, **41**(5):585–90, 3 2009.

[AW90] A. Agresti and J. Wiley. *Categorical data analysis*. Wiley New York, 1990.

[BB08] W. Bodmer and C. Bonilla. "Common and rare variants in multifactorial susceptibility to common diseases." *Nature genetics*, **40**(6):695–701, 2008.

[BDR02] Silviu A. Bacanu, Bernie Devlin, and Kathryn Roeder. "Association studies for quantitative traits in structured populations." *Genet Epidemiol*, **22**(1):78–93, 1 2002.

[BFJ08] Paul I. W. de Bakker, Manuel A. R. Ferreira, Xiaoming Jia, Benjamin M. Neale, Soumya Raychaudhuri, and Benjamin F. Voight. "Practical aspects of imputation-driven meta-analysis of genome-wide association studies." *Hum Mol Genet*, **17**(R2):R122–8, 10 2008.

[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.

[BKK94] Rogier M. Bertina, Bobby PC Koeleman, Ted Koster, Frits R. Rosendaal, Richard J. Dirven, Hans de Ronde, Pieter A. Van Der Velden, and Pieter H. Reitsma. "Mutation in blood coagulation factor V associated with resistance to activated protein C." *Nature*, **369**(6475):64–67, 1994.

[BMS06] Paul I. W. de Bakker, Gil McVean, Pardis C. Sabeti, Marcos M. Miretti, Todd Green, Jonathan Marchini, Xiayi Ke, Alienke J. Monsuur, Pamela Whittaker, Marcos Delgado, Jonathan Morrison, Angela Richardson, Emily C. Walsh, Xiaojiang Gao, Luana Galver, John Hart, David A. Hafler, Margaret Pericak-Vance, John A. Todd, Mark J. Daly, John Trowsdale, Cisca Wijmenga, Tim J. Vyse, Stephan Beck, Sarah Shaw Murray, Mary Carrington, Simon Gregory, Panos Deloukas, and John D. Rioux. “A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC.” *Nat Genet*, **38**(10):1166–72, 10 2006.

[BN95] D. J. Balding and R. A. Nichols. “A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity.” *Genetica*, **96**(1-2):3–12, 1995.

[Bog09] Clifton Bogardus. “Missing heritability and GWAS utility.” *Obesity*, **17**(2):209–10, 2 2009.

[BSS08] Lara E. Bauman, Janet S. Sinsheimer, Eric M. Sobel, and Kenneth Lange. “Mixed effects models for quantitative trait loci mapping with inbred strains.” *Genetics*, **180**(3):1743–61, 11 2008.

[BVE08] Hylke M. Blauw, Jan H. Veldink, Michael A. van Es, Paul W. van Vught, Christiaan G. J. Saris, Bert van der Zwaag, Lude Franke, J. Peter H. Burbach, John H. Wokke, Roel A. Ophoff, and Leonard H. van den Berg. “Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen.” *Lancet Neurol*, **7**(4):319–26, 4 2008.

[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.

[CA07] Wei M. Chen and Goncalo R. Abecasis. “Family-based association tests for genomewide association scans.” *Am J Hum Genet*, **81**(5):913–26, 2007.

[CGK09] Yoon Shin Cho, Min Jin Go, Young Jin Kim, Jee Yeon Heo, Ji Hee Oh, Hyo-Jeong J. Ban, Dankyu Yoon, Mi Hee Lee, Dong-Joon J. Kim, Miey Park, Seung-Hun H. Cha, Jun-Woo W. Kim, Bok-Ghee G. Han, Haesook Min, Younjhin Ahn, Man Suk Park, Hye Ree Han, Hye-Yoon Y. Jang, Eun Young Cho, Jong-Eun E. Lee, Nam H. Cho, Chol Shin, Taesung Park, Ji Wan Park, Jong-Keuk K. Lee, Lon Cardon, Geraldine Clarke, Mark I. McCarthy, Jong-Young Y. Lee, Jong-Koo K. Lee, Bermseok Oh, and Hyung-Lae L. Kim. “A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits.” *Nat Genet*, **41**(5):527–34, 5 2009.

[CKP04] Jonathan C. Cohen, Robert S. Kiss, Alexander Pertsemlidis, Yves L. Marcel, Ruth McPherson, and Helen H. Hobbs. “Multiple rare alleles contribute to low plasma levels of HDL cholesterol.” *Science*, **305**(5685):869–72, 8 2004.

[Cla09] David G. Clayton. “Prediction and interaction in complex disease genetics: experience in type 1 diabetes.” *PLoS Genet*, **5**(7):e1000540, 7 2009.

[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.

[COC54] W. G. COCHRAN. “The combination of estimates from different experiments.” *BIOMETRICS*, **10**(1):101–129, 1954.

[Con04] International Human Genome Sequencing Consortium. “Finishing the euchromatic sequence of the human genome.” *Nature*, **431**(7011):931–45, 10 2004.

[Con07] Wellcome Trust Case Control Consortium. “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.” *Nature*, **447**(7145):661–78, 6 2007.

[Con08] International Schizophrenia Consortium. “Rare chromosomal deletions and duplications increase risk of schizophrenia.” *Nature*, **455**(7210):237–41, 9 2008.

[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.

[CSS93] E. H. Corder, A. M. Saunders, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, G. W. Small, A. D. Roses, J. L. Haines, and M. A. Pericak-Vance. “Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer’s disease in late onset families.” *Science*, **261**(5123):921–3, 8 1993.

[CWW09] Yoonha Choi, Ellen M. Wijsman, and Bruce S. Weir. "Case-control association testing in the presence of unknown relationships." *Genet Epidemiol*, **33**(8):668–78, 3 2009.

[DG08] Frank Dudbridge and Arief Gusnanto. "Estimation of significance thresholds for genomewide association scans." *Genet Epidemiol*, **32**(3):227–34, 4 2008.

[DL86] R. DerSimonian and N. Laird. "Meta-analysis in clinical trials." *Control Clin Trials*, **7**(3):177–88, 9 1986.

[DR99a] B. Devlin and K. Roeder. "Genomic control for association studies." *Biometrics*, **55**(4):997–1004, 12 1999.

[DR99b] B. Devlin and Kathryn Roeder. "Genomic control for association studies." *Biometrics*, **55**(4):997–1004, 1999.

[EDB00] M. P. Epstein, W. L. Duren, and M. Boehnke. "Improved inference of relationship for pairs of individuals." *Am J Hum Genet*, **67**(5):1219–31, 11 2000.

[EMI07] Evangelos Evangelou, Demetrius M. Maraganore, and John P. A. Ioannidis. "Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease." *PLoS One*, **2**(2):e196, 2007.

[EMM06] David M. Evans, Jonathan Marchini, Andrew P. Morris, and Lon R. Cardon. "Two-stage two-locus models in genome-wide association." *PLoS Genet*, **2**(9):e157, 9 2006.

[EPD07] Douglas F. Easton, Karen A. Pooley, Alison M. Dunning, Paul D. P. Pharoah, Deborah Thompson, Dennis G. Ballinger, Jeffery P. Struewing, Jonathan Morrison, Helen Field, Robert Luben, Nicholas Wareham, Shahana Ahmed, Catherine S. Healey, Richard Bowman, SEARCH collaborators, Kerstin B. Meyer, Christopher A. Haiman, Laurence K. Kolonel, Brian E. Henderson, Loic Le Marchand, Paul Brennan, Suleeporn Sangrajrang, Valerie Gaborieau, Fabrice Odefrey, Chen-Yang Y. Shen, Pei-Ei E. Wu, Hui-Chun C. Wang, Diana Eccles, D. Gareth Evans, Julian Peto, Olivia Fletcher, Nichola Johnson, Sheila Seal, Michael R. Stratton, Nazneen Rahman, Georgia Chenevix-Trench, Stig E. Bojesen, Brge G. Nordestgaard, Christen K. Axelsson, Montserrat Garcia-Closas, Louise Brinton, Stephen Chanock, Jolanta Lissowska, Beata Peplonska, Heli Nevanlinna, Rainer Fagerholm, Hannaleena Eerola, Daehee Kang, Keun-Young Y. Yoo, Dong-Young Y. Noh, Sei-Hyun H. Ahn, David J.

Hunter, Susan E. Hankinson, David G. Cox, Per Hall, Sara Wedren, Jianjun Liu, Yen-Ling L. Low, Natalia Bogdanova, Peter Schrmann, Thilo Drk, Rob A. E. M. Tollenaar, Catharina E. Jacobi, Peter Devilee, Jan G. M. Klijn, Alice J. Sigurdson, Michele M. Doody, Bruce H. Alexander, Jinghui Zhang, Angela Cox, Ian W. Brock, Gordon MacPherson, Malcolm W. R. Reed, Fergus J. Couch, Ellen L. Goode, Janet E. Olson, Hanne Meijers-Heijboer, Ans van den Ouweland, Andr Uitterlinden, Fernando Rivadeneira, Roger L. Milne, Gloria Ribas, Anna Gonzalez-Neira, Javier Benitez, John L. Hopper, Margaret McCredie, Melissa Southey, Graham G. Giles, Chris Schroen, Christina Justenhoven, Hiltrud Brauch, Ute Hamann, Yon-Dschun D. Ko, Amanda B. Spurdle, Jonathan Beesley, Xiaoqing Chen, kConFab, AOCS Management Group, Arto Mannermaa, Veli-Matti M. Kosma, Vesa Kataja, Jaana Hartikainen, Nicholas E. Day, David R. Cox, and Bruce A. J. Ponder. "Genome-wide association study identifies novel breast cancer susceptibility loci." *Nature*, **447**(7148):1087–93, 6 2007.

[Esk08] Eleazar Eskin. "Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information." *Genome Res*, **18**(4):653–60, 4 2008.

[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.

[Ewe04] W. J. Ewens. *Mathematical population genetics*. Springer, 2 edition, 2004.

[Fis18] S. R. A. Fisher. "The correlation between relatives on the supposition of Mendelian inheritance." *Trans R Soc Edinb*, **52**:399–433, 1918.

[Fle93] J. L. Fleiss. "The statistical basis of meta-analysis." *Stat Methods Med Res*, **2**(2):121–45, 1993.

[FM96] Dogulas Scott Falconer and Trudy F.C C. Mackay. *Introduction to Quantitative Genetics*. Longman, 4, revised, illustrated edition, 1996.

[FWD12] Jingyuan Fu, Marcel G. M. Wolfs, Patrick Deelen, Harm-Jan J. Westra, Rudolf S. N. Fehrmann, Gerard J. Te Meerman, Wim A. Buurman, Sander S. M. Rensen, Harry J. M. Groen, Rinse K. Weersma, Leonard H. van den Berg, Jan Veldink, Roel A. Ophoff, Harold Snieder, David van Heel, Ritsert C. Jansen, Marten H. Hofker, Cisca Wijmenga, and Lude Franke. "Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression." *PLoS Genet*, **8**(1):e1002431, 1 2012.

[FWW04] Nicola S. Fearnhead, Jennifer L. Wilding, Bruce Winney, Susan Tonks, Sylvia Bartlett, David C. Bicknell, Ian P. M. Tomlinson, Neil J. McC Mortensen, and Walter F. Bodmer. "Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas." *Proc Natl Acad Sci U S A*, **101**(45):15992–7, 11 2004.

[GGS08] Ivan P. Gorlov, Olga Y. Gorlova, Shamil R. Sunyaev, Margaret R. Spitz, and Christopher I. Amos. "Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms." *Am J Hum Genet*, **82**(1):100–12, 1 2008.

[GLB09] Weihua Guan, Liming Liang, Michael Boehnke, and Gonalo R. Abecasis. "Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies." *Genet Epidemiol*, **33**(6):508–17, 1 2009.

[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.

[HFS12] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Goncalo R. Abecasis. "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." *Nat Genet*, **44**(8):955–9, 8 2012.

[HHH87] M. Hooper, K. Hardy, A. Handyside, S. Hunter, and M. Monk. "HPRT-deficient (Lesch-Nyhan) mouse embryos derived from germline colonization by cultured cells." *Nature*, **326**(6110):292–5, 1987.

[Hin79] D. V. Hinkley. *Theoretical statistics*. CRC Press, Boca Raton, illustrated, reprint edition, 1979.

[HKE09a] B. Han, H. M. Kang, and E. Eskin. “Rapid and accurate multiple testing correction and power estimation for millions of correlated markers.” *PLoS Genet*, **5**(4), 2009.

[HKE09b] 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.

[HKS08] B. Han, H. M. Kang, M. S. Seo, N. Zaitlen, and E. Eskin. “Efficient association study design via power-optimized tag SNP selection.” *Ann Hum Genet*, **72**(Pt 6):834–47, 11 2008.

[HSE13] Buhm. Han, Jae Hoon Sul, Eleazar Eskin, Paul I. W. de Bakker, and Soumya Raychaudhuri. “A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping.” 2013.

[HT96] R. J. HARDY and S. G. THOMPSON. “A likelihood approach to meta-analysis with random effects.” *Statistics in Medicine*, **15**(6):619–629, 1996.

[HT02] J. Higgins and S. G. Thompson. “Quantifying heterogeneity in a meta-analysis.” *Statistics in medicine*, **21**(11):1539–1558, 2002.

[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.

[IPE07a] John P. A. Ioannidis, Nikolaos A. Patsopoulos, and Evangelos Evangelou. “Heterogeneity in meta-analyses of genome-wide association investigations.” *PLoS One*, **2**(9):e841, 2007.

[IPE07b] John P. A. Ioannidis, Nikolaos A. Patsopoulos, and Evangelos Evangelou. “Uncertainty in heterogeneity estimates in meta-analyses.” *BMJ*, **335**(7626):914–6, 11 2007.

[IWS05] Rafael A. Irizarry, Daniel Warren, Forrest Spencer, Irene F. Kim, Shyam Biswal, Bryan C. Frank, Edward Gabrielson, Joe G. N. Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C. Hilmer, Eric Hoffman, Anne E. Jedlicka, Ernest Kawasaki, Francisco Martnez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye, and Wayne Yu. “Multiple-laboratory comparison of microarray platforms.” *Nat Methods*, **2**(5):345–50, 5 2005.

[JFO08] Weizhen Ji, Jia Nee Foo, Brian J. O'Roak, Hongyu Zhao, Martin G. Larson, David B. Simon, Christopher Newton-Cheh, Matthew W. State, Daniel Levy, and Richard P. Lifton. "Rare independent mutations in renal salt handling genes contribute to blood pressure variation." *Nature genetics*, **40**(5):592–9, 5 2008.

[JRV08] Eveliina Jakkula, Karola Rehnstrm, Teppo Varilo, Olli P. H. Pietilinen, Tiina Paunio, Nancy L. Pedersen, Ulf deFaire, Marjo-Riitta R. Jrvelin, Juha Saharinen, Nelson Freimer, Samuli Ripatti, Shaun Purcell, Andrew Collins, Mark J. Daly, Aarno Palotie, and Leena Peltonen. "The genome-wide patterns of variation expose significant substructure in a founder population." *Am J Hum Genet*, **83**(6):787–94, 12 2008.

[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.

[KK04] T. Kariya and H. Kurata. *Generalized least squares*. John Wiley & Sons Inc, 2004.

[KMG08] Sekar Kathiresan, Olle Melander, Candace Guiducci, Aarti Surti, Nol P. Burtt, Mark J. Rieder, Gregory M. Cooper, Charlotta Roos, Benjamin F. Voight, Aki S. Havulinna, Bjrn Wahlstrand, Thomas Hedner, Dolores Corella, E. Shyong Tai, Jose M. Ordovas, Gran Berglund, Erkki Vartiainen, Pekka Jousilahti, Bo Hedblad, Marja-Riitta R. Taskinen, Christopher Newton-Cheh, Veikko Salomaa, Leena Peltonen, Leif Groop, David M. Altshuler, and Marju Orho-Melander. "Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans." *Nat Genet*, **40**(2):189–97, 2 2008.

[KPS07] Gregory V. Kryukov, Len A. Pennacchio, and Shamil R. Sunyaev. "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies." *Am J Hum Genet*, **80**(4):727–39, 4 2007.

[KRB89] Bat-sheva . S. Kerem, Johanna M. Rommens, Janet A. Buchanan, Danuta Markiewicz, Tara K. Cox, Aravinda Chakravarti, Manuel Buchwald, and Lap-Chee . C. Tsui. "Identification of the cystic fibrosis gene: genetic analysis." *Science*, **245**(4922):1073–1080, 1989.

[KX09] Seyoung Kim and Eric P. Xing. “Statistical estimation of correlated genome associations to a quantitative trait network.” *PLoS Genet*, **5**(8):e1000587, 8 2009.

[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.

[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.

[Lan02] K. Lange. *Mathematical and statistical methods for genetic analysis*. Springer, 2002.

[LCP08] Hana Lango, UK Type 2 Diabetes Genetics Consortium, Colin N. A. Palmer, Andrew D. Morris, Eleftheria Zeggini, Andrew T. Hattersley, Mark I. McCarthy, Timothy M. Frayling, and Michael N. Weedon. “Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk.” *Diabetes*, **57**(11):3129–35, 11 2008.

[LL08] Bingshan Li and Suzanne M. Leal. “Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.” *Am J Hum Genet*, **83**(3):311–21, 9 2008.

[LLK12] Jennifer Listgarten, Christoph Lippert, Carl M. Kadie, Robert I. Davidson, Eleazar Eskin, and David Heckerman. “Improved linear mixed models for genome-wide association studies.” *Nature methods*, **9**(6):525–526, 2012.

[LMP09] Jennifer K. Lowe, Julian B. Maller, Itsik Pe’er, Benjamin M. Neale, Jacqueline Salit, Eimear E. Kenny, Jessica L. Shea, Ralph Burkhardt, J. Gustav Smith, Weizhen Ji, Martha Noel, Jia Nee Foo, Maude L. Blundell, Vita Skilling, Laura Garcia, Marcia L. Sullivan, Heather E. Lee, Anna Labek, Hope Ferdowsian, Steven B. Auerbach, Richard P. Lifton, Christopher Newton-Cheh, Jan L. Breslow, Markus Stoffel, Mark J. Daly, David M. Altshuler, and Jeffrey M. Friedman. “Genome-wide association studies in an isolated founder population from the Pacific Island of Kosrae.” *PLoS Genet*, **5**(2):e1000365, 2 2009.

[LR99] M. Lynch and K. Ritland. “Estimation of pairwise relatedness with molecular markers.” *Genetics*, **152**(4):1753–66, 8 1999.

[LS94] E. S. Lander and N. J. Schork. “Genetic dissection of complex traits.” *Science*, **265**(5181):2037–48, 9 1994.

[LS09] Dan-Yu Y. Lin and Patrick F. Sullivan. "Meta-analysis of genome-wide association studies with overlapping subjects." *Am J Hum Genet*, **85**(6):862–72, 12 2009.

[LSN95] B. A. Loiselle, V. L. Sork, J. Nason, and C. Graham. "Spatial genetic structure of a tropical understory shrub, Psychotria officinalis (Rubiaceae)." *American Journal of Botany*, pp. 1420–1425, 1995.

[LTP09] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome Biol*, **10**(3):R25, 2009.

[LW98] Michael Lynch and Bruce Walsh. *Genetics and analysis of quantitative traits*. Sinauer, Sunderland, Mass., illustrated edition, 1998.

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

[MB09] Bo Eskerod Madsen and Sharon R. Browning. "A groupwise association test for rare mutations using a weighted sum statistic." *PLoS Genet*, **5**(2):e1000384, 2 2009.

[MBC08] Teri A. Manolio, Lisa D. Brooks, and Francis S. Collins. "A HapMap harvest of insights into the genetics of common disease." *J Clin Invest*, **118**(5):1590–605, 5 2008.

[McC03] Charles E. McCulloch. *Generalized linear mixed models*. Institute of Mathematical Statistics ; Alexandria, Va. : American Statistical Association,, Beachwood, Ohio, 2003.

[MCC09a] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, and Aravinda Chakravarti. "Finding the missing heritability of complex diseases." *Nature*, **461**(7265):747–753, 2009.

[MCC09b] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. "Finding the missing heritability of complex diseases." *Nature*, **461**(7265):747–53, 10 2009.

[MDC05] Jonathan Marchini, Peter Donnelly, and Lon R. Cardon. "Genome-wide strategies for detecting multiple loci that influence complex diseases." *Nat Genet*, **37**(4):413–7, 4 2005.

[MH59] N. MANTEL and W. HAENSZEL. "Statistical aspects of the analysis of data from retrospective studies of disease." *J Natl Cancer Inst*, **22**(4):719–48, 4 1959.

[MHB10] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." *Genome Res*, **20**(9):1297–303, 9 2010.

[MHM07] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. "A new multipoint method for genome-wide association studies by imputation of genotypes." *Nat Genet*, **39**(7):906–13, 7 2007.

[Mil03] Brook G. Milligan. "Maximum-likelihood estimation of relatedness." *Genetics*, **163**(3):1153–67, 3 2003.

[MS00] M. S. McPeek and L. Sun. "Statistical tests for detection of misspecified relationships by use of genome-screen data." *Am J Hum Genet*, **66**(3):1076–94, 3 2000.

[MT07] Stephan Morgenthaler and William G. Thilly. "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST)." *Mutat Res*, **615**(1-2):28–56, 2 2007.

[NAM01] D. L. Newman, M. Abney, M. S. McPeek, C. Ober, and N. J. Cox. "The importance of genealogy in determining genetic associations with complex traits." *Am J Hum Genet*, **69**(5):1146–8, 11 2001.

[NH03] Pauline C. Ng and Steven Henikoff. "SIFT: Predicting amino acid changes that affect protein function." *Nucleic Acids Res*, **31**(13):3812–4, 7 2003.

[NHW07] Sergey Nejentsev, Joanna M. M. Howson, Neil M. Walker, Jeffrey Szeszko, Sarah F. Field, Helen E. Stevens, Pamela Reynolds, Matthew Hardy, Erna King, Jennifer Masters, John Hulme, Lisa M. Maier, Deborah Smyth, Rebecca Bailey, Jason D. Cooper, Gloria Ribas, R. Duncan Campbell, David G. Clayton, John A. Todd, and Wellcome Trust Case Control Consortium. "Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A." *Nature*, **450**(7171):887–92, 12 2007.

[NJB08] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltn Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. “Genes mirror geography within Europe.” *Nature*, **456**(7218):98–101, 11 2008.

[NS08] John Novembre and Matthew Stephens. “Interpreting principal component analyses of spatial population genetic variation.” *Nat Genet*, **40**(5):646–9, 5 2008.

[OAM01] C. Ober, M. Abney, and M. S. McPeek. “The genetic dissection of complex traits in a founder population.” *Am J Hum Genet*, **69**(5):1068–79, 11 2001.

[PBL10] Enrico Petretto, Leonardo Bottolo, Sarah R. Langley, Matthias Heinig, Chris McDermott-Roe, Rizwan Sarwar, Michal Pravenec, Norbert Hübner, Timothy J. Aitman, Stuart A. Cook, and Sylvia Richardson. “New insights into the genetic control of gene expression using a Bayesian multi-tissue approach.” *PLoS Comput Biol*, **6**(4):e1000737, 4 2010.

[PC02] Jonathan K. Pritchard and Nancy J. Cox. “The allelic architecture of human disease genes: common disease-common variant... or not?” *Hum Mol Genet*, **11**(20):2417–23, 10 2002.

[PCS06] Giuseppe Pilia, Wei-Min M. Chen, Angelo Scuteri, Marco Orr, Giuseppe Albai, Mariano Dei, Sandra Lai, Gianluca Usala, Monica Lai, Paola Loi, Cinzia Mameli, Loredana Vacca, Manila Deiana, Nazario Olla, Marco Masala, Antonio Cao, Samer S. Najjar, Antonio Terracciano, Timur Nedorezov, Alexei Sharov, Alan B. Zonderman, Gonalo R. Abecasis, Paul Costa, Edward Lakatta, and David Schlessinger. “Heritability of cardiovascular and personality traits in 6,148 Sardinians.” *PLoS Genet*, **2**(8):e132, 8 2006.

[PDC04] Petko M. Petkov, Yueming Ding, Megan A. Cassell, Weidong Zhang, Gunjan Wagner, Evelyn E. Sargent, Steven Asquith, Victor Crew, Kevin A. Johnson, Phil Robinson, Valerie E. Scott, and Michael V. Wiles. “An efficient SNP system for mouse genome scanning and elucidating strain relationships.” *Genome Res*, **14**(9):1806–11, 9 2004.

[PKB10] Alkes L. Price, Gregory V. Kryukov, Paul I. W. de Bakker, Shaun M. Purcell, Jeff Staples, Lee-Jen J. Wei, and Shamil R. Sunyaev. “Pooled association tests for rare variants in exon-resequencing studies.” *Am J Hum Genet*, **86**(6):832–8, 6 2010.

[PNT07] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. "PLINK: a tool set for whole-genome association and population-based linkage analyses." *Am J Hum Genet*, **81**(3):559–75, 9 2007.

[PPP06] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. "Principal components analysis corrects for stratification in genome-wide association studies." *Nat Genet*, **38**(8):904–9, 8 2006.

[PPR06] Nick Patterson, Alkes L. Price, and David Reich. "Population structure and eigenanalysis." *PLoS Genet*, **2**(12):e190, 12 2006.

[Pri01] J. K. Pritchard. "Are rare variants responsible for susceptibility to complex diseases?" *Am J Hum Genet*, **69**(1):124–137, 2001.

[PSR00] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. "Association mapping in structured populations." *Am J Hum Genet*, **67**(1):170–81, 7 2000.

[Ran69] P. Rantakallio. "Groups at risk in low birth weight infants and perinatal mortality." *Acta Paediatr Scand*, **193**:Suppl 193:1+, 1969.

[RD02] Nalini Ravishanker and Dey Dipak. *A first course in linear model theory*. CRC Press, illustrated edition, 2002.

[Rit09] K. Ritland. "Estimators for pairwise relatedness and individual inbreeding coefficients." *Genetics Research*, **67**(02):175–185, 2009.

[RPF07] Stefano Romeo, Len A. Pennacchio, Yunxin Fu, Eric Boerwinkle, Anne Tybjaerg-Hansen, Helen H. Hobbs, and Jonathan C. Cohen. "Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL." *Nature genetics*, **39**(4):513–6, 4 2007.

[RS09] Cyril S. Rakovski and Daniel O. Stram. "A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors." *PLoS One*, **4**(6):e5825, 2009.

[SB09] Matthew Stephens and David J. Balding. "Bayesian statistical methods for genetic association studies." *Nat Rev Genet*, **10**(10):681–90, 10 2009.

[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.

[SHH11] Jae Hoon Sul, Buhm Han, Dan He, and Eleazar Eskin. "An optimal weighted aggregated association test for identification of rare variants involved in common diseases." *Genetics*, **188**(1):181–8, 5 2011.

[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.

[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. Järvelin, 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.

[Sto02] J. D. Storey. "A direct approach to false discovery rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3):479–498, 2002.

[TDY06] S. V. Tavtigian, A. M. Deffenbaugh, L. Yin, T. Judkins, T. Scholl, P. B. Samollow, D. de Silva, A. Zharkikh, and A. Thomas. "Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral." *J Med Genet*, **43**(4):295–305, 4 2006.

[TH00] S. C. Thomas and W. G. Hill. "Estimating quantitative genetic parameters using sibships reconstructed from marker data." *Genetics*, **155**(4):1961–72, 8 2000.

[TJK09] Gilles Thomas, Kevin B. Jacobs, Peter Kraft, Meredith Yeager, Sholom Wacholder, David G. Cox, Susan E. Hankinson, Amy Hutchinson, Zhaoming Wang, Kai Yu, Nilanjan Chatterjee, Montserrat Garcia-Closas, Jesus Gonzalez-Bosquet, Ludmila Prokunina-Olsson, Nick Orr, Walter C. Willett, Graham A. Colditz, Regina G. Ziegler, Christine D. Berg, Saundra S. Buys, Catherine A. McCarty, Heather Spencer Feigelson, Eugenia E. Calle, Michael J. Thun, Ryan Diver, Ross Prentice, Rebecca Jackson, Charles Kooperberg, Rowan Chlebowski, Jolanta Lissowska, Beata Peplonska, Louise A. Brinton, Alice Sigurdson, Michele Doody, Parveen Bhatti, Bruce H. Alexander, Julie Buring, I-Min M. Lee, Lars J. Vatten, Kristian Hveem, Merethe Kumle, Richard B. Hayes, Margaret Tucker, Daniela S. Gerhard, Joseph F. Fraumeni, Robert N. Hoover, Stephen J. Chanock, and David J. Hunter. "A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1)." *Nat Genet*, **41**(5):579–84, 3 2009.

[TM07] Timothy Thornton and Mary Sara McPeek. "Case-control association testing with related individuals: a more powerful quasi-likelihood score test." *Am J Hum Genet*, **81**(2):321–37, 8 2007.

[TOB09] Sean V. Tavtigian, Peter J. Oefner, Davit Babikyan, Anne Hartmann, Sue Healey, Florence Le Calvez-Kelm, Fabienne Lesueur, Graham B. Byrnes, Shu-Chun C. Chuang, Nathalie Forey, Corinna Feuchtinger, Lydie Gioia, Janet Hall, Mia Hashibe, Barbara Herte, Sandrine McKay-Chopin, Alun Thomas, Maxime P. Vallée, Catherine Voegele, Penelope M. Webb, David C. Whiteman, Australian Cancer Study, Breast Cancer Family Registries (BCFR), Kathleen Cuningham Foundation Consortium for Research into Familial Aspects of Breast Cancer (kConFab), Suleeporn Sangrajrang, John L. Hopper, Melissa C. Southey, Irene L. Andrulis, Esther M. John, and Georgia Chenevix-Trench. "Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer." *Am J Hum Genet*, **85**(4):427–46, 10 2009.

[VP04] Teppo Varilo and Leena Peltonen. "Isolates and their potential use in complex gene mapping efforts." *Curr Opin Genet Dev*, **14**(3):316–23, 6 2004.

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

[WAH06] Bruce S. Weir, Amy D. Anderson, and Amanda B. Hepler. "Genetic relatedness analysis: modern data and new challenges." *Nat Rev Genet*, **7**(10):771–80, 10 2006.

[Wal07] Francis O. Walker. "Huntington's disease." *Lancet*, **369**(9557):218–28, 1 2007.

[Was04] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 9 2004.

[WMM08] Tom Walsh, Jon M. McClellan, Shane E. McCarthy, Anjené M. Addington, Sarah B. Pierce, Greg M. Cooper, Alex S. Nord, Mary Kusenda, Dheeraj Malhotra, Abhishek Bhandari, Sunday M. Stray, Caitlin F.

Rippey, Patricia Roccanova, Vlad Makarov, B. Lakshmi, Robert L. Findling, Linmarie Sikich, Thomas Stromberg, Barry Merriman, Nitin Gogtay, Philip Butler, Kristen Eckstrand, Laila Noory, Peter Gochman, Robert Long, Zugen Chen, Sean Davis, Carl Baker, Evan E. Eichler, Paul S. Meltzer, Stanley F. Nelson, Andrew B. Singleton, Ming K. Lee, Judith L. Rapoport, Mary-Claire C. King, and Jonathan Sebat. "Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia." *Science*, **320**(5875):539–43, 4 2008.

[Wri31] S. Wright. "Evolution in Mendelian Populations." *Genetics*, **16**(2):97–159, 3 1931.

[WT98] A. S. Whittemore and I. P. Tu. "Simple, robust linkage tests for affected sibs." *Am J Hum Genet*, **62**(5):1228–42, 5 1998.

[XRL08] B. Xu, J. L. Roos, S. Levy, E. J. Van Rensburg, J. A. Gogos, and M. Karayiorgou. "Strong association of de novo copy number mutations with sporadic schizophrenia." *Nature genetics*, **40**(7):880–885, 2008.

[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.

[ZAK07] Keyan Zhao, Mara Jos 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.

[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.

[ZS12] Xiang Zhou and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies." *Nature genetics*, **44**(7):821–824, 2012.

[ZSS08] Eleftheria Zeggini, Laura J. Scott, Richa Saxena, Benjamin F. Voight, Jonathan L. Marchini, Tianle Hu, Paul I. W. de Bakker, Gonalo R. Abecasis, Peter Almgren, Gitte Andersen, Kristin Ardlie, Kristina Bengtsson Bostrm, Richard N. Bergman, Lori L. Bonnycastle, Knut Borch-Johnsen, Nol P. Burtt, Hong Chen, Peter S. Chines, Mark J. Daly, Parimal Deodhar, Chia-Jen J. Ding, Alex S. F. Doney, William L. Duren, Katherine S. Elliott, Michael R. Erdos, Timothy M. Frayling, Rachel M. Freathy, Lauren Gianniny, Harald Grallert, Niels Grarup, Christopher J. Groves, Candace Guiducci, Torben Hansen, Christian Herder, Graham A. Hitman, Thomas E. Hughes, Bo Isomaa, Anne U. Jackson, Torben Jrgensen, Augustine Kong, Kari Kubalanza, Finny G. Kuruvilla, Johanna Kuusisto, Claudia Langenberg, Hana Lango, Torsten Lauritzen, Yun Li, Cecilia M. Lindgren, Valeriya Lyssenko, Amanda F. Marvelle, Christa Meisinger, Kristian Midthjell, Karen L. Mohlke, Mario A. Morken, Andrew D. Morris, Narisu Narisu, Peter Nilsson, Katharine R. Owen, Colin N. A. Palmer, Felicity Payne, John R. B. Perry, Elin Pettersen, Carl Platou, Inga Prokopenko, Lu Qi, Li Qin, Nigel W. Rayner, Matthew Rees, Jeffrey J. Roix, Anelli Sandbaek, Beverley Shields, Marketa Sjgren, Valgerdur Steinthorsdottir, Heather M. Stringham, Amy J. Swift, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Nicholas J. Timpson, Tiinamaija Tuomi, Jaakko Tuomilehto, Mark Walker, Richard M. Watanabe, Michael N. Weedon, Cristen J. Willer, Wellcome Trust Case Control Consortium, Thomas Illig, Kristian Hveem, Frank B. Hu, Markku Laakso, Kari Stefansson, Oluf Pedersen, Nicholas J. Wareham, Inłs Barroso, Andrew T. Hattersley, Francis S. Collins, Leif Groop, Mark I. McCarthy, Michael Boehnke, and David Altshuler. "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." *Nat Genet*, **40**(5):638–45, 5 2008.