1-1-2016

Sparse Simultaneous Signal Detection With Applications in Genomics

Julie Kobie
University of Pennsylvania, jkobie01@gmail.com

Sparse Simultaneous Signal Detection With Applications in Genomics

Abstract
Studying complex diseases, such as autoimmune diseases, can lead to the detection of pleiotropic loci with otherwise small effects. Through the detection of pleiotropic loci the genetic architecture of these complex diseases can be better defined, allowing for subsequent improvements in their treatment and prevention efforts. Here, we investigate the genetic relatedness of complex diseases through the detection and quantification of simultaneous disease-associated genetic variants using genome-wide association study (GWAS) data. We propose two max-type statistics, with and without an added level of dependency on the directions of the genetic effects, that globally test whether a pair of complex diseases shares at least one disease-associated genetic variant. The proposed global tests are based on the simultaneity of complex disease-associated genetic variants, allowing for the determination of exact p-values from a permutation distribution assuming independence. While an independence assumption is often imposed on genetic variants, we propose a perturbation procedure for evaluating the statistical significance of one of the proposed global tests, preserving the inherent dependency structure among genetic variants. We extend that global test beyond the detection of genetic relatedness at identical genetic variants to the detection of genetic relatedness within dependency-defined windows across the genome. With the proposed methods we identify pairs of pediatric autoimmune diseases (pAIDs) that exhibit evidence of genetic sharing, such as Crohn's disease and ulcerative colitis.

We then characterize the detected genetic sharing between a pair of complex diseases through the quantification of shared disease-associated genetic variants using GWAS data. We develop a quantification measure as a function of standardized variant effect sizes, adjusted for the total number of genetic variants and varied GWAS sample size. The quantification measure acts as an estimate of the genetic correlation among shared disease-associated genetic variants. We use a bootstrapping procedure to estimate the properties of our quantification measure. In applying the developed measure to pAID GWAS we observe similar trends in relatedness among pAIDs pairs.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Epidemiology & Biostatistics

First Advisor
Hongzhe Li

Second Advisor
Nandita Mitra

SPARSE SIMULTANEOUS SIGNAL DETECTION WITH APPLICATIONS IN GENOMICS

Julie Kobie

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

_____

Hongzhe Li, Professor of Biostatistics

Graduate Group Chairperson

_____

John H. Holmes, Professor of Medical Informatics in Epidemiology

Dissertation Committee

Nandita Mitra, Associate Professor of Biostatistics

Sarah J. Ratcliffe, Associate Professor of Biostatistics

Hakon Hakonarson, Associate Professor of Pediatrics

SPARSE SIMULTANEOUS SIGNAL DETECTION WITH APPLICATIONS IN GENOMICS

ACKNOWLEDGEMENT

First and foremost, I want to thank my advisor Hongzhe Li, without whom this work would not have been possible. I am deeply grateful for his patience and willingness to devote his time and expertise to me throughout this dissertation, and my graduate education. I also want to thank my committee chair Nandita Mitra for her invaluable mentorship throughout my graduate education and for providing me with collaborative opportunities I wouldn't have had otherwise. I would like to thank committee member Sarah Ratcliffe for her thoughtful comments on the methodology and interpretation of this dissertation and for her role in facilitating the sharing of computing resources, including this LaTeX dissertation template.

I would like to thank our clinical collaborator and committee member Hakon Hakonarson for his clinical expertise, especially with respect to the pediatric autoimmune disease (pAID) genome-wide association study (GWAS) data motivating, and used throughout, this dissertation. I would also like to thank Yun Rose Li for providing the pAID GWAS data and sharing her extensive knowledge of the data itself, and beyond, when needed. I want to thank Sihai Dave Zhao for his input in getting this dissertation off the ground, his positive attitude, and enthusiasm in statistical genomics research. I would also like to thank Iuliana Ionita-Laza for providing her functional annotation data and taking the time to answer any questions that came up along the way.

I must thank the biostatistics faculty and other biomedical faculty members, from whom I have learned a great deal. I especially want to thank Benjamin French and Jinbo Chen who were engaging and effective teachers, largely impacting my graduate education. I also must thank the biostatistics staff, namely Marissa Fox and Cathy Vallejo for all of their help behind the scenes. I without a doubt must thank the computing staff, specifically computing wizards Curt Calafut, Anand Srinivasan and Clay Wells for their personal attention and sometimes constant email correspondence. I also want to thank past and present biostatistics graduate students for their unwavering support and strong sense of community. I am especially thankful for the mentorship and guidance of Kay See Tan and Jarcy Zee, the statistical expertise of Edward Kennedy and the computing expertise of Cheng Jia.

Lastly, this wouldn't have been possible without the love and support of my family and friends. My parents, Linda and Gary Kobie, and fiancé, Eric Etnier, have never stopped believing in me and

have always encouraged me to keep going.

I also want to acknowledge the Ophthalmic Statistical Genetics and Bioinformatics Training Grant, T32-EY021451, and NIH R01-GM097505, both of which funded this dissertation research.

ABSTRACT

SPARSE SIMULTANEOUS SIGNAL DETECTION WITH APPLICATIONS IN GENOMICS

Julie Kobie

Hongzhe Li

Studying complex diseases, such as autoimmune diseases, can lead to the detection of pleiotropic loci with otherwise small effects. Through the detection of pleiotropic loci the genetic architecture of these complex diseases can be better defined, allowing for subsequent improvements in their treatment and prevention efforts. Here, we investigate the genetic relatedness of complex diseases through the detection and quantification of simultaneous disease-associated genetic variants using genome-wide association study (GWAS) data. We propose two max-type statistics, with and without an added level of dependency on the directions of the genetic effects, that globally test whether a pair of complex diseases shares at least one disease-associated genetic variant. The proposed global tests are based on the simultaneity of complex disease-associated genetic variants, allowing for the determination of exact p-values from a permutation distribution assuming independence. While an independence assumption is often imposed on genetic variants, we propose a perturbation procedure for evaluating the statistical significance of one of the proposed global tests, preserving the inherent dependency structure among genetic variants. We extend that global test beyond the detection of genetic relatedness at identical genetic variants to the detection of genetic relatedness within dependency-defined windows across the genome. With the proposed methods we identify pairs of pediatric autoimmune diseases (pAIDs) that exhibit evidence of genetic sharing, such as Crohn's disease and ulcerative colitis.

We then characterize the detected genetic sharing between a pair of complex diseases through the quantification of shared disease-associated genetic variants using GWAS data. We develop a quantification measure as a function of standardized variant effect sizes, adjusted for the total number of genetic variants and varied GWAS sample size. The quantification measure acts as an estimate of the genetic correlation among shared disease-associated genetic variants. We use a bootstrapping procedure to estimate the properties of our quantification measure. In applying the developed measure to pAID GWAS we observe similar trends in relatedness among pAIDs pairs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION

Understanding the shared genetic architecture of complex diseases is key for improving the efficiency and effectiveness of treatments by aiding in the detection and identification of common therapeutic mechanisms. Complex diseases are multifactorial disorders with largely unknown etiologies. Unlike Mendelian diseases caused by a single genetic mutation, complex diseases are likely caused by a combination of genetic mutations, affecting multiple genes, coupled with various lifestyle and environmental factors. The detection and identification of shared genetic risk factors have become important strategies in the study of complex disease groups, such as autoimmune diseases and psychiatric disorders.

Autoimmune diseases affect 8% of Americans and represent a leading cause of death and chronic disability, further burdening our health care system (Li et al., 2015a). Their high rates of familial clustering and comorbidities are evidence of a shared genetic architecture, underlying the disease etiologies (Cooper, Bynum, and Somers, 2009; Li et al., 2015b). The shared genetic architecture of autoimmune diseases is thought to be driven by both pleiotropic disease-associated single nucleotide polymorphisms (SNPs), acting via shared mechanisms, and the inherent polygenic nature of complex diseases (Chung et al., 2014). Pleiotropic SNPs are genetic mutations at a single locus with the ability to contribute to multiple disease phenotypes, like the clinically-distinct disease subtypes observed and classified as autoimmune diseases, while polygenicity is the influence of many genes on the observed disease phenotype. Pleiotropic effects and the polygenic nature of complex diseases present challenges in the detection and identification of disease-associated SNPs.

Genome-wide association studies (GWAS) have proven to be effective in identifying thousands of complex disease-associated SNPs (Hindorff et al., 2009), though the identified SNPs only explain a small proportion of complex disease heritability (Manolio et al., 2009). This phenomenon, dubbed “the missing heritability” (Manolio et al., 2009), can be explained by the limited sample size of most GWAS (Yang et al., 2010). Most GWAS are not powered to detect disease-associated SNPs of a polygenic architecture, as these SNPs typically have small effects and, given the limited sample size, are too weak to pass genome-wide significance (Chung et al., 2014; Yang et al., 2010). That

being said, the value of GWAS in the study of complex diseases is not lost. With increased sample size GWAS have the potential to detect and identify complex disease-associated SNPs that go on to explain a larger proportion of complex disease heritability (Lee et al., 2012a; Yang et al., 2011b). Pooling independent GWAS with standard approaches for meta-analysis increase sample size and result in improved power, but the standard approaches are not optimal for the study of complex disease sets whose GWAS are heterogeneous (Bhattacharjee et al., 2012).

In this dissertation we develop methods to investigate the genetic relatedness within complex disease sets through the detection and characterization of shared disease-associated SNPs among pairs of clinically-distinct disease subtypes of the set. We take an integrative approach, combining complex disease GWAS pairwise and collapsing over variants spanning the entire genome. In studying complex disease sets, rather than restricting our analyses to a single disease subtype, we have the opportunity to discover how SNPs associate with more than one complex disease subtype likely through pleiotropic effects and shared genetic architectures (Bhattacharjee et al., 2012). Our methods are largely motivated by real GWAS data from the Center for Applied Genomics of the Children's Hospital of Philadelphia of ten clinically-distinct pediatric autoimmune diseases: thyroiditis (THY), spondyloarthropathy (SPA), psoriasis (PSOR), celiac disease (CEL), systemic lupus erythematosus (SLE), common variable immunodeficiency (CVID), ulcerative colitis (UC), type I diabetes (T1D), juvenile idiopathic arthritis (JIA) and Crohn's disease (CD) (Li et al., 2015a,b). Genome-wide data analyses of this dissertation exclude variants within the major histocompatibility complex (MHC), already found to be shared among pAID pairs (Li et al., 2015b).

1.1. Background

Statisticians, geneticists and biologists alike have contributed to the development of methods for the study of complex disease sets. Here we review methods in the detection, identification and quantification of shared genetic variants across complex diseases and traits. Many of these methods share a common goal, that is to gain a better understanding of complex disease genetic etiology with real GWAS applications in autoimmune diseases, psychiatric disorders and other complex traits. Psychiatric disorders affect more than 4% of American adults and, like autoimmune diseases, their overlapping, nonspecific symptom patterns, leading to blurred boundaries of clinical diagnosis, point to the presence of pleiotropic effects (Cross-disorder Group of the Psychiatric Genomics

Consortium, 2013). As we review existing methods in the detection, identification and quantification of shared genetic SNPs across complex diseases we will also address the many unique features of GWAS data including but not limited to genetic sparsity in a high dimension, a complex dependency structure among genetic variants and differences in the direction of genetic association.

1.1.1. Detection

Methods in detection are often integrative, combining information to give a global assessment. Problems in detection have close ties to methods in quantification, or estimation, explored in Section 1.1.3 and Chapter 4. Here we review the simultaneous signal detection problem that underlies the methods of Chapters 2 and 3.

The simultaneous signal detection problem is a generalization of the one-sample normal mixture detection problem (Jin and Donoho, 2004)

$$H_0 : \ X_i \sim N(0,1)$$
$$H_1 : \ X_i \sim (1-\epsilon_n)N(0,1) + (\epsilon_n)N(\mu_n, 1),$$

where X_i is, for example, GWAS Z-score for the i^{th} SNP, ϵ_n is the proportion of disease-associated SNPs on the high dimensional order of n and μ_n is some nonzero mean. Though detecting the proportion of nonull SNPs, testing whether $H_0 : \ \epsilon_n \neq 0$, is difficult under GWAS data conditions termed the "rare and weak" setting by Donoho and Jin, 2008. GWAS data is sparse, in that the number of disease-associated SNPs make up a small proportion, ϵ_n, of the total number of SNPs, n. And of that small proportion of disease-associated SNPs, most have small to moderate effects, or magnitude of disease association, making the detection and estimation of ϵ_n challenging.

In the one-sample problem there are two integrative-like approaches to testing $H_0 : \ \epsilon_n \neq 0$, one being the sum of square-type test statistic, $\sum_i = 1^n X_i^2$, and the other being the max-type test statistic, $max|X_i|$. Jin and Donoho, 2004 find these tests are suboptimal when compared to likelihood ratio test alternatives, though they often require complete specification of the null and alternative distributions which contain difficult-to-estimate, unknown parameters in this setting. Jin and Donoho, 2004 discuss an alternative procedure called higher criticism that performs as well as the likelihood ratio test, comparing the empirical distribution of Z-scores goodness-of-fit to the standard normal

distribution.

Cai and Jeng, 2011 extended the above to heteroscedastic normal mixtures and mixtures of arbitrary distributions (Cai and Wu, 2014). We expand the one-sample normal mixture detection to a two-sample mixture detection problem in the context of simultaneously detecting a shared genetic architecture between a pair of complex diseases. Our method powerfully streamlines the commonly used two-sample enrichment integration approach which relies on strict significance thresholds and SNP identification steps.

1.1.2. Identification

Methods in the identification of shared disease-associated SNPs between a pair of complex diseases provide a more specific account of a detected shared genetic architecture. In identifying specific disease-associated SNPs shared between a pair of complex diseases, researchers can better grasp the underlying biological systems of complex disease etiology (Cai and Tan, 2015), pinpointing shared pathways with the potential for use as therapeutic targets (Li et al., 2015b).

Several methods have introduced meta-like statistics combining summary-level GWAS data of complex disease sets to identify SNPs associated with a subset of the diseases (Bhattacharjee et al., 2012; Cotsapas et al., 2011). Cotsapas et al., 2011 proposed CPMA, Cross Phenotype Meta-Analysis, which detects the association of a SNP to a subset of heterogeneous GWAS. Specifically, CPMA determines evidence for the hypothesis that each SNP has multiple disease-associations, shown by deviations from an expected uniform p-value distribution. CPMA measures a deviation in p-value behavior for one disease conditional on other diseases, instead of testing all possible subsets of complex disease subtypes (Cotsapas et al., 2011). Though in doing so, the CPMA method does not account for the direction of the disease association across complex disease sets. Cotsapas et al., 2011 finds evidence that 44% of immune-mediated disease risk SNPs are association to multiple immune-mediated disease subtypes of the set. The CPMA method is performed at the SNP level and is more practical in applications with a clearly defined set of candidate SNPs.

Similarly, Bhattacharjee et al., 2012 proposes ASSET, *ass*ociation analysis based on sub*sets*, a subset-based association-testing framework that pools the analyses of multiple heterogeneous GWAS at a given SNP. ASSET explores all possible subsets of a complex disease set, or set of traits, to identify the subset with the strongest association signal with a max-type statistic (Bhat-

tacharjee et al., 2012). ASSET has clear advantages over standard meta-analysis approaches with the ability to incorporate prior information and accommodate SNP effects in differing directions. ASSET, like CPMA, is performed at the SNP level, and thus could be computationally intensive as the number of possible subsets grows exponentially as the number of studies in a complex disease set increases.

Evidence of a shared genetic architecture between pairs of complex diseases has prompted the development of pleiotropy-informed, or pleiotropy-enriched, methods (Andreassen et al., 2013; Chung et al., 2014). Pleiotropy-informed methods use GWAS data of one disease as leverage to improve the power of detecting and identifying disease-associated SNPs in another, genetically related, disease. Andreassen et al., 2013 exploited evidence of genetic sharing between schizophrenia (SCZ) and bipolar disorders (BPD) to improve the power of detecting SCZ-associated SNPs. Similarly, Chung et al., 2014 proposed a pleiotropy-informed statistical framework, GPA, building off of a two-group mixture model (Efron, 2008) while allowing for the incorporation of functional annotation, and using the GPA framework, Chung et al., 2014 identified likely polygenic SNPs associated with attention deficit-hyperactivity disorder (ADHD), autism spectrum disorder (ASD), BPD, major depressive disorder (MDD) and SCZ that were not identified in with standard, single-subtype analysis methods. Both pleiotropy-informed methods utilize conditional false discovery rates (cFDR) to prioritize SNPs.

GWAS are subject to multiple testing and SNPs are often prioritized based on whether they exceed a Bonferroni-corrected threshold, also known as genome-wide significance. This method of SNP prioritization is highly conservative, and the false discovery rate (FDR), first introduced by Benjamini and Hochberg, 1995, was introduced as a less conservative prioritization method while still adjusting for multiple comparisons. FDR seeks to reduce the probability of the expected proportion of false discoveries rather than seeking to reduce the probability of at least one false discovery, which is true of making the Bonferroni-correction. The use of a cFDR, conditioning on additional, genetically related disease data, allows for a reduction in FDR and an enrichment in disease-associated SNPs as a function of their association with the conditional disease (Andreassen et al., 2013). Though these methods are not without limitations, they rely on strong distributional assumptions and the performance of GPA is not well documented in a sparse setting (Chung et al., 2014). Andreassen et al., 2013 uses an *ad hoc* pleiotropy-informed approach to show an enrichment of SNPs

associated with SCZ as a function of their association with cardiovascular-disease (CVD).

1.1.3. Quantification

After detecting a shared genetic architecture, the characterization of the shared disease-associated SNPs follows naturally. In addition to obtaining a more specific account of the detected shared disease-associated SNPs with identification methods of Section 1.1.2, methods in the quantification of shared disease-associated SNPs give a more specific account of how much is shared relative to other complex disease pairs. A standardized quantification measure allows for comparison across pairs and subsequent treatment prioritization of complex diseases with a larger overlapping genetic architecture.

The genetic correlation between a pair of complex diseases, or genome-wide aggregate of shared disease-associated SNP effects without imposing thresholding restraints, quantifies a complex disease pair's genetic relatedness. Coheritability is a commonly used estimate of the genetic correlation, estimated using a restricted maximum likelihood approach (REML). Specifically, coheritability is an estimate of the genetic covariance from a bivariate linear mixed model framework divided by the product of the single-disease genetic standard deviation estimates (Lee et al., 2012a). The concept of coheritability as an estimate of the genetic correlation is an extension of the concept of single-disease heritability, or the proportion of variance in a disease phenotype explained by the genetic variation in the population (Yang et al., 2010). Estimates of coheritability rely on individual-level genotype data for the derivation of a genomic similarity relationship matrix (Lee et al., 2012a; Yang et al., 2010). Individual-level genotype data are often difficult to obtain, making the implementation of coheritability estimation challenging for a wide array of complex disease sets.

Polygenic risk scores, developed to predict the risk of complex disease (Wray, Goddard, and Visscher, 2007), have also been used to estimate the genetic correlation between genetically related complex diseases and traits (Dudbridge, 2013). Though like coheritability estimation, the estimation of the genetic correlation in the polygenic risk score framework, too, relies on individual-level genotype data. And while powerful in the presence of null SNPs, the polygenic risk score framework requires some thresholding optimization for SNP inclusion in the risk score model (Dudbridge, 2013). Both the coheritiability and polygenic risk score methods accommodate binary disease outcomes with a liability threshold model, which assumes all individuals have an unobserved, normally

distributed liability trait and links that to the observed binary disease outcomes (Dempster and Lerner, 1950; Dudbridge, 2013; Lee et al., 2012a).

More recently Bulik-Sullivan et al., 2015 proposed a method for estimating the genetic correlation between a pair of complex diseases relying only on readily available summary-level GWAS data rather than individual-level genotype data. Using the linear relationship between SNPs in high linkage disequilibrium (LD) and their corresponding effect sizes, Bulik-Sullivan et al., 2015 models the product of the marginal Z-scores for SNPs of a pair of complex diseases, or traits, as a linear function of their corresponding SNP LD scores to estimate the genetic covariance. The genetic correlation is then the estimated genetic covariance divided by the square root of the product of the single-disease heritability estimates, mirroring the previously discussed estimate of coheritability (Bulik-Sullivan et al., 2015). Coheritability and the estimate of genetic correlation by Bulik-Sullivan et al., 2015 take the direction of the genetic effects into account. That is, both estimates of the genetic correlation are bounded between -1 and 1, where a negative value represents the genetic correlation in opposing directions.

Cai and Tan, 2015 link methods in quantification to methods in detection with the estimation and testing of a quadratic functional under a two-sequence Gaussian model. While their method does not estimate the genetic correlation, nor does it account for differing directions of genetic effects, their estimated quadratic functional is directly motivated by the simultaneous signal detection problem discussed in Section 1.1.1 (Cai and Tan, 2015). Cai and Tan, 2015 devise optimal estimation methods, estimating their quadratic functional assuming sparse mean vectors typical of GWAS applications. In this dissertation we utilize their sparse estimation methods (Cai and Tan, 2015) for estimating a statistic developed to represent the genetic correlation among shared complex disease-associated SNPs.

1.2. Novel Developments

In this dissertation we develop methods for the analysis of complex diseases, specifically in the detection and quantification of shared complex disease-associated SNPs between complex disease pairs, indicative of an overlapping genetic architecture. The work consists of 3 parts, each part returning to the pAID GWAS data (Li et al., 2015a,b) referenced at the beginning of this Chapter. In Chapter 2, we propose a global detection test to combine GWAS results of complex disease pairs,

testing whether a pair of complex diseases shares at least one disease-associated SNP. The test can be posed as a simultaneous signal detection problem, equivalent to testing whether a pair of complex diseases exhibits at least one simultaneous signal, or SNP with nonzero effect size across the pair. We then extend the test to account for the direction of the shared genetic effects, only detecting simultaneous signals existing in the same direction. Intuitively, if a pair of clinically-distinct diseases, classified within the same complex disease set, shows evidence of a shared genetic architecture or overlapping disease pathway through the detection of shared disease-associated SNPs, the shared SNPs are expected to have effects in the same direction.

Unlike most tests requiring an arbitrary threshold to identify SNPs associated with disease, our proposed test statistics are easy to implement without requiring the use of thresholds or tuning parameters. The proposed test statistics take an integrative approach. First, combining the GWAS data of a pair of complex diseases by taking the pairwise minimum of the absolute value of the marginal Z-score for each SNP. Then, collapsing over SNPs genome-wide, taking the maximum score across all SNPs. The proposed global detection tests are based on the simultaneity of signals, scanning aligned pairs of complex disease GWAS Z-scores for at least one shared signal. The simultaneity of signals can be destroyed by permuting the locations of the Z-scores, emulating the null distribution of no signals shared between the complex disease pair. We propose a procedure for determining the exact p-value under permutation, assessing the statistical significance of our global test statistics without making any distributional assumptions. With simulations we show the power and type I error of our test of the simultaneous detection of disease-associated SNPs is dependent on both the magnitude of the effect sizes and the sparsity of the signal. We go on to apply these methods to GWAS data of ten clinically-distinct pAIDs, identifying a shared genetic architecture, accounting for the directions of the shared associations, in disease pairs UC-CD and CVID-JIA.

While no distributional assumptions are made in assessing the statistical significance of our global tests under permutation, we must assume the location of SNPs is exchangeable, or that SNPs are independent of one another. Most methods with applications in genetics and genomics assume independence among genetic variants, ignoring the assumption's lack of validity. The assumed independence between SNPs is 'achieved' through a LD pruning procedure which selects SNPs based on arbitrarily constructed LD blocks, unnecessarily throwing away potentially valuable data.

In Chapter 3 we address the assumption of independence among SNPs, building off of our proposed methods in Chapter 2. We develop a method for assessing the statistical significance of the global detection test presented in Chapter 2 while preserving the inherent dependency structure across the genome. Specifically, we implement a perturbation procedure proposed by Lin and Zou, 2004, and expanded on by Zou et al., 2004, that exploits independent standard normal random variables to emulate the null distribution. The perturbation method (Lin, 2005; Lin and Zou, 2004; Zou et al., 2004) utilizes the individual-level complex disease GWAS data rather than the summary-level Z-scores used in Chapter 2. As such, we redefine the global detection test statistic with respect to score statistics. By allowing for dependency among SNPs in the statistical evaluation of our global detection test statistic, we alleviate data restrictions, enabling the use of multiple imputation to create matching sets of SNPs across disease pairs for complete SNP alignment.

We extend the global detection test of Chapter 2 to detecting shared disease-associated variants at identical SNPs to detecting disease-associated variants within a LD-defined window. In Chapter 2 we assume complex diseases with some shared genetic architecture will have disease-associated genetic variants at identical SNPs. Though with the inherent dependency between genetic variants, power for detecting a shared genetic architecture between a pair of diseases is likely lost in limiting our analysis to the detection of shared disease-associated variants at identical SNPs. Eliminating the need for an independence assumption is necessary for the extension of the global detection test to detection within a window because SNPs within a window are dependent and the studied windows overlap with one another in areas of the genome with high LD. With simulations we show assessing the significance of our global test statistic assuming independence via permutation is conservative, while assessing the significance conserving the inherent dependency structure within the genome via perturbation controls the type I error. We also show an improvement in the power of detecting a shared genetic architecture between a pair of complex diseases using the proposed perturbation method. We go on to apply these methods to a subset of the pAID GWAS data, and our results are consistent with those of Chapter 2 suggesting ignoring the dependency structure is not an issue in this particular data set.

In Chapter 4, we move from the detection of genetic sharing (of Chapters 2 and 3) to the quantification of genetic sharing between a pair of complex diseases. We propose a statistic to quantify the genetic relatedness between a pair of complex diseases again using summary-level GWAS data.

The statistic is a function of SNP effect sizes and acts as an estimate of the genetic correlation among shared disease-associated variants. Our proposed statistic adjusts for both the number of SNPs studied and the respective sample sizes of each of the GWAS pairs. The varied sample size across different GWAS must be accounted for, as the magnitude of the observed effect size is driven in part by the sample size. We utilize the estimation method in a sparse regime by Cai and Tan, 2015 to estimate our proposed statistic and use a bootstrapping procedure to obtain an estimate of the variability in our statistic. We then apply these methods to a subset of the pAID GWAS data, comparing our estimate of the genetic correlation to naively taking the correlation of the pair of marginal Z-scores for each SNP without accounting for the sparsity of the data.

Such improvements in the study of complex diseases, such as pAIDs, will better our understanding of the genetic basis of these diseases, improving the efficiency and effectiveness of their treatment.

CHAPTER 2

STATISTICAL TEST FOR THE DETECTION OF SHARED COMMON GENETIC VARIANTS BETWEEN COMPLEX DISEASES BASED ON GWAS

2.1. Introduction

Genome-wide association studies (GWAS) have proven to be effective in identifying common (minor allele frequency (MAF) $> 5\%$) disease-associated single nucleotide polymorphisms (SNPs) with moderate effects. Though GWAS are often underpowered, requiring larger sample sizes, and the identified SNPs explain only a small proportion of disease heritability (Hindorff et al., 2009). Recent studies reinforce the value of GWAS, specifically in the study of complex diseases, suggesting with increased sample size, GWAS-identified SNPs capture a larger proportion of heritability for complex diseases than previously reported (Lee et al., 2012b; Yang et al., 2011b). Meta-analyses are powerful approaches commonly used for pooling the results of independent GWAS, increasing the total sample size, but the use of standard meta-analysis approaches are not optimal when studying complex diseases (Bhattacharjee et al., 2012). More specifically, standard meta-analysis approaches are not optimal when potential disease-associated variants only have an effect in a subset of the complex diseases of interest or when the direction of the effect differs among the complex diseases (Bhattacharjee et al., 2012). Pooling inferences across a set of complex diseases enables a gain in statistical power and allows for a stronger scientific statement (Benjamini and Heller, 2008). In this paper complex diseases are related but clinically-distinct diseases. Studying complex diseases can lead to the detection of pleiotropic effects while, when restricting analyses to clinically-identical diseases, the opportunity to discover how variants associate with complex diseases through pleiotropic effects is lost (Bhattacharjee et al., 2012).

Pleiotropy is the phenomenon of a single locus' ability to influence multiple traits or diseases. For example, a single disease-associated variant could cause a disease, or multiple related but clinically-distinct diseases, with a wide range of symptoms. Thus, understanding pleiotropy would lead to a better understanding of the genetic etiology and nosology of complex diseases, ultimately improving treatment and prevention efforts (Bhattacharjee et al., 2012; Cross-disorder Group of the Psychiatric Genomics Consortium, 2013). Related but clinically-distinct diseases and disorders, such

as autoimmune diseases and psychiatric disorders, are typically differentiated based on observed symptom patterns, though symptom patterns among diseases often overlap making it difficult to characterize the differences between these sets of complex diseases (Cross-disorder Group of the Psychiatric Genomics Consortium, 2013). For example, the clinical boundaries of psychiatric disorders are blurred, due largely in part to overlapping symptom patterns and relatively unknown pathogenic mechanisms of diseases (Cross-disorder Group of the Psychiatric Genomics Consortium, 2013). Similarly, autoimmune diseases present their own diagnostic challenges as their symptoms span many body organs and are typically nonspecific. Autoimmune diseases show evidence of genetic overlap with more than half of all GWAS-identified autoimmune disease-associated variants shared by at least two other, clinically-distinct, autoimmune diseases (Cotsapas et al., 2011). Thus, the susceptibility of disease is thought to be influenced by a strong genetic predisposition as evidenced by high rates of familial clustering and co-occurrence of disease (Cooper, Bynum, and Somers, 2009). Particularly, the diagnosis of early-onset autoimmune diseases in children may be associated with a higher risk for those children to develop secondary or tertiary clinically-distinct autoimmune diseases. Identifying genetic risk factors, especially those shared across multiple diseases, has become an important strategy in assessing the genetic architecture of complex disease sets and disorders (Cross-disorder Group of the Psychiatric Genomics Consortium, 2013).

The goal of this paper is to investigate the genetic relatedness of complex disease sets through the detection of shared common genetic variants. This paper proposes a global test to combine GWAS results of complex diseases, ultimately testing whether a pair of complex diseases shares at least one common genetic variant. The test can be posed as a simultaneous signal detection problem, in that, the test is equivalent to testing whether two diseases exhibit at least one simultaneous disease-associated variant, or signal. We further extend the test to account for the direction of the shared variants' disease association. Intuitively, if a pair of related but clinically-distinct diseases shows evidence of genetic sharing through the detection of shared disease-associated variants, the shared variants are expected to be associated with both diseases in the same direction. Whereas, if the direction of the shared variant's disease association differs across the pair (i.e., a disease-associated variant is positively associated with, or detrimental to, one disease and negatively associated with, or protective against, the other), that particular shared variant would not be indicative of genetic sharing between the pair of complex diseases. Biologically relevant shared SNPs have been shown to confer risk in type I diabetes (T1D), while protecting against Crohns disease (CD)

(Wang et al., 2010), thus understanding the direction of shared SNPs is important in understanding the shared genetic etiology of disease. The simultaneous detection of disease-associated variants is dependent on both the strength (magnitude) and sparsity of the associated variants.

The chapter is organized as follows. We first formulate the problem of detecting shared, disease-associated genetic variants between a pair of diseases as a simultaneous signal detection problem. We then propose a global test in the setting of combining pairs of heterogeneous disease GWAS and extend the test to take the directions of the genetic effects into account. We evaluate the power and type I error of the proposed tests through simulation studies. We apply these two tests to GWAS data of a set of 10 clinically-distinct, pediatric autoimmune diseases (pAIDs), thyroiditis (THY), spondyloarthropathy (SPA), psoriasis (PSOR), celiac disease (CEL), systemic lupus erythematosus (SLE), common variable immunodeficiency (CVID), ulcerative colitis (UC), type I diabetes (T1D), juvenile idiopathic arthritis (JIA) and Crohn's disease (CD), with shared controls, in order to investigate the genetic sharing among these diseases. We utilize the global detection tests in a sequential procedure to identify the shared genetic variants detected between a given pAID pair. We conclude the paper with a discussion of the methods.

2.2. Statistical Formulation and Tests for Detection of Shared Genetic Variants

Assume GWAS data for all pairs of complex diseases, under question of the existence of genetic sharing, are readily available. Let U_i be the Z-score of the marginal association between disease A and the i^{th} SNP ($i = 1, \cdots, n$) and let V_i be the Z-score of the marginal association between disease B and the i^{th} SNP, where disease A and disease B are a pair of complex diseases. Under the null hypothesis of no association between disease A and the i^{th} SNP $U_i \sim N(0,1)$, while under the alternative hypothesis $U_i \sim N(\mu_i, \sigma_i^2)$ in the presence of an association. Similarly, for disease B, under the null $V_i \sim N(0,1)$, while under the alternative $V_i \sim N(\nu_i, \tau_i^2)$. U_i are assumed to be independent, as is assumed of V_i, which is achieved by selecting SNPs that are not in linkage disequilibrium (LD) with one another, or through LD-pruning. Ideally, U_i is assumed to be independent of V_i, which is achieved if U_i and V_i are calculated from different, non-overlapping datasets. For GWAS with shared controls, when the sample size of the controls is large, these statistics are nearly independent.

2.2.1. Test of simultaneous signal detection

As previously mentioned, the question of whether two diseases share at least one common, disease-associated genetic variant can be posed as a simultaneous signal detection problem testing

$$\begin{aligned} &H_0 : |\mu_i| \wedge |\nu_i| = 0,\ i = 1, 2, \cdots, n \\ &H_A : \text{there is at least one } i \text{ such that } |\mu_i| \wedge |\nu_i| \neq 0, \end{aligned} \tag{2.1}$$

the 'signal' being the mean value of Z-score(s) U_i, or V_i. Testing this set of hypotheses is also known as testing the global null hypothesis, or the conjunction of the null (Benjamini and Heller, 2008). When H_0 is rejected, the pair of complex diseases, A and B, are concluded to have some degree of genetic sharing. In order to develop a test for hypothesis (2.1), U_i and V_i can be summarized into a single statistic, $T_i = |U_i| \wedge |V_i|$, which is always positive because T_i depends only on the magnitude of each Z-score.

Under a random effects framework, the distribution of T_i can be written as a mixture of two components,

$$T_i \sim \epsilon G + (1 - \epsilon)F,$$

where $\epsilon \in [0, 1]$ is the mixture proportion and distribution functions G and F can be written as

$$\begin{aligned} G \sim & |N(\mu, \sigma^2)| \wedge |N(\nu, \tau^2)|, \\ F \sim & p_{1n}|N(0,1)| \wedge |N(0,1)| + \\ & p_{2n}|N(\mu, \sigma^2)| \wedge |N(0,1)| + p_{3n}|N(0,1)| \wedge |N(\nu, \tau^2)|, \end{aligned}$$

where $p_{1n} + p_{2n} + p_{3n} = 1$. Here, in the setting of combining pairs of clinically-distinct GWAS, p_{1n} is the proportion of SNPs not associated with either disease A or disease B, while p_{2n} and p_{3n} are the proportions of SNPs associated with either disease A or disease B, respectively. Therefore, of the distribution functions considered above, F represents the null distribution and G represents the alternative distribution of the simultaneous signal detection problem. In this random effects

framework, the simultaneous signal detection problem (2.1) above becomes the test of

$$H_0 : \epsilon = 0$$

$$H_A : \epsilon > 0,$$

which is a generalization of the normal mixture detection problem testing

$$H_0 : \ \epsilon = 0 \text{ vs. } H_A : \ \epsilon > 0$$

for $X_i \sim (1-\epsilon)N(0,1) + \epsilon N(\mu, 1)$ (Jin and Donoho, 2004). Unlike the normal mixture distribution of X_i, the null distribution of T_i, F, contains unknown parameters. We consider a max test statistic collapsing over all genetic variants, genome-wide

$$M_n = \max_{i=1,\cdots,n} T_i.$$

If M_n is greater than a predefined critical value, H_0 will be rejected by the max test suggesting diseases A and B share at least one common genetic variant.

With finite samples it is useful to calculate p-values for the max test statistic, M_n, but obtaining accurate p-values can be difficult because the null distribution, F, contains unknown parameters. Instead, we formulated an analytical p-value based on the distribution of the permutation of the location of the Z-scores, U_i and V_i, relative to each other. Recall the basis of the global max test is to detect simultaneous disease-association signals between a pair of complex diseases so, by destroying the simultaneity between disease association signals, the null distribution, F, can be mimicked. In utilizing the permutation distribution, the distributions of the individual Z-scores are preserved and the calculation of the p-value itself is independent of correlation between U_i and V_i. The exact p-value is defined to be the proportion of permuted test statistics exceeding the observed test statistic, M_n, and can be easily derived as

$$P(M \geq M_n) = 1 - \frac{\binom{m}{0}\binom{n-m}{k}}{\binom{n}{k}},$$

where

$$m = \sum_{i}^{n} I(|U_i| \geq M_n),$$
$$k = \sum_{i}^{n} I(|V_i| \geq M_n).$$

This provides an exact p-value determination without actually performing any permutations.

2.2.2. Test of simultaneous signal detection with effects in the same direction

The null hypothesis (2.1) and the global test reviewed in the previous Section do not take the directions of the genetic effects into account. In real applications, it is also of interest to test whether two diseases share the same genetic variants with effects in the same direction, i.e., whether the shared disease-associated genetic variants are detrimental to or protective against both diseases. To accommodate differing directional effects of SNPs among pairs of complex diseases, the null hypothesis is given by

$$H_0 : |\mu_i| \wedge |\nu_i| = 0,\ i = 1, 2, \cdots, n$$
$$\text{or } |\mu_i| \wedge |\nu_i| \neq 0 \text{ and } sg(\mu_i) \neq sg(\nu_i) \tag{2.2}$$
$$H_A : \text{there is at least one } i \text{ such that } |\mu_i| \wedge |\nu_i| \neq 0 \text{ and } sg(\mu_i) = sg(\nu_i), \tag{2.3}$$

where $sg(x)$ is the direction of the signal of x. More specifically, $sg(x)$ is 1 if the sign of x is positive and -1 if the sign of x is negative. When H_0 is rejected, the pair of complex diseases, A and B, are still concluded to have some degree of genetic sharing. Though, in using the proposed test, an added level of complexity can be inferred from the test's conclusions that cannot be inferred from the test formulated in (2.1). That is, when rejecting H_0, disease A and disease B's shared signals are guaranteed to have effects in the same direction. In order develop a test for hypothesis (2.2), U_i and V_i can be summarized into a single statistic $W_i = |U_i| \wedge |V_i| sg(U_i) sg(V_i)$, which, again, is similar to the summary statistic in previous section, but adapted to depend on the direction of the associations in addition to their magnitudes. W_i can take both positive and negative values, and W_i is only positive when U_i and V_i have associations in the same direction, $sg(U_i) = sg(V_i)$.

Using a random effects framework similar to that of the previous Section, consider the following

distribution functions

$$
\begin{aligned}
J \sim & p_{1n}|N(\mu,\sigma^2)| \wedge |N(\nu,\tau^2)|(1)(1) + p_{2n}|N(\mu,\sigma^2)| \wedge |N(\nu,\tau^2)|(-1)(-1), \\
H \sim & p_{3n}|N(0,1)| \wedge |N(0,1)| + \\
& p_{4n}|N(\mu,\sigma^2)| \wedge |N(0,1)|(1) + p_{5n}|N(\mu,\sigma^2)| \wedge |N(0,1)|(-1) + \\
& p_{6n}|N(0,1)| \wedge |N(\nu,\tau^2)|(1) + p_{7n}|N(0,1)| \wedge |N(\nu,\tau^2)|(-1) + \\
& p_{8n}|N(\mu,\sigma^2)| \wedge |N(\nu,\tau^2)|(-1)(1) + p_{9n}|N(\mu,\sigma^2)| \wedge |N(\nu,\tau^2)|(1)(-1),
\end{aligned}
$$

where $p_{1n}+p_{2n} = 1$ and $p_{3n}+\cdots+p_{9n} = 1$. Here, p_{1n} is the proportion of SNPs positively associated with both disease A and disease B and similarly, p_{2n} is the proportion of SNPs negatively associated with both disease A and disease B. A nonzero value of the proportions p_{1n} or p_{2n} is representative of the alternative hypothesis introduced previously and suggests some degree of genetic sharing between diseases A and B. p_{3n} is the proportion of SNPs not associated with either disease A or disease B. p_{4n} and p_{5n} are the proportions of SNPs positively or negatively associated with only disease A, while p_{6n} and p_{7n} are the proportions of SNPs positively or negatively associated with only disease B. p_{8n} and p_{9n} are the proportions of SNPs associated with both disease A and disease B but in differing directions. H represents the null distribution and J represents the alternative. The distribution of the summary statistic, $W_i = |U_i| \wedge |V_i| sg(U_i) sg(V_i)$, is given as

$$W_i \sim \epsilon J + (1-\epsilon)H,$$

where $\epsilon \in [0,1]$ is the mixture proportion. The simultaneous signal detection problem again becomes the test of $H_0 : \epsilon = 0$ vs. $H_A : \epsilon > 0$. To test the null hypothesis (2.2), we define the test statistic

$$M_n = \max_{i=1,\cdots,n} W_i.$$

Again, if M_n is greater than a predefined critical value, H_0 will be rejected by the max test suggesting diseases A and B share at least one common genetic variant.

Now, the analytical p-value, still formulated based on the permutation distribution, must account for the direction of a particular SNP's association in addition to its magnitude. The analytical p-value is defined to be the probability at least one U_i, with a magnitude of at least M_n, is permuted such

that it aligns with at least one V_i, also with a magnitude of at least M_n, and $sg(U_i) = sg(V_i)$. This probability can be written out explicitly as

$$P(M \geq M_n) = 1 - \left(\frac{\sum_{m=0}^{min(a^+,\, b^-)} \binom{b^-}{m} \binom{n-b^--b^+}{a^+-m} \binom{n-a^+-b^-+m}{a^-}}{\binom{n}{a^+}\binom{n-a^+}{a^-}} \right),$$

where

$$\begin{aligned} a^+ &= \sum_i^n I(|U_i| \geq M_n) I(sg(U_i) > 0), \\ a^- &= \sum_i^n I(|U_i| \geq M_n) I(sg(U_i) < 0), \\ b^+ &= \sum_i^n I(|V_i| \geq M_n) I(sg(V_i) > 0), \\ b^- &= \sum_i^n I(|V_i| \geq M_n) I(sg(V_i) < 0). \end{aligned}$$

Again, providing a simplified determination of the exact p-value for the proposed test. In practice, Stirling's approximation is used to approximate the exact probability above (see Section 2.3).

2.3. Simulation Studies

Simulations were carried out to verify the accuracy of the approximated analytical p-value when considering the direction of effects with respect to the p-value obtained from actually carrying out the permutations. Similarly, simulations were used to compare the power and type I error of the direction-dependent test under various conditions.

The permuted and analytical p-values were compared in two datasets: one generated under the null and the other generated under the alternative distribution. In each dataset $n = 500000$ pairs of Z-scores, U_i and V_i, were generated such that

$$\begin{pmatrix} U_i \\ V_i \end{pmatrix} \sim p_1 N_2\left(\begin{pmatrix}\mu_i \\ \mu_i\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right) + p_2 N_2\left(\begin{pmatrix}-\mu_i \\ -\mu_i\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right) + p_3 N_2\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right) +$$
$$p_4 N_2\left(\begin{pmatrix}\mu_i \\ 0\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right) + p_5 N_2\left(\begin{pmatrix}-\mu_i \\ 0\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right) + p_6 N_2\left(\begin{pmatrix}0 \\ \mu_i\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right) +$$
$$p_7 N_2\left(\begin{pmatrix}0 \\ -\mu_i\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right) + p_8 N_2\left(\begin{pmatrix}-\mu_i \\ \mu_i\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right) + p_9 N_2\left(\begin{pmatrix}\mu_i \\ -\mu_i\end{pmatrix}, \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}\right),$$

where μ_i, the Z-score mean, and $p_1, \cdots, p_9$, the proportions of SNPs, were varied as detailed below. N_2 indicates that the pairs of Z-scores were generated from a mixture of bivariate normal distributions with correlation equal to zero. Without loss of generality, it is assumed that $|\mu_i| = |\nu_i|$ and $\sigma_i^2, \tau_i^2 \geq 1$.

The Z-score means, μ_i, represent the strength of a particular SNP's association, or signal. μ_i were varied at $2, 3, 4$ and 4.5, which correspond to p-values of $0.02, 0.001, 3 \times 10^{-5}$ and 3×10^{-6}, respectively. The proportions, $p_1, \cdots, p_9$, were varied to change the approximate number of association signals, $\mu_i \neq 0$, while still allowing the number of signals to remain sparse. Table 2.1 presents the sparsity simulation settings of varied proportions and uses the following notation to indicate the approximate number of signals present in the respective dataset: $X(x)$, where X is the approximate number of shared signals, or disease-associated variants, generated from the alternative distribution, J, and x is the approximate number of signals generated from the null distribution, H. Thus, approximately x total signals exist in the dataset when generated under the null distribution and approximately $(X + x)$ total signals exist in the dataset when generated under the alternative distribution. Without loss of generality $p_{1n} = p_{2n}$ and $p_{4n} = p_{5n} = p_{6n} = p_{7n} = p_{8n} = p_{9n}$. Notice, the proportions of SNPs differ depending on whether the data was generated under the null or alternative distribution. Recall, under the null distribution the proportions p_{1n} and p_{2n} are equal to 0 indicating no simultaneous signals or genetic sharing between the pair of heterogeneous diseases, and thus $X = 0$ of the $X(x)$ notation previously described.

After generating a dataset of $n = 500000$ pairs of Z-scores, U_i and V_i, the observed test statistic, M_n, was calculated. To determine the permuted p-value, the magnitudes of the $n = 500000$ generated U_i were permuted 2500 times while fixing the locations of the signs of U_i and fixing the locations

Table 2.1: Three sparsity simulation settings considered. Each setting is represented by the two numbers $X(x)$, where X is the approximate number of shared signals generated from the alternative distribution, J, and x is the approximate number of signals generated from the null distribution, H

	I: 75 (600)		**II:** 40 (300)		**III:** 40 (150)	
Probability	Alternative	Null	Alternative	Null	Alternative	Null
p_1	0.000075	0	0.0000375	0	0.0000375	0
p_2	0.000075	0	0.0000375	0	0.0000375	0
p_3	0.99865	0.9988	0.999325	0.9994	0.999625	0.9997
p_4	0.0002	0.0002	0.0001	0.0001	0.00005	0.00005
p_5	0.0002	0.0002	0.0001	0.0001	0.00005	0.00005
p_6	0.0002	0.0002	0.0001	0.0001	0.00005	0.00005
p_7	0.0002	0.0002	0.0001	0.0001	0.00005	0.00005
p_8	0.0002	0.0002	0.0001	0.0001	0.00005	0.00005
p_9	0.0002	0.0002	0.0001	0.0001	0.00005	0.00005

of both the signs and magnitudes of V_i, with respect to each other. In permuting one of the two generated Z-scores, the simultaneity of the generated association signals, and corresponding direction of the signal, is lost. Intuitively, as the generated signals become less sparse, the ability to emulate the null distribution through permutation diminishes because, in increasing the number of signals, the chances that a signal will align with another signal of the same sign after permutation increase. With each permutation, the permuted test statistic is compared to the observed test statistic, M_n, and the permuted p-value is equal to the proportion of permuted test statistics exceeding M_n. To then determine the analytical p-value, a^+, a^-, b^+, b^- are computed as indicated above using the observed test statistic, M_n. Note, Stirling's approximation, $log(n!) \approx nlog(n) - n + O(log(n))$ where $O(log(n)) = 0.5log(2\pi n)$, was used in the calculation of the analytical p-value. In this case, n can be small and thus for the most accurate results the $O(log(n))$ term cannot be ignored.

2.3.1. Comparison of permuted and analytical p-values

Table 2.2 compares the permuted (based on 2500 permutations) and analytical p-values for datasets generated under varying strengths of association and the three sparsity simulation settings (Table 2.1). The permuted p-value is well emulated by the analytical p-value, and the use of the analytical p-value is a computationally efficient alternative to carrying out the permutations to obtain the permuted p-value.

Table 2.2: Comparison of permuted (based on 2500 permutations) and analytical permutation p-values for direction-dependent max test calculated from datasets generated under the alternative and null distributions for different association strengths and sparsity settings

		Alternative		Null	
μ	Setting	Permuted	Analytical	Permuted	Analytical
2	**I**	0.3056	0.2893	0.1828	0.1621
	II	0.3784	0.3615	0.2036	0.1872
	III	0.7916	0.8110	0.7020	0.6973
3	**I**	0.0164	0.0192	0.3484	0.3544
	II	0.0012	0.0014	0.3676	0.3743
	III	0.0008	0.0010	0.8880	0.8959
4	**I**	0.0036	0.0022	0.5012	0.4981
	II	0.0048	0.0033	0.9072	0.9089
	III	0.0004	0.0015	0.2720	0.3103
4.5	**I**	0	0	0.3208	0.3258
	II	0.0024	0.0028	0.7488	0.7654
	III	0.0008	0.0015	0.1424	0.1894

2.3.2. Power and type I error

In comparing the power and type I error of the proposed test across various conditions, 1000 datasets were generated, as described above. That is, 1000 replications were generated under the null distribution to determine the type I error and 1000 replications were generated under the alternative distribution to determine the power. Z-score means, μ_i, varied from 2 to 4.5 as before and the proportions of SNPs varied by sparsity simulation settings, indicated in Table 2.1. Analytical p-values were computed for each of the 1000 datasets and the power (or type I error) is equal to the number of datasets with an analytical p-value less than the corresponding nominal level, $\alpha = 0.05$ or 0.01, divided by the total number of datasets.

Table 2.3 shows the type I error and power of the proposed test for datasets generated under varying strengths of association and three sparsity simulation settings. At both the nominal $\alpha = 0.05$ level and $\alpha = 0.01$, we observe that the type I error is well controlled under various simulation settings. When signal strengths are high, but not shared between two diseases (under the null), the analytical approximation of the adaptive test is slightly conservative, especially when the signals are sparse.

We also observe that the power increases as the strength of the association and sparsity of the signals increase (Table 2.3). The power of the test is consistently greater than 90% when $\mu_i > 2$ for $\alpha = 0.05$ and when $\mu_i > 3$ for $\alpha = 0.01$ (Table 2.3).

Table 2.3: Power and type I error of direction-dependent max test estimated from 1000 replications at $\alpha = 0.05$ or 0.01 of datasets generated under different association strengths and sparsity simulation settings

		α = 0.05		α = 0.01	
μ	Setting	'Power'	'Type I Error'	'Power'	'Type I Error'
2	**I**	0.170	0.053	0.079	0.010
	II	0.163	0.048	0.061	0.011
	III	0.273	0.050	0.113	0.009
3	**I**	0.956	0.050	0.819	0.015
	II	0.946	0.039	0.752	0.003
	III	0.990	0.026	0.846	0.001
4	**I**	1	0.047	1	0.006
	II	1	0.026	0.996	0.003
	III	1	0.008	0.964	0.000
4.5	**I**	1	0.034	1	0.005
	II	1	0.016	0.999	0.000
	III	1	0.004	0.977	0.001

2.4. Analysis of Genetic Sharing of 10 Pediatric Autoimmune Diseases

Autoimmune diseases affect approximately 8% of all Americans and are a leading cause of death in women up to age 64 (Cooper, Bynum, and Somers, 2009). Medical professionals are still learning about autoimmune diseases, now with over 80 clinically-distinct autoimmune diseases identified. Little is known of the relationships between different autoimmune diseases, and the degree to which genetic variants associated with one autoimmune disease influence the risk of developing a second, clinically-distinct autoimmune disease has not been well characterized. The methods described above are applied to GWAS data for all possible pairs of 10 pediatric autoimmune diseases (pAIDs) to determine whether pairs of diseases exhibit some evidence of genetic sharing. pAIDs are particularly good candidates for studying the existence of genetic sharing between any two diseases because genetic risk factors are thought to have a stronger contribution in early-onset disease.

Over 5200 pediatric cases across 10 pAIDs: THY, SPA, PSOR, CEL, SLE, CVID, US, T1D, JIA and CD, and over 11000 population-based controls without known autoimmune, inflammatory or immunodeficiency disorders were genotyped at Children's Hospital of Philadelphia (CHOP) on two comparable GWAS platforms. The effects due to artifacts introduced by use of multiple genotyping platforms or multiple study sites are likely small because all the samples were genotyped on comparable genotyping platforms at a single site.

2.4.1. Detection of Shared Genetic Variants

Analytical p-values were calculated for all possible pairs of the 10 pAIDs, using both proposed tests: with and without accounting for the direction of SNPs' effects. Z-scores for each SNP were calculated marginally for each disease. Each of the 10 autoimmune diseases had a varying number of cases: 99 (THY), 111 (SPA), 113 (PSOR), 183 (CEL), 256 (SLE), 309 (CVID), 895 (UC), 1139 (T1D), 1165 (JIA) and 2039 (CD). It is important to note, only autosomal SNPs were included in the analytical p-value calculation. That is, mitochondrial SNPs and SNPs found on the X or Y chromosomes were excluded. Similarly, SNPs within the major histocompatibility complex (MHC) region, defined as $25,500,000-34,000,000$ base pairs (bp) of chromosome 6, were excluded as they are already known to be highly associated with autoimmune diseases. Each pairwise p-value was calculated in a 'complete' sense, that is only complete pairs of autoimmune diseases contributed to the calculation of their respective p-value. Thus, each pairwise p-value was calculated based on a different number of SNPs. For any given pair of autoimmune diseases, the percentage of missing data was no more than 5% with the number of complete-case SNPs in each pair totaling just over $480,000$.

Table 2.4 presents the pairwise analytical p-values calculated using the max tests with and without accounting for the direction of SNP effects (second and first row, respectively, for each pair). It is important to note, the p-values presented in Table 2.4 have not been adjusted for multiple testing. There are many differences in the p-values obtained from the two tests, suggesting accounting for the direction of the effect tells a different story than not accounting for it at all. Notably, the pairwise p-value between diseases JIA and UC is no longer highly significant at the nominal $\alpha = 0.05$ level when accounting for the direction of the simultaneous associations. This observation also holds true for the following disease pairs: CD and JIA, T1D and JIA and CEL and PSOR.

Figure 2.1 demonstrates the observed differences noted above when accounting for the direction of the association. Disease pairs CD and JIA, T1D and JIA and UC and JIA exhibit relatively strong simultaneous signal(s) (SNP association p-value(s) $\leq 10^{-5}$), of which, exist in opposing directions. That is, the simultaneous SNP (indicated by a blue star in Figure 2.1) is positively associated with one disease (indicated by a black vertical line) and negatively associated with the other disease of the pair (indicated by a red vertical line). The max test ignoring the direction of effects identifies

Table 2.4: Pairwise analytical p-values using max tests with (second rows) and without (first rows) accouting for the direction of effects for the 10 studied pAIDs

	SPA	PSOR	CEL	SLE	CVID	UC	T1D	JIA	CD
THY	0.5666	0.5119	0.0567	0.0181	0.4886	0.4941	0.0646	0.0555	0.1497
	0.4059	0.3553	0.0414	0.0117	0.325	0.3035	0.0358	0.0302	0.7848
SPA		0.0365	0.9405	0.6382	0.2134	0.7246	0.9915	0.2617	0.4733
		0.0262	0.7998	0.7148	0.1306	0.4853	0.9164	0.1474	0.2686
PSOR			0.0290	0.1259	0.0848	0.9706	0.5193	0.5815	0.7004
			0.9091	0.0754	0.0532	0.9525	0.3253	0.3653	0.4450
CEL				0.3363	0.5241	0.3803	0.6817	0.5667	0.9375
				0.2064	0.3485	0.2261	0.4610	0.3616	0.7402
SLE					0.9927	0.0528	0.1188	0.8816	0.2862
					0.9493	0.0290	0.0631	0.6627	0.3545
CVID						0.0549	0.0527	3.7e-05	0.1238
						0.0300	0.2263	2.9e-05	0.9782
UC							0.0003	5.0e-05	2.7e-05
							0.0001	0.1261	8.3e-06
T1D								0.0015	0.0032
								0.6105	0.0014
JIA									0.0003
									0.1021

disease pairs CD and JIA, T1D and JIA and UC and JIA as exhibiting genetic sharing with p-values below the nominal $\alpha = 0.05$ level (first rows, Table 2.4), but the modified max test, requiring shared disease-associated effects to be in the same direction, does not identify these pairs (second rows, Table 2.4) because their simultaneous signals exit in opposite directions (opposing colored lines at the blue star in Figure 2.1). The plot of the disease pair CEL and PSOR (lower right of Figure 2.1) is less intuitive. While the test ignoring the direction of effects also identifies this pair as exhibiting genetic sharing at the nominal $\alpha = 0.05$ level (first rows, Table 2.4), this is likely an artifact of multiple testing. Recall $\binom{10}{2}$ tests are being done and, in accounting for this, disease pairs UC and JIA, CD and JIA and T1D and JIA are still identified as exhibiting genetic sharing using the direction-independent test but the disease pair CEL and PSOR is no longer identified.

Figure 2.2 further demonstrates the differences between the proposed tests. Figure 2.2 presents complete-case, pairwise scatter-plots of marginal SNP Z-scores for a subset of the 10 pAIDs studied, specifically CVID, UC, T1D, JIA and CD of which have a larger number of cases in comparison to the remaining studied pAIDs. Large Z-scores for disease pairs UC and JIA, CD and JIA and T1D and JIA are discordant, or their Z-scores are of opposing signs. Disease pairs CVID and JIA and UC and CD are still identified as exhibiting genetic sharing after accounting for the direction of

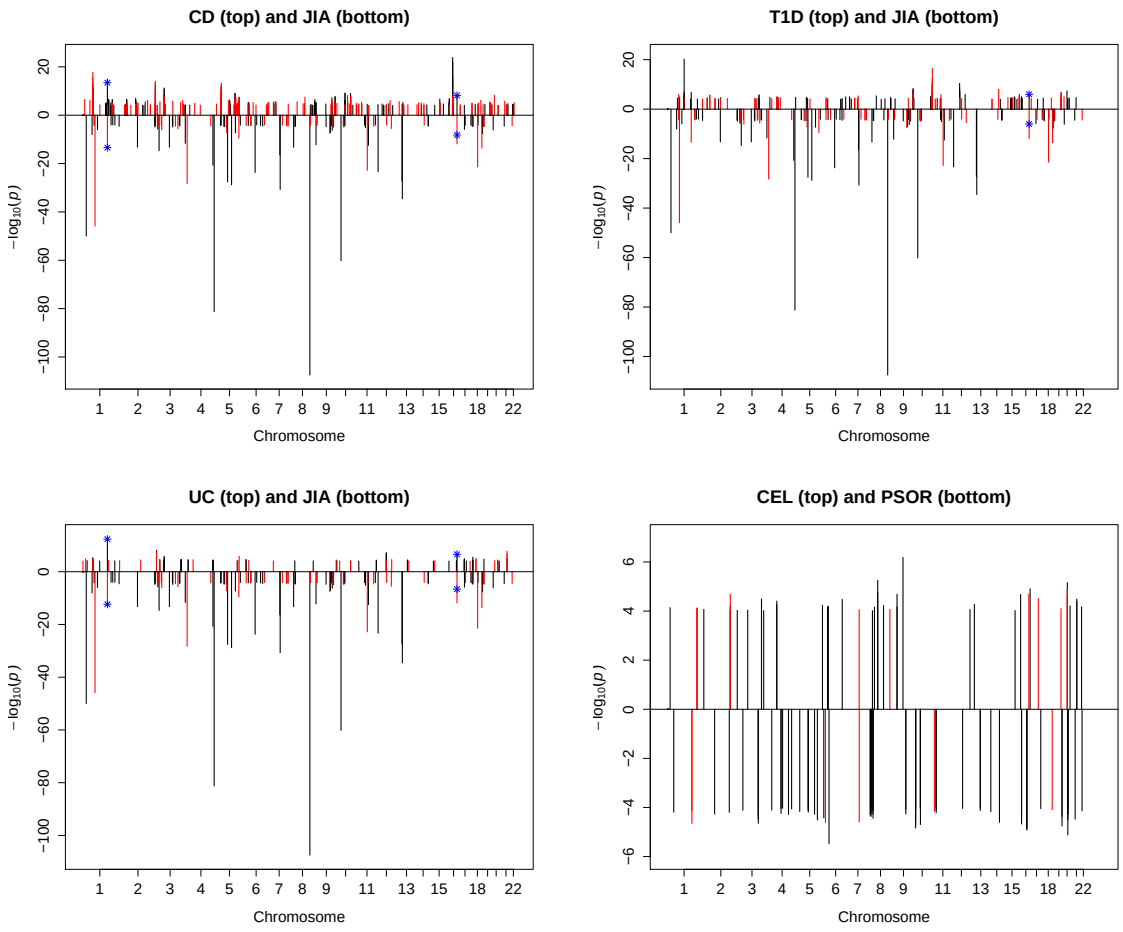

Figure 2.1: Complete-case, paired Manhattan-like plots of marginal SNP association p-values less than 10^{-4} from GWAS for the following disease pairs: CD and JIA, T1D and JIA, UC and JIA and CEL and PSOR. Each vertical line is $-log_{10}(p\text{-value})$, where black indicates that particular SNP was positively associated with disease and red indicates that SNP was negatively associated with disease. The blue star(s) indicate the maximum p-value(s) of simultaneous association signal(s) within the pair of diseases equaling less than 10^{-5}

the association (Table 2.4). The complete-case Z-score scatterplot of these two pairs shows large Z-scores are concordant, or of the same signs (Figure 2.2), further demonstrating the abilities of the direction modified test in detecting the shared genetic variants with the same effect directions.

2.4.2. Sequential Identification Procedure of Shared Genetic Variants

We utilize the global detection tests in a sequential procedure to identify the shared genetic variants detected between pAID disease pairs in Section 2.4.1. The procedure starts with a 'complete' set of SNPs from the corresponding GWAS of a disease pair. In applying the global detection

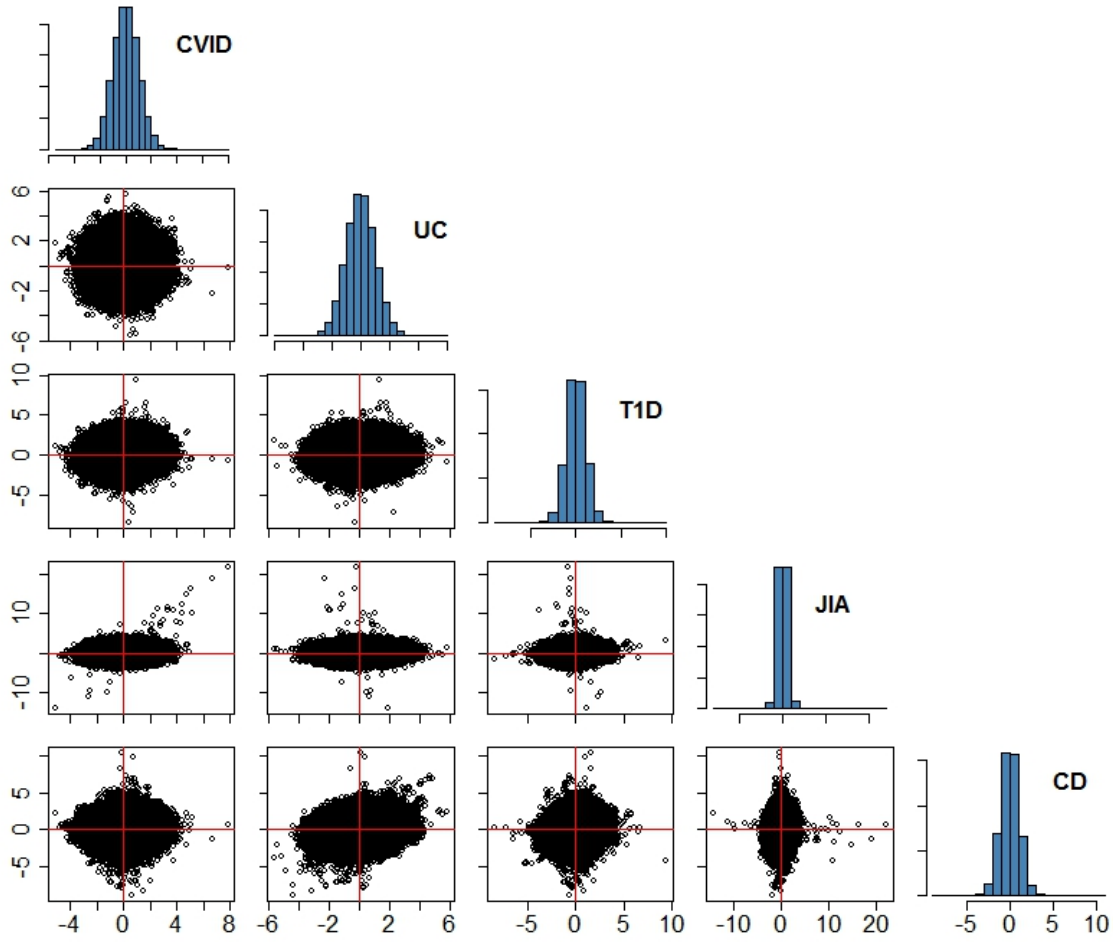

Figure 2.2: Histograms and complete-case, pairwise scatterplots with coordinate axes in red of marginal SNP Z-scores for a subset of the 10 pAIDs studied

test proposed in Section 2.2.1 to the 'complete' SNP set, an exact permutation-based p-value is calculated as in Section 2.4.1. Then the top signal, or SNP with largest pairwise minimum of the absolute value of corresponding Z-scores, $max(|U_i| \wedge |V_i|)$, is removed and the permutation-based p-value is recalculated. This removal procedure is continued, sequentially, until the recalculated permutation-based p-value falls outside a specified significance threshold.

Figure 2.3 graphically displays the implementation of the sequential identification procedure for disease pairs JIA-CVID and CD-UC, plotting -log of recalculated permutation-based p-values by the respective number of SNPs removed from the set. As SNPs are removed, the global detection permutation-based p-values become less significant. Eventually all drivers of genetic sharing are removed from the tested SNP set causing the permutation p-values to level off as additional SNPs, with no shared relationship, are removed (Figure 2.3). The rightmost plots of Figure 2.3 show the chromosomal locations of the top shared SNPs between disease pairs JIA-CVID (Figure 2.3, top) and CD-UC (Figure 2.3, bottom). Note, few of the top drivers of genetic sharing are on the same chromosome within a disease pair, suggesting, in this dataset, SNPs are likely independent. Without loss of generality, the p-values of Figure 2.3 are calculated without considering the direction of effects.

Table 2.5 provides the chromosomal location (CHR), SNP rsID, base pair position (BP) and associated gene, if documented (Sherry et al., 2001), for the sequentially-identified top drivers of genetic sharing between disease pairs JIA-CVID and CD-UC (Figure 2.3). Many of the top SNPs driving genetic sharing between CD-UC map to protein-edcoding genes (Table 2.5).

Table 2.5: Chromosomal location (CHR), SNP rsID, base pair position (BP) and associated gene (Sherry et al., 2001) for the sequentially-identified top drivers of genetic sharing highlighted in Figure 2.3 for disease pairs JIA-CVID and CD-UC

JIA-CVID				CD-UC			
CHR	SNP	BP	gene	CHR	SNP	BP	gene
8	rs3019885	118025645	SLC30A8	1	rs12039194	164537228	PBX1
4	rs4862110	183751029	-	16	rs2221705	79362411	MAF
6	rs6928830	84219312	-	5	rs10045431	158814533	-
1	rs2066363	82237577	ADGRL2	3	rs4625	49572140	DAG1
10	rs7100025	37592538	-	3	rs9858280	49597737	BSN

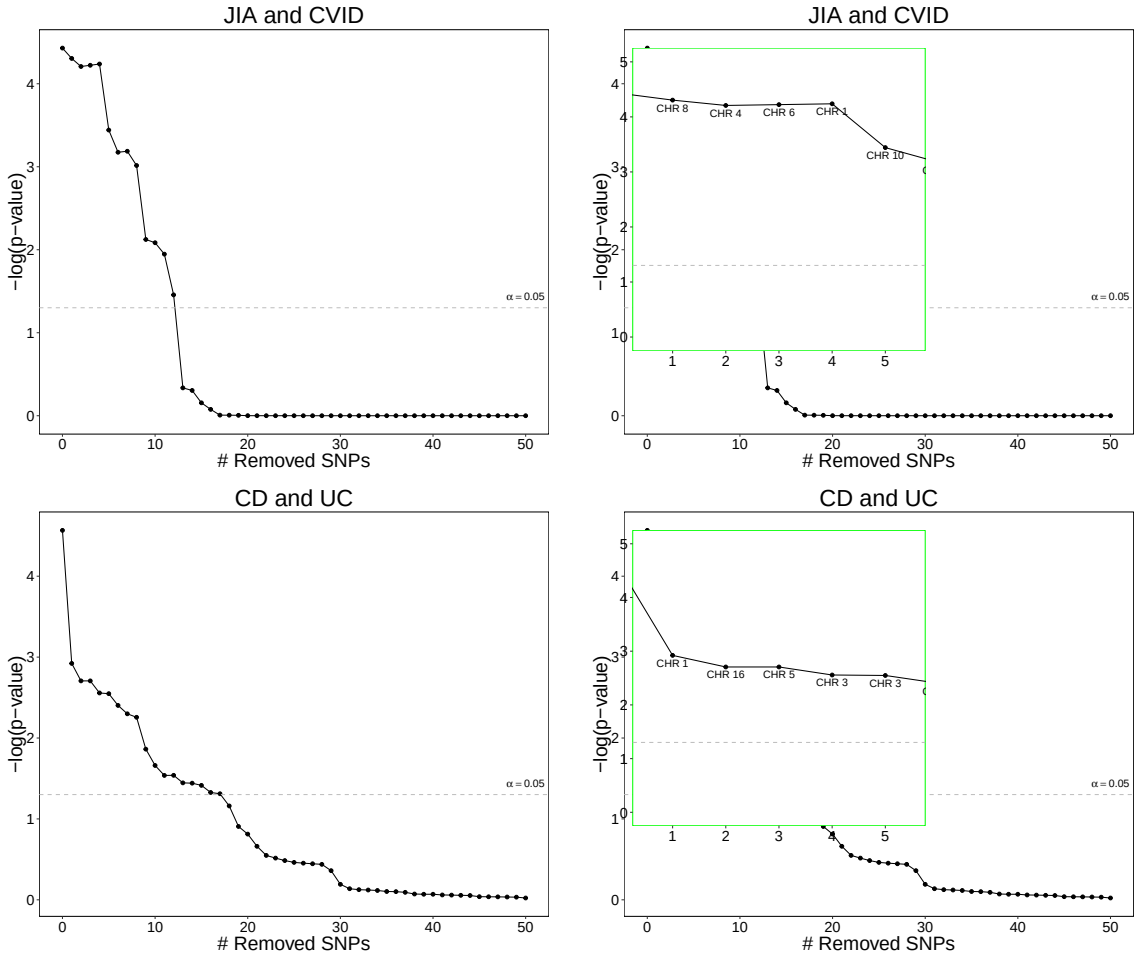

Figure 2.3: Plot of sequential identification procedure: $-log_{10}$(permutation-based p-value) by the number of top SNPs removed for disease pairs JIA-CVID (top) and CD-UC (bottom) with respective zoomed-in plots with chromosomal location of top shared SNPs (rightmost, green box), with nominal significance threshold ($\alpha = 0.05$)

2.5. Conclusion

This chapter introduces two tests for the detection of simultaneous signals based on GWAS data of two genetically related diseases. The first test tests for the sharing of common genetic variants between two diseases without considering the directions of the effects. Specifically, the statistical test detects the absolute value of simultaneous signals, allowing the simultaneous signals to exist in opposite directions. The second test, an extension, includes an added dependency on the direction of the association signal and only detects simultaneous signals existing in the same direction. This new test is biologically relevant in the context of how two complex diseases are genetically related, as shared SNPs have been identified to confer risk in type I diabetes (T1D) while protecting against Crohns disease (CD) (Wang et al., 2010). Both the test statistics are of a max-type, in that they take the maximum score across all SNPs. By destroying the simultaneity of the signals through permutation, the null distribution can be mimicked. A procedure for obtaining the analytical permutation p-value was developed for these two global tests and was shown to be close to the permuted p-value (Table 2.2). The simulations show the proposed analytical p-value has a well-defined power and type I error for various simulation settings. The power increases and the type I error is reduced as the strength and sparsity of the signals increase (Table 2.3). Because the tests are based on the simultaneity of disease association signals, as signals become less sparse the likelihood of destroying their simultaneity with permutations decreases. Thus, the global tests discussed here are suitable for sparse data, as is typical of genome-wide association studies.

The proposed tests were applied to GWAS for a set of 10 clinically-distinct, pediatric autoimmune diseases (pAIDs) with shared controls. When accounting for the directions of the diseases association signals, a different set of results were obtained in comparison to the results obtained using the max test that ignores the effect directions. Thus, in assessing the genetic relatedness between pairs of complex diseases, using the adaptation presented in this paper will provide a more specific account as to what those genetic similarities represent. That is, if the null hypothesis is rejected between any pair of complex diseases, it is concluded that the pair of complex diseases has at least one simultaneous association signal with effects in the same direction. A major advantage of the adaptation presented in this paper is its ease of implementation on readily available summary-level data.

Our analysis of the pAID data has clearly indicated sharing of common variants among these pAIDs. The logical next step is to develop methods that effectively utilize such a sharing in identifying additional SNPs that are associated with such autoimmune diseases. We propose a sequential identification procedure utilizing the proposed global detection tests. The sequential identification procedure identifies the top drivers of genetic sharing between disease pairs and gives us reason to believe SNPs of the applied pAID GWAS are independent. Thus, the dependency among SNPs genome-wide has no effect on our results in this pAID application. Another approach is to apply the conditional false discovery rate procedure to improve detection of common variants associated with these diseases, as shown to be quite effective in joint analysis of Schizophrenia and bipolar disorder (Andreassen et al., 2013).

CHAPTER 3

DETECTION OF SHARED GENETIC VARIANTS BETWEEN COMPLEX DISEASES WHILE PRESERVING DEPENDENCY STRUCTURE

3.1. Introduction

Genome-wide association studies (GWAS) have identified thousands of complex disease-associated single nucleotide polymorphisms (SNPs) (Hindorff et al., 2009). Although, these identified SNPs only explain a small proportion of complex disease heritability (Manolio et al., 2009). Unlike Mendelian diseases, the genetic architecture of complex diseases is largely unknown. Complex disease groups, such as autoimmune diseases, are hypothesized to have overlapping genetic etiologies driven by pleiotropic disease-associated SNPs, or SNPs with the ability to contribute to multiple disease phenotypes. Complex diseases are also thought to be polygenic in nature, in that many genetic variants with small effect sizes influence the observed disease phenotype rather than fewer genetic variants with large effect sizes, as is typical of Mendelian traits and diseases (Chung et al., 2014). The polygenicity of complex diseases has been supported by recent GWAS (Morris, Voight, and Teslovich, 2012), and it is the polygenic nature of these complex diseases that makes detection and identification of disease-associated SNPs particularly difficult (Manolio et al., 2009). Improving methods for the detection and identification of genetic risk factors, whether pleiotropic or polygenic, shared across complex diseases is key for bettering our understanding of complex disease genetic architecture (Cross-disorder Group of the Psychiatric Genomics Consortium, 2013).

A diverse set of methods have been developed for studying the genetic relatedness of complex diseases. Several methods have introduced meta-like statistics combining summary-level GWAS data of clinically-distinct complex diseases to identify SNPs associated with a subset of the diseases (Bhattacharjee et al., 2012; Cotsapas et al., 2011). More recently, methods have focused on the detection and quantification of genetic sharing between pairs of complex diseases (Bulik-Sullivan et al., 2015; Kobie et al., 2015). In Chapter 2 we proposed a global test as a means of combining GWAS summary-level Z-scores across complex diseases to detect whether a pair of complex diseases exhibited evidence of genetic sharing. Specifically, the proposed global test scans aligned pairs of complex disease GWAS Z-scores in search of at least one genetic variant, or identical SNP,

associated with both diseases of the pair. The test proposed by Kobie et al., 2015 detected, with strong statistical evidence, two pairs of pediatric autoimmune diseases (pAIDs): common variable immunodeficiency (CVID) - juvenile idiopathic arthritis (JIA) and ulcerative colitis (UC) - Crohn's disease (CD), to have at least one identical disease-associated SNP in both diseases. These results are consistent with those of Bhattacharjee et al., 2012 and Cotsapas et al., 2011, also showing evidence of genetic sharing among immune-mediated diseases.

Bulik-Sullivan et al., 2015, too, utilized GWAS summary-level Z-scores to identify pairs of complex traits and diseases that exhibited evidence of genetic sharing. Though, rather than testing for the existence of genetic sharing as Kobie et al., 2015, Bulik-Sullivan et al., 2015 quantified genetic sharing by estimating pairwise genetic correlations. While each of the aforementioned methods provide us with some information about the genetic relatedness of complex disease pairs, each method relies on different sets of biological assumptions and has varied data restrictions.

The test proposed by Kobie et al., 2015 assumes sparsity, in that the power of the test improves as the number of disease-associated variants becomes sparser. And, of particular interest, the test proposed by Kobie et al., 2015 assumes independence, and thus, the statistical significance can be evaluated using permutation. A majority of genetic tests assume SNPs are independent of one another, and while the validity of this assumption is often ignored, we know SNPs are not independent when in linkage disequilibrium (LD) (Lin, 2006). The assumed independence between SNPs is typically 'achieved' through a LD pruning procedure which selects SNPs based on arbitrarily constructed LD blocks, unnecessarily throwing away potentially valuable data. Bulik-Sullivan et al., 2015 may escape the independence assumption, modeling the product of the marginal Z-scores for a pair of diseases by a previously defined LD score in their pairwise estimation of the genetic correlation. Though Bulik-Sullivan et al., 2015 does make some SNP independence assumptions in their simulations, suggesting their method is not robust to highly dependent SNPs.

In this chapter we aim to evaluate the statistical significance of the global detection test proposed by Kobie et al., 2015 while preserving the inherent dependency structure across the genome. Specifically, we implement a perturbation method proposed by Lin and Zou, 2004, and expanded on by Zou et al., 2004, that exploits independent standard normal random variables to emulate the null distribution, as the standard normal random variables are independent of the observed SNP data. The perturbation method (Lin, 2005; Lin and Zou, 2004; Zou et al., 2004) utilizes the raw,

individual-level GWAS data of multiple complex diseases, rather than the summary-level GWAS Z-scores used in the global detection tests proposed by Kobie et al., 2015. Thus, we redefine the global detection statistics of Kobie et al., 2015. Unless genotyped on the same platform, GWAS of multiple diseases have differing sets of SNPs presenting an analysis challenge, in that SNP sets across complex disease pairs cannot be perfectly aligned. Allowing for dependency among SNPs in the statistical evaluation alleviates such data restrictions, enabling the use of multiple imputation to create matching sets of SNPs across disease pairs for complete SNP alignment.

In using the perturbation method, we eliminate the need for an independence assumption for valid statistical evaluation of the test statistics proposed by Kobie et al., 2015. With such, we can alter the global tests proposed by Kobie et al., 2015 to address an additional limitation: Kobie et al., 2015 assumes diseases with some shared genetic architecture will have disease-associated genetic variants at identical SNPs. Though with the inherent dependency between genetic variants, as previously discussed with respect to LD, and the thought that only one causal variant exists within independent blocks (Pickrell, 2014; Veyrieras et al., 2008), power for detecting a shared genetic architecture between a pair of diseases is likely lost by limiting our analysis to the detection of simultaneous disease-associated variants at identical SNPs. To increase the power of detecting a shared genetic architecture, we extended the simultaneous detection of disease-associated variants at identical SNPs to the detection of disease-associated variants within a conservative, LD-defined window.

This paper is organized as follows: We first briefly review the global detection test statistic proposed by Kobie et al., 2015 and redefine the statistic such that, the correlated nature of SNPs is incorporated into the evaluation of its statistical significance. We then extend the test statistic from detecting shared genetic variants at identical SNPs to detecting shared genetic variants within a LD-defined window. With simulations we evaluate the power and type I error of the global detection test statistic, and its extension to detection within a LD-defined window, using the perturbation procedure. We then apply the perturbation procedure in the reevaluation of the global detection test statistic proposed by Kobie et al., 2015, and evaluation of its extension, for GWAS data of 4 pediatric autoimmune diseases (pAIDs): common variable immunodeficiency (CVID), ulcerative colitis (UC), type I diabetes (T1D) and Crohn's disease (CD), with shared controls, in order to investigate the genetic relatedness among these diseases without assuming independence. We conclude the

paper with a discussion of the methods and the results.

3.2. Statistical Evaluation of Simultaneous Detection via Perturbation Procedure

Assume individual-level genotype GWAS data are readily available for a pair of complex diseases, disease A and disease B, of which are hypothesized to have some shared genetic architecture. Let Y_j be a binary indicator of whether the j^{th} individual ($j = 1, \cdots, N_A$) has disease A, and let Z_k be a binary indicator of whether the k^{th} individual ($k = 1, \cdots, N_B$) has disease B. Let X_{ji} be the genotype for the i^{th} SNP ($i = 1, \cdots, n$) of the j^{th} individual from the GWAS of disease A. Similarly, let W_{ki} be the genotype for the i^{th} SNP ($i = 1, \cdots, n$) of the k^{th} individual from the GWAS of disease B. Here, the genotypes, X_{ji} and W_{ki}, are coded under an assumed additive model. Also notice, the number of individuals across studies, N_A and N_B, can vary while the number, n, and identity of SNPs across studies should be identical.

The global detection test proposed by Kobie et al., 2015 was formulated utilizing GWAS summary-level marginal Z-scores rather than individual-level genotype GWAS data. Summary-level Z-scores can be obtained directly from the individual-level genotype data. As in Kobie et al., 2015, let U_i be the Z-score of the marginal association between disease A and the i^{th} SNP and, similarly, let V_i be the Z-score of the marginal association between disease B and the i^{th} SNP. Specifically, $U_i = \hat{\beta}/\text{SE}(\hat{\beta})$, where β and SE(β) are estimated from the marginal logistic regression model, logit($P(Y_j = 1)$) = $\gamma + \beta X_{ji}$, and where $P(Y_j = 1)$ is the probability of the j^{th} individual having disease A. Similarly, $V_i = \hat{\alpha}/\text{SE}(\hat{\alpha})$, estimated from logit($P(Z_k = 1)$) = $\delta + \alpha W_{ki}$, where $P(Z_k = 1)$ is the probability of the k^{th} individual having disease B. Under the null hypothesis of no association between the i^{th} SNP and disease A, or disease B, U_i and V_i are approximately normal with mean 0 and variance 1. Whereas, in the presence of an association, $U_i \sim N(\mu_i, \sigma_i^2)$ and $V_i \sim N(\nu_i, \tau_i^2)$. Kobie et al., 2015 proposed the max test statistic summarizing across both diseases of the pair and over all SNPs,

$$M_n = \max_{i=1,\cdots,n} T_i,$$

where $T_i = |U_i| \wedge |V_i|$, to test the following set of hypotheses

$$H_0 : \text{for all } i,\ |\mu_i| \wedge |\nu_i| = 0$$

$$H_A : \text{there is at least one } i \text{ such that, } |\mu_i| \wedge |\nu_i| \neq 0.$$

That is, under the null hypothesis SNP i is not simultaneously associated with both diseases, or for all n SNPs, at least one Z-score of the pair for SNP i, U_i and V_i, has a mean of 0. Under the alternative hypothesis at least one SNP i has a pair of Z-scores, U_i and V_i, with nonzero means, μ_i and ν_i. When H_0 is rejected, disease A and disease B are concluded to have some degree of overlapping genetic architecture. Kobie et al., 2015 evaluated the statistical significance of M_n with a formula derived utilizing a permutation distribution assuming independence between SNPs.

3.2.1. Redefine U_i and V_i for Evaluation of M_n with Perturbation Method

To evaluate the statistical significance of M_n while preserving the inherent dependency structure across the genome, or evaluate the significance of M_n without assuming independence, we apply the perturbation method by Lin and Zou, 2004, Zou et al., 2004 and Lin, 2005. The perturbation method is formulated using the efficient score function, thus requiring individual-level data (Lin, 2005; Lin and Zou, 2004; Zou et al., 2004). We redefine U_i and V_i as score statistics for the i^{th} SNP

$$U_i = R_i^T Q_i^{-1} R_i$$
$$V_i = P_i^T S_i^{-1} P_i,$$

where R_i and P_i are score functions,

$$R_i = \sum_{j=1}^{N_A} R_{ji}$$
$$P_i = \sum_{k=1}^{N_B} P_{ki},$$

and where Q_i and S_i are the respective covariance matrices of R_i and P_i,

$$Q_i = \sum_{j=1}^{N_A} R_{ji} R_{ji}^T$$
$$S_i = \sum_{k=1}^{N_B} P_{ki} P_{ki}^T.$$

R_{ji} and P_{ki} are efficient score functions and represent the data contributions from the j^{th} and k^{th} individuals of disease A and disease B, respectively. Here,

$$R_{ji} = (Y_j - \gamma_y)(X_{ji} - \pi_i)$$
$$P_{ki} = (Z_k - \delta_z)(W_{ki} - \omega_i),$$

where γ_y and δ_z are the proportion of diseased A and diseased B and π_i and ω_i are the population means of X_{ji} and W_{ki}, respectively. $R_{ji} \overset{iid}{\sim} N(0, R_{ji}R_{ji}^T)$ and $P_{ki} \overset{iid}{\sim} N(0, P_{ki}P_{ki}^T)$, thus under the null hypothesis of no association between the i^{th} SNP and disease A, or disease B respectively, $R_i \sim N(0, Q_i)$ and $P_i \sim N(0, S_i)$. Since the j individuals of disease A GWAS are independent, $var(R_i) = var(\sum_{j=1}^{N_A} R_{ji}) = \sum_{j=1}^{N_A} var(R_{ji})$. The same holds for $var(P_i)$ of disease B. Additionally, $(R_1, \cdots, R_n)$ and $(P_1, \cdots, P_n)$ are approximately multivariate normal with mean zero and covariance matrices $cov(R_i, R_f) = \sum_{j=1}^{N_A} R_{ji}R_{jf}^T$ and $cov(P_i, P_f) = \sum_{k=1}^{N_B} P_{ki}P_{kf}^T$, where $f \neq i$. That is,

$$\begin{pmatrix} R_1 \\ \vdots \\ R_n \end{pmatrix} \sim N_n \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{bmatrix} Q_1 & \cdots & cov(R_1, R_n) \\ \vdots & \ddots & \vdots \\ cov(R_n, R_1) & \cdots & Q_n \end{bmatrix} \right)$$

$$\begin{pmatrix} P_1 \\ \vdots \\ P_n \end{pmatrix} \sim N_n \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{bmatrix} S_1 & \cdots & cov(P_1, P_n) \\ \vdots & \ddots & \vdots \\ cov(P_n, P_1) & \cdots & S_n \end{bmatrix} \right).$$

Note, here, $cov(R_i, R_f)$ and $cov(P_i, P_f)$ can be nonzero because SNPs are not assumed to be independent. With U_i and V_i redefined, we can let $M_n = \underset{i=1,\cdots,n}{max} T_i$, where $T_i = U_i \wedge V_i$, as previously defined, to test whether a pair of complex diseases, disease A and disease B, exhibit some evidence of genetic sharing. Note because U_i and V_i are redefined as score statistics, and are thus always positive, T_i is no longer defined wtih the absolute value of U_i and V_i as in Kobie et al., 2015. The null distribution of M_n cannot be explicitly defined as SNPs, and thus U_i (and V_i), are not independent of one another. As Lin and Zou, 2004, Zou et al., 2004 and Lin, 2005, we regard the large sample distributions of U_i and V_i as stochastic processes in n SNPs, and we

aim to approximate the null distributions of U_i and V_i with a Monte Carlo approach. We generate perturbed replicates under the null defined as

$$\tilde{U}_i = \tilde{R}_i^T Q_i^{-1} \tilde{R}_i$$
$$\tilde{V}_i = \tilde{P}_i^T S_i^{-1} \tilde{P}_i,$$

where

$$\tilde{R}_i = \sum_{j=1}^{N_A} R_{ji} G_j$$
$$\tilde{P}_i = \sum_{k=1}^{N_B} P_{ki} G_k,$$

where G_j, G_k are independent standard normal random variables, $N(0,1)$. Because G_j and G_k are independent of R_{ji} and P_{ki}, perturbations have no effect on the variance or covariance of R_i and P_i,

$$\begin{pmatrix} \tilde{R}_1 \\ \vdots \\ \tilde{R}_n \end{pmatrix} \sim N_n \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{bmatrix} Q_1 & \cdots & cov(R_1, R_n) \\ \vdots & \ddots & \vdots \\ cov(R_n, R_1) & \cdots & Q_n \end{bmatrix} \right)$$

$$\begin{pmatrix} \tilde{P}_1 \\ \vdots \\ \tilde{P}_n \end{pmatrix} \sim N_n \left(\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{bmatrix} S_1 & \cdots & cov(P_1, P_n) \\ \vdots & \ddots & \vdots \\ cov(P_n, P_1) & \cdots & S_n \end{bmatrix} \right).$$

From the above, we can compute a perturbed $\tilde{M}_n$ under the null distribution such that

$$\tilde{M}_n = \underset{i=1,\cdots,n}{max} \tilde{T}_i,$$

where

$$\tilde{T}_i = max \begin{cases} \tilde{U}_i \wedge \tilde{V}_i \\ U_i \wedge \tilde{V}_i \\ \tilde{U}_i \wedge V_i \end{cases} .$$

The null hypothesis is composed of three possible scenarios: (1) SNP i is not associated with either disease A or disease B, (2) SNP i is associated with disease A and not associated with disease B and (3) SNP i is associated with disease B and not disease A. Perturbations are carried out to emulate all possible null scenarios, though taking the maximum over all three is a conservative approach. The intuition for emulating the null distribution is reminiscent of He et al., 2013 and Liu et al., 2010, where they use a multivariate normal random vector to approximate the null distribution of swapping cases and controls. To evaluate the statistical significance of M_n, we compute several perturbations, 1000 to 100000, $\tilde{M}_n$ and let the p-value be the proportion of perturbed $\tilde{M}_n$ greater than the observed M_n. Unlike permutation, the perturbation method involves only the simulation of standard normal random variables and does not require shuffling of the data, allowing its implementation in a variety of data structures (Lin, 2005).

3.2.2. Extension of Simultaneous Detection to LD-defined Window

With the ability to evaluate the statistical significance of M_n without making an independence assumption, we extend Kobie et al., 2015 from detecting simultaneous disease-associated variants at identical SNPs to detecting simultaneous disease-associated variants within a LD-defined window. For the i^{th} SNP we define a LD window of size C_i, such that the l^{th} SNP within the window, $l \in C_i$, and the i^{th} SNP have a $r^2 \geq 0.5$, where r^2 is the coefficient of determination, or square of the correlation coefficient. The exact specification of the LD window is arbitrary, though it is worth noting all r^2 were derived from the control population void of known autoimmunity or immunodeficiency diagnoses. Also note, LD windows for various SNPs will be overlapping, reinforcing the need for a method of evaluating the statistical significance in the absence of an independence assumption. The potential problems posed by the overlapping of windows, and varying of window sizes, from

SNP-to-SNP are discussed in the sections that follow. Now consider, for the i^{th} SNP,

$$O_i = \max_{l=1,\cdots,C_i} U_l$$
$$V_i = P_i^T S_i^{-1} P_i,$$

where U_l and V_i are defined identically to U_i and V_i in section 3.2.1, and $M_n = \max_{i=1,\cdots,n} T_i$, where now

$$T_i = O_i \wedge V_i.$$

We can evaluate the statistical significance of M_n for detecting at least one simultaneous disease-associated variant within a LD-defined window with the perturbation method described previously, in section 3.2.1. We again generate perturbed replicates $\tilde{M}_n$ under the null, where $\tilde{M}_n = \max_{i=1,\cdots,n} \tilde{T}_i$, and

$$\tilde{T}_i = max \begin{cases} \tilde{O}_i \wedge \tilde{V}_i \\ O_i \wedge \tilde{V}_i \\ \tilde{O}_i \wedge V_i \end{cases}.$$

The p-value is again the proportion of perturbed $\tilde{M}_n$ greater than the observed M_n. It is important to note when perturbing O_i, we are actually perturbing a window of C_i efficient score functions. Recall, each SNP i has a corresponding window of size C_i, which varies from SNP to SNP. SNPs in high LD with surrounding SNPs will have a larger corresponding window size. Thus, the size of the window likely impacts the effectiveness of the perturbation procedure in emulating the null distribution of no disease-associated genetic variants.

3.3. Simulation Studies by Resampling

Datasets were generated under the null and alternative hypotheses to investigate the type I error and power of the proposed perturbation evaluation method and its extension from simultaneous detection at identical SNPs to detection within a LD-defined window. To maintain the inherent dependency structure observed among genetic variants in the human genome, a prediction-based resampling procedure was used to simulate case-control status and generate datasets of genotypes for one chromosome. The prediction-based resampling procedure implemented here used control

data from the described real data application (Section 3.4) to independently generate datasets for a pair of diseases. Specifically, genotypes of 36760 typed SNPs on chromosome 1 from 10718 pediatric controls were used.

To simulate case-control status, first, 20 uncorrelated, common (minor allele frequency (MAF) $>$ 0.05) causal variants were selected from the 36760 SNPs of chromosome 1. The number of causal variants selected, 20, mirrors the sparsity of causal variants in the entire genome. The sparse set of causal variants was then used in an additive, multivariable logistic regression prediction model to obtain predicted probabilities of disease for each of the 10718 pediatric controls. Causal variant effect sizes (e^{β}), used in the prediction model, were randomly generated between $1.2 - 1.5$ and the prevalence of disease was assumed to be approximately 15%. Case-control status was assigned to each of the 10718 pediatric controls using binomial random variables generated from the predicted probabilities. To generate a disease dataset, the genotypes of a sample of 3000 individuals, of the 10718, were selected with 2 controls for every case.

The resampling procedure detailed above was then completed again for the second disease of the pair, with a new set of causal variants, causal effect sizes and resulting predicted probabilities. Of the 20 causal variants, a proportion were allowed to overlap between the 2 diseases of the pair. In particular, we are interested in the null scenario with no overlap (0%) of causal variants between the 2 diseases and the alternative scenario with 100% overlap of causal variants between the 2 diseases. Under the null the generated pair of diseases has no shared disease-associated genetic variants, indicating the diseases are not genetically related. While under the alternative the generated pair of diseases shares all disease-associated genetic variants, indicating the diseases are genetically related.

We simulated pairs of disease datasets 1000 times, and for each pair of datasets, calculated p-values based on 1000 perturbations. The QQ-plot in Figure 3.1 shows the permutation method is conservative, while the perturbation method appears to control the type I error at $\alpha = 0.05$ for small p-values, becoming more conservative for larger p-values. Figure 3.2 shows both 'no window' (detection at identical SNPs) and 'window' (detection within a LD-defined window) perturbation methods have a slightly inflated type I error at 0% shared causal SNPs. The permutation method, assuming independence, appears to be conservative with type I errors lower than the set α-levels indicated in Figure 3.2. Note, the 'no window' perturbation analysis and the permutation analysis

are detecting exactly the same thing, disease-associated variants at identical SNPs, but with different assumptions, while the 'window' perturbation analysis is an extension of detection at identical SNPs.

Figure 3.2 compares the power of the perturbation method, and its extension, to the power of the permutation method assuming independence for $5 - 25\%$ causal SNP overlap between the pair of diseases. As the proportion of shared variants between a pair of diseases exceeds 25%, the power gains among the 3 methods (permutation, 'no window' perturbation and 'window' perturbation) are indistinguishable. While the perturbation and permutation methods appear to have comparable power for the detection of shared genetic variants, Figure 3.2 demonstrates the perturbation methods to be more powerful in identifying genetic relatedness when the proportion of shared disease-associated variants is less than 25%. At both $\alpha = 0.05$ (Figure 3.2, leftmost) and $\alpha = 0.10$ (Figure 3.2, rightmost), the 'no window' and 'window' perturbation methods perform similarly, 'no window' showing some increased power over the 'window' perturbation method.

Figure 3.3 considers an alternative overlap scenario that defines SNP overlap as SNPs shared within a dependency-defined window rather than shared identically. This alternative scenario can be generated similarly to that described above, but rather than allowing the percentage of shared SNPs to match identically across the pair of diseases, a shared SNP is a randomly chosen SNP within the corresponding causal SNP's LD-defined window. LD-defined windows include SNPs with $r^2 \geq 0.5$ within a 1MB block (500kB upstream and 500kB downstream) of the SNP in question. Under the alternative definition of genetic overlap, the 'window' perturbation method outperforms the 'no window' perturbation method, thus there is a potential for improved power of detection with the 'window' perturbation method (Figure 3.3). One could consider increasing the number of causal SNPs to better observe separation between the 'no window' and 'window' perturbation methods.

3.4. Analysis of Genetic Sharing of 4 Pediatric Autoimmune Diseases

Following the analyses of Kobie et al., 2015, we aim to evaluate whether particular pairs of pediatric autoimmune diseases (pAIDs) exhibit evidence of genetic sharing without making any independence assumptions. Our analyses focus on pairs of the previously investigated pAIDs: common variable immunodeficiency (CVID), ulcerative colitis (UC), type I diabetes (T1D) and Crohn's disease (CD) (Kobie et al., 2015; Li et al., 2015b). Genotypes of 473228 SNPs genome-wide were

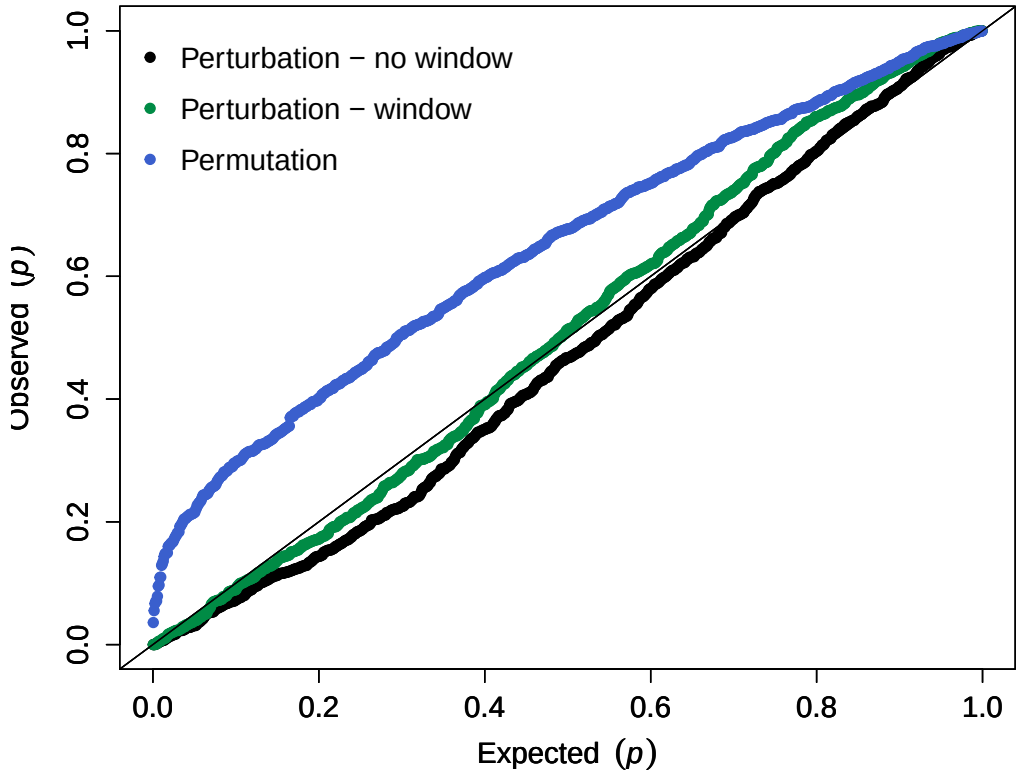

Figure 3.1: QQ-plot for the perturbation method (based on 1000 perturbations) in the detection of shared genetic variants at an identical SNP ('no window') and the detection of shared genetic variants within a LD-defined window ('window') compared to the permutation method estimated from 1000 replications at $\alpha = 0.05$ for 0% causal SNP overlap

obtained for 308 CVID cases, 865 UC cases, 1086 T1D cases, 1922 CD cases and 10718 shared controls (Li et al., 2015b). With individual-level genotype data, perturbation-based p-values can be calculated, assessing the statistical significance of the max statistic, M_n, detecting shared disease-associated variants at identical SNPs and within a LD-defined window, as described in Sections 3.2.1 and 3.2.2 respectively. It is important to note, only autosomal SNPs were included in the analyses to follow, excluding mitochondrial SNPs and SNPs found on either of the two sex chromosomes, X and Y. SNPs within the major histocompatibility complex (MHC) region were also excluded (Kobie et al., 2015).

Table 3.1 presents pAIDs pairwise perturbation p-values, based on 10000 perturbations, for detect-

ing at least one simultaneous disease-associated variant at identical SNPs and for detecting at least one simultaneous disease-associated variant within a LD-defined window. LD-defined windows include SNPs with $r^2 \geq 0.5$ within a 1MB block (500kB upstream and 500kB downstream) of the SNP in question. Table 3.1 also includes all possible pAIDs pairwise permutation-based p-values, assuming independence, for detecting at least one simultaneous disease-associated variant at identical SNPs (Kobie et al., 2015) for comparison.

Note similar conclusions are drawn of disease pairs when evaluating the significance of the max test statistic, M_n, with permutation or perturbation methods. For example, disease pairs CVID-UC, CVID-T1D and CVID-CD were not found to exhibit evidence of genetic sharing by Kobie et al., 2015 at the nominal $\alpha = 0.05$ level, and the same conclusions are drawn using perturbation methods, 'no window' and 'window', with comparable 'no window' p-values of 0.2582, 0.6924 and 0.5575, respectively (Table 3.1). Similarly, the disease pair UC-CD still shows strong evidence of genetic sharing after adjusting for multiple testing. The precision of the perturbation p-value is dependent on the number of perturbations performed. This is not true of the permutation method implemented for evaluation by Kobie et al., 2015, as an analytical approximation of the exact permutation p-value was proposed. And, while similar conclusions are drawn using both the permutation and perturbation evaluation methods for this particular dataset, the same may not hold in other datasets, especially those with a large proportion of SNPs in high LD with one another.

Table 3.1: Perturbation p-values (based on 10000 perturbations) for the detection of shared disease-associated variants at identical SNPs and within a LD-defined window compared to permutation-based p-values by Kobie et al., 2015

Disease Pair	Kobie et al., 2015	Identical SNP Detection	LD Window Detection
CVID-UC	0.8168	0.2582	0.3956
CVID-T1D	0.0337	0.6924	0.8417
CVID-CD	0.0391	0.5575	0.2194
UC-T1D	0.4277	0.0612	0.1049
UC-CD	3.76e-04	<10e-04	<10e-04

3.5. Conclusion

This chapter implements a perturbation method (Lin, 2005; Lin and Zou, 2004; Zou et al., 2004) to evaluate the statistical significance of a max test statistic, M_n, for the detection of sharing between pairs of diseases. The perturbation method, unlike the commonly used permutation method, does not make any independence assumptions, allowing the inherent dependency structure among

SNPs to remain. Specifically, the perturbation method mimics the null distribution of M_n, exploiting standard normal random variables to preserve the covariance between score functions at the SNP-level. In accounting for the inherent dependency among genetic variants in the evaluation of our max test statistic, M_n, our test can be applied to varied data structures, such as large imputed datasets, and scenarios, including extending the max test to detect shared genetic variants within a LD-defined window. In extending the max test to detect shared genetic variants within a LD-defined window, the number and identity of SNPs are not required to be identical across the pair of GWAS.

By accounting for the dependency structure of the genome, we can improve the power of detecting whether a pair of complex diseases shares sparse disease-associated genetic variants. In simulations, using the proposed 'no window' perturbation method is more powerful for detecting identical shared SNPs than the permutation method and 'window' perturbation method when the proportion of sharing is low. When altering our definition of genetic sharing from identical SNPs to SNPs within a dependency-defined window, 'no window' and 'window' perturbation methods perform more similarly, while still outperforming the permutation method when the proportion of sharing is low. In defining sharing within a window rather than at identical SNPs in simulation, the 'window' perturbation method has the potential to perform better than 'no window' perturbation method. The lack of separation is likely due to the complexity of the window analysis, in that window sizes vary SNP-to-SNP and windows are highly dependent, overlapping with one another. Thus, the perturbation method may not be the most accurate representation of the null distribution of the window extension (Section 3.2.2). To further investigate the separation between the 'no window' and 'window' perturbation methods, we can consider increasing the number of causal SNPs.

The 'no window' and 'window' perturbation methods were applied to all possible pairs of 4 clinically-distinct pediatric autoimmune diseases (pAIDs) and compared the results to the results obtained from permutation methods. The perturbation results are consistent with the permutation-based results. Unlike the permutation method which has an analytical expression for determining the exact p-value (Kobie et al., 2015), the accuracy of perturbation p-values is dependent on the number of perturbations. With 10000 perturbations we observe perturbation p-values of 0 for genetically related disease pair UC-CD.

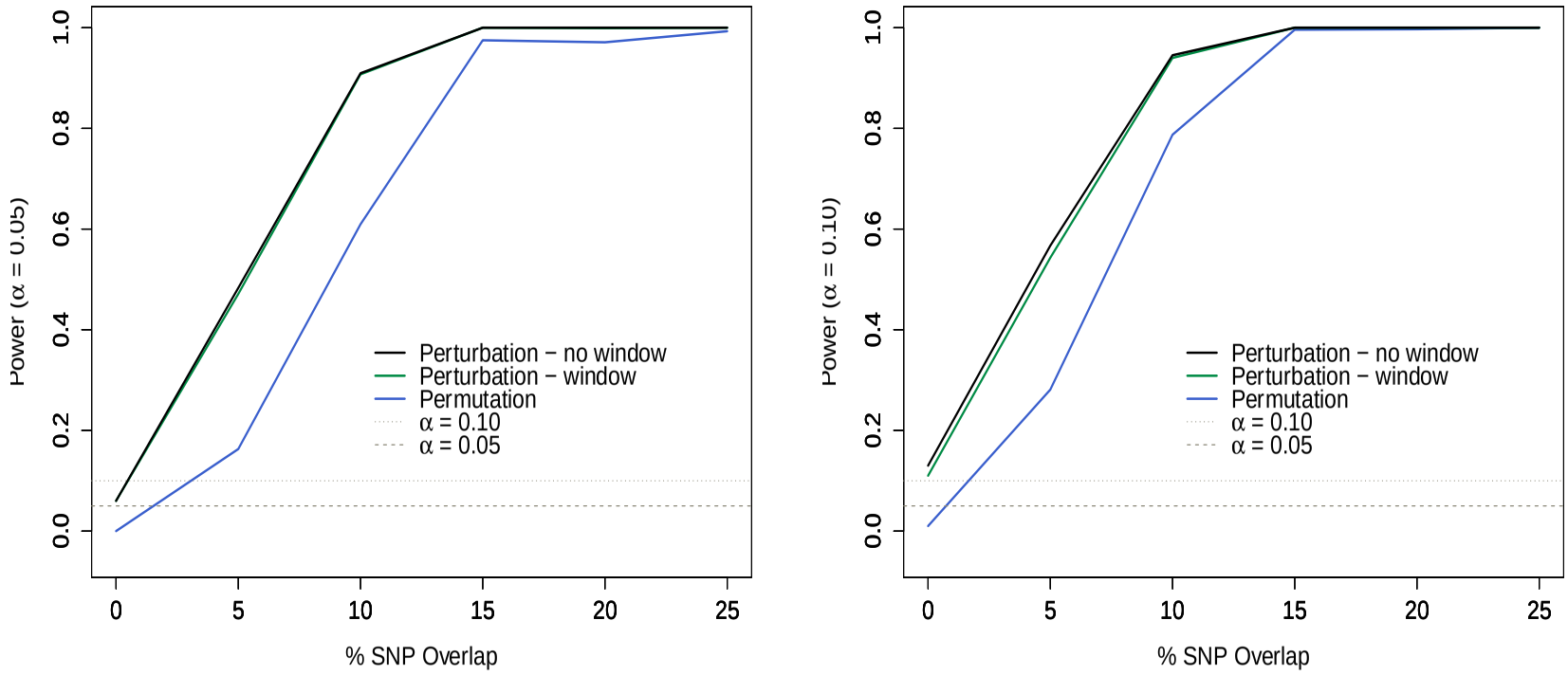

Figure 3.2: Power curve for the perturbation method (based on 1000 perturbations) in the detection of shared genetic variants at an identical SNP ('no window') and the detection of shared genetic variants within a LD-defined window ('window') compared to the permutation method estimated from 1000 replications at $\alpha = 0.05$ (leftmost) and $\alpha = 0.10$ (rightmost) for $0, 20, 40, 60, 80$ and 100% causal identical SNP overlap

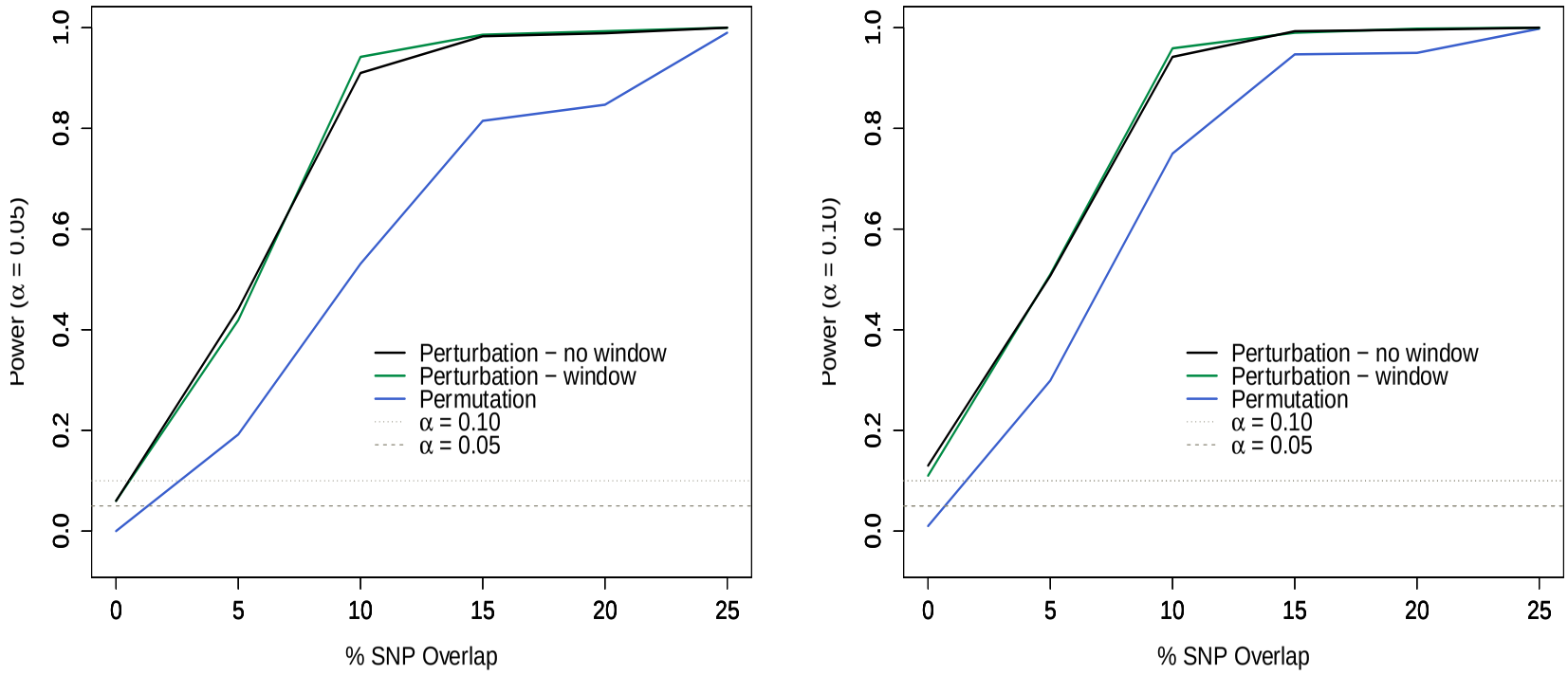

Figure 3.3: Power curve for the perturbation method (based on 1000 perturbations) in the detection of shared genetic variants at an identical SNP ('no window') and the detection of shared genetic variants within a LD-defined window ('window') compared to the permutation method estimated from 1000 replications at $\alpha = 0.05$ (leftmost) and $\alpha = 0.10$ (rightmost) for $0, 20, 40, 60, 80$ and 100% causal overlap within a window

CHAPTER 4

STATISTIC QUANTIFYING SHARED GENETIC VARIANTS BETWEEN COMPLEX DISEASES

4.1. Introduction

Understanding the shared genetic architecture of complex diseases has important epidemiological implications, including improved treatment and prevention efforts. While genome-wide association studies (GWAS) have effectively identified thousands of complex disease-associated genetic variants (Hindorff et al., 2009; Manolio et al., 2009), pooling inferences across clinically-distinct complex disease GWAS enables statistical and scientific gains (Benjamini and Heller, 2008). By integrating clinically-distinct complex disease GWAS, the power of detecting a shared genetic architecture is improved while also providing a united view of the underlying biological systems (Cai and Tan, 2015). Several methods have made strides in the detection (Cai and Tan, 2015; Kobie et al., 2015) and identification (Andreassen et al., 2013; Bhattacharjee et al., 2012; Chung et al., 2014; Cotsapas et al., 2011) of shared disease-associated genetic variants between pairs of complex diseases and sets of complex traits, though the degree to which these pairs of diseases, or sets of traits, have an overlapping genetic architecture in comparison to other pairs and sets is not well defined.

Improving methods in the detection, identification and quantification of shared genetic variants across complex diseases is vital for bettering our understanding of complex disease genetic etiology. A standardized quantification of complex disease pairs' genetic relatedness, for subsequent comparison to other disease pairs, is of particular interest in treatment prioritization. Because GWAS allow for the determination of variant-specific effect sizes, an overlapping genetic architecture can be tested through the detection of shared disease-associated genetic variants and characterized through the quantification of correlated effect sizes across complex diseases (Bulik-Sullivan et al., 2015). A common approach in characterizing the level of genetic relatedness, or quantifying genetic sharing, between a pair of complex diseases or traits is to estimate the genetic correlation. Genetic correlation is the genome-wide aggregate of shared disease-associated variant effect sizes without imposing any thresholding restraints, thus, including variants that do not

reach genome-wide significance. The existing methods for estimating the genetic correlation with GWAS data face limitations in wide applicability.

Coheritability is a commonly used estimate of the genetic correlation between a pair of binary complex diseases, or pair of quantitative traits. The concept of coheritability as an estimate of the genetic correlation is an extension of the concept of single-disease heritability, or the proportion of variance in a disease phenotype explained by the genetic variation in the population (Yang et al., 2010). Coheritability is estimated using a restricted maximum likelihood (REML) approach and is, specifically, an estimate of the genetic covariance in a linear mixed model framework divided by the product of the estimates of the respective single-disease genetic standard deviations (Lee et al., 2012a). Estimating the genetic correlation using the REML approach relies on individual-level genotype data, which is often difficult to obtain (Lee et al., 2012a).

More recently Bulik-Sullivan et al., 2015 proposed a method for estimating pairwise genetic correlations, relying only on readily available summary-level GWAS data rather than individual-level genotype data. Bulik-Sullivan et al., 2015 exploit the documented relationship between single nucleotide polymorphisms (SNPs) in high LD and their corresponding effect sizes to estimate the genetic correlation, specifically by modeling the product of the marginal Z-scores for SNPs of a pair of diseases, or traits, with respect to those SNPs' LD scores (Bulik-Sullivan et al., 2015). The estimation technique proposed by Bulik-Sullivan et al., 2015 is dependent on prior information of the diseases', or traits', underlying genetic architecture, as represented by single-disease heritability estimates.

In this chapter we propose a statistic to quantify the genetic relatedness between a pair of complex diseases using summary-level GWAS data. The statistic acts as an estimate of the genetic correlation among shared disease-associated genetic variants and was largely motivated by the quadratic functional proposed by Cai and Tan, 2015. Cai and Tan, 2015 propose a quadratic functional under a two-sequence Gaussian model for the detection and quantification of simultaneous signals, or, in our application, shared disease-associated variants. In addition, Cai and Tan, 2015 devise an optimal estimation method of the quadratic functional assuming both normal mean vectors are sparse, as is typical of genomics applications. They standardize their estimator by the number of SNPs studied (Cai and Tan, 2015), while our proposed quantity adjusts for both the number of SNPs studied and the respective sample sizes of GWAS pairs. Because our quantity is a function

of SNP effect sizes, the varied sample size across studies must be accounted for, as the magnitude of the observed effect size is driven in part by the sample size.

In addition to quantifying the genetic relatedness of complex disease pairs, we consider weighting that quantification with respect to the functional annotations of genetic variants. That is, giving a higher quantification of genetic relatedness to disease pairs with shared disease-associated variants of a larger functional importance. Often SNPs are assumed to be exchangeable with one another, but that is certainly not the case when working with genome-wide data, especially with respect to a SNP's functional annotation. We use a transformed Eigen score as a functional weight of our quantification of genetic relatedness. Ionita-Laza et al., 2015 proposed the Eigen score as a meta-score integrating various functional annotations into one. The Eigen score is an all-encompassing functional score that outperforms any single individual functional annotation, thus eliminating the need to choose a particular annotation a priori (Ionita-Laza et al., 2015).

This chapter is organized as follows. We first propose a standardized quantification measure utilizing summary-level GWAS data to characterize the genetic relatedness between a pair of complex diseases, adjusting for varied GWAS sample sizes. We then apply a functional weight to the quantification measure in the form of an Eigen score. We obtain the proposed quantification measures for all possible pairs of GWAS data of 4 clinically-distinct, pediatric autoimmune disease (pAIDs): common variable immunodeficiency (CVID), ulcerative colitis (UC), type I diabetes (T1D) and Crohn's disease (CD), and compare the results to previously reported detection p-values.

4.2. Statistical Formulation of Genetic Sharing Quantification Measure

Assume GWAS data are readily available for a pair of complex diseases, disease A and disease B, from which summary-level marginal Z-scores are obtained. Let U_i be the Z-score of the marginal association between disease A and the i^{th} SNP ($i = 1, \cdots, n$) and, similarly, let V_i be the Z-score of the marginal association between disease B and the i^{th} SNP. Specifically,

$$U_i = \hat{\beta}_i / SE(\hat{\beta}_i)$$
$$V_i = \hat{\alpha}_i / SE(\hat{\alpha}_i),$$

where β_i, SE(β_i), α_i and SE(α_i) are estimated from their corresponding marginal logistic regression models

$$logit(P(Y_j = 1)) = \beta_0 + \beta X_{ji}$$
$$logit(P(Z_k = 1)) = \alpha_0 + \alpha W_{ki},$$

where P(Y_j = 1) is the probability the j^{th} individual ($j = 1, \cdots, N_A$) has disease A, P(Z_k = 1) is the probability the k^{th} individual ($k = 1, \cdots, N_B$) has disease B and X_{ji} and W_{ki} are the respective genotypes for the i^{th} SNP. Note, each study's sample size, N_A and N_B, can vary while the number, n, and identity of the SNPs should be identical.

4.2.1. A New Quantification Measure

The varied sample size across studies must be accounted for when quantifying the genetic relatedness between any given pair of diseases. The standard error (SE) includes a factor of the square root of the corresponding sample size, that is SE(β_i) = $\sqrt{\frac{var(\beta_i)}{N_A}}$ (or SE(α_i) = $\sqrt{\frac{var(\alpha_i)}{N_B}}$), which impacts the significance of the respective effect size. For a given effect size, the significance level, and the magnitude of the effect size itself, increases with an increasing sample size.

Under the alternative hypothesis, in the presence of an association, U_i and V_i are approximately normally distributed

$$U_i \sim N(\sqrt{N_A}\mu_i, 1)$$
$$V_i \sim N(\sqrt{N_B}\nu_i, 1),$$

where μ_i and ν_i are the standardized effect sizes for the i^{th} SNP of disease A and disease B respectively, independent of corresponding study sample sizes. To quantify the genetic relatedness between disease A and disease B, consider the quantity

$$Q(\mu, \nu) = \frac{\sum_{i=1}^{n} |\mu_i \nu_i|}{\sqrt{\sum_{i=1}^{n} \mu_i^2 \sum_{i=1}^{n} \nu_i^2}},$$

where $Q(\mu, \nu) \in (0, 1)$ by the Cauchy-Schwarz inequality. The quantity $Q(\mu, \nu)$ is a function of the

true effect sizes for the i^{th} SNP's association with disease A or disease B, μ_i and ν_i respectively, so the resulting magnitude of $Q(\mu, \nu)$ is independent of study sample size. More specifically, $Q(\mu, \nu)$ represents the correlation among SNPs associated with both diseases of the pair ($|\mu_i| \wedge |\nu_i| \neq 0$) while adjusting for the varied study sample sizes for diseases A and B. When $Q(\mu, \nu) = 0$, disease A and disease B bear no genetic relationship. That is, $Q(\mu, \nu) = 0$ when for all i $|\mu_i| \wedge |\nu_i| = 0$, or not one of the n SNPs studied is associated with both disease A and disease B. $Q(\mu, \nu) = 1$ when for all i $|\mu_i| \wedge |\nu_i| \neq 0$ and $|\mu_i| = |\nu_i|$.

To estimate $Q(\mu, \nu)$ we consider the estimation method by Cai and Tan, 2015, which optimally estimates a quadratic functional in a "sparse regime". A "sparse regime" is defined such that the proportion of nonzero signals (SNPs associated with disease) is bounded by $\frac{1}{\sqrt{n}}$ (Cai and Jeng, 2011; Cai and Tan, 2015; Jin and Donoho, 2004), as is typical of GWAS. Specifically, the numerator of $Q(\mu, \nu)$ is estimated by

$$\sum_{i=1}^{n} |\mu_i \nu_i| = \sum_{i=1}^{n} \left[(|U_i| - \sqrt{\log n})_+ - \mu_0 \right] \left[(|V_i| - \sqrt{\log n})_+ - \nu_0 \right],$$

where $\mu_0 = E_0(|U_i| - \sqrt{\log n})_+$ and $\nu_0 = E_0(|V_i| - \sqrt{\log n})_+$, and the denominator of $Q(\mu, \nu)$ is estimated by

$$\sum_{i=1}^{n} \mu_i^2 = \sum_{i=1}^{n} \left[(U_i^2 - 2 \log n)_+ - \mu_{0^2} \right]$$
$$\sum_{i=1}^{n} \nu_i^2 = \sum_{i=1}^{n} \left[(V_i^2 - 2 \log n)_+ - \nu_{0^2} \right],$$

where $\mu_{0^2} = E_0(U_i^2 - 2 \log n)_+$ and $\nu_{0^2} = E_0(V_i^2 - 2 \log n)_+$. μ_0, ν_0, μ_{0^2} and ν_{0^2} can be approximated with integration and because U_i and V_i have the same distribution under the null, $\mu_0 = \nu_0$ and $\mu_{0^2} = \nu_{0^2}$.

$$\begin{aligned} \mu_0 &= E_0(|U_i| - \sqrt{\log n})_+ \\ &= \int_0^\infty (|U_i| - \sqrt{\log n}) 2 f(|U_i| - \sqrt{\log 2n}) dU_i \end{aligned}$$

$$\mu_{0^2} = E_0(U_i^2 - 2\log n)_+$$
$$= \int_0^\infty (U_i^2 - 2\log n) g(U_i^2 - 2\log n) dU_i^2,$$

where $f(\cdot)$ is the probability density function (*pdf*) of the standard normal distribution and $g(\cdot)$ is the *pdf* of the central chi-squared distribution. Recall, the variance of U_i and V_i is assumed to be 1.

The sparsity of the data cannot be ignored, especially when estimating the magnitudes of SNP-specific effect sizes. The number of nonzero effect sizes is too small to be noticed in any sum of order n (Jin and Donoho, 2004). Cai and Tan, 2015 utilize a soft thresholding procedure to shrink effect sizes of a small magnitude toward the null, or 0. This allows the estimate of the genetic correlation, $\hat{Q}(\mu, \nu)$, to be driven by the sparse nonzero effect sizes, rather than the null effect sizes, which make up the majority. Their estimation procedure relies on the optimal effect size soft threshold for denoising: $logn$, in the two-sequence simultaneous signal detection case between a piar of diseases, and $2\log n$, in the one-sequence signal detection case. Cai and Tan, 2015 have an additional thresholding step, subtracting μ_0 or ν_0, was used for debiasing estimates of zero coordinates of μ and ν. Notice, the estimator $\hat{Q}(\mu, \nu)$ is a function of U and V rather than the unknown, true effect sizes μ and ν, thus the variability in our estimator is dependent on the sample size of each study.

4.2.2. Functional Weighting

Consider the functionally weighted quantity

$$Q_w(\mu, \nu) = \frac{\sum_{i=1}^n w_i |\mu_i \nu_i|}{\sqrt{\sum_{i=1}^n \mu_i^2 \sum_{i=1}^n \nu_i^2}},$$

where $w_i = \frac{nE_i}{\sum_{i=1}^n E_i}$, and E_i is a transformed Eigen score for the alternative allele of the i^{th} SNP. Ionita-Laza et al., 2015 provides allele-specific Eigen scores, from which an Eigen score for the alternative allele of the i^{th} SNP was transformed from the original scale $(-\infty, \infty)$ to $[0, 1]$ scale.

4.3. Quantification of Genetic Sharing of 4 Pediatric Autoimmune Diseases

Our analyses follow from Chapter 2. Kobie et al., 2015 detects genetic sharing among pediatric autoimmune disease (pAID) pairs. Here we quantify the shared genetic variants for all possible pairs of the previously investigated pAIDs: common variable immunodeficiency (CVID), ulcerative colitis (UC), type I diabetes (T1D) and Crohn's disease (CD) (Kobie et al., 2015; Li et al., 2015b), with the standardized quantification measure $Q(\mu, \nu)$. $Q(\mu, \nu)$ was estimated for all possible pAID pairs using Z-scores for each SNP, U_i and V_i. Z-scores for each SNP were calculated marginally from individual-level pAID GWAS data. Only autosomal SNPs were included in the analyses, excluding mitochondrial SNPs and SNPs found on either of the two sex chromosomes. SNPs within the major histocompatibility complex (MHC) region of chromosome 6 were also excluded (Kobie et al., 2015). GWAS for each pAID GWAS has an identical set of $473,220$ SNPs.

While the estimation of $Q(\mu, \nu)$ relies on summary-level GWAS data, we used individual-level GWAS data to aid in the estimation of the standard error (SE) of $Q(\mu, \nu)$. The SE of $Q(\mu, \nu)$ was estimated with a bootstrapping procedure: selecting individuals and their corresponding genome-wide genotypes from each pAID GWAS with replacement, re-calculating SNP Z-scores and re-calculating $Q(\mu, \nu)$ for 100 replications. Table 4.1 shows pairwise estimates of $Q(\mu, \nu)$ with corresponding SE estimates. Table 4.1 compares our quantification measure to published genetic correlation estimates (Li et al., 2015a) for specified disease pairs. The published estimates of genetic correlation, or coheritability ($REML - coh^2$), are calculated using restricted maximum likelihood (REML) estimation methods in a bivariate linear mixed effect model framework (Lee et al., 2012a; Lee et al., 2011; Li et al., 2015a; Yang et al., 2011a,b). $REML - coh^2$ estimates account for the direction of effects while estimates of $Q(\mu, \nu)$ do not. Table 4.1 also compares estimates of $Q(\mu, \nu)$ with permutation-based p-values for the detection of genetic sharing among pAID pairs (Kobie et al., 2015) and with naive estimates of the genetic correlation, r, without accounting for the sparsity of the data. Naive estimates of the genetic correlation, r, are calculated using the summary-level SNP Z-scores of the disease pair and, like $REML - coh^2$, account for the direction of SNP effects. Table 4.1 presents the Spearman correlation, which is robust to non-linear relationships between the set of Z-scores.

Table 4.1 also presents the functionally weighted quantification measure, $Q_w(\mu, \nu)$. $Q_E(\mu, \nu)$ gives

Z-scores a higher weight if their corresponding SNP is more functionally relevant.

Table 4.1: Pairwise estimates of the genetic correlation $Q(\mu, \nu)$, and weighted genetic correlation $Q_w(\mu, \nu)$, with standard error (SE) compared to pairwise permutation-based p-values, coheritability estimates $REML - coh^2$ (Li et al., 2015a) and Spearman correlation estimates, r

Disease Pair	Permutation	$Q(\mu, \nu)$ (SE)	$Q_w(\mu, \nu)$ (SE)	$REML - coh^2$ (SE)	r
CVID-UC	0.8168	0.0012 (0.0056)	0.0012 (0.0058)	0.134 (0.170)	0.10
CVID-T1D	0.0337	0.0079 (0.0068)	0.0072 (0.0067)	0.207 (0.167)	0.03
CVID-CD	0.0391	0.0030 (0.0051)	0.0027 (0.0049)	0.115 (0.116)	0.16
UC-T1D	0.4277	0.0065 (0.0068)	0.0068 (0.0070)	-0.095 (0.086)	0.09
UC-CD	3.8e-04	0.3786 (0.0148)	0.3960 (0.0147)	0.674 (0.072)	0.30
T1D-CD	0.0262	0.0126 (0.0069)	0.0105 (0.0070)	0.142 (0.064)	0.11

Estimates of $Q(\mu, \nu)$ are consistent with permutation-based p-values detecting genetic sharing. Genetically related disease pair UC-CD has an estimated $Q(\mu, \nu)$ closest to 1, 0.3786 (0.0148), while CVID-UC, CVID-T1D, CVID-CD, UC-T1D, T1D-CD disease pairs with little to no evidence of genetic sharing have $Q(\mu, \nu)$ estimates much closer to 0 (Table 4.1). For disease pairs CVID-UC, CVID-T1D, CVID-CD, UC-T1D with little to no evidence of genetic sharing, $Q(\mu, \nu)$ estimates appear more comparable than the permutation-based method with p-values ranging from 0.0337 to 0.8168 (Table 4.1). Similar conclusions are drawn from $REML - coh^2$ and Spearman r, with estimates closest to 1 for the genetically related disease pair UC-CD (Table 4.1). Note, $REML - coh^2$ estimates have much larger SE estimates, and conclusions drawn from Spearman r estimates are less consistent across disease pairs with no evidence of genetic sharing (Table 4.1).

Weighted genetic correlation estimates, $Q_w(\mu, \nu)$, are comparable to the unweighted estimates (Table 4.1). In this case, integrating annotation information has little to no effect on the quantification measure across the studied disease pairs.

4.4. Conclusion

This chapter proposes a quantification measure acting as an estimate of the genetic correlation among shared genetic variants between a pair of complex diseases. The quantification measure, $Q(\mu, \nu)$, is a function of SNP effect sizes, independent of the varied sample size across disease GWAS. We utilize an estimation procedure by Cai and Tan, 2015 for the estimation of $Q(\mu, \nu)$, which is optimal under sparse conditions typical of GWAS.

In applying our quantification measure to all possible pairs of 4 pediatric autoimmune diseases

(pAIDs), the conclusions drawn from estimates of $Q(\mu, \nu)$ are consistent with permutation-based p-values and coheritability estimates, $REML - coh^2$. Our quantification measure only relies on the summary-level GWAS data rather than the individual-level data required by coheritability estimates under the random effects linear mixed model framework. Our quantification measure also adjusts for the varied sample size across GWAS. $Q(\mu, \nu)$ estimates appear more uniform across disease pairs with no evidence of genetic sharing than the permutation-based method with diverse p-values. Naive Spearman correlation estimates, r, tend to overestimate the genetic correlation of disease pairs that show no evidence of genetic sharing. r is calculated using the summary-level SNP Z-scores without adjusting for the varied sample size across GWAS. Non-zero SNP Z-scores from disease GWAS with larger samples sizes are likely inflated in magnitude with respect to the corresponding Z-scores of disease GWAS of smaller sample sizes. Also, estimates of r do not take into account the sparisty of GWAS data.

We also consider adding a functional weight to our estimate of the genetic correlation, $Q_w(\mu, \nu)$. Intuitively, disease pairs that share variants with larger functional importance are more likely to have a shared genetic etiology, or overlapping genetic pathways causing disease. Though in this case, integration of annotation information with the transformed Eigen score has little effect on $Q(\mu, \nu)$ across disease pairs.

CHAPTER 5

DISCUSSION

In this dissertation we develop methods to investigate the genetic relatedness within complex disease sets through the detection and quantification of shared disease-associated single nucleotide polymorphisms (SNPs). In Chapter 2, we developed a global test to integrate complex disease genome-wide association studies (GWAS) and detect whether a pair of diseases exhibits evidence of genetic sharing through the detection of shared SNPs. We then added a level of dependency on the direction of SNP association to the global test, allowing for a more specific account of what the detected genetic similarities represent. We also proposed a sequential identification procedure, utilizing our global test of detection, to identify the top drivers of genetic sharing.

In Chapter 3, we implemented a perturbation method to evaluate the statistical significance of the global detection test proposed in Chapter 2 without assuming independence among SNPs. The perturbation method exploits independent standard normal random variables to emulate the null distribution of while preserving the inherent dependency among SNPs. We then extended the test of detection from detecting SNPs at identical SNPs to detecting SNPs within a dependency-defined window. In Chapter 4 we propose a quantification measure to quantify the detected genetic sharing between complex disease pairs using summary-level GWAS data. We consider a functional weight, giving more weight to SNPs with more functional relevance.

In Chapters 2-4 we apply our methods to a set of clinically-distinct pediatric autoimmune disease (pAID) GWAS. With the proposed detection methods we were able to identify pAID disease pairs that show evidence of genetic sharing. Our quantification measure identified the disease pair ulcerative colitis (UC)-Crohn's disease (CD) as exhibiting the most genetic sharing of all studied pAID pairs, which is consistent with our detection results. While incorporating the functional annotation information in our quantification measure of Chapter 4 does not alter our conclusions in its absence, the development of integrative methods is crucial for epidemiological advances.

BIBLIOGRAPHY

Andreassen, OA, Thompson, WK, Schork, AJ, Ripke, S, Mattingsdal, M, Kelsoe, JR, Kendler, KS, O'Donovan, MC, Rujescu, D, Werge, T, Sklar, P, Roddey, JC, Chen, CH, McEvoy, L, Desikan, RS, Djurovic, S, and Dale, AM (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genetics* 9.4, e1003455.

Benjamini, Y and Hochberg, Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, 289–300.

Benjamini, Y and Heller, R (2008). Screening for partial conjunction hypotheses. *Biometrics* 64.4, 1215–22.

Bhattacharjee, S, Rajaraman, P, Jacobs, KB, Wheeler, WA, Melin, BS, Hartge, P, Yeager, M, Chung, CC, Chanock, SJ, and Chatterjee, N (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *American Journal of Human Genetics* 90.5, 821–35.

Bulik-Sullivan, B, Finucane, HK, Anttila, V, Day, FR, Reprogen Consortium, Genomics Consortium, P, and Neale, BM (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* 47, 1236–1241.

Cai, TT and Jeng, XJ (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.5, 629–662.

Cai, TT and Tan, XL (2015). Optimal estimation of a quadratic functional and detection of simultaneous signals. available online. URL: `http://www-stat.wharton.upenn.edu/~tcai/paper/QF-2-Sample.pdf`.

Cai, T and Wu, Y (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory* 60.4, 2217–2232.

Chung, D, Yang, C, Li, C, Gelernter, J, and Zhao, H (2014). GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. *PLoS Genetics* 10.11, e1004787.

Cooper, GS, Bynum, MLK, and Somers, EC (2009). Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *Journal of Autoimmunity* 33.3-4, 197–207.

Cotsapas, C, Voight, BF, Rossin, E, Lage, K, Neale, BM, Wallace, C, Abecasis, G, Barrett, J, Behrens, J, De Jager, P, Elder, J, Graham, R, Gregersen, P, Klareskog, L, Siminovitch, K, Heel, D van, Wijmenga, C, Worthington, J, Todd, J, Hafler, D, Rich, S, and Daly, M (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genetics* 7, e1002254.

Cross-disorder Group of the Psychiatric Genomics Consortium (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 381, 1371–9.

Dempster, ER and Lerner, IM (1950). Heritability of threshold characters. *Genetics* 35.2, 212–236.

Donoho, D and Jin, J (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *PNAS* 105.39.

Dudbridge, F (2013). Power and predictive accuracy of polygenic risk scores. *PLoS genetics* 9.3, e1003348.

Efron, B (2008). Microarrays, Empirical Bayes and the two-groups model. *Statistical Science* 23.1, 1–22.

He, X, Fuller, CK, Song, Y, Meng, Q, Zhang, B, Yang, X, and Li, H (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics* 92.5, 667–80.

Hindorff, LA, Sethupathy, P, Junkins, HA, Ramos, EM, Mehta, JP, Collins, FS, and Manolio, TA (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106.23, 9362–7.

Ionita-Laza, I, McCallum, K, Xu, B, and Buxbaum, J (2015). A spectral approach integrating functional genomic annotations for coding and noncoding variants. available online. URL: http://www.columbia.edu/~ii2135/Eigen_11_24.pdf.

Jin, J and Donoho, D (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* 32.3, 962–994.

Kobie, J, Zhao, SD, Li, YR, Hakonarson, H, and Li, H (2015). Statistical tests for the detection of shared common genetic variants between complex diseases based on GWAS. in preparation.

Lee, SH, Yang, J, Goddard, ME, Visscher, PM, and Wray, NR (2012a). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28.19, 2540–2.

Lee, SH, DeCandia, TR, Ripke, S, Yang, J, Sullivan, PF, Goddard, ME, Keller, MC, Visscher, PM, and Wray, NR (2012b). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* 44.3, 247–50.

Lee, SH, Wray, NR, Goddard, ME, and Visscher, PM (2011). Estimating missing heritability for disease from genome-wide association studies. *American Journal of Human Genetics* 88.3, 294–305.

Li, YR, Zhao, SD, Li, J, Bradfield, JP, Mohebnasab, M, Steel, L, Kobie, J, Abrams, DJ, Mentch, FD, Glessner, JT, Guo, Y, Wei, Z, Connolly, JJ, Cardinale, CJ, Bakay, M, Li, D, Maggadottir, SM, Thomas, Ka, Qui, H, Chiavacci, RM, Kim, CE, Wang, F, Snyder, J, Flatø, B, Fø rre, O, Denson, LA, Thompson, SD, Becker, ML, Guthery, SL, Latiano, A, Perez, E, Resnick, E, Strisciuglio, C, Staiano, A, Miele, E, Silverberg, MS, Lie, BA, Punaro, M, Russell, RK, Wilson, DC, Dubinsky, MC, Monos, DS, Annese, V, Munro, JE, Wise, C, Chapel, H, Cunningham-Rundles, C, Orange, JS, Behrens, EM, Sullivan, KE, Kugathasan, S, Griffiths, AM, Satsangi, J, Grant, SFA, Sleiman, PMA, Finkel, TH, Polychronakos, C, Baldassano, RN, Luning Prak, ET, Ellis, JA, Li, H, Keating, BJ, and Hakonarson, H (2015a). Genetic sharing and heritability of paediatric age of onset autoimmune diseases. *Nature Communications* 6.8442.

Li, YR, Li, J, Zhao, SD, Bradfield, JP, Mentch, FD, Maggadottir, SM, Hou, C, Abrams, DJ, Chang, D, Gao, F, Guo, Y, Wei, Z, Connolly, JJ, Cardinale, CJ, Bakay, M, Glessner, JT, Li, D, Kao, C, Thomas, KA, Qiu, H, Chiavacci, RM, Kim, CE, Wang, F, Snyder, J, Richie, MD, Flatø, B, Fø rre, Oy, Denson, LA, Thompson, SD, Becker, ML, Guthery, SL, Latiano, A, Perez, E, Resnick, E, Russell, RK, Wilson, DC, Silverberg, MS, Annese, V, Lie, BA, Punaro, M, Dubinsky, MC, Monos, DS, Strisciuglio, C, Staiano, A, Miele, E, Kugathasan, S, Ellis, JA, Munro, JE, Sullivan, KE, Wise, CA, Chapel, H, Cunningham-Rundles, C, Grant, SFA, Orange, JS, Sleiman, PMA, Behrens, EM, Griffiths, AM, Satsangi, J, Finkel, TH, Keinan, A, Prak, ETL, Polychronakos, C, Baldassano, RN, Li, H, Keating, BJ, and Hakonarson, H (2015b). Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine* 21, 1018–1027.

Lin, DY (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics (Oxford, England)* 21.6, 781–7.

Lin, DY and Zou, F (2004). Assessing genomewide statistical significance in linkage studies. *Genetic Epidemiology* 27.3, 202–14.

Lin, D (2006). Evaluating statistical significance in two-stage genomewide association studies. *The American Journal of Human Genetics*, 505–509.

Liu, JZ, McRae, AF, Nyholt, DR, Medland, SE, Wray, NR, Brown, KM, Hayward, NK, Montgomery, GW, Visscher, PM, Martin, NG, and Macgregor, S (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics* 87.1, 139–45.

Manolio, TA, Collins, FS, Cox, NJ, Goldstein, DB, Hindorff, LA, $\cdots$, Mackay, TFC, McCarroll, SA, and Visscher, PM (2009). Finding the missing heritability of complex diseases. *Nature* 461.7265, 747–53.

Morris, A, Voight, B, and Teslovich, T (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics* 44.9, 981–990.

Pickrell, JK (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *American Journal of Human Genetics* 94.4, 559–73.

Sherry, S, Ward, M, Kholodov, M, Baker, J, Phan, L, Smigielski, E, and Sirotkin, K (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29.1, 308–11.

Veyrieras, JB, Kudaravalli, S, Kim, SY, Dermitzakis, ET, Gilad, Y, Stephens, M, and Pritchard, JK (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics* 4.10, e1000214.

Wang, K, Baldassano, R, Zhang, H, Qu, HQ, Imielinski, M, Kugathasan, S, Annese, V, Dubinsky, M, Rotter, JI, Russell, RK, Bradfield, JP, Sleiman, PMA, Glessner, JT, Walters, T, Hou, C, Kim, C, Frackelton, EC, Garris, M, Doran, J, Romano, C, Catassi, C, Van Limbergen, J, Guthery, SL, Denson, L, Piccoli, D, Silverberg, MS, Stanley, CA, Monos, D, Wilson, DC, Griffiths, A, Grant, SFa, Satsangi, J, Polychronakos, C, and Hakonarson, H (2010). Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Human Molecular Genetics* 19.10, 2059–67.

Wray, NR, Goddard, ME, and Visscher, PM (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research* 17.10, 1520–1528.

Yang, J, Benyamin, B, McEvoy, BP, Gordon, S, Henders, AK, Nyholt, DR, Madden, PA, Heath, AC, Martin, NG, Montgomery, GW, Goddard, ME, and Visscher, PM (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* 42.7, 565–9.

Yang, J, Lee, SH, Goddard, ME, and Visscher, PM (2011a). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics* 88.1, 76–82.

Yang, J, Manolio, TA, Pasquale, LR, Boerwinkle, E, Caporaso, N, Cunningham, JM, Andrade, M de, Feenstra, B, Feingold, E, Hayes, MG, Hill, WG, Landi, MT, Alonso, A, Lettre, G, Lin, P, Ling, H, Lowe, W, Mathias, RA, Melbye, M, Pugh, E, Cornelis, MC, Weir, BS, Goddard, ME, and Visscher, PM (2011b). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* 43.6, 519–25.

Zou, F, Fine, JP, Hu, J, and Lin, DY (2004). An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. *Genetics* 168.4, 2307–16.