Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of this thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Kelsy Alaine Broadaway	Date

Development of Novel Analytic Methods for Improved Gene Mapping of Complex Human Traits

By

Kelsy Alaine Broadaway
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Program in Population Biology, Ecology, and Evolution

Michael P. Epstein, Ph.D.
Advisor

Stephanie L. Sherman, Ph.D.
Advisor

Karen N. Conneely, Ph.D.
Committee Member

Michael E. Zwick, Ph.D.
Committee Member

Michele Marcus, Ph.D.
Committee Member

Kerry Ressler, M.D., Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Development of Novel Analytic Methods for Improved Gene Mapping of Complex Human Traits

By

Kelsy Alaine Broadaway
B.S., Montana State University, 2009

Advisors:
Michael P. Epstein, Ph.D.
Stephanie L. Sherman, Ph.D.

An abstract of a dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Division of Biological and Biomedical Sciences, Program in Population Biology, Ecology, and Evolution, in 2015

Abstract

Development of Novel Analytic Methods for Improved Gene Mapping of Complex Human Traits

by Kelsy Alaine Broadaway

Discovering the genetic contributors to complex human traits and diseases is a central goal in genetic epidemiology. Although tremendous advancements in high-throughput genotyping and sequencing technology have allowed genetic analyses on a scale undreamt of just 15 years ago, most genetic contributors to complex trait variation remain hidden. By using population genetics theory, we can create new analytic approaches that make better use of the wealth of genetic data that are now available to us. This dissertation investigates three such analytic approaches, with each employing a powerful and flexible high-dimensional modeling kernel framework for inference. This type of approach is valuable in genetic analyses because it allows simultaneous consideration of multiple genetic variants and multiple phenotypes within a single analysis. While there are already several kernel approaches implemented for genetic studies of complex traits, this dissertation introduces three new kernel methods with each method probing a different type of hypothesis. Specifically, I develop kernel methods in which gene-environment interaction effects on complex traits are suspected, kernel methods for detecting cross-phenotype effects when pleiotropy is suspected, and kernel methods for analysis of multivariate questionnaire data that are being used as a proxy for an underlying phenotype. In subsequent chapters we derive these three methods, and then conduct in-depth simulations to illustrate the statistical validity and power of the approaches compared with existing methods. For each method, we then used our approaches to analyze real genetic data from existing complex-trait studies. We found that each of our approaches offers more power than the corresponding competing methods across a broad range of analyses. The power of our three approaches indicates that they might be useful in elucidating the complicated genetic underpinnings of human traits and disease.

Development of Novel Analytic Methods for Improved Gene Mapping of Complex Human Traits

By

Kelsy Alaine Broadaway
B.S., Montana State University, 2009

Advisors:
Michael P. Epstein, Ph.D.
Stephanie L. Sherman, Ph.D.

A dissertation submitted to the Faculty of the James T. Laney School of Graduate Studies of Emory University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Division of Biological and Biomedical Sciences, Program in Population Biology, Ecology, and Evolution, in 2015

Acknowledgements

First, I would like express gratitude to my advisors, Drs. Michael Epstein and Stephanie Sherman. They gave me freedom to explore on my own, advice when I needed it, and encouragement throughout the process. More importantly, they are proof that scientific rigor and heartfelt compassion need never be at odds.

I would like to give thanks to the members of my dissertation committee: Drs. Karen Conneely, Kerry Ressler, Michele Marcus, and Michael Zwick. Their insightful comments and encouragement helped expand the breadth of this dissertation.

A great many other members of the Emory community also deserve thanks. Dr. David Cutler gave support and mentorship when I needed it the most. Dr. Richard Duncan provided invaluable computational assistance. Molecules to Mankind gave me an academic home.

Finally, none of this would have been possible without the love of my family (irrespective of our kinship coefficient). They have been my constant source of support and strength all these years. Without their patience, humor, and long walks, I would have been lost.

Table of Contents

Tables and Figures

Tables

Figures

CHAPTER 1:
Introduction

Gene Mapping of Complex Traits

The etiology of nearly all diseases and disease-related quantitative traits includes a substantial genetic component. One of the central goals of genetic epidemiology is to map the location of genomic variation that influences these complex traits. Gene mapping is impactful since it potentially could lead to important medical advances [1]. Treatment of many complex traits—particularly psychiatric disorders—remains almost entirely symptomatic. Determination of genetic variants involved in a particular disease should lead to more etiology-driven, targeted medical care. Increasing our understanding of how complex traits arise could lead to improved disease prevention through genetic screening as well as development of new and more effective therapies and pharmaceutical targets. And yet, although tremendous advancements in high-throughput genomic technology have allowed genome-wide analysis of complex traits for tens to hundreds of thousands of participants, most of genetic contributors to complex trait variation remain hidden.

Current designs for genetic studies of complex traits are based on theoretical expectations from population genetics models. For over thirty years, the quantitative genetics community has debated the very underlying nature of genetic variation that contributes to complex traits [2]. Two popular models are the Neo-Darwinian and infinitesimal models. The Neo-Darwinian hypothesis predicts that trait-influencing variants will be at relatively high frequencies, few in number, and have moderate effect sizes [3-5]. Diametrically opposed to the Neo-Darwinian view is

the infinitesimal model, which predicts that trait-influencing variants will mostly be at very low frequencies, extremely numerous throughout the genome, and with exceedingly small effect sizes [6; 7]. In subsequent sections, I will describe general strategies for gene mapping under each theory, existing results, limitations, and potential solutions to the limitations that form the work in this thesis.

Neo-Darwinian Model of Genetic Variation

The Neo-Darwinian hypothesis is that common alleles of large effect will explain a substantial fraction of trait heritability [5]. This viewpoint predicts that common causal alleles will be maintained through balancing selection, most likely in the form of fluctuating selection in space and time [8]. The Neo-Darwinian view suggests that genome-wide association studies (GWAS) are valuable tools to detect genetic associations that predict trait and disease outcome. GWAS genotype hundreds of thousands to millions of common single nucleotide polymorphisms (SNPs) across the genome, and then investigate each genotyped locus for correlation between allele and trait. GWAS rely on the two interconnected assumptions that are fundamental to the Neo-Darwinian view: first, individual causal polymorphisms have sizeable effect, and second, those polymorphisms occur at moderate frequencies [9; 10]. If the genetic variation underlying a trait follows the Neo-Darwinian view, a GWAS is well powered to detect genetic effects: causal variants will be common enough to be either directly genotyped by the GWAS array or in close linkage disequilibrium (LD) with a genotyped SNP, and effect sizes will also be large enough to be detectible in sample sizes of a few thousand subjects [11].

Conversely, if causal variants are rare or have small effect sizes, GWAS are not powered nor designed to detect a small effect variant. Therefore, GWAS fundamentally test the Neo-Darwinian hypothesis [3; 12; 13].

Exploring the Neo-Darwinian Hypothesis

Evidence suggests that common genetic variation does play an important role in many complex traits and diseases. For example, common SNPs are estimated to explain 21% of the heritability of depression[14], 14-17% of the heritability of body mass index[15], and approximately 45% of the heritability of height [16]. To date, nearly two thousand GWAS have been documented by the National Human Genome Research Institute (NHGRI) catalog, many of which reported compelling associations between common genetic variants and a variety of diseases [17; 18].

However, even after GWAS involving tens to hundreds of thousands of study subjects, identification of trait-influencing genetic variation is far from complete [19-21]. The common variants that have been identified by GWAS tend to have modest effect sizes and account for very little overall heritability. For example, a prodigious meta-analysis of human height—analyzing over 250,000 individuals— recently reported 697 common genetic variants associated with the trait. And yet, altogether, the nearly 700 identified variants account for only 20% of the heritability of height [22]. If we expect common variants to explain approximately 45% of height heritability [16], we must conclude that the majority of common variants associated with height remain unidentified. To remain undetected by a GWAS approach with such a large sample size, the remaining causal common variants must have

vanishingly small effect sizes. Similar conclusions have been reached with other phenotypes [23-25].

The fundamental conclusion to be drawn from the GWAS era is that common alleles of large effect are exceedingly rare. However, common variants likely influence trait etiology, albeit by imposing small effects. GWAS approaches show value and should not be discarded. Rather, we should delineate barriers to detecting additional common causal variants through GWAS, and in doing so direct development of novel statistical techniques.

Effects of Gene-Environment Interactions

The inherent purpose of most genomic studies is to find evidence for additive allelic effects; the much-discussed "missing heritability" problem plaguing human genetics research refers to the additive or "narrow-sense" definition of heritability [26]. Statistically, narrow-sense heritability is assessed via testing for a main genetic effect. For the objective of explaining missing heritability, finding evidence of non-additive effects, such as gene-environment interaction, is inconsequential. The additive effect of an individual allele is defined as the average effect of substituting one allele for another; variation due to environmental differences is taken into account in this average [27]. Therefore, in absence of main genetic effect, finding evidence of gene-environment interaction will provide no information into explaining narrow-sense heritability.

However, while gene-environment interaction cannot directly account for missing heritability, the ability to detect main effects of causal variants might be

hindered by heterogeneity of genetic effect sizes due to environment. For example, if a genetic variant influences a trait, but that effect is contingent on exposure to an environmental factor, the main genetic effect of that variant will be reduced. Given smaller main effect sizes, the causal SNP could be discarded as non-significant. The differences in the distribution of the environmental factor between initial and validation studies could also impede replication of the initial SNP finding. In both cases, considering gene-environment interaction might be beneficial. However, including interaction terms comes at a cost: researchers who choose to model gene-environment interaction effects contend with a host of issues, including defining and measuring relevant environmental exposures, increased testing burden, and insufficient power to detect interactive effects [28; 29].

The challenges in gene-environment analyses are particularly daunting if we anticipate main effect sizes that are too small to detect in large GWAS. The interaction effect could be larger than the main effect, particularly in cases of crossover interaction, in which the opposite effects are present for opposite alleles. However, for many cases, interaction between gene and environment should be expected to be even smaller than genetic main effects, resulting in limited statistical power to detect the interaction. Unless the effect size of the interaction is much larger than the main genetic effect, typical regression approaches will not gain any statistical power through tests of interaction instead of main effect tests [28-30]. When there is interest in modeling the modifying effects of genotype on phenotype in the presence of interaction with environment, an optimal statistical test would be one

that remains powerful even when interaction is much smaller than the main genetic effect.

In **Chapter 2**, I introduce a gene-based approach for association mapping of multiple SNPs to consider joint tests of gene- and gene-environment interaction. We implement this approach using a popular machine-learning technique called kernel-machine regression (KMR). At the core of KMR is a kernel function, which transforms high-dimensional information for a pair of subjects into a scalar quantitative measure representing their similarity. Using kernel functions, we can compare phenotypic pairwise similarity to genotypic pairwise similarity in the presence of environmental interactions. The approach incorporates LD information from multiple SNPs simultaneously in analysis and permits flexible modeling of interaction effects. Using simulated data, we show that our approach typically outperforms the traditional joint test under strong gene-environment interaction models and further outperforms the traditional main-effect association test under models of weak or no gene-environment interaction effects. We illustrate our test using genome-wide association data from the Grady Trauma Project, a cohort of highly traumatized, at-risk individuals, which has previously been investigated for interaction effects. This work was recently published in *Genetic Epidemiology*.

Imperfect Phenotyping

In population genetics, we tend to envision a phenotype as a quantifiable entity, often approximately following a Gaussian distribution [6]. While this model is profoundly useful—and in theory such underlying phenotypes should exist—we are

frequently unable to directly measure the true phenotypes of interest. Instead, we attempt to capture the latent phenotype of interest from multiple angles, via several interrelated yet distinct measurements. Specifically, most psychiatric phenotypes are measured via a multi-question survey or test, with each question aimed at reflecting different aspects of International Statistical Classification of Diseases (ICD) or the Diagnostic and Statistical Manual of Mental Disorders (DSM) definitions of the disorder of interest [31; 32].

Modern definitions of psychiatric disorders explicitly acknowledge the high degree of symptom heterogeneity found among patients with the same psychiatric diagnosis. For example, for diagnosis of a major depressive episode, the DSM-IV-TR queries on depressed mood, anhedonia, weight or appetite change, insomnia or hypersomnia, agitation, fatigue, feelings of worthlessness or guilt, diminished ability to concentrate, and recurrent thoughts of death [33]. Diagnosis requires that the patient have at least five of the symptoms (one of which must either be depressed mood or anhedonia), for at least two weeks. The Beck Depression Inventory II (BDI) queries severity of all these symptoms over a 21-item questionnaire [34]. As a result, a completed BDI provides nuanced information on many depressive symptoms. However, most analytic techniques only allow testing of a single outcome. Neuropsychiatric researchers must therefore collapse a multifaceted set of measures into a single composite score, which can serve as a proxy for the phenotype of interest. For this reason, most psychiatric questionnaires used in genetic studies (including the BDI) are designed to be collapsible, typically using a simple cumulative score approach. The cumulative score can then be treated either

as a continuous outcome, or cutoffs can be applied to indicate presence/absence of disease symptoms.

Reducing multivariate questionnaire information into a univariate outcome is an often-lamented statistical necessity, as it nearly always comes at a cost. Coarse phenotyping can result in a diluted association between gene and trait. For example, a gene might be associated with only a certain subset of depressive symptoms, such as somatic symptoms (e.g. fatigue), but not affective symptoms (e.g. self-dislike). In this case, the effect size of the gene would be biased toward zero if individuals with different types of depressive symptoms were classified as sharing the same phenotype. Researchers are interested in incorporating multivariate questionnaire data in a broader and more flexible manner than a simple cumulative score. Unfortunately, to date, there remains a paucity of statistical genetic alternatives that can handle the complicated ordinal data arising from questionnaires.

In **Chapter 4**, I introduce a genetic association test for analysis of multivariate outcomes collected in questionnaires, surveys, and scores that provides a more flexible and powerful strategy for gene mapping compared to using the typical cumulative score. Our approach, which we call the Gene Association Method for Broader Integration of Tests and Scores (GAMBITS), allows researchers to model questionnaire data in a nonparametric manner. GAMBITS is designed to test whether pairwise similarity in questionnaire responses is independent of pairwise genotypic similarity in a gene or region of interest. GAMBITS relies on a novel class of machine-learning methods, termed kernel distance-covariance (KDC) methods, for inference. Like KMR methods, KDC methods rely on kernel functions to model

data similarity among different pairs of samples. However, while KMR assumes only univariate phenotype data for each subject, KDC allows each subject to possess multivariate phenotype data (such as multivariate questionnaire data).

The GAMBITS framework allows for an arbitrary number of questions within the questionnaire as well as an arbitrary number of genotypes, thereby permitting gene-based testing of both rare and common variants. We show that GAMBITS is robust to the ordinal outcomes frequently collected in questionnaires, and further can correct for important covariates. We show that GAMBITS is markedly more powerful than the traditional methods, particularly when not all questions in the survey are associated with the causal variant. We demonstrate our approach using GWAS data from the Grady Trauma Project.

Infinitesimal Model of Genetic Variation

Lande popularized an alternate hypothesis to the Neo-Darwinian view, which has been termed the infinitesimal model[6; 7]. Unlike the Neo-Darwinian model, which predicts relatively few causal variants with moderate effect size, the infinitesimal model predicts a very large number of causal loci for a given trait, each with a nearly infinitesimal effect size [6]. The theoretical argument hinges on the idea that an allele with a large effect on disease will likely be deleterious, and thus kept at low frequency by purifying selection[35; 36]. Only alleles of exceedingly small effect size should drift to moderate frequency in the population. The infinitesimal model indicates that to find genetic variants of large effect, we must query rare genetic variants. However, since GWAS do not assay rare genetic variation, the tool is not useful in testing for associations under the infinitesimal model.

Empirical evidence suggests that heritabilities of most traits are indeed driven by hundreds to thousands of causal variants, with each variant having only a minor effect on overall phenotypic variance [22; 37-40]. The advent of lower-cost next generation sequencing platforms, as well as new genotyping arrays that include rare alleles, have recently made rare-variant association studies a reality. Although fewer rare variant association studies have been performed than GWAS, multiple rare variants have been successfully linked to a variety of phenotypes and diseases [41; 42]. A particularly exciting success story in rare variant associations comes in from the *PCSK9* gene, the results of which prompted novel translational therapy for hypercholesterolemia [43; 44]. Unfortunately, despite promising initial results, many rare variant analyses to date have either failed to uncover any associations, or reveal associations with only modest effect sizes [41].

Ramifications of the Infinitesimal Model: Pleiotropy

The neo-Darwinian school stressed a pair of important observations that must be true under the infinitesimal model, given the estimated mutation rate per nucleotide to new alleles [45-47]. First, a very large number of genes are capable of being mutated to rare, high-effect alleles for a very large number of traits; second, nearly every rare allele of large effect must be contributing to a large number of different traits [5]. These make an uncomfortable set of criticisms. However, if we embrace the infinitesimal model, we must also embrace its Neo-Darwinian critique: pleiotropy, and particularly rare-variant pleiotropy, should be ubiquitous in the human genome.

Pleiotropy, the phenomenon of a single genetic variant affecting two or more distinct traits, has profound implications in human genetics. Pleiotropy may underlie many fundamental principles in biology, including senescence [48], evolution [49; 50], and human health [51; 52]. Strongly pleiotropic genes are expected to be under strong stabilizing selection, since a gene with multiple biological functions is likely to be more limited in its possible variations than a gene with a single function [51; 53]. Additionally, alleles that are advantageous to one trait might be deleterious to another [49; 53]. Perhaps the most extreme example was proposed by Williams in 1957: senescence might be a by-product of genetic variants that are advantageous to development and reproduction [48].

Empirically, many genetic loci do appear to harbor variants that are associated with multiple traits. A frequently cited Mendelian example is phenylketonuria (PKU). PKU is an autosomal recessive disorder caused by deficiency of the phenylalanine hydroxylase enzyme, which leads to excess levels of phenylalanine. This in turn leads to a range of phenotypes, including severe intellectual disability, reduced pigmentation, and eczema [54]. Empirical evidence for pleiotropy among complex traits is growing rapidly, although such evidence is sometimes labeled a "cross-phenotype" (CP) association, to differentiate between observations arising from statistical tests and the true causal relationships described by pleiotropy. Evidence of common-variant CP associations has already been reported in numerous human traits, including psychiatric disorders, autoimmune diseases, and inflammatory bowel diseases [52; 55-57].

Also supporting the prediction of pervasive pleiotropy is the observation that many phenotypes are highly comorbid. Neuropsychiatric disorders display high comorbidity: in total, the National Institute of Mental Health (NIMH) estimates that as much as 45% of individuals diagnosed with a mental disorder meet criteria for two or more disorders, with severity strongly correlated with comorbidity [58]. Likewise, nearly 75% of adults with diabetes also have hypertension [59], and patients with rheumatoid arthritis are about twice as likely to suffer from myocardial infarction as individuals without arthritis [60]. While some comorbidities are likely artifacts of phenotypic measurements or selection bias, many are almost certainly biological in nature.

When pleiotropy exists, considering each phenotype separately ignores information provided by CP correlation [61; 62], which might result in decreased power to detect the genetic effect. However, while there are several excellent statistical methods appropriate for pleiotropic analysis of common genetic variants [63-68], there is a lack of analogous statistical approaches to assess CP associations of rare genetic variants. Existing CP association methods are designed to assess the effect of a single polymorphism at a time; however, in rare variant analysis, a test typically requires aggregation of information from multiple rare variants within a gene simultaneously. Although a few rare-variant pleiotropy approaches have been proposed, they tend to be computationally intensive as the number of phenotypes increases, and are limited to normally distributed outcomes [67; 69].

In **Chapter 3**, I introduce a new statistical method for CP analysis of rare variants. As with the previously mentioned GAMBITS approach, this new method

relies on the KDC framework to create a nonparametric test of independence between high-dimensional rare-variant genotype data and high-dimensional phenotype data. This approach, which we refer to as the Gene Association with Multiple Traits (GAMuT) test, can accommodate both binary and continuous phenotypes and can adjust for covariates. We use simulated data to demonstrate that GAMuT provides increased power over standard univariate methods of rare-variant testing for individual traits when pleiotropy exists. We also illustrate our approach using exome-chip data from the Genetic Epidemiology Network of Arteriopathy.

Scope of the Thesis

In this dissertation, I use KMR and KDC techniques to address important problems in gene-mapping studies of complex disease and disease-related quantitative traits. In the subsequent three chapters, I will explore a KMR method to model gene-environment interaction, a KDC method to test for pleiotropic effects, and a KDC method to allow for a broader genetic association testing of questionnaires, tests, and scores. Finally, in Chapter 5, I will discuss the conclusions drawn from the kernel approaches introduced here, limitations of the approaches outlined in this thesis, and provide recommendations for future research.

CHAPTER 2:
Kernel Approach for Modeling Interaction Effects in Genetic Association Studies of Complex Quantitative Traits

Kernel Approach for Modeling Interaction Effects in Genetic Association Studies of Complex Quantitative Traits

K. Alaine Broadaway[1], Richard Duncan[1], Karen N. Conneely[1], Lynn M. Almli[2], Bekh Bradley[2,3], Kerry J. Ressler[2], Michael P. Epstein[1]

[1]Department of Human Genetics, Emory University, Atlanta, GA

[2]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA

[3]Clinical Psychologist, Mental Health Service Line, Department of Veterans Affairs Medical Center, Atlanta, GA

Disclaimer: The contents of this manuscript do not reflect the views of the Department of Veterans Affairs, or the United States Government

Published in *Genetic Epidemiology*. 2015; 39(5) 366-75.

ABSTRACT

The etiology of complex traits likely involves the effects of genetic and environmental factors, along with complicated interaction effects between them. Consequently, there has been interest in applying genetic association tests of complex traits that account for potential modification of the genetic effect in the presence of an environmental factor. One can perform such an analysis using a joint test of gene and gene-environment interaction. An optimal joint test would be one that remains powerful under a variety of models ranging from those of strong gene-environment interaction effect to those of little or no gene-environment interaction effect. To fill this demand, we have extended a kernel-machine based approach for association mapping of multiple SNPs to consider joint tests of gene and gene-environment interaction. The kernel-based approach for joint testing is promising, since it incorporates linkage disequilibrium information from multiple SNPs simultaneously in analysis and permits flexible modeling of interaction effects. Using simulated data, we show that our kernel-machine approach typically outperforms the traditional joint test under strong gene-environment interaction models and further outperforms the traditional main-effect association test under models of weak or no gene-environment interaction effects. We illustrate our test using genome-wide association data from the Grady Trauma Project, a cohort of highly traumatized, at-risk individuals, which has previously been investigated for interaction effects.

INTRODUCTION

In recent years, many genetic studies of complex human traits have employed genome-wide association studies (GWAS) to enable near-comprehensive assessment of common genetic variation across the genome. Empirical evidence suggests that common genetic variation plays an important role in many complex traits and diseases, with common variants estimated to explain 25-33% of risk to schizophrenia [70; 71], 40% of risk for bipolar disorder [72], and 50% of risk for autism spectrum disorder [73], among other traits. However, even in studies involving tens of thousands of study subjects, the identification of specific common trait-influencing variation remains elusive. One potential reason for the lack of replicable GWAS hits is that a single-nucleotide polymorphism (SNP) may influence a trait but the effect is modified by an interaction with an environmental factor [74] such as age [75-77]. If one ignores the gene-environment interaction effect and considers only the marginal effect of the SNP, the causal SNPs might regrettably be disregarded. Additionally, the differences in the distribution of the environmental factor between the initial and validation studies could impede replication of the initial SNP finding. This possibility, along with the observation of gene-environment interactions in various genetic studies [78-85] has spurred interest in performing genome-wide association studies of complex traits that accounts for possible genetic modification of effect by environment [86-88].

To account for possible modification of genetic effects by environment in candidate-gene and GWAS projects, one can apply a joint test of SNP main effect and SNP-environment interaction effect on phenotype. Such a joint test can be more

powerful than a test of SNP main effect alone if an interaction exists [30; 89]. A typical joint test involves fitting a regression model that accounts for the main effect of a single SNP, main effect of the environment, and a two-way interaction between the SNP and environment. One then constructs from the fitted regression model a two degree-of-freedom test of the joint null hypothesis that there is no SNP and no SNP-environment interaction effect, typically using a Wald or likelihood-ratio statistic [30; 90].

While some recent interaction findings using methods like the joint test as well as other procedures have been reported (e.g. [85; 86; 88; 91-93]), by and large the field has not matured in a way to match its propitious beginnings [82; 94]. A possible explanation lies with an inherent motivation behind interaction studies: one reason to include a modifying effect of environment within a genetic analysis is to find subgroups of individuals where the genetic effects are of larger magnitude than the overall group as a whole, and thus gain power over genetic studies that fail to account for the environmental modifier. This gain in power occurs when the interaction effect is much larger than the main effect, such as when a genotype has an effect on phenotype in the presence of environmental effect but no effect in the absence of the exposure (termed "complete" interaction: see model M1 in Figure 1). However, when the interaction effect is of equal magnitude or smaller than the main SNP effect (see models M3-M5 in Figure 1), a main effect test (which has 1 degree of the freedom less than the joint test) might perform similarly or even better than the joint test [28; 30].

When there is interest in considering the modifying effects of genotype on phenotype in the presence of interaction with environment, an optimal joint association test would then be one that remains powerful under a variety of interaction models, ranging from those of strong interaction effect to little or no interaction effect [28; 95; 96]. To fill this demand, we present an approach to performing a joint test of gene and gene-environment interaction for common SNPs that builds upon the kernel-based methods introduced by Kwee et al. [97] and Wu et al. [98] to test for genetic main effects. Our kernel-based approach for joint analysis begins by grouping SNPs into SNP sets based on prior biological knowledge. We then apply a kernel function that quantifies the pairwise similarity between subjects based on the genotypes of the SNPs falling within the set, as well as environmental exposure. By introducing a garrote parameter into the kernel function (as considered in Maity and Lin [99] for microarray analysis), we can then construct a score statistic to assess whether pairwise genetic similarity in the presence of possibly modifying effects of environment correlates with phenotypic similarity.

The kernel-based approach to joint gene and gene-environment interaction testing is promising for three reasons. First, examining sets of SNPs rather than each SNP independently (as done in the methods of Kraft et al. [30] and Wang et al. [90]) will greatly reduce multiple-testing burden. For example, in a GWAS, while the traditional single-SNP regression approach could result in millions of tests [100], grouping all typed SNPs into genes and then implementing the kernel-based approach will result in ~20,000 tests [101]. Second, since multiple typed markers are likely to be in linkage disequilibrium (LD) with the causal variant, joint

consideration of these markers will capture the effect of a true causal variant more effectively than independent marker testing. Third, the kernel approach readily allows for inclusion of prior information (such as biological plausibility or association signals from prior association studies) in the form of weights to assist in the formation of the kernel matrix. SNP set methods have proved to be more powerful than univariate testing of main genetic effects [97; 98; 102] and we anticipate similar trends when considering joint tests of gene and gene-environment effects.

The remainder of this manuscript is organized as follows. We first describe our joint SNP set analysis framework, including how to form SNP sets and how to test SNP sets for association using a kernel framework that allows for potential modifying effects by an environmental factor. Next, we present simulation results comparing our joint approach both to traditional joint tests of gene and gene-environment interaction as well as to traditional tests of main genetic effects only. We then illustrate the kernel-machine approach using quantitative measures of post-traumatic stress disorder and depression collected as part of the Grady Trauma Project. We finish with concluding remarks and discuss potential extensions of our approach.

MATERIALS AND METHODS

Assumptions and Notation**:**

Assume a population-based study that samples N unrelated subjects. For each subject $j = 1,...,N$, we let Y_j denote the continuous phenotype and $\boldsymbol{X_j}$ be a vector of covariates. We further define E_j to be a continuous or categorical environmental

exposure of interest. Assume also that each subject has been genotyped at a collection of M common SNPs in a genetic region of interest. Define $\boldsymbol{G_j} = (G_{j,1}, G_{j,2}, \ldots, G_{j,m})$ as the genotypes at the M SNPs for subject j, where $G_{j,m}$ is coded as the number of copies of the minor allele that subject j possesses at variant m. The SNPs included in $\boldsymbol{G}$ will be referred to as the "SNP set." Wu et al. [98] suggest several ways for constructing SNP sets. A natural strategy is to group together all genetic variants that are located on or near a gene. However, we note that this strategy is reliant on the quality of the database used to define the SNPs that fall within the gene and may also result in the set harboring SNPs that are not necessarily in LD. Consequently, it may be advantageous to consider other SNP sets such as haplotype blocks or sliding windows. For illustration purposes in this manuscript, we will form SNP sets based on genes and consider all genotyped SNPs between the start and end of transcription, as well as variants within 2kb up- and down-stream from the gene to capture nearby regulatory regions.

Traditional Single-SNP Tests

We first describe two traditional tests that consider the analyses on the level of an individual SNP. First, if we believe a SNP has modest-to-no interaction with environment to influence outcome (see Models M4-M5 in Figure 1), we would typically apply a main-effects only model that implements a linear regression of the form

$$\boldsymbol{Y} = \boldsymbol{X}\gamma + \beta_{SNP}\boldsymbol{G_m} + \beta_{ENV}\boldsymbol{E} + \boldsymbol{e} \quad (1)$$

where $\boldsymbol{Y}$ is an N x 1 vector of phenotypes, $\boldsymbol{X}$ is an N x c vector of c covariates (including an intercept) with regression parameter vector γ, **E** denotes an N x 1 vector of the environmental exposure (considered a covariate and not an effect modifier) with regression coefficient β_{ENV}, and $\boldsymbol{G}_{\mathbf{m}}$ denotes an Nx1 vector of SNP genotypes at SNP m with regression parameter β_{SNP}. Finally, the residual error e follows a MVN distribution, $e \sim MVN(0, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{I}$ denotes the NxN identity matrix. We then implement a likelihood ratio test to assess the null hypothesis of $\text{H}_0 : \beta_{\text{SNP}} = 0$ for each SNP m. To adjust for multiple testing of M correlated SNPs, we could apply procedures like P_{ACT} [103] or use a permutation procedure that randomly shuffles the M genotypes of each subject as a unit (preserving the LD structure).

If we instead suspect sizable SNP-environment interaction (see models M1-M2 in Figure 1), we might then apply a joint test of SNP and SNP-environment interaction using the following modified model from (1)

$$\boldsymbol{Y} = \boldsymbol{X}\gamma + \beta_{\boldsymbol{SNP}}\boldsymbol{G}_{\boldsymbol{m}} + \beta_{\boldsymbol{ENV}}\boldsymbol{E} + \beta_{\boldsymbol{SNP*ENV}}\boldsymbol{G}_{\boldsymbol{m}}\boldsymbol{E} + \boldsymbol{e} \qquad (2)$$

where the notation is the same as defined in equation (1). The difference between model (2) and model (1) is the inclusion of a two-way interaction between $\mathbf{G_m}$ and **E** with regression parameter $\beta_{\boldsymbol{SNP*ENV}}$. Using a likelihood ratio test, we can assess the null hypothesis $\text{H}_0 : \beta_{\boldsymbol{SNP}} = \beta_{\boldsymbol{SNP*ENV}} = 0$. We repeat model (2) and obtain likelihood-ratio tests for each of the M genotyped SNPs. To adjust for multiple testing we require permutations, since P_{ACT} is only applicable to studies of main SNP effect. In

performing permutations under the null hypothesis, one must take care to preserve the relationship between phenotype ***Y*** with the covariates ***X*** and the environmental predictor ***E***; failure to preserve this relationship can lead to invalid inference [104]. We preserve this relationship while also maintaining LD structure among SNPs by randomly permuting the *M* genotypes of each subject as a unit, acknowledging that such permutation assumes that the genotypes are uncorrelated with the environmental predictors in the population.

Two-way Interaction Kernels

Using kernel regression, Kwee et al. [97] and Wu et al. [98] implemented mixed models for testing the effect of variant sets on complex human phenotypes. These approaches use a kernel function $K(\boldsymbol{G}_j, \boldsymbol{G}_k)$ to quantify the genetic similarity between subjects *j* and *k* across the *M* SNPs in the SNP set. We modify the methods described by Kwee et al. [97] and Wu et al. [98] to permit joint gene and gene-environment interaction testing as follows. First, we select a kernel that appropriately models interactions. While many kernels are available [105], we explore the use of the joint weighted 2-way interaction kernel (W2WK) in this work. We define $\mathbf{Z_j} = (\mathbf{G_j}, \mathrm{E_j})$ as the combined genetic and environmental information on each subject. We then define the weighted 2-way interaction kernel for subjects *j* and *k* as

$$\mathrm{K}(\mathrm{Z}_j, \mathrm{Z}_k) = (1 + \sum_{\mathrm{m=1}}^{M} w_m G_{jm} G_{km})(1 + \mathrm{E}_j \mathrm{E}_k) \quad (3)$$

Under this kernel, weight for the m^{th} variant, w_m, reflects the relative contribution of that variant to our estimate of local genetic similarity between subjects *j* and *k.* Ideally, causal variants would receive a large weight, and noncausal SNPs would receive a weight close to zero, making the weight of these SNPs

negligible. Although by nature we do not know which SNPs are causal, a careful weighting scheme can result in more power. Wu et al. [98] and Schifano et al [106] provide nice discussions on relevant weighting approaches for common SNP analyses. For all simulations and analyses reported here, we implement a weighting scheme based on the minor-allele frequency (MAF) of each assayed SNP that weights rarer variants over more common ones; the particular weight we apply for the m^{th} variant is $w_m = 1/\sqrt{\text{MAF}_m}$.

Based on the chosen kernel function, we can then define the kernel matrix $\boldsymbol{K}$ as the NxN matrix, where the $(j,k)^{\text{th}}$ element is equal to $K(\boldsymbol{Z}_j, \boldsymbol{Z}_k)$. The resulting $\boldsymbol{K}$ matrix represents genomic and environmental likeness, as well as interaction between genotype and environment, between all pairs of individuals across the M variants in the SNP set. Once we construct $\boldsymbol{K}$, we incorporate this kernel matrix within a mixed model that, for each pair of subjects, compares the genetic and environmental similarity to phenotypic similarity, adjusting for covariates. For continuous phenotypes, we can fit the mixed model as

$$\mathbf{Y} = \mathbf{X}\gamma + \mathbf{U} + e \qquad (4)$$

As with the traditional models described above, γ denotes a vector of regression parameters for fixed-effect covariates $\mathbf{X}$ and e is a vector of independent random errors that follows a normal distribution. $\mathbf{U}$ denotes a random effect affiliated with the variant set that follows the multivariate normal (MVN) distribution with a mean 0 and covariance matrix $\tau\mathbf{K}$. Within this random effect, τ denotes the component of variance due to the effects of the environment, variants within the variant set, and

the interactions between these factors.

When an interaction kernel is applied to the linear mixed model in equation (4) the environmental risk factors are a component of the random effects portion (**U**) of the model. Therefore, a null hypothesis of $\tau = 0$ would correspond to testing if none of the genetic, environmental, or interaction factors within the kernel influences the trait. This null is not particularly interesting in genetic studies, since the test would be significant if only the environmental factors, and no genetic or gene-environment interaction factors, were associated with the phenotype of interest. We therefore modify the kernel such that a significant finding is due only to a genetic effect in the presence of a potential interaction with the modeled environmental factors. To do so, we use a strategy employed by Maity and Lin [99] for microarray analysis and attach an extra "garrote" parameter, δ, to the genetic effects in the kernel function such that the weighted two-way interaction kernel becomes

$$\mathrm{K}(\mathrm{Z}_j, \mathrm{Z}_k) = (1 + \delta \sum_{m=1}^{M} w_m \mathbf{G}_{jm} \mathbf{G}_{km})(1 + \mathrm{E}_j \mathrm{E}_k). \quad (5)$$

With this reparameterization, we can then test for the effect of the gene in the presence of potential interactions with the environmental factors by considering the null hypothesis H_0: $\delta=0$ [107]. Maity and Lin [99] demonstrate that the appropriate score test is

$$\mathrm{S} = \hat{\tau}_0 (\mathbf{Y} - \mathbf{X}\hat{\beta}_0)^T \mathbf{V}^{-1} [d\mathbf{K} / d\delta_0] \mathbf{V}^{-1} (\mathbf{Y} - \mathbf{X}\hat{\beta}_0) / 2 \quad (6)$$

where $\mathbf{V} = \hat{\sigma}_0^2 \mathbf{I} + \hat{\tau}_0^2 \mathbf{K}_{(\delta=0)}$, and $d\mathbf{K}/d\delta_0$ denotes the derivative of $\mathbf{K}$ with respect to δ under the null hypothesis. ($\hat{\beta}_0$, $\hat{\tau}_0$, and $\hat{\sigma}_0^2$) are estimators of (β, τ, σ^2) under the null hypothesis, which can be estimated by applying restricted maximum likelihood

(REML) procedures to the reduced form of the linear mixed model [108]. The asymptotic distribution of the test *S* follows a complicated mixture of χ^2 distributions. We approximate the distribution using Welch-Satterthwaite's method [109], although we could also use Davies' method [110].

Simulations

To validate our method in terms of appropriate type I error and to assess its power compared to traditional joint and main-effect tests, we carry out simulation studies under a range of configurations. We perform simulations based on SNPs and LD patterns found 2 kb up- and down-stream from *signal transducer and activator of transcription 3 (STAT3)*, a gene on chromosome 17q21.31. We show the pairwise LD structure of SNPs in *STAT3* in Supplementary Figure 1. To incorporate observed LD patterns from HapMap samples, we used the HAPGEN package [111] to generate simulated SNP data. HAPGEN generates simulated genotype information for all SNPs identified in HapMap within the *STAT3* gene; however, to better replicate real GWAS conditions, we applied the testing approaches only to those SNPs that would be typed on standard genotyping arrays. Although 27 common SNPs fall within the *STAT3* gene, only 14 of the 27 are genotyped on the Illumina HumanOmni1-Quad genotyping platform. Thus, the 14 typed SNPs form the SNP set for the kernel approach, and only the 14 typed SNPs are tested for association using the traditional main and joint tests. Under simulations where the causal SNP is not genotyped, power to detect an association relies on LD between the causal SNP and typed SNPs.

Size and Statistical Power

We conducted simulations under four types of null linear models to verify that the joint W2WK approach properly controls the type I error rate. We assumed a model of $Y_j = \beta_{ENV} E_j + e_j$, where the error term, e_j, follows a standard normal distribution. E_j models an environmental exposure under a Bernoulli(0.5) distribution. We let β_{ENV}, the main effect size of the environment, be set to 0, 0.33, 0.67, and 1 (corresponding to R^2 values of approximately 0, 0.03, 0.11, and 0.25 respectively). For null simulations, we set sample size to *N*=250, 500, and 1000. For each of the four null models, we evaluated size using 5000 replicates of the data.

We next performed power calculations to compare the kernel approach to the traditional joint and main-effect tests under different levels of SNP-environment interaction. We simulate data for subject *j* under the model

$$Y_j = 0.33E_j + \zeta G_{j,m} + \zeta(0.2 - \zeta)G_{j,m}E_j + e_j \quad (7)$$

where E_j again models the environmental exposure under a Bernoulli(0.5) distribution, e_j follows a standard normal distribution, and $G_{j,m}$ is the allele count of the causal SNP in subject *j*. We set the values of ζ to the 5 different values (0.00, 0.05, 0.10, 0.15, and 0.20), which corresponds to simulation Models M1-M5, respectively (see Table 1). As ζ (and the model number) increases, the interaction effect decreases while the main effect increases. For instance, in Model 1, where $\zeta = 0$, we assume 'complete' interaction: the causal SNP affects phenotype exclusively through gene-environment interaction. Model 3 assumes that the genotypic effect is twice as large in exposed individuals compared with unexposed individuals. Model 5 assumes that the subgroups have the same mean genotypic effect; that is, there is no

gene-environment interaction. For each of the 5 models, we allowed each of the 27 common SNPs in *STAT3* to be causal in turn. We model genotypic effect acting first in an additive then in a dominant manner. For power simulations, we set sample size to N=500. As with the size simulations, we assumed only genotype information for the 14 SNPs on the Illumina HumanOmni1-Quad platform was available, and used only these SNPs to compute the test statistics. Power was estimated as the proportion of P-values <0.05 and was evaluated based on 500 replicates of the data per model.

RESULTS

Table II shows the empirical size for common variant analyses at α=0.05. Our simulations confirm that our 2-way interaction kernel approach maintains appropriate type-I error, regardless of main effect size of environment. The type-I error of the traditional main effect and traditional joint test were also appropriate.

Figures 2-4 show the power results for models M1, M3, and M5, the effect of the SNP on outcome originated under either an additive genetic model (left) or a dominant model (right). Similar power results for models M2 and M4 are shown in Supplemental Figures 2 and 3, respectively. Power is plotted as a function of causal SNP, where the causal SNPs are ordered by genomic location. The genotyped SNPs (denoted by the 'x' on the bottom of the plots) were used to compute the test statistics, but each HapMap SNP (regardless of whether it is typed) is treated as causal in turn. Thus, in situations where the causal SNP is not typed, we rely on the correlation of the causal SNP with observed typed SNPs in the set to gain statistical

power. The MAF of the SNPs is plotted below the power plot in the grey line. For example, SNP 9 (rs9909659) has a MAF 0.21 and is not genotyped on the Illumina array. However, as shown in Supplementary Figure 1, it is in strong LD with several SNPs that are typed (R^2>0.9 for SNP 3 (rs3198502), SNP 4 (rs1053005), and SNP 6 (rs3744483)). Power to detect an effect of SNP 9 relies on LD among these genotyped SNPs.

In our simulations, the traditional joint test did not always outperform the main-effect test, even when a significant interaction effect was present. Implementation of the traditional joint test resulted in considerable increases in power relative to the traditional main test only under models of complete interaction (model M1, Figure 2). Under the M2 model shown in Table I, despite the fact that the interaction effect is 4-fold larger than the genetic main effect in this model, the traditional joint test provides only a modest power gain over the traditional main effect test (Supplementary Figure 2). When the interaction effect is equal to or smaller than the genetic main effect (model M3-M5, Figures 3 & 4 and Supplementary Figure 3), the traditional main effect tests are consistently more powerful than the traditional joint tests.

Across all five models, for both additive and dominant assumptions, the joint W2WK approximately matches or outperforms the optimal traditional test. Under the complete and strong interaction models (models M1 and M2, Figure 2 and Supplementary Figure 2), the joint W2WK kernel matches or outperforms the traditional joint test across all SNPs. In models M3 and M4, although the main effect test outperforms the traditional joint test, the joint W2WK outperforms the main

effect test (Figures 3 and Supplementary Figure 3). Even under the assumption of no gene-environment interaction occurring (M5, Figure 4), our joint W2WK approach remains somewhat more powerful or of approximately equal power when compared with the traditional main-effect test for the large majority of SNPs.

The power of our joint W2WK approach relies on LD existing between the causal SNP and genotyped SNPs in the sample. To examine the relationship between LD and power in our simulated datasets, we calculated the median squared correlation (median R^2) of the causal SNP with genotyped SNPs in our SNP set across simulated datasets for a specific model. As shown in Supplementary Figure 4, our joint W2WK approach is least powerful compared with the traditional approaches when median R^2 between causal SNP and genotyped SNPs is close to 0 but becomes increasingly more powerful than these other approaches as the median R^2 increases. Our findings regarding the relationship between median R^2 and power for our joint W2WK test yield similar conclusions to those reported by Wu et al. [98] and Schifano et al. [106] for SNP set analysis of main effects on phenotype.

The joint W2WK approach offers more power than the optimal traditional approach across a considerable range of causal SNP minor allele frequencies. Although all approaches are more powerful as the causal SNP's MAF increases, there is no clear relationship between MAF and relative strength of our approach (Supplementary Figure 5). The weighting scheme we selected for the bulk of our simulations is an inverse relationship with MAF; this weighting scheme is most beneficial when the MAF of the causal SNP is rarer relative to the genotyped SNPs in the set. To examine if power of the joint W2WK approach is affected by the choice

of MAF weight, we also performed an unweighted analysis that assumes equal contribution from all genotyped SNPs in the SNP set. We present these results in Supplementary Figure 5. Overall, power using the unweighted version of our approach was somewhat lower than from using the joint W2WK approach, except for causal SNPs that were quite common (MAF > 0.35). For these SNPs, the unweighted approach offered slightly more power to detect an effect than the joint W2WK.

Application to Grady Trauma Project Data

Depression is a moderately heritable disorder ($h^2 \approx 0.30$), yet, despite substantial interest in identifying genetic causes of the disorder, its genetic underpinnings remain largely unidentified [112]. Research indicates a potential association between depression and genes in the cannabinoid receptor 1 (CB_1) pathway [113-116]. The relationship between depression and CB_1 may be modified, however, by gender [117; 118].

We applied our joint W2WK approach to a GWAS study of depression to assess the relationship between CB_1 genes and outcome, allowing for interaction with gender, and contrasted our results with those found under the traditional single-SNP tests. Data used in our analysis were collected as part of a larger study, called the Grady Trauma Project (GTP), which investigates the role of genetic risk factors for psychiatric disorders such as post-traumatic stress disorder and depression [119; 120]. Participants in the GTP are served by the Grady Hospital in Atlanta, Georgia, and are predominantly urban, African American, and of low

socioeconomic status. GTP staff approach subjects in the waiting rooms of Grady Primary Care and Obstetrics and Gynecology and obtain their written consent to participate. GTP staff conduct an extensive verbal interview, which includes demographic information, a history of stressful life events, and several psychological surveys. The GTP queries participants on the Beck Depression Inventory (BDI), a 21-item multiple-choice questionnaire that assesses symptoms of depression [34]. Summing the responses yields a score ranging from 0-63, with scores higher than 28 being indicative of moderate to severe depression. We selected this score as a continuous outcome variable, transforming each individual's BDI scores to $y_i = \ln(BDI_i + 1)$, where ln is the natural log, to uphold the normality assumption required for the traditional tests.

The GTP genotyped participants on the Illumina HumanOmni1-Quad array to permit GWAS analyses. For this work, we studied the *Cannabinoid Receptor 1* gene (*CNR1)*, on chromosome 6q14-q15, which encodes for the CB_1 receptor, and the *Fatty Acid Amide Hydrolase* gene (*FAAH)*, on chromosome 1p35-p34, which breaks down the primary endocannabinoid in humans, as genes of interest based on prior reports of the cannabinoid pathway involvement in depression and negative affect. The HumanOmni1-Quad array genotypes 11 common SNPs within 2kb up- and downstream from *CNR1*, and 7 common polymorphisms in and near the *FAAH* gene. Additionally, we obtained the top 10 principal components (PCs) from the GWAS data, which we included as covariates in all models to account for population stratification. We obtained BDI scores, genotype, PCs, and gender information on 3475 subjects.

We applied our joint W2WK test, along with traditional joint and traditional main-effect tests, to the dataset. We observed that both the traditional main-effect and traditional joint tests indicated a nominal association between *CNR1* and BDI scores, whether gender was included only as a covariate or as an effect modifier (*P*-values 0.007 and 0.024 respectively). The *P*-value testing this association using the joint W2WK approach yielded similar trends as the two other tests but yielded a p-value that was at least 4-fold smaller than either traditional approach (*P*-value 0.0016). Evidence suggests that the association between *CNR1* and BDI scores might be due to a blend of genetic main effect and interaction between *CNR1* and gender, but is unlikely to be an example of complete or very strong interaction, since the P-values of the traditional main effect test is smaller than that of the traditional joint test). None of the three tests found a significant association between variants in the *FAAH* gene and BDI scores (Table III).

DISCUSSION

We have presented a kernel machine based framework for SNP set analysis for continuous outcomes when an interaction between genotype and an environmental insult is suspected. The proposed test is a variance component score test, which relies on fitting the null linear regression model to compute the test statistic. Since the *P*-values are computed analytically, our method allows faster analyses on a genomewide scale than the traditional regression approaches, which might rely on permutation procedures to establish significance. Analysis of simulated *STAT3* data for 500 and 1000 subjects takes 30 seconds and 3.5 minutes,

respectively, on a MacBook Pro possessing a 2.2 GHz processor and 8 GB of memory. If one were to parallelize the approach across 50 CPUs, one could complete a GWAS analysis of 20,000 gene sets with a sample size of 1000 in approximately one day. We provide R software implementing the approach on our website (see Web Resources) which can be run through PLINK, if desired.

In general, our joint W2WK approach has more power than the either the joint or main-effect traditional approaches. Since the magnitude and prevalence of interactions is largely unknown, we considered several models of gene and gene-environment interaction effects. When the underlying model is one of complete interaction, the joint W2WK outperforms the traditional joint test across a range of minor allele frequencies and LD patterns. Our approach performs particularly well relative to either traditional test when the underlying causal model involves a blend of both interaction and main genetic effects. We consider the power gains under these models to be especially noteworthy, since they are considered to be more biologically plausible than a model of complete interaction [28]. We lastly considered the scenario that no gene-environment interaction is occurring; that is, all the genotypic effect occurs through genetic main effect. Using traditional joint testing under this model would result in costly loss of power relative to the traditional main-effect test. However, across all modeled causal SNPs, the joint W2WK approach maintains power that rivals or even modestly outperforms the traditional main-effect test. We have also demonstrated that the joint W2WK approach has more power relative to the traditional approaches across all interaction models

when the genetic effects are acting in a dominant fashion, but tested assuming an additive effect.

Our joint W2WK is explicitly designed to test for a joint effect of genetic main effect and interaction between gene and environment. If instead one is interested in testing exclusively for a gene-environment interaction effect using SNP sets, one can apply a related kernel procedure created by Lin et al. called GESAT. [121]. Like W2WK, GESAT is a variance-component score test that utilizes a kernel function for analysis. However, while our joint W2WK procedure models genetic, environment, and gene-environment interaction effects as random effects via a kernel function in the mixed-model framework, GESAT only models the gene-environment interaction term as random and models the main genetic and environmental effects parametrically as fixed effects (estimated under the null using ridge regression). GESAT is useful when an interaction-only test is more desirable than a joint test, such as for detecting a crossover interaction (i.e. when the genetic effects change as a function of environment, such that the genotype conferring lowest risk in one environment confers highest risk in another environment).

Although the results presented here are focused on interaction between environment and common SNPs, the approach is readily extendible to rare variant gene-environment interaction analysis. The approach can also be used to simultaneously model multiple environmental exposures, which would be useful in cases where several environmental measurements might be expected to correlate with a true latent exposure that is interacting with genotype to influence outcome. We will explore these ideas more in future work.

ACKNOWLEDGEMENTS

This work was support by NIH R01-HG007508 from the National Human Genome Research Institute R01-MH071537 from the National Institute of Mental Health, and National Institute of Arthritis and Musculoseketal and Skin Diseases R01-AR060893. The contents of this manuscript do not reflect the views of the Department of Veterans Affairs or the United States Government. For purposes of disclosing duality of interest, Michael Epstein is a consultant for Amnion Laboratories.

WEB RESOURCES

Epstein Software: http://www.genetics.emory.edu/labs/epstein/software

Table I: Models of gene-environment interaction

		Total Genetic Effect	
Model	ζ	Exposed Subgroup	Unexposed Subgroup
M1	0.00	0.20	0.00
M2	0.05	0.20	0.05
M3	0.10	0.20	0.10
M4	0.15	0.20	0.15
M5	0.20	0.20	0.20

The five models of interaction considered in our simulations. Under M1, we assume that unexposed individuals have no genotypic effect. In M2, the genotypic effect in unexposed individuals is ¼ that of exposed individuals. In M3, unexposed individuals have on average half the genotypic effect of exposed individuals. In M4, unexposed individuals have ¾ the mean genotypic effect as the exposed subgroup. Finally, in M5, mean genotypic effect is equal across subgroups; no interaction effect is occurring.

Table II: Controlling for Type I Error Rate: Empirical Size

Sample Size	250				500				1000			
Environmental Effect	0.00	0.33	0.67	1.00	0.00	0.33	0.67	1.00	0.00	0.33	0.67	1.00
W2WK	.043	.045	.050	.049	.050	.050	.051	.047	.049	.046	.051	.052
Traditional Joint	.051	.052	.056	.053	.051	.048	.052	.049	.052	.051	.050	.047
Traditional Main	.055	.053	.058	.053	.051	.050	.048	.053	.049	.051	.049	.049

Empirical sizes for the W2WK, traditional joint, and traditional main-effect tests at the α=0.05 level. Sample sizes were set to N=250, 500, and 1000. Environmental main effect was allowed to vary from 0, 0.33, 0.67, and 1 (corresponding to R^2 values of approximately 0, 0.03, 0.11, and 0.25 respectively).

Table III: Analysis of the Grady Trauma Project Data

	Gene	
	CNR1	FAAH
W2WK	0.002	0.406
Traditional Joint	0.024	0.407
Traditional Main	0.007	0.603

P-values using the joint W2WK, traditional joint, and traditional main-effect only tests on the Grady Trauma Project dataset. The joint W2WK and traditional joint analyses considered interactions with gender.

FIGURE LEGENDS

Figure 1: A range of possible gene and gene-environment models, from M1 (complete interaction) to M5 (no interaction). The red line represents genotypic effect among individuals who have been exposed to the environmental insult. The 5 blue lines represent possible genotypic effects for individuals who were not exposed to the environmental insult, resulting in models that range from complete interaction (M1) to no interaction (M5).

Figure 2: Power for the W2WK (red), traditional joint (blue), and traditional main effect (green) approaches by causal SNP under the M1 model. An "x" marks the 14 SNPs that were modeled as genotyped in our simulations. The MAF of each SNP (grey line) is along the right Y-axis. Plot on the left assumes an additive model; the plot on the right assumes the underlying model is dominant, but was tested as additive. Inset plot shows the underlying model (M1): solid black line represents genotypic effect among individuals who have been exposed to the environmental insult. Dotted black line shows genotypic effect of individuals who were not exposed to the environmental insult under the M1 model. Dotted grey lines indicate alternate models that are considered elsewhere in this manuscript.

Figure 3: Power for the W2WK (red), traditional joint (blue), and traditional main effect (green) approaches by causal SNP under the M3 model. An "x" marks the 14 SNPs that were modeled as genotyped in our simulations. The MAF of each SNP (grey line) is along the right Y-axis. Plot on the left assumes an additive model; the plot on the right assumes the underlying model is dominant, but was tested as additive. Inset plot shows the underlying model (M3): solid black line represents genotypic effect among individuals who have been exposed to the environmental insult. Dotted black line shows genotypic effect of individuals who were not exposed to the environmental insult under the M3 model. Dotted grey lines indicate alternate models that are considered elsewhere in this manuscript.

Figure 4: Power for the W2WK (red), traditional joint (blue), and traditional main effect (green) approaches by causal SNP under the M5 model. An "x" marks the 14 SNPs that were modeled as genotyped in our simulations. The MAF of each SNP (grey line) is along the right Y-axis. Plot on the left assumes an additive model; the plot on the right assumes the underlying model is dominant, but was tested as additive. Inset plot shows the underlying model (M5): solid black line represents genotypic effect among individuals who have been exposed to the environmental insult. Dotted black line shows genotypic effect of individuals who were not exposed to the environmental insult under the M5 model. Dotted grey lines indicate alternate models that are considered elsewhere in this manuscript.

Figure 1: Models of Gene and Gene-Environment Interaction on Phenotype.

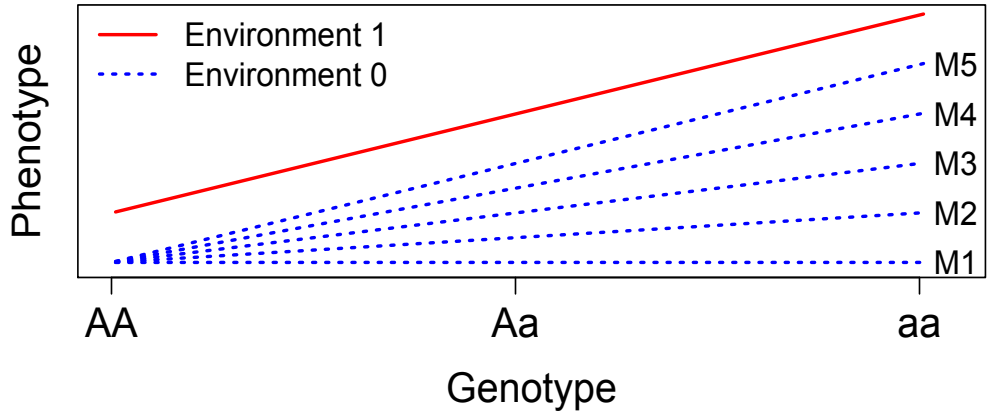


Figure 2: Power under M1 Model

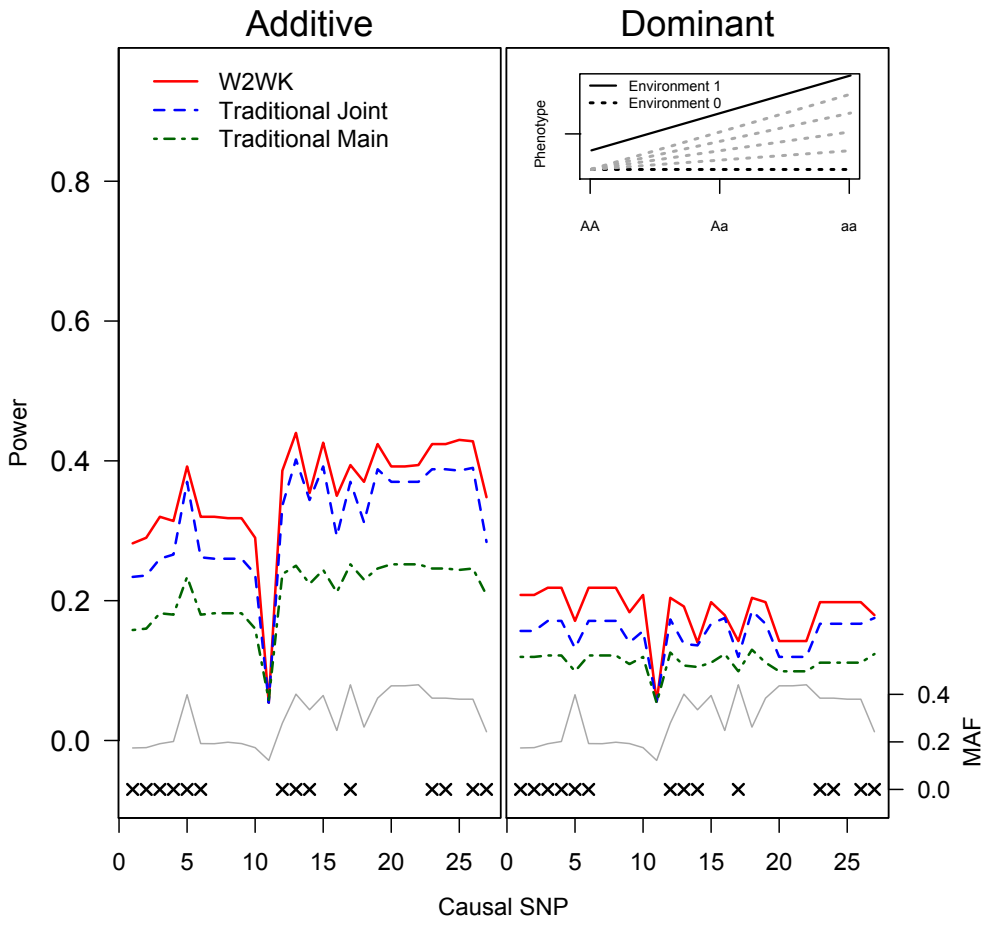


Figure 3: Power under M3 Model

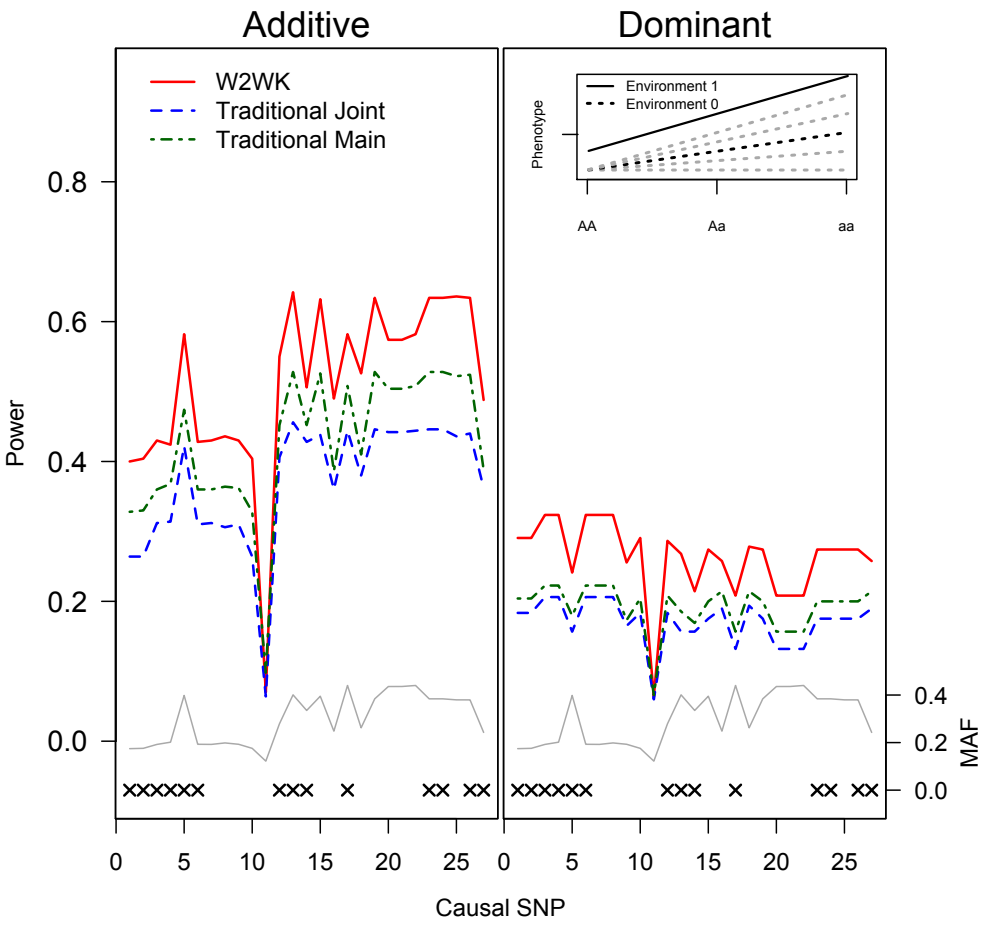
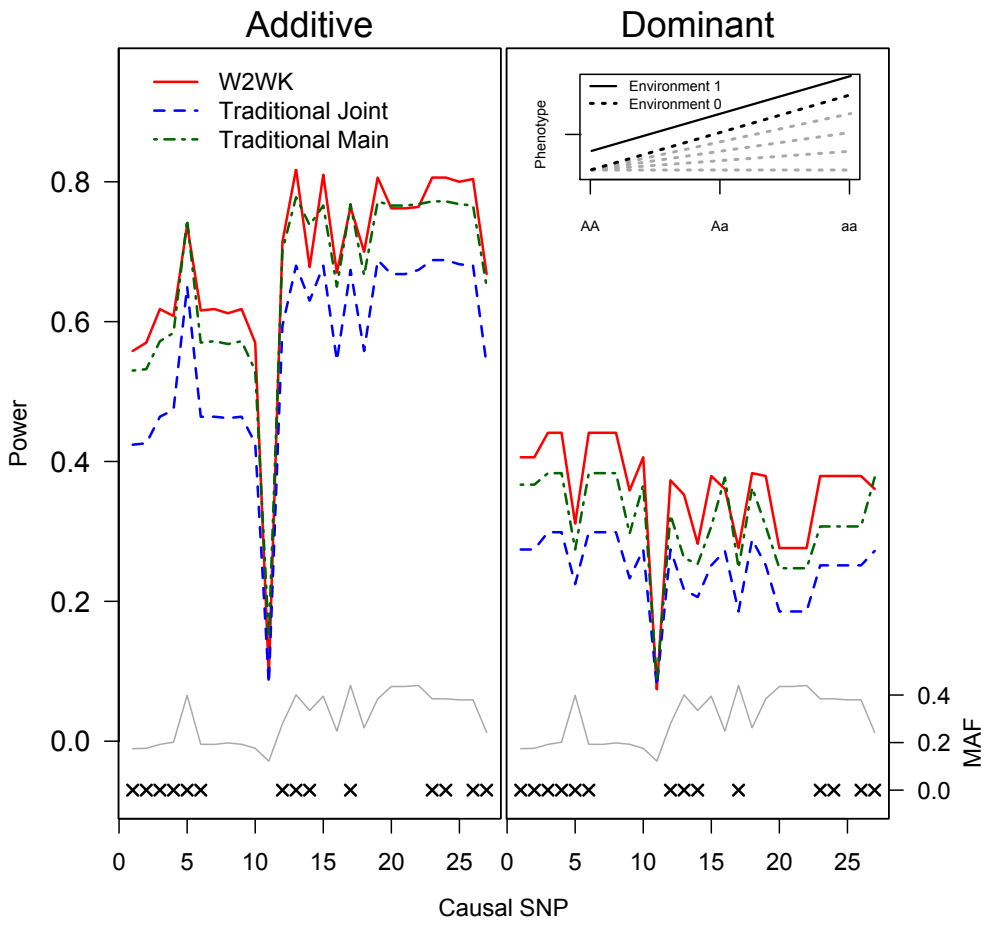


Figure 4: Power M5 Model



Supplementary Figure Legends:

Supplementary Figure 1: Pairwise LD (R^2) heatmap for all SNPs reported in HapMap ±2kb of the *STAT3* gene.

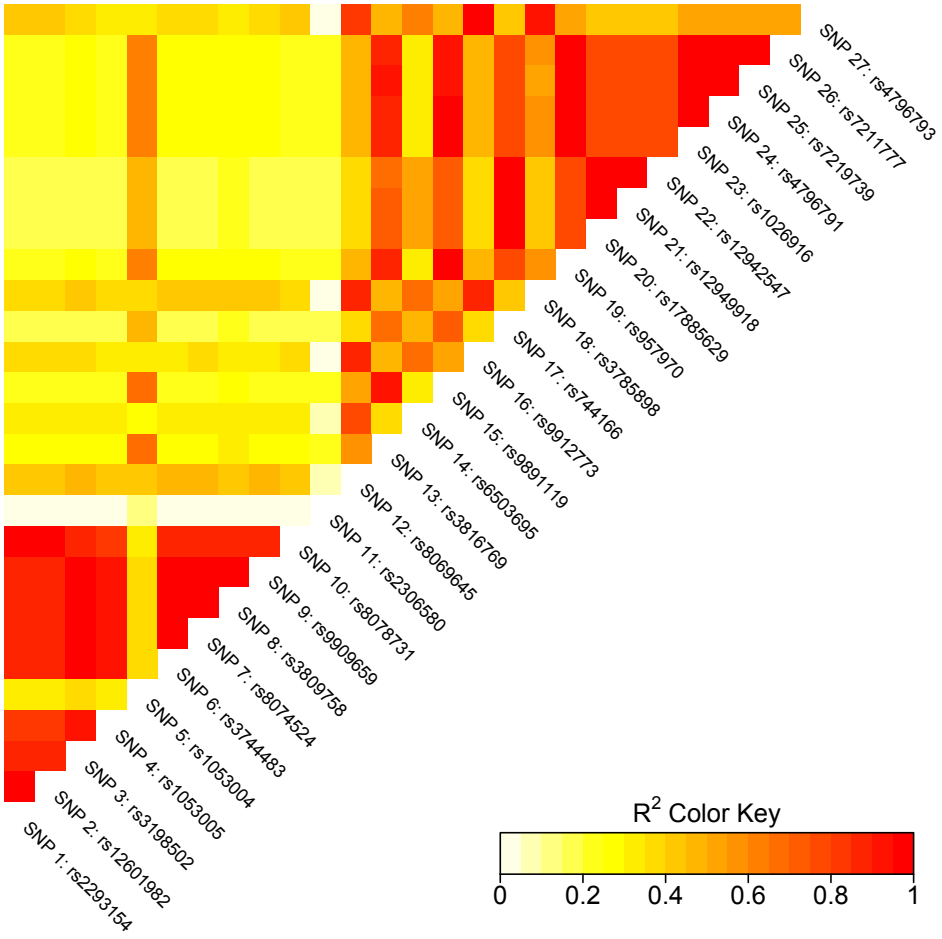
Supplementary Figure 2: Power for the W2WK (red), traditional joint (blue), and traditional main effect (green) approaches by causal SNP under the M2 model. An "x" marks the 14 SNPs that were modeled as genotyped in our simulations. The MAF of each SNP (grey line) is along the right Y-axis. Plot on the left assumes an additive model; the plot on the right assumes the underlying model is dominant, but was tested as additive. Inset plot shows the underlying model (M2): solid black line represents genotypic effect among individuals who have been exposed to the environmental insult. Dotted black line shows genotypic effect of individuals who were not exposed to the environmental insult under the M2 model. Dotted grey lines indicate alternate models that are considered elsewhere in this manuscript.

Supplementary Figure 3: Power for the W2WK (red), traditional joint (blue), and traditional main effect (green) approaches by causal SNP under the M4 model. An "x" marks the 14 SNPs that were modeled as genotyped in our simulations. The MAF of each SNP (grey line) is along the right Y-axis. Plot on the left assumes an additive model; the plot on the right assumes the underlying model is dominant, but was tested as additive. Inset plot shows the underlying model (M4): solid black line represents genotypic effect among individuals who have been exposed to the environmental insult. Dotted black line shows genotypic effect of individuals who were not exposed to the environmental insult under the M4 model. Dotted grey lines indicate alternate models that are considered elsewhere in this manuscript.

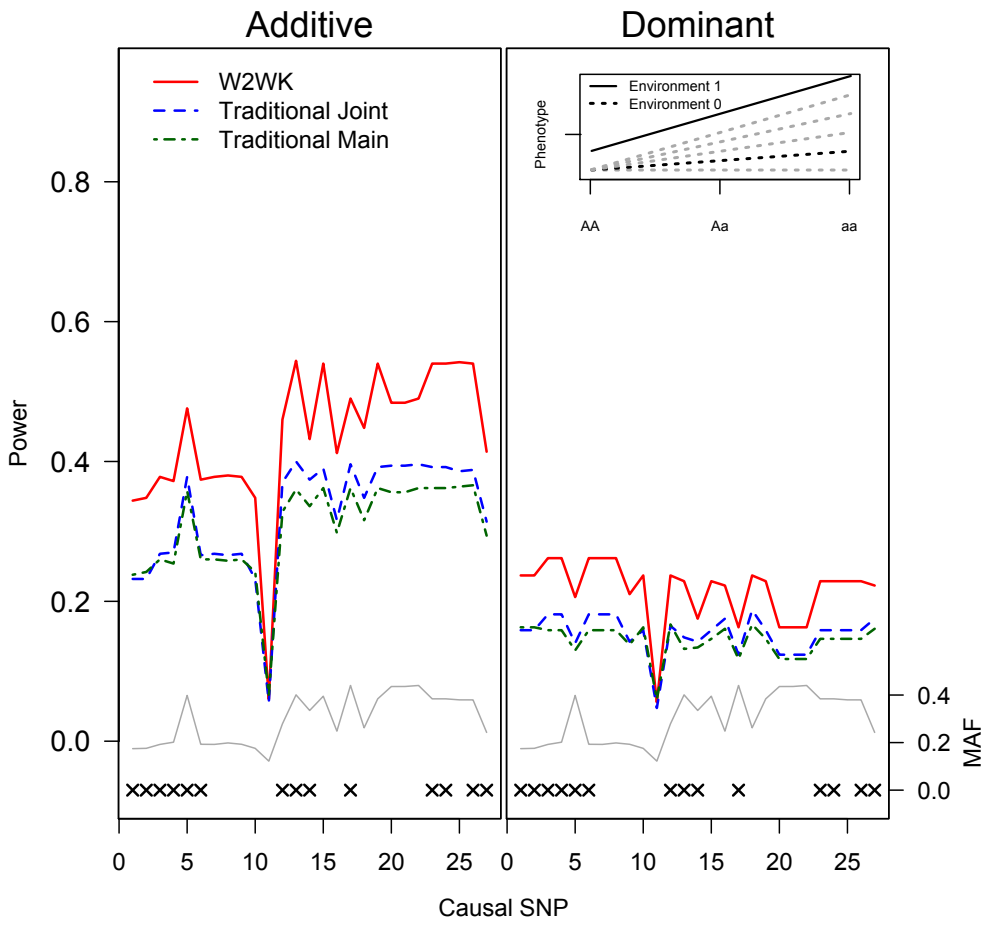
Supplementary Figure 4: Power is plotted as a function of median R^2, assuming additive effects. We define median R^2 as the median squared correlation of the causal SNP with the genotyped SNPs in our SNP set. Our W2WK is shown in red, traditional joint test is shown in blue, and the traditional main effect test is shown in green.

Supplementary Figure 5: Power for the W2WK (red), traditional joint (blue), and traditional main effect (green) approaches by causal SNP under the M1-M5 additive models. In addition, we show the power to detect an effect using an unweighted two-way interaction kernel (cyan). When MAF of the causal SNP is greater than ~0.35, an unweighted kernel approach is slightly more powerful than the weighting scheme we selected for our W2WK approach.

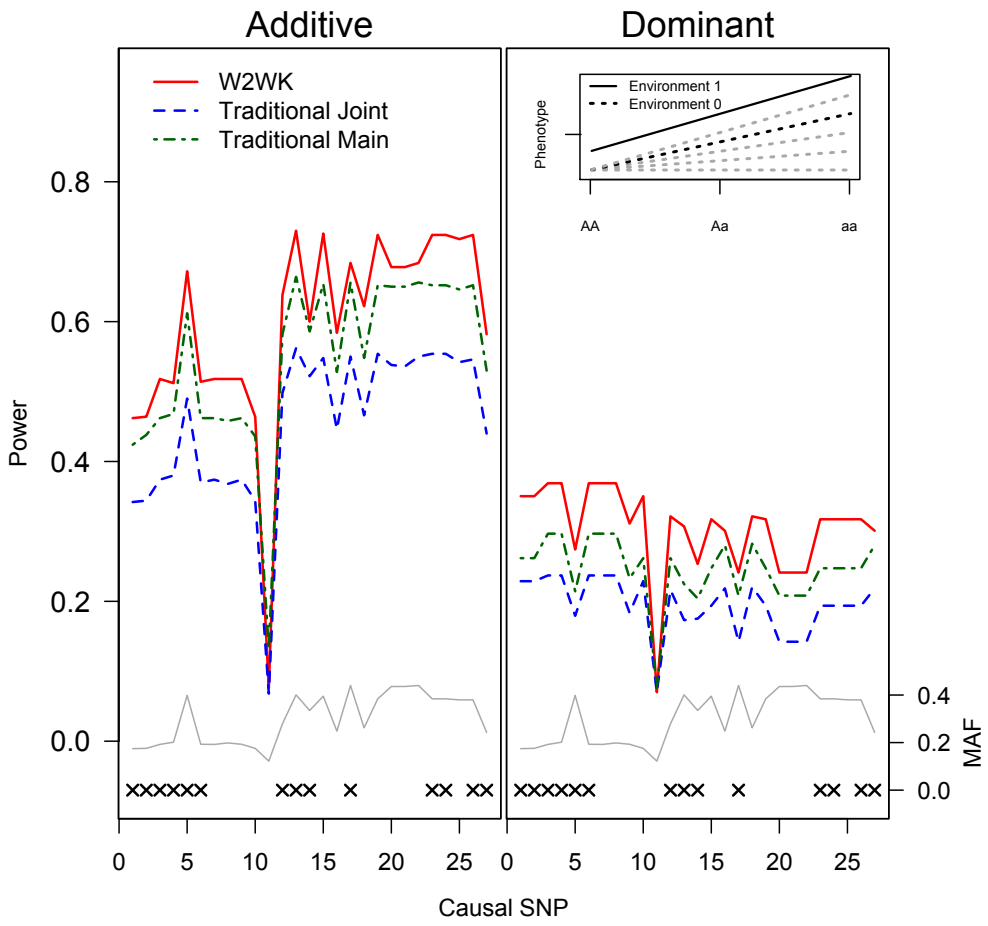
Supplementary Figure 1: Pairwise LD of *STAT3*



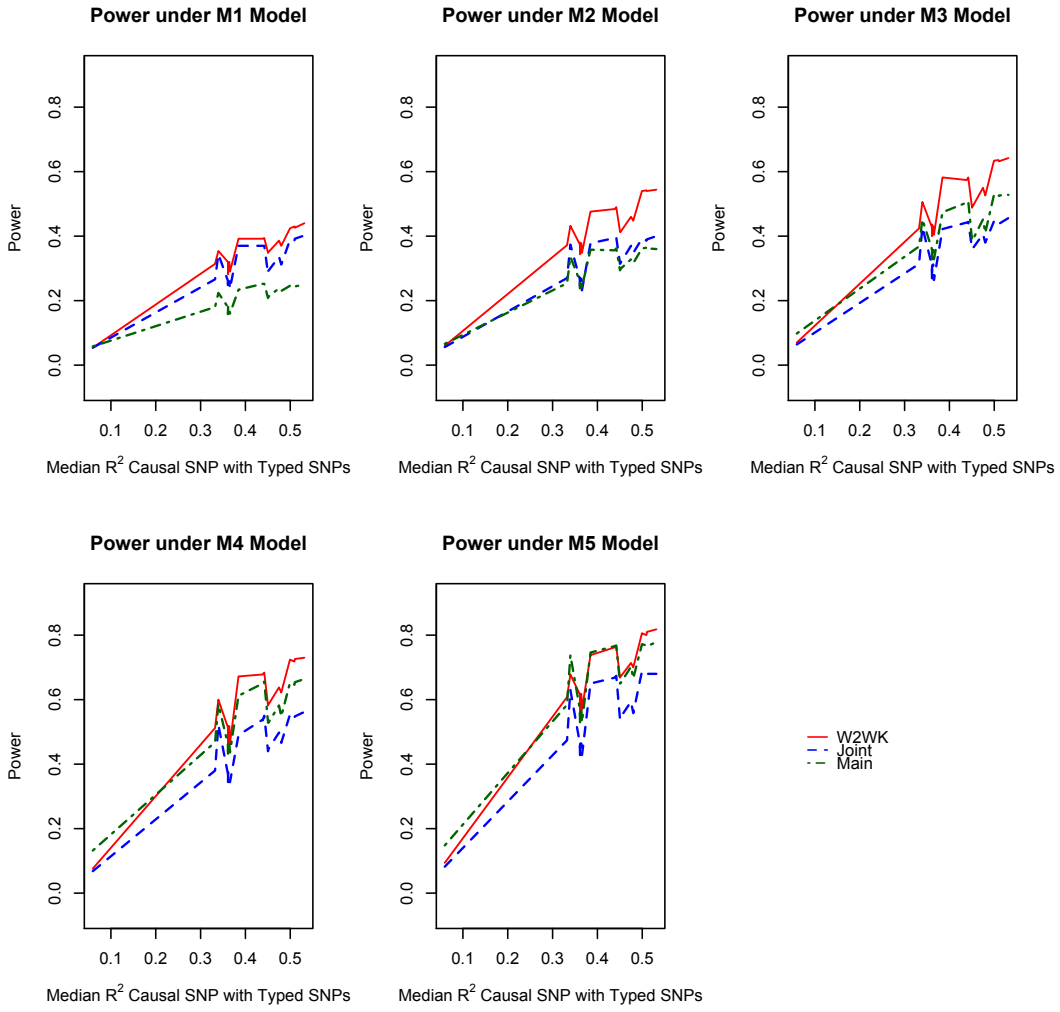
Supplementary Figure 2: Power under M2 Model



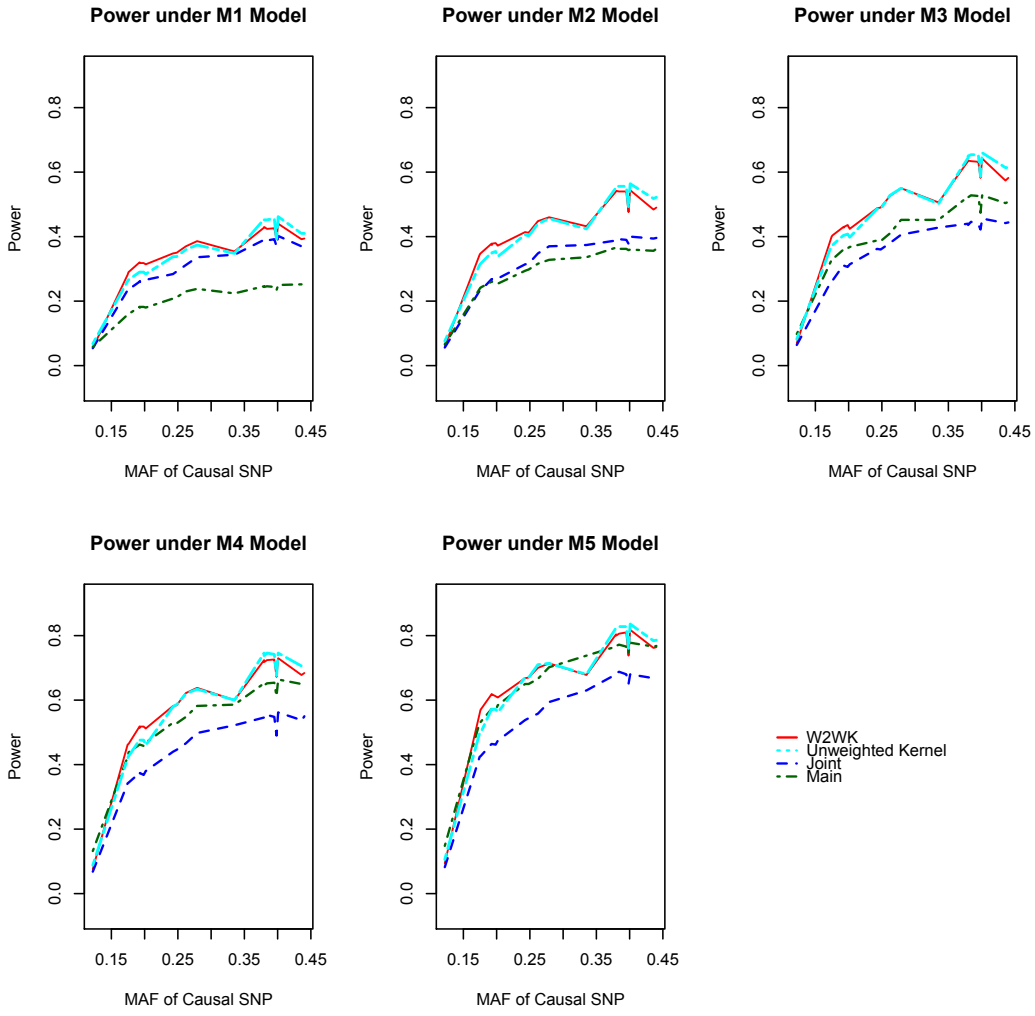
Supplementary Figure 3: Power under M4 Model



Supplementary Figure 4: Power as a Function of Median R^2



Supplementary Figure 5: Power as a function of causal SNP MAF



CHAPTER 3:
A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants

A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants

K. Alaine Broadaway[1], David J. Cutler[1], Richard Duncan[1], Jacob L. Moore[2], Erin B. Ware[3,4], Min A Jhun[3], Lawrence F. Bielak[3], Wei Zhao[3], Jennifer A. Smith[3], Patricia A. Peyser[3], Sharon L.R. Kardia[3], Debashis Ghosh[5], Michael P. Epstein[1]

[1]Department of Human Genetics, Emory University, Atlanta, GA

[2]Department of Evolution and Ecology, University of California, Davis, Davis, CA

[3]Department of Epidemiology, University of Michigan, Ann Arbor, MI

[4]Institute for Social Research, University of Michigan, Ann Arbor, MI

[5]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO

ABSTRACT

Increasing empirical evidence suggests that many genetic variants influence multiple distinct phenotypes. When cross-phenotype effects exist, multivariate association methods that model pleiotropy are often more powerful than univariate methods that model each phenotype separately. While several statistical approaches exist for testing pleiotropy for common variants, there is a lack of cross-phenotype tests for gene-based analysis of rare variants. In order to fill this important gap, we introduce a new statistical method for cross-phenotype analysis of rare variants using a nonparametric distance-covariance approach that compares similarity in multivariate phenotypes to similarity in rare-variant genotypes across a gene. The approach can accommodate both binary and continuous phenotypes and further can adjust for covariates. Our approach yields a closed-form test whose significance can be evaluated analytically, thereby improving computational efficiency and permitting application on a genome-wide scale. We use simulated data to demonstrate that our method, which we refer to as the Gene Association with Multiple Traits (GAMuT) test, provides increased power over competing approaches. We also illustrate our approach using exome-chip data from the Genetic Epidemiology Network of Arteriopathy.

INTRODUCTION

The 1980s were an era of debate in the theoretical quantitative genetics community between two competing schools of thought [2]. The question of interest was "What is the nature of genetic variation contributing to complex traits?" On one hand there was the infinitesimal school [7], which argued that complex traits were the result of mutation/selection balance under stabilizing selection. The variants that contributed to traits were a combination of very rare alleles of potentially large effect combined with many common alleles of exceedingly small effect. The opposing camp, sometimes called Neo-Darwinian [5], argued that a substantial fraction of genetic variation was contributed by high frequency alleles of large effect, whose frequency was maintained through balancing selection [8]. The neo-Darwinian's school leveled two interrelated and potentially fatal criticisms at the infinitesimal camp: believing in the infinitesimal model requires one to simultaneously accept that 1) much of the standing genetic variation is due to extremely rare alleles of large effect, and 2) a large fraction of the genome of an organism is contributing to nearly every phenotype [5]. That means that nearly every rare, large-effect allele must simultaneously be contributing to a large number of different traits. The neo-Darwinian school argued the only alternative to believing in this worldview was to suppose that a substantial fraction of the variation in complex traits was contributed to by common alleles of large effect.

Perhaps without explicitly acknowledging it [3; 12; 13], the genome-wide association study (GWAS) era was fundamentally testing the predictions of the neo-Darwinian school. We now know that, by and large, common alleles of large effect

do not exist. When considered collectively, common variants can explain a sizable proportion of the heritability for many complex traits like height, body-mass index, and cardiovascular disease [15; 16; 37]. However, common trait-influencing variants identified and replicated by GWAS tend to have very modest effect sizes. Much of the genetic contributors to complex traits still remain undiscovered, and are presumably due to very rare variation. Thus, while it may be time to reject the neo-Darwinian worldview in favor of the infinitesimal model, we cannot logically do so without simultaneously embracing the central Neo-Darwinian critique of the infinitesimal school: most traits should be affected by a large fraction of the genome, and rare alleles of large effect should be generally highly pleiotropic for seemingly unrelated phenotypes. Moreover, if we adopt this worldview whole-heartedly, it suggests a paradigm shift in how we should approach genetic association studies.

If rare alleles of large effect are both ubiquitous and generally highly pleiotropic, we can leverage this to discover genes involved in complex traits. When pleiotropy exists, an analysis that models multiple phenotypes simultaneously in a multivariate or "cross-phenotype" framework will provide greater statistical power than a standard univariate method that considers each phenotype separately [61; 62]. Because underlying genetic pleiotropy will induce phenotypic correlation, a genetic association that exists with multiple traits will be more readily detectible through cross-phenotype analyses due to the extra information provided by cross-phenotype correlation. This information is ignored in univariate analyses. Additionally, when pleiotropy is suspected, allowing for cross-phenotype

associations may yield a more biologically plausible statistical model, and potentially help to explain shared pathogenesis[61; 122].

Cross-phenotype association tests for common variants using single-nucleotide polymorphisms (SNPs) have demonstrated considerable success [52; 123]. For example, common-variant cross-phenotype association has been reported among Crohn's disease and ulcerative colitis [124], different facial morphology measures [125], and among bipolar disorder, autism spectrum disorder, ADHD, major depressive disorder, and schizophrenia[55]. However, while there are several excellent statistical methods appropriate for pleiotropic analysis of common genetic variants [63-68], theory tells us that rare alleles cannot be ignored, and pleiotropy due to rare alleles should be more pronounced. Unfortunately, there is a shortage of analogous statistical approaches to assess cross-phenotype associations of rare genetic variants.

Currently, most cross-phenotype association methods are designed to assess the effect of a single polymorphism at a time; however, in rare variant analysis, a test typically requires aggregation of information from multiple rare variants within a gene simultaneously. One possible rare-variant pleiotropic test is a modification of the common-variant method of Maity et al. [67]. While the Maity approach was developed to study the relationship between multiple SNPs in a gene and multiple correlated phenotypes using mixed models, it could be adapted to consider rare variants rather than common SNPs. Additionally, Wang et al. proposed an alternative gene-level test of pleiotropy that uses multivariate functional linear models (MFLM) [69]. However, we note that the approaches of Maity and Wang only

allow for continuous phenotypes and thus cannot be applied to important categorical phenotypes like presence or absence of a disease. Ideally, a pleiotropic test of rare variation should be able to handle both continuous and categorical phenotypes and be able to scale efficiently to handle an arbitrary number of phenotypes. Here, we present a method that meets both these criteria.

We propose a method called Gene Association with Multiple Traits (GAMuT) for association testing of high-dimensional phenotype data with high-dimensional genotype data. GAMuT relies on a machine-learning framework called kernel distance-covariance (KDC) [126-130] to provide a nonparametric test of independence between a set of phenotypes and a set of genetic variants. The KDC framework used by GAMuT assesses whether pairwise phenotypic similarity in a sample is independent of pairwise rare-variant genotypic similarity in a gene or region of interest. The framework allows for an arbitrary number of phenotypes that can be both continuous and/or categorical in nature, and similarly allows for an arbitrary number of genotypes, thereby permitting gene-based testing of rare variants. GAMuT can correct for important covariates, such as measures of ancestry to account for population stratification. Furthermore, GAMuT is a closed form test that yields analytic *P*-values, thus scaling easily to genome-wide analysis.

This manuscript is organized as follows. First, we develop GAMuT using the KDC framework and show how we derive analytic *P*-values for this test. We also describe how we can adjust for covariates in GAMuT. Additionally, we describe an efficient resampling strategy that can be used if one wishes to construct a GAMuT test multiple times using different similarity measures for phenotypes and/or

genotypes. This resampling strategy appropriately corrects for multiple testing but is far less computationally intensive than standard permutations. Next, we present simulation work comparing GAMuT to MFLM and univariate SKAT [131] analysis of rare variants under various trait-influencing models and demonstrate that our analytic strategy can be considerably more powerful than these competing approaches; both when pleiotropy truly exists but also when variants influence only one of the phenotypes under consideration. Finally, we apply GAMuT to perform exome-chip analysis of multivariate phenotypic measures of cardiovascular health using data from the Genetic Epidemiology Network of Arteriopathy (GENOA) [132].

MATERIALS AND METHODS

Assumptions and Notation: We assume a sample of N subjects who have been measured for multiple phenotypes of interest and possess sequencing or exome-chip data in a target gene or region. For subject j (j=1,…,N), we define $\boldsymbol{P}_j = (P_{j,1}, P_{j,2}, \ldots, P_{j,L})$ as the L phenotypes of the subject and allow such phenotypes to be continuous and/or categorical in nature. We then define a matrix of phenotypes for the entire sample $\boldsymbol{P} = (\boldsymbol{P}_1^T, \boldsymbol{P}_2^T, \ldots, \boldsymbol{P}_N^T)^T$, which is of dimension N x L. Similarly, we define $\boldsymbol{G}_j = (G_{j,1}, G_{j,2}, \ldots, G_{j,V})$ to be the genotypes of subject j at V rare-variant sites in the gene of interest, where $G_{j,v}$ is coded as the number of copies of the minor allele that the subject possesses at variant v. We then construct the matrix of rare-variant genotypes for the sample as $\boldsymbol{G} = (\boldsymbol{G}_1^T, \boldsymbol{G}_2^T, \ldots, \boldsymbol{G}_N^T)^T$ which is of dimension N x V.

GAMuT Test of Cross-Phenotype Associations: We create GAMuT to examine the relationship between phenotypes $\boldsymbol{P}$ and rare-variant genotypes $\boldsymbol{G}$. GAMuT is

based on a KDC machine-learning technique [126-130], which allows nonparametric tests of independence between two distinct sets of multivariate variables. For each set of multivariate variables, KDC constructs an *N x N* matrix with individual elements of the matrix corresponding to similarity (or dissimilarity) in the variables among different pairs of subjects. KDC then evaluates whether the pairwise elements in the similarity matrix of one set of multivariate variables is independent of the pairwise elements in the similarity matrix for the other set of multivariate variables.

Leveraging the KDC framework, we create a rare-variant test of pleiotropy to test for independence between $\boldsymbol{P}$ (*N x L* matrix of multivariate phenotypes) and $\boldsymbol{G}$ (*N x V* matrix of multivariate rare-variant genotypes). To do this, we first develop an *N x N* phenotypic-similarity matrix $\boldsymbol{Y}$ (based on $\boldsymbol{P}$) and an *N x N* genotypic-similarity matrix $\boldsymbol{X}$ (based on $\boldsymbol{G}$). The choice of how to model pairwise similarity or dissimilarity for a set of multivariate outcomes is quite flexible. For example, for phenotypes $\boldsymbol{P}$, we can model the matrix $\boldsymbol{Y}$ using a projection matrix [133; 134], such that $\boldsymbol{Y} = \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T$. We can also construct the model $\boldsymbol{Y}$ using user-selected kernel functions [97; 98; 105; 131]. Denote the kernel function $y(\boldsymbol{P}_i, \boldsymbol{P}_j)$ as the measure of similarity between subjects *i* and *j* across the *L* phenotypes. We can model $y(\boldsymbol{P}_i, \boldsymbol{P}_j)$ using kernel similarity functions like the linear kernel, $y(\boldsymbol{P}_i, \boldsymbol{P}_j) = \sum_{l=1}^{L} P_{i,l} P_{j,l}$; a quadratic kernel, $y(\boldsymbol{P}_i, \boldsymbol{P}_j) = (1 + \sum_{l=1}^{L} P_{i,l} P_{j,l})^2$; or a Gaussian kernel, $y(\mathbf{P}_i, \mathbf{P}_j) = \exp(-\sum_{l=1}^{L} (P_{i,l} P_{j,l})^2 / \delta)$, where δ is a tuning parameter.

For genotypes $\boldsymbol{G}$, we model the corresponding matrix $\boldsymbol{X}$ using kernel functions $x(\boldsymbol{G}_i,\boldsymbol{G}_j)$ that can take the same form (e.g. linear, quadratic, or Gaussian) used to construct $y(\boldsymbol{P}_i,\boldsymbol{P}_j)$. A few genetic-specific kernel functions also exist, like the identity-by-state (IBS) kernel, $x(\boldsymbol{G}_i,\boldsymbol{G}_j)=\sum_{v=1}^{V}\mathrm{IBS}(G_{i,v}G_{j,v})/2V$, where $\mathrm{IBS}(G_{i,v}G_{j,v})$ denotes the number of alleles (0,1, or 2) shared IBS by subjects *i* and *j* at variant *v*. Also, we may wish to further augment $x(\boldsymbol{G}_i,\boldsymbol{G}_j)$ to preferentially upweight the contributions of particular rare variants in $\boldsymbol{G}$ over others in the gene. For example, we may wish to give more weight to variants that are more rare in the population or to variants that are predicted to be deleterious in nature[135-137]. We can do this by creating a diagonal weight matrix $\boldsymbol{W}$= diag($w_1, w_2, \ldots, w_V$), where w_v, reflects the relative weight for the v^{th} variant in the gene. Using $\boldsymbol{W}$, we can then create a weighted linear kernel function as $\boldsymbol{X}=\boldsymbol{GWG}^T$. Derivation of other weighted kernel functions is straightforward.

Once we construct the similarity matrixes $\boldsymbol{Y}$ and $\boldsymbol{X}$, we derive our GAMuT approach as a test of independence between the elements of these two matrices. We first center each matrix as $\boldsymbol{Y_c}=\boldsymbol{HYH}$ and $\boldsymbol{X_c}=\boldsymbol{HXH}$. Here, $\boldsymbol{H} = (\boldsymbol{I} - 1_N 1_N^T / N)$ is a centering matrix with property $\boldsymbol{HH}=\boldsymbol{H}$, $\boldsymbol{I}$ is an identity matrix of dimension *N*, and 1_N is an *Nx1* vector with each element equal to 1. Using $\boldsymbol{Y_c}$ and $\boldsymbol{X_c}$, we construct our GAMuT test of independence of the two matrices as

$$T_{GAMuT}=\frac{1}{\mathrm{N}}\mathrm{trace}(\boldsymbol{Y}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}) \quad (1)$$

Under the null hypothesis where the two matrices are independent, T_{GAMuT} follows the same asymptotic distribution as

$$\frac{1}{N^2}\sum_{i,j}\lambda_{X,i}\lambda_{Y,i}z_{ij}^2 \quad (2)$$

where $\lambda_{X,i}$ is the ith ordered non-zero eigenvalue of $\mathbf{X_c}$, $\lambda_{Y,j}$ is the j^{th} ordered non-zero eigenvalue of $\mathbf{Y_c}$, and z_{ij}^2 are independent and identically-distributed χ_1^2 variables [130]. Given *L* phenotypes and *V* rare-variant sites, and further assuming sample size *N* is larger than both *L* and *V*, the maximum number of elements in the summation will be *L*V*.

Based on the KDC literature, we could derive the *P*-value of the GAMuT test approximately using a gamma distribution [126], or instead using permutation techniques [128; 130]. In our experience, the gamma approximation is accurate for *P*-values as small as 0.01 but becomes less accurate in the more extreme tails of the distribution (results not shown). Given large-scale genetic studies require *P*-values much smaller than 0.01 to declare significance in the presence of multiple testing, the gamma approximation is not suitable in this setting. The derivation of *P*-values using permutations is a valid alternative, but computationally demanding and difficult to scale to genome wide analyses. Consequently, we instead derive *P*-values for GAMuT using Davies' exact method [110], which is a computationally efficient method to provide accurate *P*-values in the extreme tails of tests that follow mixtures of chi-square variables [131]. An implementation of Davies' method is available in the R package *CompQuadForm* [138].

Relationship of GAMuT to Other Multivariate Association Tests: While the form of the GAMuT test is quite general, we note that specific choices of $\boldsymbol{Y}$ and $\boldsymbol{X}$ can lead to test statistics that have similar forms to other multivariate association tests previously published in the literature. If we assume a projection matrix $\boldsymbol{Y}$ for the phenotypes (with each phenotype mean centered prior to analysis) and assume $\boldsymbol{X}$ is the Gower distance (or some other measure of genetic dissimilarity as opposed to similarity), the GAMuT test has a form similar to the numerator of existing multivariate distance matrix regression (MDMR) tests [133; 134; 139]. We note however that MDMR procedures typically require permutations for inference whereas we can derive analytic p-values of GAMuT directly using Davies' method.

In addition to MDMR, we also note that applying GAMuT using a linear kernel to model the phenotype similarity matrix $\boldsymbol{Y}$ and to further model the genotype similarity matrix $\boldsymbol{X}$ results in a test that becomes a rare-variant version of the multivariate kernel-machine test of Maity et al. [67; 127] created for the analysis of common variants. The approach of Maity, however, required perturbations to calculate p-values of individual tests where again GAMuT can derive p-values analytically using Davies' method.

GAMuT Testing Assuming Multiple Candidate Matrices: The GAMuT test in the previous section requires *a priori* selection of the functions used to construct the phenotypic similarity matrix $\boldsymbol{Y}$ and genotypic similarity matrix $\boldsymbol{X}$. In practice though, it is often unclear what the optimal choices for $\boldsymbol{Y}$ and $\boldsymbol{X}$ should be. For example, an investigator may want to model phenotypes $\boldsymbol{P}$ in the matrix $\boldsymbol{Y}$ using both the projection matrix and the linear kernel function. Also, an investigator may want to

construct the genotype-similarity matrix $\boldsymbol{X}$ under different kernel functions (e.g. linear and IBS) and assuming different weight functions (e.g. minor allele frequency (MAF) weights, functionality weights). If we construct GAMuT tests under multiple different phenotypic and genotypic similarity matrices, we then need to adjust for the additional tests that were performed. To adjust for additional tests, one could use a Bonferroni correction or apply permutations. However, a Bonferroni correction likely will lead to conservative inference as these tests are correlated, while permutations are computationally demanding and unappealing on a genome-wide level.

Rather than use Bonferroni or permutations, we follow the ideas of Zhang et al. [130] and Wu et al.[140] to develop a perturbation (resampling) approach to correct for testing of multiple candidate matrices in GAMuT that is more computationally efficient than standard permutations. Assume we test M different combinations of $\boldsymbol{Y}$ and $\boldsymbol{X}$. For combination m (m=1,…,M), we let $p^{(m)}$ denote the uncorrected GAMuT P-value and further let $\lambda_Y^{(m)}$ and $\lambda_X^{(m)}$ denote the vectors of all non-zero eigenvalues for $\boldsymbol{Y_c}$ and $\boldsymbol{X_c}$, respectively, for that combination. We then implement the following perturbation procedure to obtain a P-value that accounts for the testing of all M combinations.

1) Calculate the minimum observed p-value as $p^o = \min_{1 \le m \le M} p^{(m)}$.

2) For perturbation l *(l=1,…L)*, generate a set of independent χ_1^2 variables $\boldsymbol{z^*_l}$ of length equal to $L*V$.

3) For each combination *m*, calculate the test $T_l^{(m)} = \frac{1}{N^2}\sum_{i,j}\lambda_{X,i}^{(m)}\lambda_{Y,j}^{(m)}z_{ij,l}^{*}$ and obtain a new *P*-value $p_l^{(m)}$ using Davies' method.

4) Evaluate the minimum *P*-value for perturbation *l* as $p_l^{*} = \min_{1\le m\le M} p_l^{(m)}$

5) Repeat Steps 2-4 a total of *L* times and obtain $p_1^{*}, p_2^{*}, ..., p_L^{*}$

6) Derive the final *P*-value as $p = L^{-1}\sum_{l=1}^{L} I\left[p_l^{*} \le p^0\right]$

Adjusting for Covariates: Pleiotropic tests must adjust for important covariates, such as principal components of ancestry, to avoid potential confounding of results. We can control for confounders before applying GAMuT by regressing each phenotype separately on covariates of interest and then using the residuals to form the phenotypic similarity matrix $\boldsymbol{Y}$. Although residualizing binary phenotypes is not standard, studies have suggested that this procedure does not affect the validity of genetic association tests in case-control studies [141; 142]. As we describe in the Results section, the residualizing procedure provides an effective correction for confounders in the analysis of binary outcomes within our simulated datasets.

Simulations: We conducted simulations to verify that GAMuT properly preserves type I error and to assess power of GAMuT relative to competing approaches for genetic analysis of multiple phenotypes. To create genetic data for these simulations, we generated 20,000 haplotypes of 30 kb in size using COSI, a coalescent model that mimics LD pattern, local recombination rate, and population history for individuals of European descent [143]. To create multivariate phenotype

data, we assume either six or ten phenotypes for each subject generated from a multivariate normal distribution with mean vector 0 and *LxL* residual correlation matrix **Σ**. To model the residual correlation matrix, we considered scenarios of low residual correlation among phenotypes (pairwise correlation among phenotypes selected from a uniform (0,0.3) distribution), moderate residual correlation (pairwise correlation selected from a uniform (0.3,0.5) distribution), and high residual correlation (pairwise correlation selected from a uniform (0.5, 0.7) distribution). To generate binary traits, we defined phenotype measurements for the top quartile as affected ($P_{i,l}$=1), and defined 1-3rd quartile measurements as controls ($P_{i,l}$=0). We considered sample size *N* of either 1000 or 2500 subjects.

To investigate the performance of GAMuT under confounding and to assess whether the approach can successfully adjust for relevant covariates in this setting, we also simulated phenotypes under a confounding model where phenotypes were independent of genotypes, but both phenotypes and genotypes are associated with a normally-distributed covariate Z. We simulated phenotypes correlated with the covariate Z under the model $\mathbf{P} \sim MVN(0.2\mathbf{Z}, \Sigma)$, where **Z** denotes the *N* x 1 sample vector of covariates. To simulate correlation between rare-variant genotypes and covariate, we let 5% of the rare variants in our haplotypes be causal. We set effect size, $\beta_{Z,r}$, of each causal genetic variant *r* on **Z,** as $\beta_{Z,r} = (0.3 + N(0,0.1) * \left|\log_{10}(MAF_r)\right|$, where MAF_r is the minor allele frequency of causal variant *r.* Evaluating type I error under this model allows us to verify that our approach to controlling for confounders is valid.

We also performed type I error calculations to examine the validity of our resampling approach to adjust for multiple similarity matrices when applying GAMuT. For a given null dataset, we applied GAMuT using three combinations of phenotype similarity matrices $\boldsymbol{Y}$ and genotype similarity matrices $\boldsymbol{X}$:

1) Model phenotypes using a projection matrix, model genotypes using a weighted linear kernel
2) Model phenotypes using a linear kernel, model genotypes using a weighted linear kernel
3) Model phenotypes using a projection matrix, model genotypes using an unweighted linear kernel

We then implement the perturbation procedure described above to obtain a *P*-value accounting for testing the three combinations of similarity matrices. For both continuous and binary null simulations, we applied GAMuT to 10,000 simulated datasets.

For power models, we simulated data sets in which 5% of the rare variants in our haplotypes were modeled as causal. We set effect size of each causal variant, *r,* for phenotype *l*, $\beta_{r,l}$, as $\beta_{r,l} = (0.4 + N(0,0.1)) * \left| log_{10}(MAF_r) \right|$. This formulation sets mean effect size of causal variant *r* as inversely proportional to its MAF, such that very rare variants have on average a larger effect size than less rare variants. Allowing $\beta_{r,l}$ to vary around a normal distribution maintains the relationship between MAF and effect size, while allowing the variant to have a slightly different effect size for each phenotype.

We performed power simulations both in situations where there was no pleiotropy (i.e. only 1 of the 6 or 10 phenotypes were associated with the rare causal variants) and also when there was pleiotropy. Under pleiotropy, we varied the number of phenotypes associated with the rare variants, such that not all of the tested phenotypes will be dependent on the gene of interest. Under models assessing ten phenotypes, we consider situations where one, two, four, six, or eight phenotypes are actually associated with the gene. Under models assessing six phenotypes, we consider situations where only one, three, or five phenotypes are associated. We control correlation among phenotypes through consideration of the relative variance of phenotype explained by the *R* causal variants. We define this relative variance for phenotype *l* as $h_l = \sum_{r=1}^{R} \beta_{r,l}^2 * 2MAF_r(1\text{-}MAF_r)$. As in Galesloot et al. [61], we define the overall correlation between phenotypes *l* and *l'* as $E_{l,l'} = \sqrt{1-h_l}\sqrt{1-h_{l'}} * \boldsymbol{\Sigma}_{l,l'}$ where $\boldsymbol{\Sigma}_{l,l'}$ is (l,l') element of the *LxL* residual phenotypic correlation matrix. This allows the residual correlation structure among phenotypes to stay at the defined values.

For demonstration purposes, we also estimated power for limited simulations where we considered multiple combinations of phenotypic/genotypic similarity matrices for analyses. For such simulations, we considered a weighted linear kernel to form $\boldsymbol{X}$, and either the projection matrix or linear kernel to form $\boldsymbol{Y}$. We then implement the perturbation procedure described above to obtain a *P*-value accounting for the testing of the two similarity matrices.

For all simulations and analyses reported here, unless specified otherwise, we implement a weighting scheme based on the MAF of each variant that weights very rare variants more heavily than less rare variants. We selected the weighting scheme recommended by Wu et al. [131], setting $w_v = Beta(MAF_v, 1, 25)/Beta(0, 1,25)$.

We evaluate GAMuT using the simulated data, and compare our approach to competing strategies. For the analysis of continuous phenotypes, we compared GAMuT to the MFLM approach of Wang et al. [69]. Our implementation of MFLM used the B-spline basis based on Pillai-Bartlett trace, selecting the default parameters suggested by the authors for data analysis. Additionally, we compared GAMuT to a standard rare-variant association approach that ignored pleiotropy. Here, we consider the standard approach to be application of the popular SKAT [131] test, a powerful, kernel-based univariate test for sequencing data. We applied SKAT to each of the simulated phenotypes and then based inference on the minimum SKAT *P*-value across phenotypes analyzed. Since we perform SKAT testing on each of our *L* phenotypes, we must correct for multiple hypothesis testing. Although a permutation-based procedure is the gold standard for multiple test correction, it is computationally intensive and unlikely to scale to genome-wide analysis. Instead, we perform multiple testing correction using two approaches. First, we implement a simple Bonferroni correction of $\alpha_{\text{BONFERRONI}}=\alpha_{\text{e}}/L$, where α_e is the experimental-wise error rate. Unfortunately, this approach can be conservative, especially for tightly correlated phenotypes. We therefore also consider a more liberal threshold by estimating the effective number of independent tests, L_{eff}, where L_{eff} is the number of principal components necessary to explain either 98% or 90% of

phenotypic variance in *L* phenotypes [144]. We can then calculate a more liberal correction of $\alpha_{EFFECTIVE} = \alpha_e / L_{eff}$. While thresholds of 90-98% of phenotypic variance are more liberal than 99.5% threshold recommended by Gao et al.[144], we wanted to estimate the upper bounds of power to detect an effect using SKAT. Correction using the permutation approach should therefore fall somewhere between the conservative Bonferroni approach and the liberal principal component approaches.

For the analysis of binary phenotypes, we are unaware of existing methods for testing cross-phenotype effects of rare variants. Hence, we only compared GAMuT to univariate SKAT testing as described in the previous paragraph.

Analysis of GENOA Study: High body mass index (BMI), low high-density lipoprotein (HDL), and high blood pressure are interrelated conditions that increase risk of developing cardiovascular disease, stroke, kidney disease, and Type 2 diabetes. These conditions are moderately heritable. The heritability of BMI has been estimated to be between 17%[15] and 34%[145] depending on methods used for the estimation. Similarly, heritability of HDL is estimated at 40-48%[145; 146], while the estimates of heritability of blood pressure range from 30%[145] to 48-67%[147]. Understanding genetic factors underlying these conditions is of considerable clinical importance. Several GWAS, including pleiotropic analyses of common variants, have been performed on one or more of the conditions [148-154]. These studies have been tremendously successful in identification of common genetic variants; however, much of the genetic underpinnings of the conditions remains unexplained[155].

The GENOA study[132; 156] seeks to identify genetic variants that influence risk for hypertension and arteriosclerotic complications of hypertension. The GENOA resources include a cohort of African American sibships from Jackson, Mississippi. In the initial phase of the GENOA study, all members of sibships containing ≥2 individuals diagnosed with hypertension prior to age 60 were invited to participate, including both hypertensive and normotensive siblings. GENOA investigators collected extensive phenotypic information on each participant, including BMI, HDL, systolic blood pressure (SBP) and diastolic blood pressure (DBP). We selected these continuous measures for analysis. Additionally, GENOA investigators genotyped 1,429 subjects on the Illumina HumanExome Beadchip. We used the HumanExome-12 support files provided by Illumina to identify 48,712 non-singleton, rare or less-common autosomal genetic variants (MAF<3%; hereafter referred to as "rare-variant") that fell within known genes. We further excluded genes with fewer than 5 rare-variant sites within the GENOA dataset, leaving 3,277 genes in our analysis. Although GENOA collects data on sibs, GAMuT assumes study subjects are unrelated. Therefore, we randomly selected one sibling from each family for inclusion in our analysis.

We performed standard data cleaning, removed subjects who did not fast for at least 10 hours prior to phenotype collection, and removed related subjects that were identified as relatives either via pedigree information or identified as first-degree cryptic relatives identified using the program RELPAIR [157]. The final sample for analysis consisted of 539 unrelated subjects with measures of all four phenotypes. For each of the study participants, we also obtained gender, age, and

smoking status (ever smoked at least 100 cigarettes), use of anti-hypertension or lipid-lowering medication, and calculated the top ten genetic principal components using ancestry informative markers included on the Illumina array. We applied GAMuT using both a projection matrix and a linear kernel to measure pairwise phenotypic similarity. We also ran univariate SKAT on each of the four phenotypes and adjusted for multiple testing. For all GAMuT and SKAT tests, we used a weighted linear kernel (selecting the weighting scheme recommended by Wu et al. [131], described above, as we used in our simulation work) to measure pairwise genotypic similarity. Additionally, we applied MFLM to the GENOA dataset.

RESULTS

Type-I Error Simulations: Figure 1 shows the quantile-quantile (QQ) plots based on application of GAMuT to null datasets consisting of 1000 subjects assayed for ten phenotypes. We present QQ plots both for binary and continuous phenotypes assuming low, moderate, or high residual phenotypic correlation. We provide additional QQ plots of the GAMuT test for other combinations of phenotypes considered and sample size in Supplementary Figures S1-S3. For all models tested, GAMuT properly controls for type I error, even at the extreme tails of the test. We further investigated the type I error of GAMuT in the presence of confounding due to a continuous covariate (see Methods section) where we adjusted for confounding by residualizing the phenotypes on the covariate prior to analysis. Our QQ plots in Supplementary Figure S4 show that this residualization effectively controls for the

confounding for both binary and continuous phenotypes that, unadjusted, would yield inflated results.

Table 1 shows type I error at $\alpha \geq 0.001$ of GAMuT, MFLM, and univariate SKAT analyses of ten phenotypes for N=1000 and N=2500, while Supplementary Table 1 shows similar results when analyzing six phenotypes. As expected based on the QQ plots in Figure 1 and Supplementary Figures S1-S3, the GAMuT approach maintains appropriate type I error across a range of assumptions and significance thresholds. Meanwhile, we observed appropriate type I error rates of the MFLM as well as SKAT tests after multiple-testing correction. The difference in type I error between the three SKAT approaches was minor, particularly at smaller significance thresholds. This finding is consistent with previous publications[144; 158], particularly given the small number of tests performed (either six or ten phenotypes).

Figure 2 shows GAMuT QQ plots for binary and continuous phenotypes where we adjusted for multiple candidate matrices (see Methods section). The perturbation procedure properly accounts for testing three combinations of $\boldsymbol{Y}$ and $\boldsymbol{X}$, and properly controls for false positive rate for a range of assumptions. By contrast, as we show in Supplementary Figures S5 (binary outcomes) and S6 (continuous outcomes), using the minimum *P*-value of GAMuT across matrices tested (i.e. without multiple-testing correction) yields inflated results, while the Bonferroni correction yields deflated results.

Power Simulations: Next we compared the power of GAMuT with MFLM for continuous traits and univariate SKAT analysis (using three different multiple-testing corrections) for both continuous traits and binary traits. For these power

simulations, we set sample size to 1000. Power was estimated as the proportion of *P*-values < 2.5×10^{-6} (reflecting a genome-wide correction for 20,000 genes) and was evaluated based on 500 replicates of the data per model. Figure 3 shows the power results when we analyze continuous phenotypes. We plot power as a function of the number of phenotypes associated with the causal variants. The figure clearly shows that GAMuT outperforms both MFLM and the standard univariate SKAT approach for all models considered. The difference in power between the three SKAT approaches was negligible; therefore, we show only 90% cutoff to determine the effective number of independent tests, since it is the most anti-conservative correction method. As expected, GAMuT performs particularly well against SKAT and MFLM as the ratio of associated to unassociated phenotypes increases (i.e. as the gene is increasingly pleiotropic). In addition, under models of no pleiotropy where rare causal variants were only associated with 1 of the phenotypes under consideration, we observed the power of GAMuT to be approximately equal or better than SKAT.

MFLM performs poorly in all of our assumptions. We therefore simulated data that mimics the assumptions presented in the top row of Wang et al.'s Figure 4 [69]. The differing assumptions are detailed in the Supplementary Figure S7; briefly, the differences in our assumptions compared with the Wang et al. manuscript are that the latter assume smaller number of phenotypes, smaller genes, larger effect sizes, a more lenient significance threshold, and a larger percentage of causal variants. When we implement the simulation strategy of Wang et al. , we observe increases in power for MLFM versus SKAT that are similar to those in their paper.

GAMuT performance is approximately equivalent to MLFM under the simulation assumptions of Wang et al.

Figure 4 shows similar results when binary phenotypes are modeled. Since MFLM is only valid for continuous outcomes, for binary outcomes we compare GAMuT only to univariate SKAT. We observed similar improvements of power for GAMuT compared to SKAT in our binary simulations as we did for our continuous simulations. Under pleiotropic models, the improvement in power of GAMuT over SKAT grows more noticeable as the number of phenotypes associated with the gene increases. At the same time, even under power models where there is no pleiotropy (only one phenotype associated with the rare variants), our results indicate GAMuT is at least as powerful compared with the univariate SKAT approaches under models assuming low correlation, and in fact is more powerful than the univariate approach under moderate and high correlation structure.

We also implemented the perturbation approach to model phenotypic similarity using both the projection matrix and the linear kernel. For both cases, we used the weighted linear kernel to model genotypic similarity. In Figure 5 we compare power of GAMuT using the projection matrix against power when two candidate matrices are considered (projection and linear kernel), implementing the perturbation procedure to account for testing two combinations of $\boldsymbol{Y}$. Power in Figure 5 is defined as the proportion of *P*-values less than 1.5×10^{-5}, to reflect the study-wide significance threshold we will use for the GENOA data. We also show power using the linear kernel to model phenotypic similarity. While the linear kernel was not as powerful as the projection matrix on our simulated data,

simulations indicate that the perturbation procedure retains much of the power of the optimal kernel approach.

Application to GENOA Dataset: We use the GENOA dataset to test for associations between BMI, HDL, SBP, and DBP, and rare variants in 3277 genes. Prior to analysis using GAMuT, we controlled for gender, age, smoking status, use of anti-hypertension medication, use of lipid lowering medication, and ancestry on the 539 unrelated subjects. After adjusting for covariates, correlation of the four phenotypes was low to moderate with the largest pairwise correlation (0.67, Pearson's product-moment correlation *P*-value < 2.2×10^{-16}) between SBP and DBP (see Table 2). We applied GAMuT using both a projection matrix and a linear kernel to measure pairwise phenotypic similarity. For comparison, we ran MFLM as well as univariate SKAT on each of the four phenotypes and adjusted for multiple testing. For all GAMuT and SKAT tests, we used a weighted linear kernel to measure pairwise genotypic similarity. We set a stringent study-wise significance threshold of 1.5×10^{-5}, which corresponds to a Bonferroni correction based on the number of genes tested (3277): $\alpha_{\text{BONFERRONI}} = 0.05/3277$. We considered *P*-values less than $P < 1 \times 10^{-3}$ as suggestive.

Figure 6 provides genome-wide results using GAMuT and univariate SKAT analyses with top findings highlighted in Table 3. None of the methods identified any genes associated at the study-wide significance threshold. Using the linear kernel, GAMuT identified five genes of suggestive significance. Of note, the *Selectin P* (*SELP)* gene, which was identified as suggestive significance by GAMuT ($P = 1.9 \times 10^{-4}$), has previously been associated with traits related to the four GENOA phenotypes.

Haplotypes or common polymorphisms in *SELP* have been associated with myocardial infarction [159; 160] and thromboembolic stroke [161]. Expression levels of P-selectin, the protein encoded by the *SELP* gene, is increased in hypercholesterolemic patients [162] and patients with unstable angina [163]. P-selectin expression was significantly associated with carotid artery stiffness and wall thickness among Japanese individuals with type II diabetes, hypertension, or hyperlipidemia[164]. The same study found that percentage of P-selectin-positive platelets was positively associated with BMI, SBP, DBP, and inversely associated with HDL.

The projection matrix form of GAMuT identified four genes of suggestive significance. *P*-values from the two forms of GAMuT were strongly correlated (Pearson correlation = 0.90). After accounting for confounders, GAMuT did not demonstrate any systematic inflation across the genome (see QQ plots in Figure 6).

In order to correct for using two phenotypic similarity matrices for GAMuT, we performed the perturbation approach described in the methods section on the eight genes with *P*-values of less $1x10^{-3}$ for either GAMuT or SKAT. The *P*-values obtained through combined perturbation method are also shown in Table 3. Of the eight genes identified as suggestive by either or both of the GAMuT approaches, five remained suggestive after correcting for use of two GAMuT similarity matrices.

The SKAT *P*-values using the three multiple testing correction methods were identical across all genes tested. SKAT did not identify any genes at genome-wide significance. It identified four genes at the suggestive significance threshold, all of which were identified by one or both of the GAMuT tests. When we applied MFLM to the GENOA data, we observed sizeable inflation of the *P*-values. The *P*-value inflation

was not resolved by inverse-normal transforming the phenotypes, as performed in the Wang et al. [69]. See Supplementary Figure S8 for QQ plots of the untransformed and transformed analyses.

Running the GAMuT analyses on a single-threaded R script on an Intel i7-2720QM CPU took 22.3 minutes using either the linear kernel or the projection matrix to model phenotypic similarity. Implementing the perturbation approach ($1x10^6$ replicates per gene) required approximately 44.5 minutes of computing time per gene analyzed.

DISCUSSION

Some patterns in the genetic basis of complex traits have emerged in prior studies. First, common variants of relatively small individual effect located throughout the genome collectively explain a large fraction of the total genetic variance [14; 15; 22; 24; 40; 71; 165-169]. Second, for some disorders such as autism [170; 171], more than a thousand genes appear capable of harboring exceedingly rare, large effect mutations. While it is still unclear whether these two patterns are ubiquitous, they are central predictions of the infinitesimal model of allele effects. Moreover, we know from detailed theoretical analysis [5] that if the infinitesimal model is true for most phenotypes, then most rare large-effect mutations should be highly pleiotropic.

We have presented GAMuT, a framework for cross-phenotype analysis of rare variants using a nonparametric distance-covariance approach [126; 127; 130]. This approach can accommodate both binary and continuous phenotypes, and can adjust

for covariates. The GAMuT test derives analytic *P*-values based on Davies' exact method, thereby improving computational efficiency and permitting application on a genome-wide scale. Like the popular SKAT framework for univariate rare variant analysis, our approach allows for inclusion of prior information, such as biological plausibility of the variants under study, and further remains powerful when a gene harbors a mixture of rare causal variants that act in different directions on phenotype. Our approach demonstrates greater power than SKAT and MFLM when pleiotropy exists. Further, simulations indicate that even if only one phenotype is associated with the gene of interest (i.e. no pleiotropy is occurring), GAMuT is at least as powerful as univariate SKAT analyses after multiple-testing adjustment. These results hold for both continuous and binary outcomes.

GAMuT analysis of simulated datasets comprised of 1000 subjects and 10 phenotypes takes 4.5 seconds per gene for either continuous or binary phenotypes using a R script running single-threaded on an Intel i7-2720QM CPU processor. This run-time includes the formation of both the phenotype and genotype similarity matrices, as well as testing for independence. However, the phenotype similarity matrix only needs to be calculated once, even if multiple genes are being tested. We find that determining the phenotype similarity information requires 2.1 seconds for N=1000; completing the test for each gene requires 2.4 seconds. Increasing the number of phenotypes or rare-variants tested does not substantially increase GAMuT's run-time. However, increasing sample size does increase run time. For sample sizes of N=2500, 5000, and 10,000, GAMuT takes 33 seconds, 4.4 minutes, and 31.0 minutes per gene for both continuous and binary phenotypes. Thus, to

effectively scale rare-variant association testing of cross-phenotype analyses across the genome with larger sample sizes, analysis of each gene could be distributed to individual threads on a high-performance cluster. We provide R software implementing GAMuT on our website (see Web Resources) which can be run through software packages like PLINK, PLINK-SEQ, or EPACTS if desired.

We applied GAMuT to exome-chip data from the GENOA study to identify genes harboring rare variants with pleiotropic effects on four phenotypes: BMI, HDL levels, SBP, and DBP. Using the linear kernel to model phenotypic similarity and the weighted linear kernel to model genotypic similarity, we detected eight genes that were suggestively associated with our phenotypes. Of note, common variants and gene product levels of one such gene, *SELP,* have previously been associated with BMI, SBP, DBP, and HDL[162; 164].

GAMuT's KDC framework is amenable to several promising extensions that we will explore in future work. Since GAMuT is an omnibus test, an association of the gene with just one of the tested phenotypes (i.e. no pleiotropy) could result in a significant finding. While the result is valid, researchers will often wish to identify which underlying phenotype of those considered are directly associated with the gene of interest. Additionally if we identify a cross-phenotype association, a follow-up analysis could be to assess whether the cross-phenotype effect is due to biological pleiotropy (a causal locus directly affecting more than one trait) or mediation pleiotropy (a causal locus affecting only one trait, which in turn affects another trait). Existing mediation analyses are not intended to handle high-dimensional traits; we propose the creation of KDC procedures to identify whether

an observed cross-phenotype association is mediated by a different set of phenotypes. Additionally, we could also perform post-hoc GAMuT of different subgroupings of the phenotypes to identify the true phenotypes associated with the gene and adjust for multiple testing using perturbations. We will pursue these ideas in future work.

GAMuT currently assumes unrelated subjects; however, it should be reasonably straightforward to extend GAMuT to allow for case-parent trio studies. The work by Jiang et al. [172] provides a framework for transforming genotypic data for trios into data that is amenable to a kernel-based framework. Specifically, the Jiang method uses the quantitative transmission disequilibrium test introduced by Abecasis et al. [173] to decompose observed genotypes into between-family and within-family components, and then integrates within-family genetic components into a kernel-machine regression framework. While the Jiang method uses a KMR approach and is therefore only appropriate for univariate phenotype analyses, an analogous approach, using GAMuT, should allow for high-dimensional phenotype data. Finally, one might be interested in combining cross-phenotype association results from multiple studies through a meta-analysis. GAMuT is designed to test for rare variant cross-phenotype associations in a single dataset. However, the meta-analysis approach in Lee et al. [174], which is designed to combine results of multiple KMR-based studies, should be readily extendible to KDC results, such as those obtained via GAMuT.

That pleiotropy might be ubiquitous should come as no surprise. The central organismal level result of pleiotropy will be the frequent occurrence of co-morbid

diagnoses. Neuropsychiatric disorders, for instance, are particularly laden with comorbid diagnoses. The National Institute of Mental Health (NIMH) estimates that as many as 45% of individuals diagnosed with a mental disorder meet criteria for two or more disorders [58]. Likewise, nearly 75% of adults with diabetes also have hypertension [59], and patients with rheumatoid arthritis are about twice as likely to suffer from myocardial infarction as individuals without arthritis [60]. While some of these overlapping phenotypes are ultimately due to environmental risk factors, some comorbidities are almost certainly explained by common genetic pathways. Ignoring comorbidity, or worse, setting inclusion criteria that exclude individuals suffering a comorbid diagnosis, will limit biological understanding of complex traits, and may limit our ability to detect missing heritability.

ACKNOWLEDGEMENTS

This work was supported by NIH grants HG007508, HL086694, HL119443, MH071537, and AR060893. For purposes of disclosing duality of interest, Michael Epstein is a consultant for Amnion Laboratories.

WEB RESOURCES

Epstein Software: http://www.genetics.emory.edu/labs/epstein/software

OMIM: http://www.omim.org

Table 1: Empirical Type-I Error Rates Assuming Ten Phenotypes

			α = **0.05**					α = **0.001**				
					SKAT					**SKAT**		
Sample Size	**Type of Phenotypes**	**Phenotypic Correlation**	**GAMuT**	**MFLM**	**Bonf.**	**PC: 98%**	**PC: 90%**	**GAMuT**	**MFLM**	**Bonf.**	**PC: 98%**	**PC: 90%**
1000	Continuous	Low	.0453	.0503	.0455	.0455	.0545	.0007	.0009	.0010	.0010	.0011
		Moderate	.0504	.0481	.0423	.0423	.0503	.0013	.0007	.0012	.0012	.0013
		High	.0517	.0484	.0462	.0498	.0509	.0009	.0013	.0010	.0011	.0011
	Binary	Low	.0488	-	.0447	.0447	.0481	.0006	-	.0023	.0023	.0023
		Moderate	.0537	-	.0429	.0429	.0461	.0013	-	.0028	.0028	.0029
		High	.0439	-	.0474	.0487	.0509	.0003	-	.0013	.0013	.0014
2500	Continuous	Low	.0512	.0493	.0447	.0474	.0567	.0014	.0012	.0007	.0007	.0007
		Moderate	.0538	.0506	.0402	.0416	.0547	.0012	.0008	.0010	.0010	.0012
		High	.0457	.0496	.0496	.0502	.0510	.0009	.0018	.0012	.0012	.0012
	Binary	Low	.0491	-	.0360	.0480	.0529	.0015	-	.0017	.0017	.0017
		Moderate	.0524	-	.0384	.0450	.0491	.0018	-	.0015	.0015	.0015
		High	.0450	-	.0455	.0457	.0503	.001	-	.0012	.0014	.0014

Empirical size for GAMuT, MFLM, and SKAT analyses at significance thresholds of 0.05 and 0.001. Empirical size calculated from 10,000 null simulations. Simulations assume analysis of 10 phenotypes. Sample size was set at either 1000 or 2500. Phenotypes were either continuous or dichotomous. Phenotypic correlation was low (correlation <0.3), moderate (correlation 0.3-0.5), or high (correlation 0.5-0.7).

Table 2: Correlation of GENOA Phenotypes

	BMI	**HDL**	**SBP**	**DBP**
BMI	1	**-0.17**	**0.09**	0.02
HDL	-	1	-0.01	-0.03
SBP	-	-	1	**0.67**
DBP	-	-	-	1

Correlation among the four GENOA phenotypes: body mass index (BMI), high-density lipoprotein (HDL), systolic blood pressure (SBP) and diastolic blood pressure (DBP). Bolded correlations are nominally significant (Pearson's product-moment correlation *P*-value < 0.05).

Table 3: Top GENOA Results

Gene Name	Chromosome	Number Rare Variants	GAMuT			SKAT: 90% PC
			Projection Matrix	Linear Kernel	Combined (Perturbation)	
SELP	1	8	4.8E-03	1.9E-04	2.8E-04	4.9E-04
DISP1	1	8	1.0E-04	8.1E-03	1.4E-04	7.3E-03
ARHGEF10	8	14	2.8E-02	7.9E-04	1.0E-03	6.6E-04
COL17A1	10	11	6.3E-04	1.1E-03	9.2E-04	9.0E-03
STRA6	15	7	1.1E-03	9.9E-04	1.5E-03	3.4E-03
ZNF222	19	5	8.8E-04	3.6E-03	1.4E-03	4.5E-04
COL9A3	20	5	5.6E-05	2.2E-05	2.3E-05	6.7E-04
FAM83F	22	5	3.8E-03	4.4E-04	6.6E-04	9.3E-03

We identified eight genes in the GENOA dataset with *P*-values of at least suggestive significance ($P<1x10^{-3}$) using either GAMuT or SKAT, using a 90% cutoff to determine the effective number of independent tests. For the eight genes we provide gene name, chromosomal location of gene, number of rare variants (MAF < 3%) found in each gene in the GENOA dataset, and *P*-values for the four approaches.

FIGURE LEGENDS

Figure 1: The QQ plots applying GAMuT to 10,000 simulated null data sets assuming a sample size of 1000. In each simulation, 10 phenotypes are tested. Top row assumes binary phenotypes; bottom row assumes continuous phenotypes. Left column shows low residual phenotypic correlation (correlation 0-0.3), middle column shows moderate residual correlation (correlation 0.3-0.5), and right column shows high residual correlation (correlation 0.5-0.7).

Figure 2: The QQ plots applying GAMuT to 10,000 simulated null data sets assuming a sample size of 1000. *P*-values using three candidate matrices combinations were obtained for each simulation. We then implement a perturbation procedure to obtain a *P*-value accounting for testing the three combinations of similarity matrices. In each simulation, 10 phenotypes are tested. Top row assumes binary phenotypes; bottom row assumes continuous phenotypes. Left column shows low residual phenotypic correlation (correlation 0-0.3), middle column shows moderate residual correlation (correlation 0.3-0.5), and right column shows high residual correlation (correlation 0.5-0.7).

Figure 3: Power for GAMuT (red), univariate SKAT using a 90% cutoff to determine effective number of independent tests (blue), and MFLM (green) is plotted as a function of number of continuous phenotypes associated with the gene of interest. Top row assumes six continuous phenotypes are tested in each simulation, and bottom row assumes 10 continuous phenotypes are tested. Left column shows low residual phenotypic correlation (correlation 0-0.3), middle column shows moderate residual correlation (correlation 0.3-0.5), and right column shows high residual correlation (correlation 0.5-0.7).

Figure 4: Power for GAMuT (red) and univariate SKAT using a 90% cutoff to determine effective number of independent tests (blue) is plotted as a function of number of binary phenotypes associated with the gene of interest. Top row assumes six binary phenotypes are tested in each simulation, and bottom row assumes 10 binary phenotypes are tested. Left column shows low residual phenotypic correlation (correlation 0-0.3), middle column shows moderate residual correlation (correlation 0.3-0.5), and right column shows high residual correlation (correlation 0.5-0.7).

Figure 5: Power for GAMuT assuming a projection matrix (red), GAMuT assuming a linear kernel (yellow), GAMuT assuming testing of both projection matrix and linear kernel (orange), univariate SKAT using a 90% cutoff to determine effective number of independent tests (blue), and MFLM (green). In each simulation, 10 continuous phenotypes are tested. Left column shows low residual phenotypic correlation (correlation 0-0.3), while right column shows moderate residual correlation (correlation 0.3-0.5).

Figure 6: Results of the GENOA analysis. Left column shows Manhattan and QQ plots for GAMuT using a projection matrix for phenotypes. Middle column shows Manhattan and QQ plots for GAMuT using a linear kernel for phenotypes. Right column shows Manhattan and QQ plots for SKAT, using a 90% cutoff to determine the effective number of independent tests. Horizontal blue line indicates suggestive significance threshold. Horizontal red line indicates study-wide significance.

Figure 1: GAMuT QQ plots

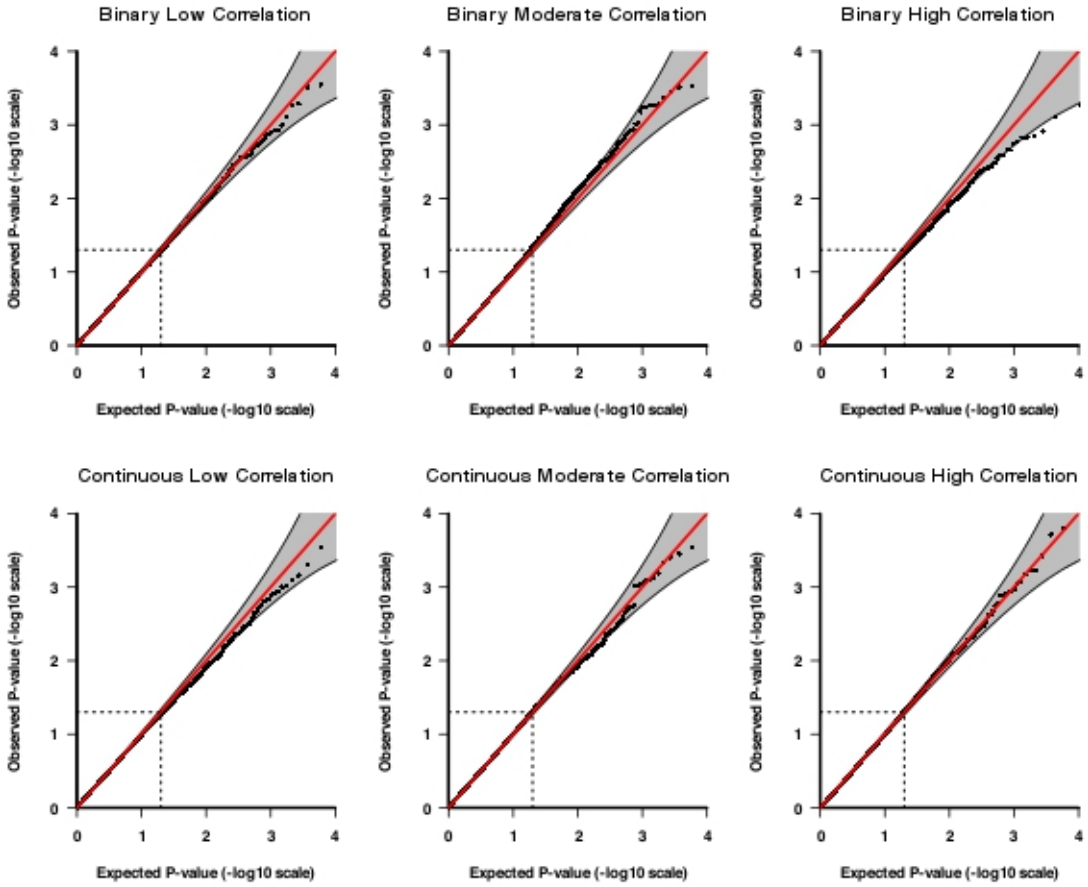


Figure 2: QQ Plots for GAMuT Assuming Multiple Matrices Tested

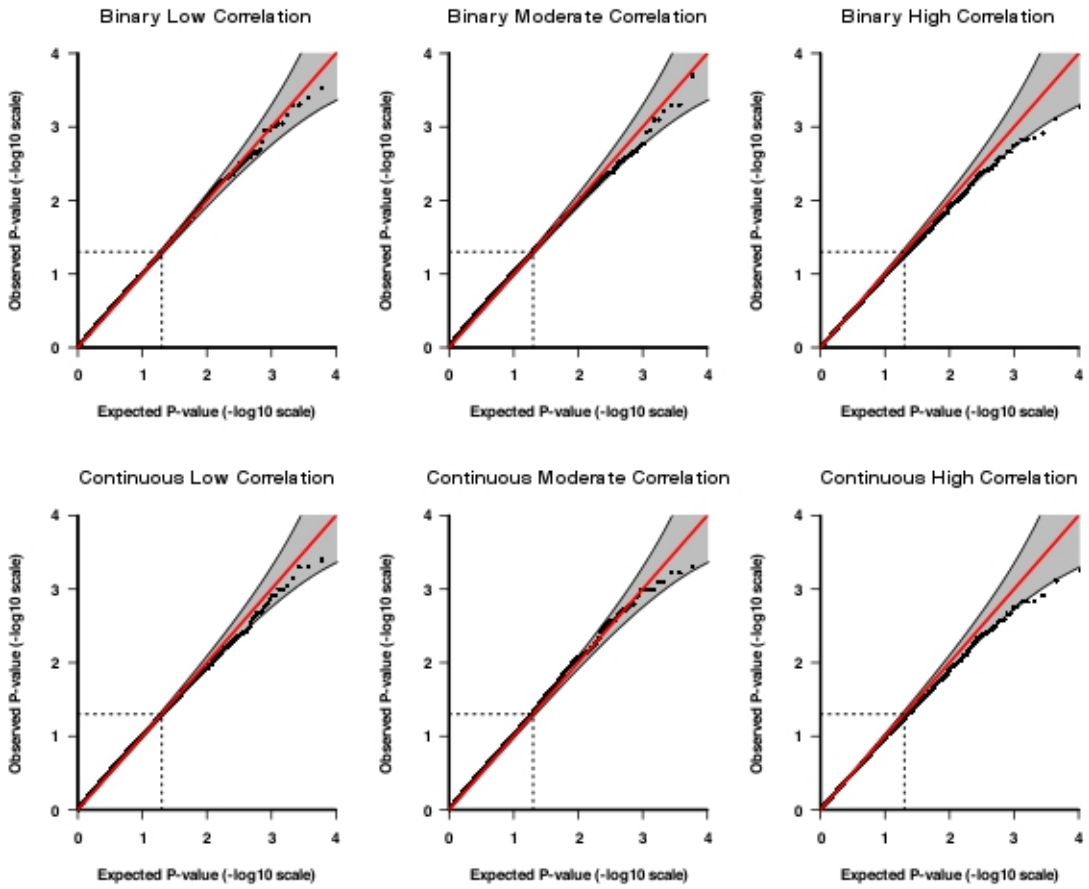


Figure 3: Power to Detect Cross-Phenotype Effects: Continuous Phenotypes

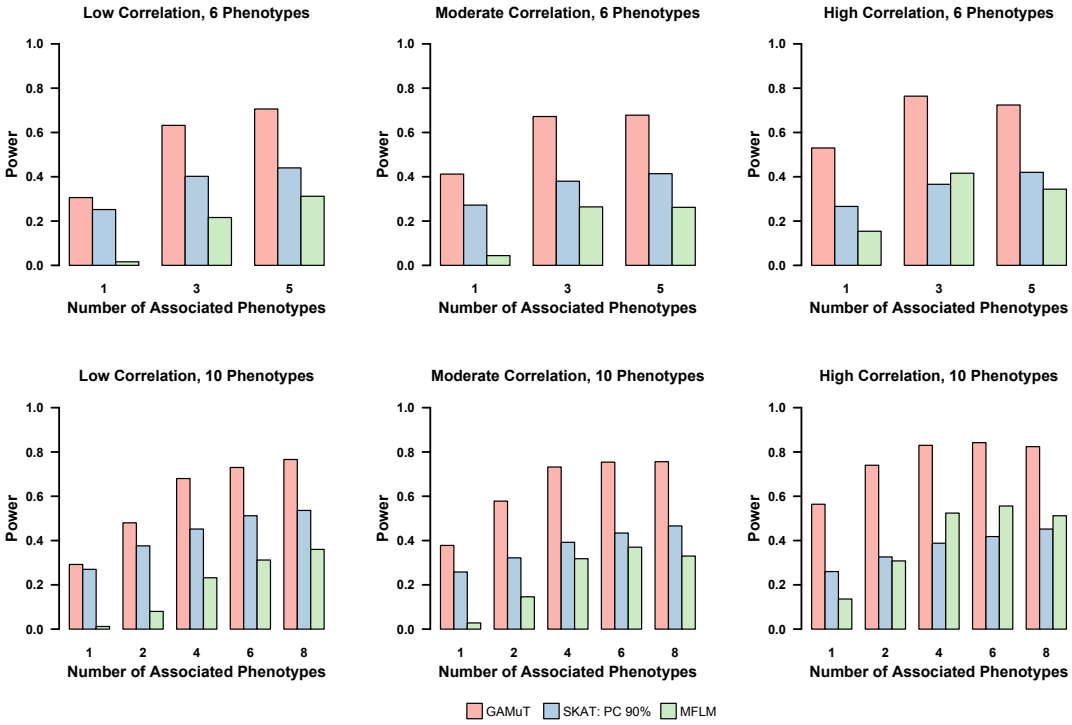


Figure 4: Power to Detect Cross-Phenotype Effects: Binary Phenotypes

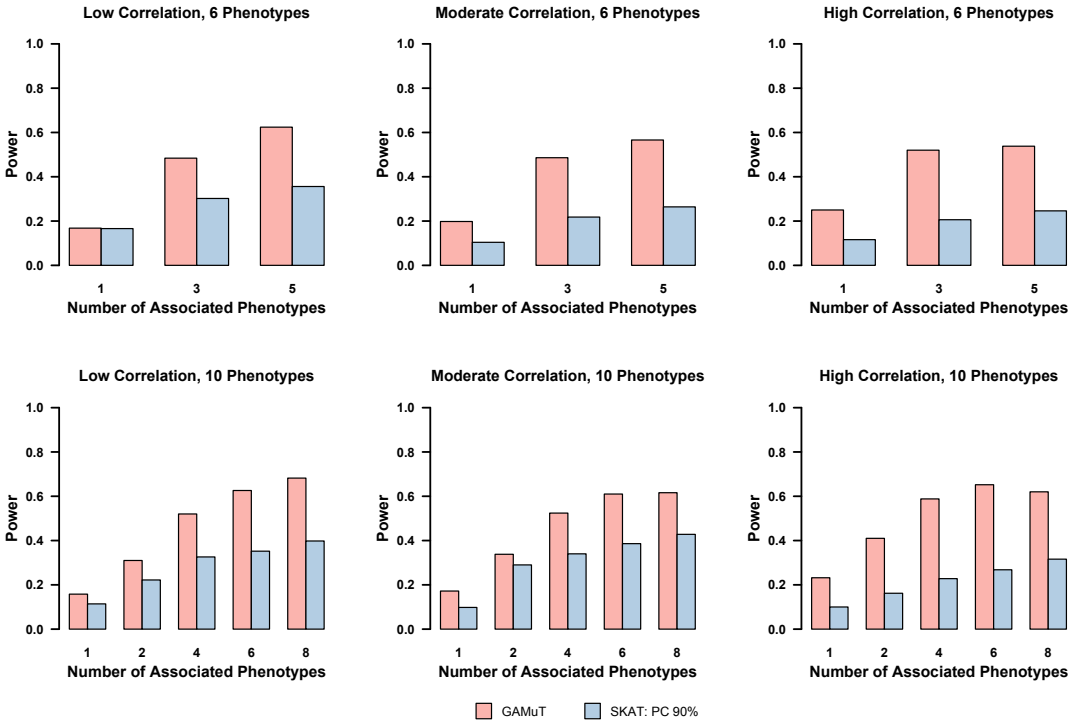


Figure 5: Power to Detect Pleiotropic Effect using Multiple Similarity Matrices

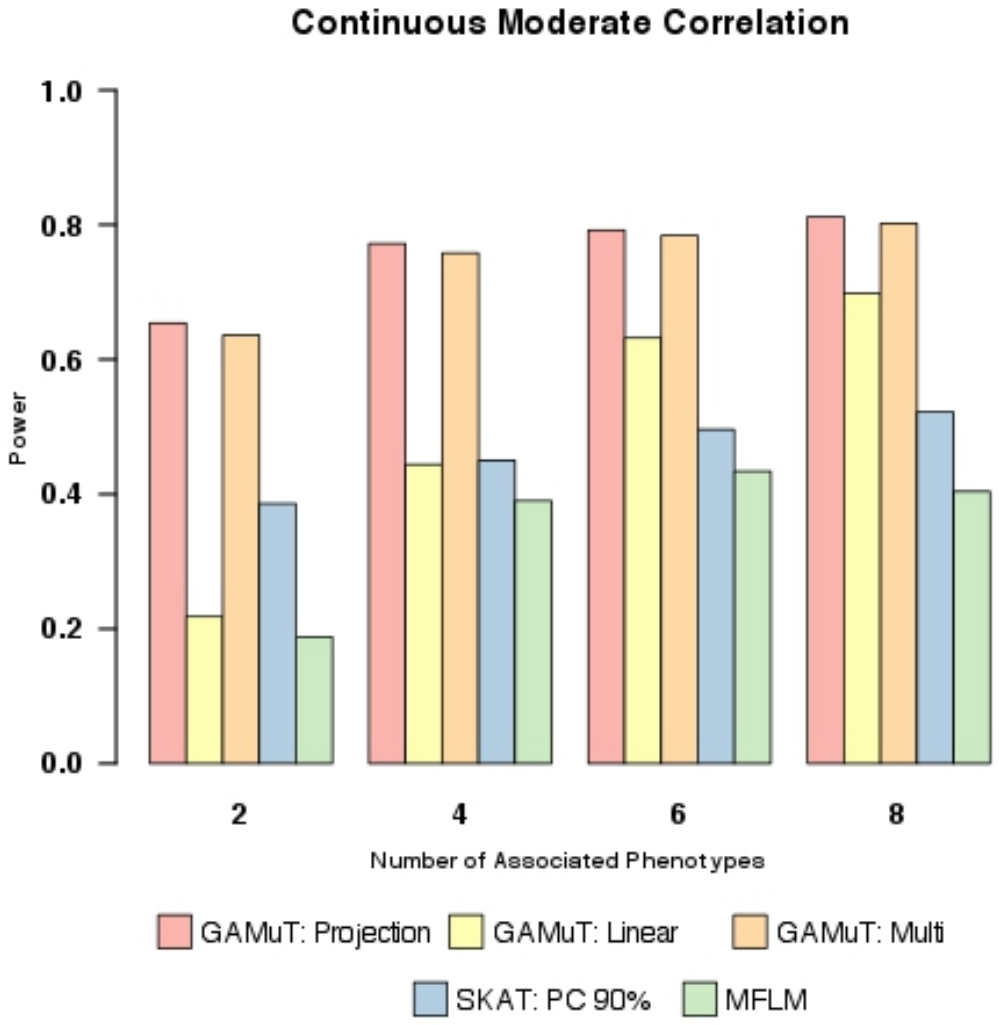
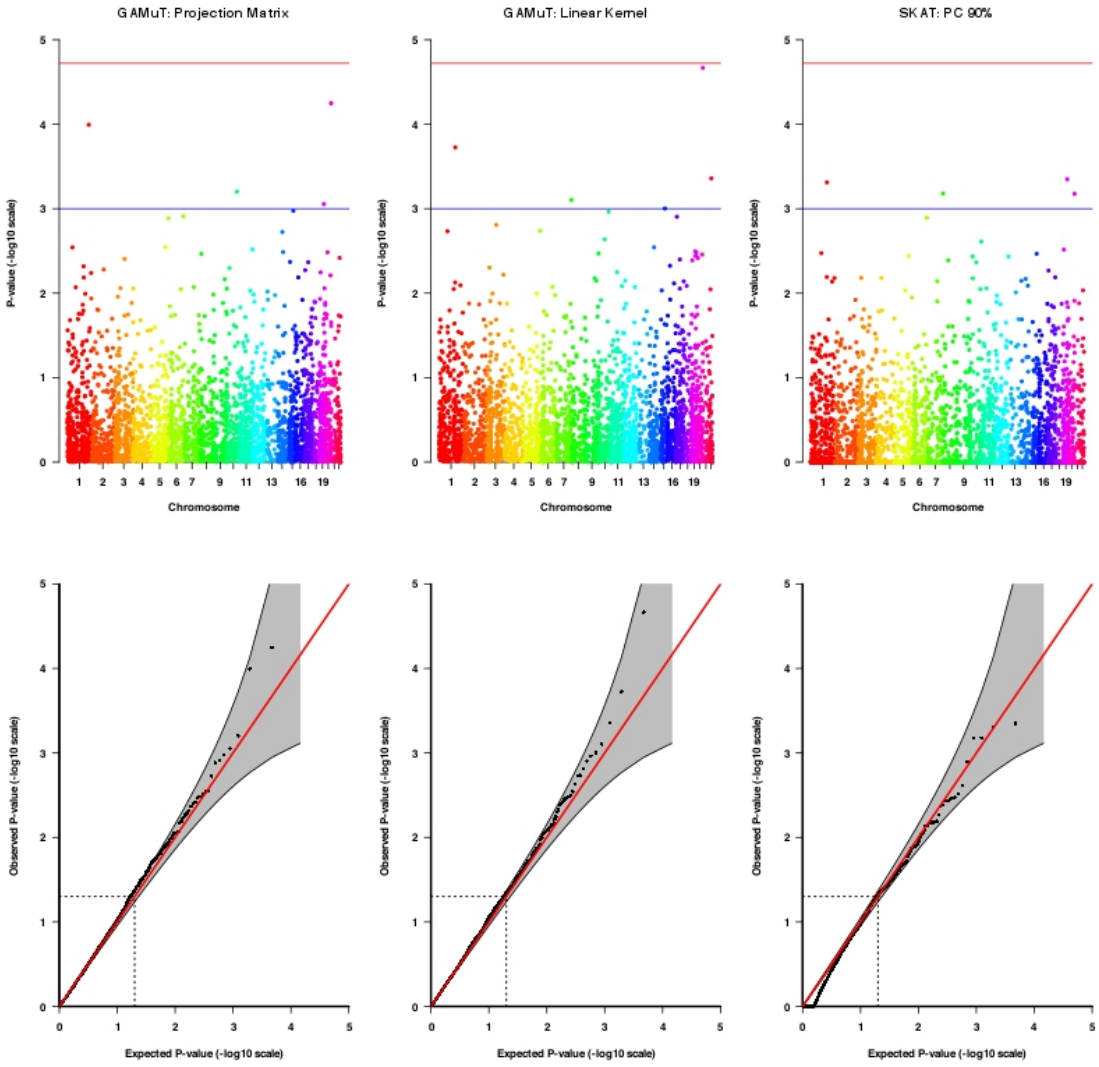


Figure 6: Results of GENOA Analyses



Supplementary Figure Legends

Supplementary Figure 1: The QQ plots applying GAMuT to 10,000 simulated null data sets assuming a sample size of 1000. In each simulation, 6 phenotypes are tested. Top row assumes binary phenotypes; bottom row assumes continuous phenotypes. Left column shows low residual phenotypic correlation (correlation 0-0.3), middle column shows moderate residual correlation (correlation 0.3-0.5), and right column shows high residual correlation (correlation 0.5-0.7).

Supplementary Figure 2: The QQ plots applying GAMuT to 10,000 simulated null data sets assuming a sample size of 2500. In each simulation, 6 phenotypes are tested. Top row assumes binary phenotypes; bottom row assumes continuous phenotypes. Left column shows low residual phenotypic correlation (correlation 0-0.3), middle column shows moderate residual correlation (correlation 0.3-0.5), and right column shows high residual correlation (correlation 0.5-0.7).

Supplementary Figure 3: The QQ plots applying GAMuT to 10,000 simulated null data sets assuming a sample size of 2500. In each simulation, 10 phenotypes are tested. Top row assumes binary phenotypes; bottom row assumes continuous phenotypes. Left column shows low residual phenotypic correlation (correlation 0-0.3), middle column shows moderate residual correlation (correlation 0.3-0.5), and right column shows high residual correlation (correlation 0.5-0.7).

Supplementary Figure 4: The QQ plots of 10,000 simulated null datasets assuming a sample size of 1000 with a confounding variable. Phenotypes are independent of genotypes, but both phenotypes and genotypes are associated with a continuous covariate. Top row shows moderately correlated binary phenotypes; bottom row shows moderately correlated continuous phenotypes. Left column shows QQ plots without adjustment for confounding; while right column shows QQ plots after adjustment for confounding by residualization.

Supplementary Figure 5: The QQ plots of 10,000 simulated null datasets assuming a sample size of 1000 and 10 binary phenotypes tested. *P*-values using three candidate matrices combinations were obtained for each simulation. Top row assumes low residual phenotypic correlation; bottom row assumes moderate correlation. We compare the perturbation procedure to account for multiple testing (left column) with GAMuT testing of multiple kernels assuming no correction, i.e. the minimum *P*-value across the three approaches (middle column), and correction using Bonferroni (right column).

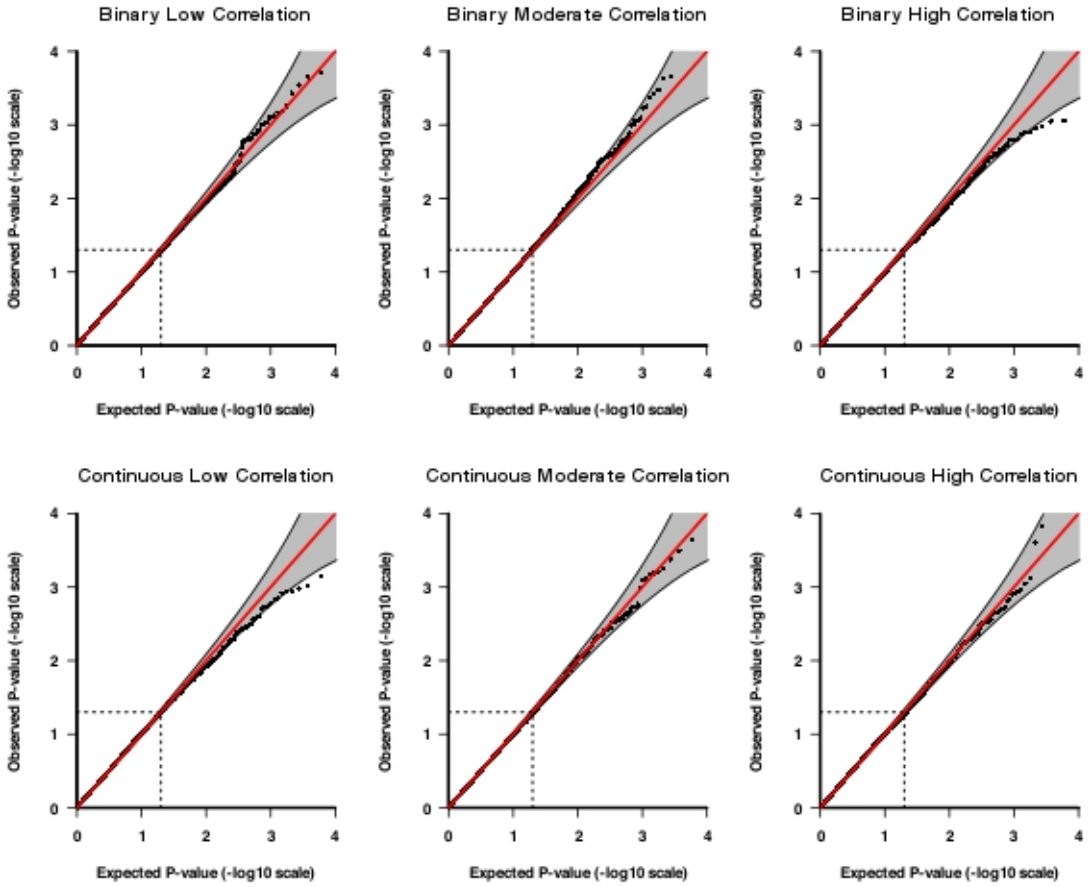
Supplementary Figure 6: The QQ plots of 10,000 simulated null datasets assuming a sample size of 1000 and 10 continuous phenotypes tested. *P*-values using three candidate matrices combinations were obtained for each simulation. Top row assumes low residual phenotypic correlation; bottom row assumes moderate

correlation. We compare the perturbation procedure to account for multiple testing (left column) with GAMuT testing of multiple kernels assuming no correction, i.e. the minimum *P*-value across the three approaches (middle column), and correction using Bonferroni (right column).

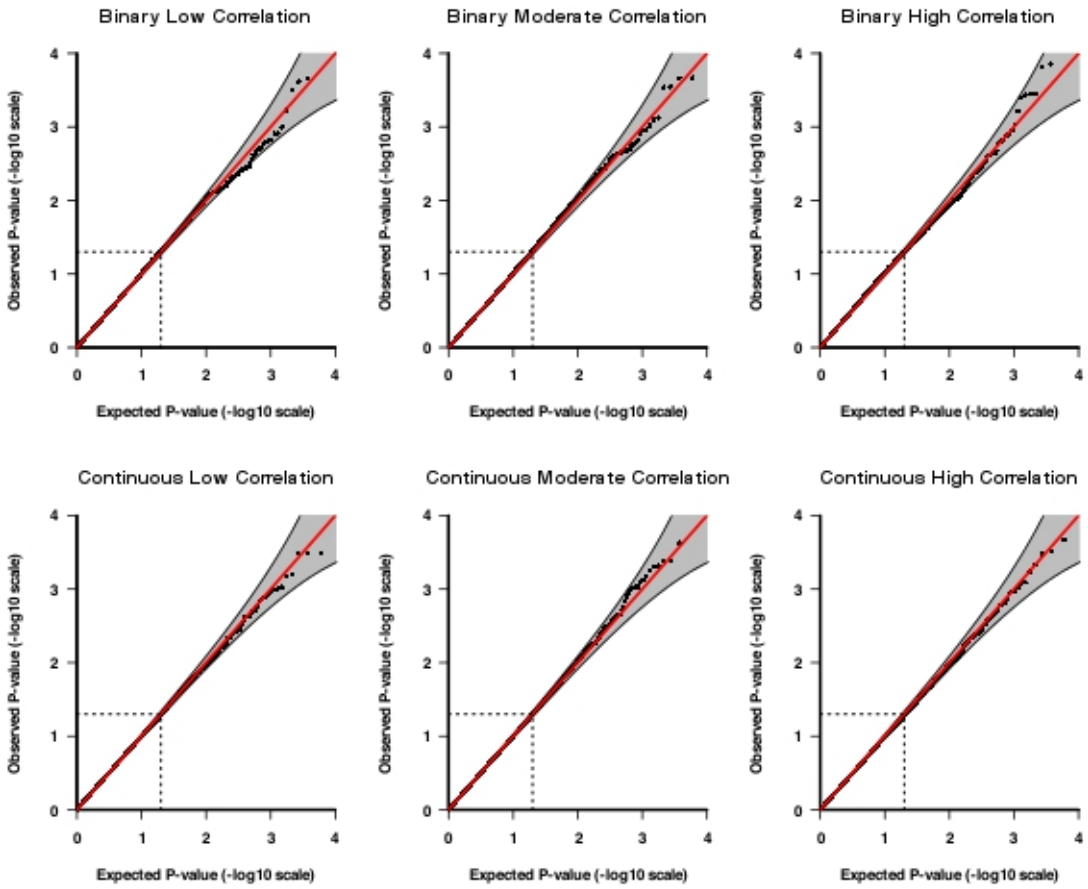
Supplementary Figure 7: Power under the assumptions similar to those presented by Wang et al (2015). Power is shown for GAMuT assuming a projection matrix (red), univariate SKAT using a 90% cutoff to determine effective number of independent tests (blue), univariate SKAT testing only the first phenotype (purple), and MFLM (green). Assumptions that were altered to replicate the Wang et al. simulations are as follows: we reduced the size of the simulated gene from 30 kb to 3 kb; we increased the genetic effect size to the parameters presented in Wang et al.; we increased the percent causal variants from 5% to 10%; we modeled only 3 continuous phenotypes with the correlation structure presented in Wang et al. (which includes negatively correlated traits); and we allowed causal variants to be protective or deleterious within the gene. Sample size varies from 500, 1000, and 1500. For all plots, causal variants had a 20%/80% positive/negative effects on trait one and 0%/100% positive/negative effects on trait 2. Left column shows results when causal variants had a 0%/100% positive/negative effects on trait three. Middle column shows results when causal variants had a 20%/80% positive/negative effects on trait three. Right column shows results when causal variants had a 50%/50% positive/negative effects on trait three.

Supplementary Figure 8: QQ plots of the GENOA analysis using MLFM. Left shows QQ plot prior to phenotype transformation. Right shows QQ plot after inverse-normal transformation of the phenotypes.

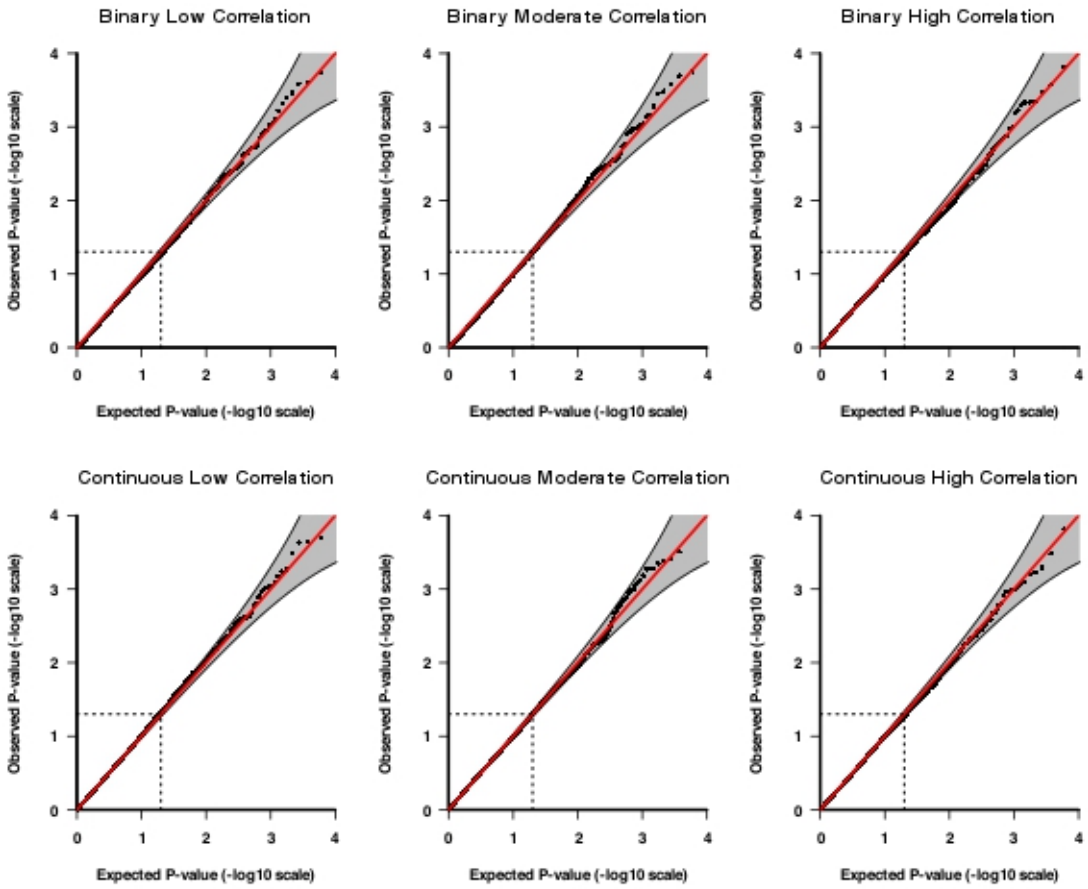
Supplementary Figure 1: GAMuT QQ plots



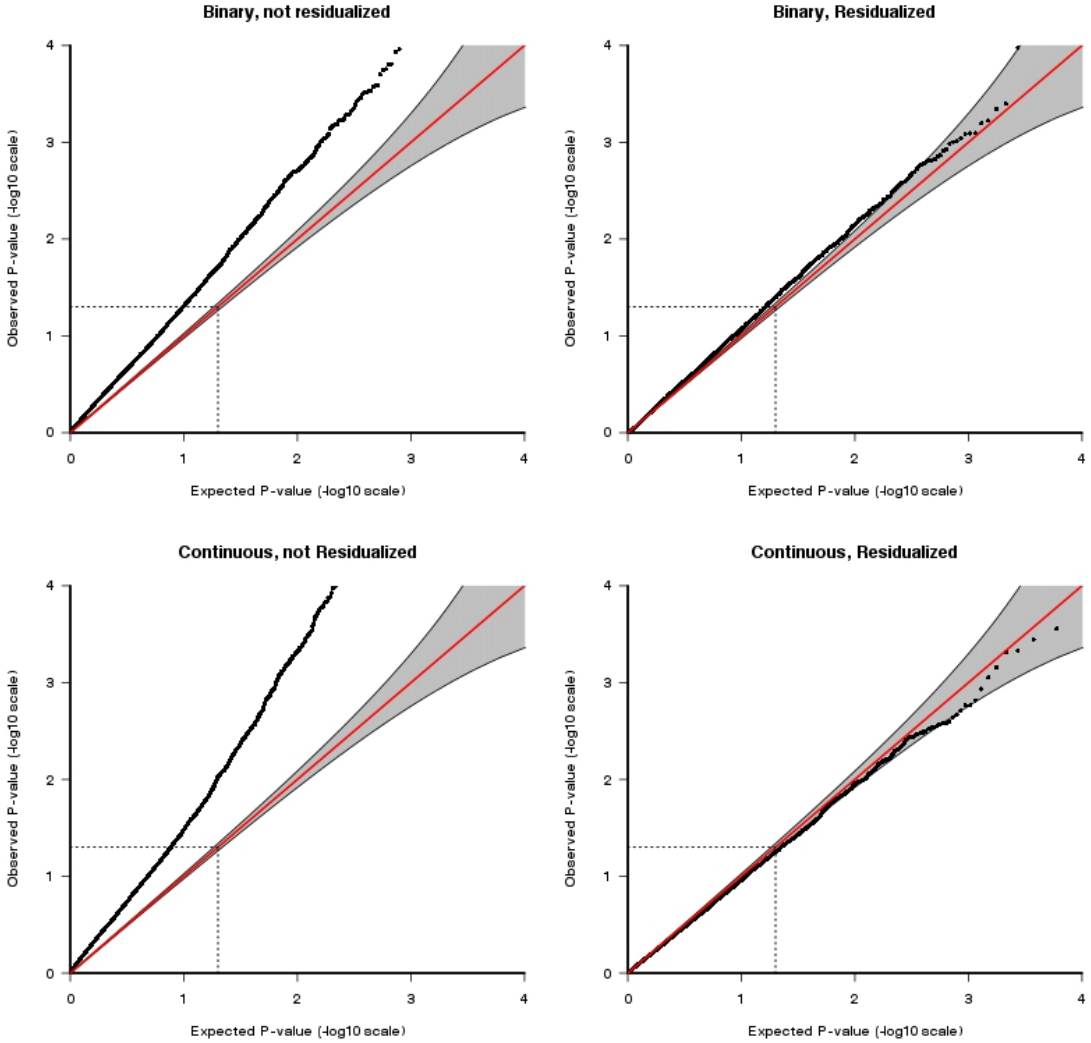
Supplementary Figure 2: GAMuT QQ plots



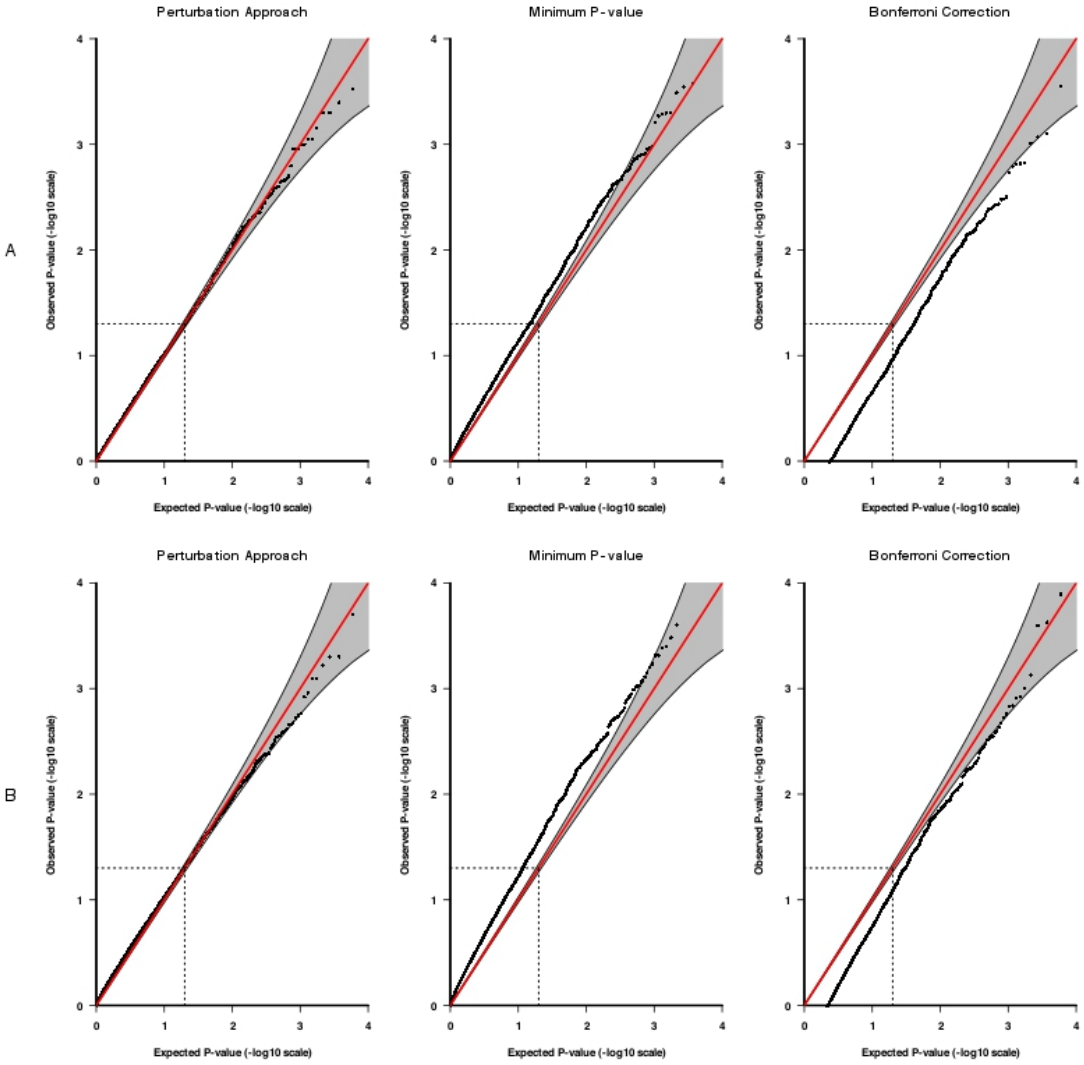
Supplementary Figure 3: GAMuT QQ plots



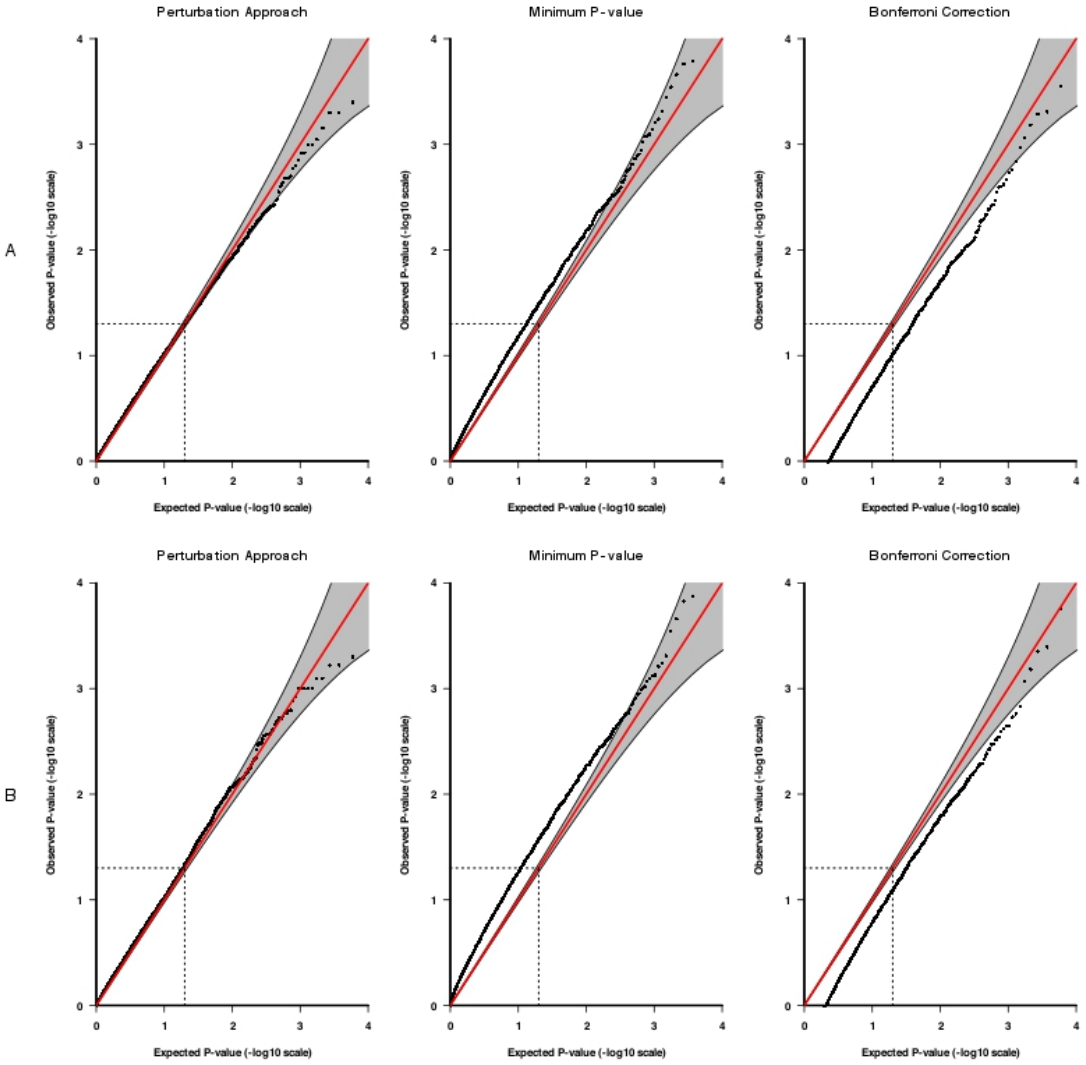
Supplementary Figure 4: QQ Plots with Confounder



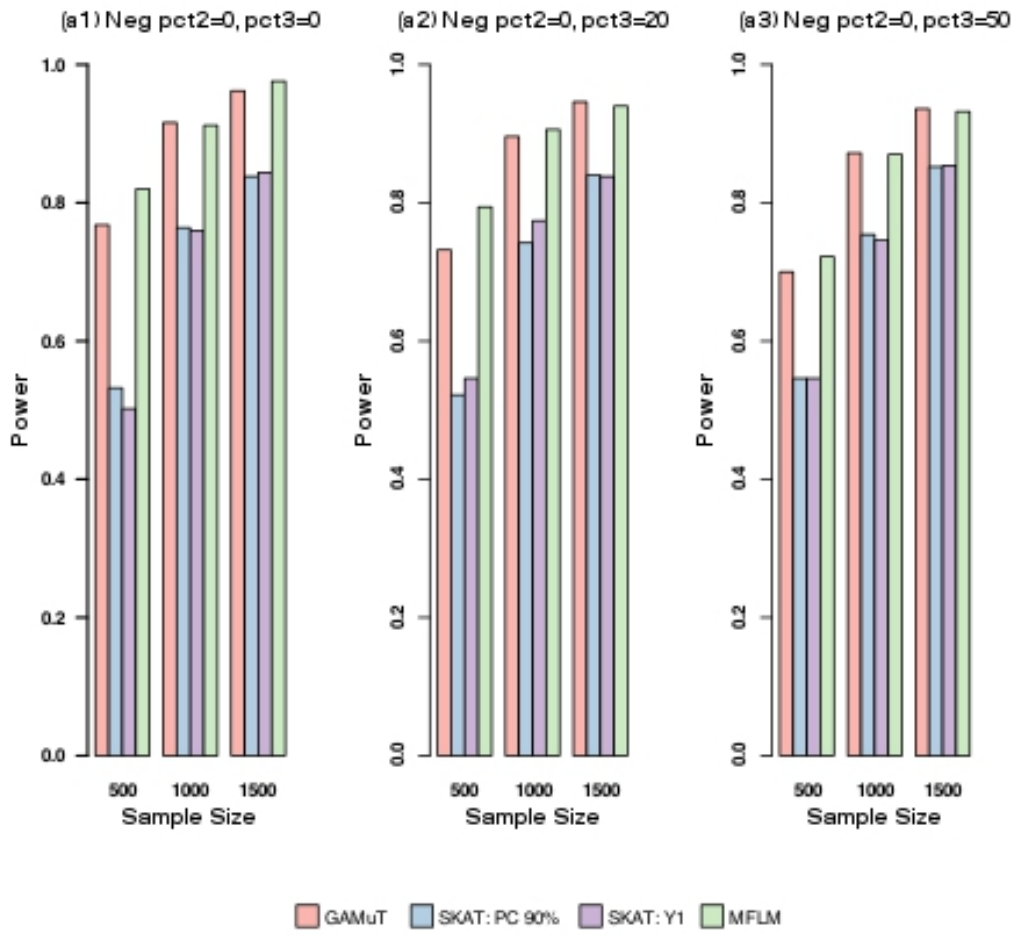
Supplementary Figure 5: QQ Plots Assuming Multiple Matrices Tested, Binary Outcomes



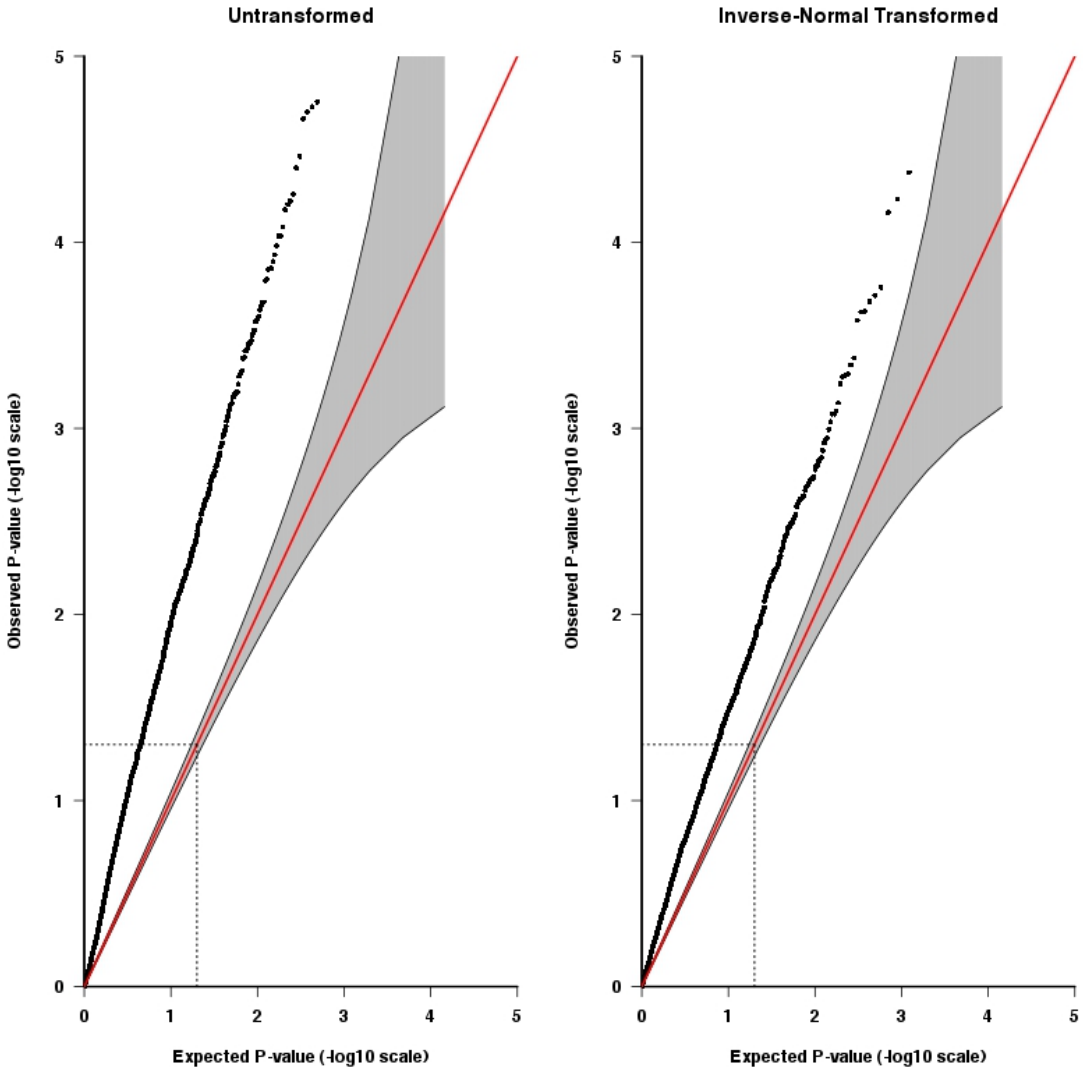
Supplementary Figure 6: QQ Plots Assuming Multiple Matrices Tested, Continuous Outcomes



Supplementary Figure 7: Power to Detect Cross-Phenotype Effects



Supplementary Figure 8: QQ Plots of GENOA Analysis using MLFM



Supplementary Table 1: Empirical Size Assuming Six Phenotypes

			α = **0.05**					α = **0.001**				
					SKAT					**SKAT**		
Sample Size	**Type of Phenotypes**	**Phenotypic Correlation**	**GAMuT**	**MFLM**	**Bonf.**	**PC: 98%**	**PC: 90%**	**GAMuT**	**MFLM**	**Bonf.**	**PC: 98%**	**PC: 90%**
1000	Continuous	Low	.0486	.0509	.0464	.0464	.0516	.0003	.0014	.0013	.0013	.0018
		Moderate	.0466	.0434	.0426	.0426	.0509	.0010	.0000	.0013	.0013	.0014
		High	.0466	.0496	.0426	.0447	.0506	.0007	.0010	.0011	.0011	.0014
	Binary	Low	.0462	-	.0413	.0413	.0473	.0014	-	.0013	.0013	.0013
		Moderate	.0520	-	.0464	.0464	.0516	.0014	-	.0013	.0013	.0018
		High	.0500	-	.0488	.0492	.0509	.0004	-	.0012	.0012	.0014
2500	Continuous	Low	.0500	.0492	.0516	.0516	.0568	.0015	.0004	.0009	.0009	.0010
		Moderate	.0485	.0494	.0462	.0462	.0547	.0008	.0009	.0008	.0008	.0009
		High	.0471	.0518	.0493	.0500	.0501	.0009	.0008	.0011	.0011	.0012
	Binary	Low	.0511	-	.0488	.0488	.0488	.0006	-	.0010	.0010	.0010
		Moderate	.0518	-	.0485	.0485	.0489	.0008	-	.0010	.0010	.0010
		High	.0508	-	.0493	.0496	.0510	.0010	-	.0013	.0014	.0015

Empirical size for GAMuT, MFLM, and SKAT analyses at significance thresholds of 0.05, 0.01, and 0.001. Empirical size calculated from 10,000 null simulations. Simulations assumed analysis of 6 phenotypes. Sample size was set at either 1000 or 2500. Phenotypes were either continuous or binary. Phenotypic correlation was low (correlation <0.3), moderate (correlation 0.3-0.5), or high (correlation 0.5-0.7).

CHAPTER 4:
A Statistical Approach for Genetic Association Testing of Symptom and Questionnaire Data

A Statistical Approach for Genetic Association Testing of Symptom and Questionnaire Data

K. Alaine Broadaway[1], Richard Duncan[1], Lynn M. Almli[2], Bekh Bradley[2,3], Kerry J. Ressler[2], Michael P. Epstein[1]

[1]Department of Human Genetics, Emory University, Atlanta, GA

[2]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, GA

[3]Clinical Psychologist, Mental Health Service Line, Department of Veterans Affairs Medical Center, Atlanta, GA

ABSTRACT

Many phenotypes are not directly measurable; instead, researchers rely on multivariate symptom data from questionnaires and surveys to indirectly assess a latent phenotype of interest. Researchers subsequently then collapse such questionnaire data into a univariate outcome to represent a surrogate for the latent phenotype. However, phenotypic heterogeneity can dilute any association between the causal gene and the univariate surrogate outcome. When a causal variant is only associated with a subset of symptoms representing the outcome, the effect will be challenging to detect using standard analytical approaches. In order to offer a more flexible method to account for multivariate ordinal data like symptom scales commonly observed in questionnaire data, we introduce a new statistical method that we refer to as the Gene Association Method for Broader Integration of Tests and Scores (GAMBITS). GAMBITS uses a nonparametric distance-covariance approach that compares similarity in multivariate symptom-scale data from questionnaires to similarity in common genetic variants across a gene. We use simulated data to demonstrate that our method provides increased power over standard approaches that collapse questionnaire data into a single surrogate outcome. We also illustrate our approach using GWAS data from the Grady Trauma Project.

INTRODUCTION

Evidence indicates that common genetic variants should explain a sizeable role of the variation in many psychiatric disorders. For example common variants are estimated to explain 40% of the heritability for bipolar disorder [72], 21% of the heritability of depression[14], and 50% of the heritability of autism spectrum disorder [73]. However, even in studies involving thousands of subjects, identification of specific common trait-influencing polymorphisms remains a challenge. To discover new associations, much attention has been spent on improving genotyping technologies to tag more variation; however, comparatively less attention has been afforded to thorough characterization of the underlying phenotypes that are considered for genetic analysis.

In genetic analyses of a trait or disease, we often envision our outcome of interest as a single, measurable entity. In practice, we are rarely able to measure the outcome of interest directly and instead attempt to capture the true, latent phenotype via several connected but discrete measurements. As an example, psychiatric genetic studies attempt to account for the heterogeneity of symptoms found in a single psychiatric disorder (e.g. major depressive disorder) by measuring the symptoms from several angles via a questionnaire or exam. In studies of depressive symptoms, many studies attempt to measure the phenotype using multiple symptom measurements from the Beck Depression Inventory-II (BDI). The BDI is a 21-item questionnaire with each question developed to correspond to DSM-IV diagnostic criteria for major depressive disorder. The answers to each question

are scored from 0 to 3, with higher scores indicating more severe depressive symptoms [34].

At the data collection state, the data captured by the BDI or other questionnaires can actually be considered a collection of interrelated multivariate phenotypes that, in the case of symptom scales, are usually ordinal in nature. However, most statistical techniques for genetic analysis are univariate and are designed to handle a single outcome at a time. To improve analytical utility, the BDI, like many questionnaires, was designed so that the multivariate symptoms are then collapsed into a univariate phenotype. The most simple and most common collapsing method is unweighted summation of each question's score [34]. The cumulative score can then be treated either as a continuous outcome, or cutoffs can be applied to indicate presence/absence of disease symptoms. However, reducing multivariate information to univariate data nearly always comes at a cost. Carefully defining a phenotype is as vital in a GWAS as reliable genotyping; any association between gene and trait may be diluted by phenotyping heterogeneity. For example, if a gene were associated with a subset of the BDI questionnaire outcomes (e.g. a somatic symptom of depression like changes in sleep patterns) but not other subsets (e.g. affective symptoms like mood or attitude), the magnitude of the overall effect size of the gene would be attenuated if the two subsets were combined into a univariate outcome measure.

A few key assumptions must be met in order for a univariate cumulative score to sufficiently summarize multivariate ordinal data. Interested readers are directed to Van der Sluis et al., who have provided thorough mathematical

definitions of these assumptions as well as thorough simulations to assess the assumptions, in a series of manuscripts [175-177]. Briefly, the three primary assumptions that must be met are: (1) the correlation between all questions in the questionnaire must be explained by a single (latent) phenotype; (2) the genetic effect must be on the latent phenotype; (3) the genetic effect—acting through the latent phenotype—must have identical effects on all of the questions in the questionnaire. For applied psychiatric phenotypes, it is more plausible that the assumptions are violated than maintained. Depressive symptoms identified by the BDI might come from multiple sources (e.g. major depressive disorder, bereavement, post-traumatic stress disorder), violating the first assumption. The causal genetic effect might directly increase somatic symptoms of depression such as changes in appetite and sleep, but not impact mood, violating the second assumption. Alternatively, a variant might in fact affect each trait identified by every question, but have slightly different effect sizes on different questions. If any of these assumptions are not met, association analysis using the cumulative score will result in a substantial loss of power [175; 177-179].

A few alternatives have been presented to model the complex multivariate data captured within questionnaires. A popular type of approach is a data reduction method like principal component analysis (PCA), which relies on identifying a linear combination of the set of questionnaire responses that maximize response variance across questions. Once the top few principal components are identified (i.e. those principal components that explain most of the questionnaire variance), association testing is performed between those top principal components and genotype [180; 181].

However, PCA-based strategies that consider only high-variance principal components were recently shown to be generally suboptimal [182].

Van der Sluis et al. [177] presented a multivariate gene-based association test by extended Simes procedure (MGAS). MGAS combines the *P*-values obtained from standard, single-SNP association test for each outcome, to produce a single multivariate gene-based *P*-value. However, MGAS relies on permutations to establish significance, which make genome-wide analyses cumbersome. Alternatively, Basu et al. [183] introduced a rapid multivariate multiple linear regression method (RMMLR), which operates on a MANOVA-based platform. However, while RMMLR establishes significance analytically, it cannot incorporate the important ordinal outcomes commonly measured in questionnaires and surveys.

With this motivation, we propose a Gene Association Method for Broader Integration of Tests and Scores (GAMBITS). GAMBITS is built on the framework of GAMuT, a test of rare variant pleiotropy [184]. Like GAMuT, GAMBITS allows for modeling high-dimensional phenotype data and high-dimensional genotype data via a machine-learning framework called kernel distance-covariance (KDC) [126-130] to provide a nonparametric test of independence between a set of phenotypes and a set of genetic variants. GAMBITS is designed to test whether pairwise similarity in questionnaire responses is independent of pairwise genotypic similarity in a gene or region of interest. The framework allows for an arbitrary number of questions within the questionnaire as well as an arbitrary number of genotypes, thereby permitting gene-based testing of common variants. We show that GAMBITS is

robust to the categorical outcomes frequently obtained by questionnaires, and further can correct for important covariates. In addition, GAMBITS is a closed form test that yields analytic *P*-values, thus scaling easily to genome-wide analysis.

The remainder of this manuscript is organized as follows. First, we derive the GAMBITS test using the KDC framework, and describe how to derive analytic *P*-values for this test as well as how to adjust for covariates in GAMBITS. We then present simulation work comparing GAMBITS to univariate kernel machine regression (KMR) [97; 98] and univariate linear regression, to demonstrate that GAMBITS can be considerably more powerful than a univariate test based on a cumulative score. We then illustrate the approach using Beck Depression Inventory scores collected as part of the Grady Trauma Project. We finish with concluding remarks and discuss potential extensions to our approach.

MATERIALS AND METHODS

Assumptions and Notation: We assume an inventory, test, or questionnaire with Q questions. The response to each question q ($q=1,\ldots,Q$) is an ordinal response ranging from 0 to F, where F is the maximum score possible. We assume a population-based sample of N subjects have responded to the questionnaire and possess common-variant data in a target gene or region. For subject j ($j=1,\ldots,N$), we define $\boldsymbol{P}_j = (P_{j,1}, P_{j,2}, \ldots, P_{j,Q})$ as subject j's responses to the Q questions. We then define a matrix of questionnaire responses for the entire sample

$\boldsymbol{P} = \left(\boldsymbol{P_1^T}, \boldsymbol{P_2^T}, \ldots, \boldsymbol{P_N^T}\right)^{\boldsymbol{T}}$, which is of dimension *N*x*Q*. We further define the cumulative score $\boldsymbol{S_j}$ as $\mathbf{S}_j = \sum_{q=1}^{Q} P_{j,q}$, which we be used in univariate analyses.

Similarly, we define $\boldsymbol{G_j}$ = ($G_{j,1}$, $G_{j,2}$, …, $G_{j,V}$) to be the genotypes of subject *j* at *V* SNPs, where $G_{j,v}$ is coded as the number of copies of the minor allele that the subject possesses at SNP *v*. The SNPs included in $\boldsymbol{G_j}$ will be referred to as the "SNP set." We then construct the matrix of genotypes for the sample as $\boldsymbol{G} = \left(\boldsymbol{G_1^T}, \boldsymbol{G_2^T}, \ldots, \boldsymbol{G_N^T}\right)^{\boldsymbol{T}}$, which is of dimension *N x V*. Several approaches to constructing a SNP set have previously been described [98; 106]. For demonstration purposes in this manuscript, we will define a SNP set as common variants (minor-allele frequency [MAF] > 5%) that fall within 2kb of a gene of interest.

GAMBITS Test: We create GAMBITS to examine the relationship between questions ***P*** and genotypes ***G***. GAMBITS is based on a KDC machine-learning technique [126-130], which allows nonparametric tests of independence between two distinct sets of multivariate variables. For each set of multivariate variables, KDC constructs an *NxN* matrix with individual elements of the matrix corresponding to similarity (or dissimilarity) in the outcomes among different pairs of subjects. KDC then evaluates whether the pairwise elements in the similarity matrix of one set of multivariate variables is independent of the pairwise elements in the similarity matrix for the other set of multivariate variables.

Leveraging the KDC framework previously described in Chapter 3 [184], we create a SNP-set test to test for independence between ***P*** (*NxL* matrix of multivariate responses to a questionnaire) and ***G*** (*NxV* matrix of multivariate genotypes). To do

this, we first develop an *NxN* questionnaire-similarity matrix ***Y*** (based on ***P***) and a *N x N* genotypic-similarity matrix ***X*** (based on ***G***). The choice of how to model pairwise similarity or dissimilarity for a set of multivariate outcomes is quite flexible. For example, for ***P***, we can model the matrix ***Y*** using a projection matrix, as suggested by Zapala and Schork[133], such that $\boldsymbol{Y} = \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}^T$. We can also construct the model ***Y*** using user-selected kernel functions [97; 98; 105; 131] such as the linear kernel, $y(\boldsymbol{P}_i,\boldsymbol{P}_j)=\sum_{l=1}^{L} P_{i,l}P_{j,l}$ or a quadratic kernel, $y(\boldsymbol{P}_i,\boldsymbol{P}_j)=(1+\sum_{l=1}^{L} P_{i,l}P_{j,l})^2$. For genotypes ***G***, we model its corresponding matrix ***X*** using kernel functions $x(\boldsymbol{G}_i, \boldsymbol{G}_j)$ that can take the same form (e.g. linear, quadratic, Gaussian, Euclidean distance) used to construct $y(\boldsymbol{P}_i, \boldsymbol{P}_j)$. A few genetic-specific kernel functions also exist, like the identity-by-state (IBS) kernel, $x(\boldsymbol{G}_i\boldsymbol{G}_j)=\sum_{v=1}^{V} IBS(G_{i,v}G_{j,v})/2V$, where $IBS(G_{i,v}G_{j,v})$ denotes the number of alleles (0,1, or 2) shared IBS by subjects *i* and *j* at SNP *v*. Also, we may wish to further augment $x(\boldsymbol{G}_i, \boldsymbol{G}_j)$ to preferentially upweight the contributions of particular SNPs in ***G*** over others in the gene. Wu et al. [98] and Schifano et al [106] provide nice discussions on relevant weighting approaches for common SNP analyses. For all simulations and analyses reported here, we implement a weighting scheme based on the minor-allele frequency (MAF) of each assayed SNP that weights rarer variants over more common ones; the particular weight we apply for the v^{th} variant is $w_v = 1/\sqrt{\mathrm{MAF}_v}$. We can do this by creating a diagonal weight matrix $\boldsymbol{W}$= diag($w_1, w_2, \ldots, w_V$), where w_v, reflects the relative weight for the v^{th} variant in the gene. Using ***W***, we can then create a weighted linear kernel function as $\boldsymbol{X}=\boldsymbol{GWG}^T$. Derivation of other weighted kernel functions is

straightforward. For all results in this manuscript, we construct $\boldsymbol{Y}$ using the projection matrix, and $\boldsymbol{X}$ using the weighted linear kernel.

Once we construct the similarity matrixes $\boldsymbol{Y}$ and $\boldsymbol{X}$, we derive our GAMBITS approach as a test of independence between the elements of these two matrices. The approach is equivalent to the test presented in detail in Chapter 3 [184]. Briefly, we center each matrix as $\boldsymbol{Y_c=HYH}$ and $\boldsymbol{X_c=HXH}$. Here, $\boldsymbol{H} = (\boldsymbol{I} - 1_N 1_N^T / N)$ is a centering matrix with property $\boldsymbol{HH=H}$, $\boldsymbol{I}$ is an identity matrix of dimension N, and 1_N is an $Nx1$ vector with each element equal to 1. Using $\boldsymbol{Y_c}$ and $\boldsymbol{X_c}$, we construct our test of independence of the two matrices as

$$T_{GAMBITS} = \frac{1}{N}\text{trace}(Y_c X_c) \quad (1)$$

Under the null hypothesis of independence of the two matrices, $T_{GAMBITS}$ follows the same asymptotic distribution as

$$\frac{1}{N^2}\sum_{i,j=1}^{N} \lambda_{X,i}\lambda_{Y,i} z_{ij}^2 \quad (2)$$

where $\lambda_{X,i}$ is the i^{th} ordered eigenvalue of $\mathbf{X_c}$, $\lambda_{Y,j}$ is the j^{th} ordered eigenvalue of $\mathbf{Y_c}$, and z_{ij}^2 are independent and identically-distributed χ_1^2 variables [130]. As presented in Chapter 3 [184] we derive *P*-values for our GAMuT test using Davies' exact method [110], which is a computationally efficient method to provide accurate *P*-values in the extreme tails of tests that follow mixtures of chi-square variables [131]. An implementation of Davies' method is available in the R library *CompQuadForm*.

Adjusting for Covariates: Genetic association tests must adjust for important covariates, such as principal components of ancestry, to avoid potential confounding

of results. We can control for confounders before applying GAMBITS by regressing each phenotype separately on covariates of interest and then using the residuals to form the phenotypic similarity matrix ***Y***. Although residualizing categorical phenotypes is not standard, studies have suggested that this procedure does not affect the validity of genetic association tests in case-control studies [141; 142]. As we describe in the Results section, such residualization provides an effective correction for confounders within our simulated ordinal datasets.

Simulations: We conducted simulations to verify that GAMBITS properly preserves type I error (i.e. empirical size) and to assess power of GAMBITS relative to standard association tests that treat questionnaire responses as a scalar outcome variable resulting from summing the responses into a continuous score. We perform simulations based on SNPs and LD patterns located within 2 kb up- and down-stream from *signal transducer and activator of transcription 3 (STAT3)*, a gene on chromosome 17q21.31. We show the MAF and pairwise LD structure of SNPs in *STAT3* in Figure 2. To incorporate observed LD patterns from HapMap samples, we used the HAPGEN package [111] to generate simulated SNP data. HAPGEN generates simulated genotype information for all SNPs identified in HapMap within the *STAT3* gene; however, to better replicate real GWAS conditions, we applied the testing approaches only to those SNPs that would be typed on standard genotyping arrays. Although 27 common SNPs fall within the *STAT3* gene, only 14 of the 27 are genotyped on the Illumina HumanOmni1-Quad genotyping platform. Thus, the 14 typed SNPs form the SNP set for the kernel approach, and only the 14 typed SNPs are tested for association. Under simulations where the causal SNP is not

genotyped, power to detect an association relies on LD between the causal SNP and typed SNPs.

We simulate multivariate questionnaire data to mimic the BDI questionnaire results obtained from GTP participants. The BDI consists of 21 groups of statements that reflect various symptoms and attitudes associated with depression. Each group includes 4 statements, which correspond to a scale of 0 to 3 in terms of intensity. The 21 groups are sadness, pessimism, past failure, loss of pleasure, guilty feelings, punishment feelings, self-dislike, self-criticalness, suicidal thoughts or wishes, crying, agitation, loss of interest, indecisiveness, worthlessness, loss of energy, changes in sleep patterns, irritability, changes in appetite, concentration difficulty, fatigue, and loss of libido. The BDI is generally self-administered or self-reported, and is scored by summing the ratings given to each of the 21 items. Summing the responses yields a score ranging from 0-63, with scores higher than 28 being indicative of moderate to severe depression.

To simulate data, we first generated 21 outcomes for each subject using a multivariate normal distribution with mean vector 0 and QxQ correlation matrix $\mathbf{\Sigma}$. We calculated $\mathbf{\Sigma}$ based on observed Spearman rank correlation calculations from the GTP BDI questionnaire responses shown in Figure 1. The observed correlations between questions ranged from 0.22 to 0.57. Next, we generated ordinal responses from the normally distributed variables to match the ordinal responses observed in GTP data. Frequency of scores by each of the 21 BDI questions is shown in Figure 3. We found the percent of GTP participants who answered 0, 1, 2, and 3 for each question. We then matched the percentages of each BDI question for each of the 21

normally distributed variables. For example, in BDI Question 1 ("Sadness"), 56% of participants answered 0 ("I do not feel sad"), 34% answered 1 ("I feel sad much of the time"), 6% answered 2 ("I feel sad all of the time"), and 4% answered 3 ("I am so sad or unhappy that I can't stand it."). To simulate ordinal responses to question 1, the lowest 56% of the continuous outcomes were assigned a score of 0, values falling in the 57-90 percentile were assigned a score of 1, 91-96 percentiles were assigned a score of 2, and values in the 97th percentile and above were assigned a score of 3. We set sample size *N* of either 1000 or 2500 subjects. We applied GAMBITS to 10,000 null simulated datasets to estimate empirical size.

To investigate the performance of GAMBITS under confounding and to assess whether the approach can successfully adjust for relevant covariates in this setting, we also simulated questions under a confounding model where question responses were independent of genotype, but both questions and genotype are associated with a continuous covariate Z. We simulated questions correlated with the covariate Z under the model $\mathbf{P} \sim MVN(0.2\mathbf{Z}, \Sigma)$, where $\mathbf{Z}$ denotes the *N* x 1 sample vector of covariates. We arbitrarily selected SNP 9 (rs9909659) as causal for the confounder. We simulated correlation between SNP 9 and the covariate by generating the effect size of SNP on confounder as $\beta_Z=0.2$. Testing empirical size under this model allows us to verify that our approach to control for confounders is valid.

For power models, we simulated data sets in which each of the 27 SNPs were modeled causal in turn. We model effect size of the causal SNP on each question, $\beta_{q,}$ as $\beta_q = N(0.1, 0.03)$. This formulation sets mean effect sizes with modest effect on the overall cumulative score; for causal SNP with MAF=0.3, this formulation

corresponds to an R^2=0.009 when the SNP is associated with all questions in the questionnaire. Allowing β_q to vary around a normal distribution allows the variant to have a slightly different effect size for each question. We also vary the number of questions that are associated with the causal SNP, such that not all of questions will be dependent on the gene of interest. We consider situations where 18/21, 9/21, and 6/21 questions are actually associated with the causal SNP. We control residual correlation among questions through consideration of trait-specific heritability (i.e. the relative variance of $\mathbf{P}_q$ explained by the causal SNP). We define trait-specific heritability for question *q* as $h_q = \beta^2_{SNP,q} * 2MAF_{SNP}(1\text{-}MAF_{SNP})$, where MAF_{SNP} is the MAF of the causal SNP. The correlation between questions *q* and q' is defined as $E_{q,q'} = \sqrt{1-h_q}\sqrt{1-h_{q'}} * \mathbf{\Sigma}_{q,q'}$ where $\mathbf{\Sigma}$ is the *LxL* residual correlation matrix shown in Figure 1. This allows the residual correlation structure among phenotypes to stay at the defined values.

We evaluate GAMBITS using the simulated data, and compare our approach to two standard approaches that use the scalar cumulative questionnaire score. First, we consider a linear regression model that follows the form

$$\boldsymbol{S} = \mathbf{Z}\gamma + \beta_{SNP}\mathbf{G}_m + e$$

where $\boldsymbol{S}$ is the Nx1 vector of cumulative scores, $\boldsymbol{Z}$ is an *N* x *c* vector of *c* covariates (including an intercept) with regression parameter vector γ, $\boldsymbol{G}_\mathbf{m}$ denotes an *N*x1 vector of SNP genotypes at SNP *m* with regression parameter β_{SNP}, and the residual error *e* follows a MVN distribution, $e \sim MVN(0, \sigma^2\boldsymbol{I})$, where $\boldsymbol{I}$ denotes the *NxN*

identity matrix. We then implement a likelihood ratio test to assess the null hypothesis of $H_0 : \beta_{SNP} = 0$ for each SNP *m*. To adjust for multiple testing of *M* correlated SNPs, we apply P_{ACT} [103] to the smallest observed *P-value*. However, GAMBITS differs from the univariate linear regression model in two main ways: first, it treats the outcome as a multivariate score, but second, it tests at the gene, rather than SNP, level. While both differences should increase power of our approach, we were interested in teasing apart the relative importance of those two differences. We therefore also contrast our approach to a popular KMR [97] test. While the KMR approach assesses the effect of all genotyped SNPs within a gene simultaneously, it is a univariate test that requires a collapsed univariate score. Therefore, comparison of GAMBITS against the univariate KMR test should help highlight the benefit of considering a multivariate questionnaire phenotype over the benefit of a gene-based analysis.

Analysis of the Grady Trauma Project:

Depression is a moderately heritable disorder ($h^2 \approx 0.30$), yet, despite substantial interest in identifying genetic causes of the disorder, its genetic underpinnings remain largely unidentified [112]. Data used in our analysis were collected as part of a larger study, called the Grady Trauma Project (GTP), which investigates the role for psychiatric disorders such as post-traumatic stress disorder and depression [119; 120]. Participants in the GTP are served by the Grady Hospital in Atlanta, Georgia, and are predominantly urban, African American, and of low socioeconomic status. GTP staff approach subjects in the waiting rooms of Grady Primary Care and Obstetrics and Gynecology and obtain their written consent to

participate. GTP staff conduct an extensive verbal interview, which includes demographic information, a history of stressful life events, and several psychological surveys. The GTP queries participants on the BDI, a 21-item multiple-choice questionnaire that assesses symptoms of depression [34].

The GTP genotyped 4,607 participants on the Illumina HumanOmni1-Quad array to permit GWAS analyses. Upon removing subjects who did not report at least one past trauma, subjects with missing BDI scores, or subjects whose genetic data failed quality control, 3,627 subjects remained. We used the support files provided by Illumina to identify 736,462 common genetic variants (MAF > 5%) that fell within 18,280 known genes. We further excluded genes with fewer than 5 or more than 1000 common SNPs, leaving 15,175 genes containing 725,558 SNPs. For each of the study participants, we also obtained age, gender, and the top ten principal components to account for ancestry. We applied GAMBITS using both a projection matrix and a linear kernel to measure pairwise phenotypic similarity. We also ran univariate KMR and linear regression on the cumulative BDI scores. For all GAMuT and KMR tests, we used a weighted linear kernel (selecting the weighting scheme recommended by Wu et al. [131], described above, as we used in our simulation work) to measure pairwise genotypic similarity. For linear regression, we adjusted for multiple testing using P_{ACT} [103].

RESULTS

Type-I Error Simulations: Figure 4 shows the quantile-quantile (QQ) plots based on application of GAMBITS, KMR, and linear regression to null datasets consisting of 1000 or 2500 subjects assayed for 21 BDI questions. For both sample

sizes tested, GAMBITS properly controls for type I error, even at the extreme tails of the test. KMR and linear regression, using the cumulative score approach, also demonstrated appropriate empirical size. Figure 5 shows that this residualization of questionnaire data prior GAMBITS analysis effectively controls for confounding that, unadjusted, would yield inflated results.

Power Simulations: Next we compared the power of GAMBITS with univariate KMR and linear regression analyses in a series of simulation studies. For these power simulations, we set sample size to 1000. Power was estimated as the proportion of *P*-values < 2.5×10^{-6} (reflecting a genome-wide correction for 20,000 genes) and was evaluated based on 500 replicates of the data per model. Figure 6 shows the power results. We plot power as a function of the causal SNP, where the causal SNPs are ordered by genomic location. The 14 genotyped SNPs (denoted by 'x' on the bottom of the plots) were used to calculate test statistics, but all 27 SNPs were treated as causal in turn. Therefore, in situations where the causal SNP is not typed, we rely on correlation of the causal SNP with observed typed SNPs in STAT3 to gain statistical power. GAMBITS offers considerable more power than the two competing univariate methods for each of the three assumptions tested. When approximately half of the questions (12/21) are associated with the causal SNP, both KMR and linear regression observe nearly zero power to detect the effect; by comparison, GAMBITS maintains power greater than 50% for 23 of the 27 causal SNPs. We observe a drop in power using GAMBITS when nearly all of the questions (left column Figure 6) are associated with the causal variant compared with a more modest number of questions are associated (middle column Figure 6). This pattern

has been observed in other multivariate approaches, including multivariate analysis of variance (MANOVA) [185].

Application to GTP: We use the GTP dataset to test for associations between the BDI questionnaire and common variants in 15,175 genes. Prior to analysis using GAMBITS, we controlled for gender, age, and ancestry on the 3,627 unrelated subjects. We applied GAMBITS using both a projection matrix and a linear kernel to measure pairwise phenotypic similarity. For comparison, we ran univariate KMR and univariate linear regression on the cumulative BDI score. We adjusted for multiple testing of SNPs in each gene in the linear regression approach using P_{ACT} [103].

For all GAMBITS and KMR tests, we used a weighted linear kernel to measure pairwise genotypic similarity. We set a stringent study-wise significance threshold of 3.3×10^{-6}, which corresponds to a Bonferroni correction based on the number of genes tested (15,175): $\alpha_{\text{BONFERRONI}} = 0.05/15175$. We considered *P*-values less than $P < 1 \times 10^{-4}$ as suggestive. Figure 7 provides genome-wide results using GAMBITS, KMR, and linear regression, with top findings highlighted in Table 1. No method identified a gene that met genome-wide significance. The projection matrix form of GAMBITS failed to identify any genes suggestive significance, while the linear kernel form of GAMBITS identified one gene of suggestive significance. Genome-wide the correlation between the two forms of GAMBITS *P*-values was 0.49 (Pearson's product-moment correlation *P*-value < 1×10^{-15}). Results using KMR were very similar to the linear kernel form of GAMBITS: correlation among *P*-values was 0.91 (Pearson's product-moment correlation *P*-value < 1×10^{-15}). Linear regression

identified two genes of suggestive significance, and results were most strongly correlated with KMR (correlation = 0.61, Pearson's product-moment correlation *P*-value < $1x10^{-15}$). None of the 4 genes has previously been associated with depressive symptoms.

DISCUSSION

We have presented GAMBITS, a KDC framework for SNP set analysis for multivariate ordinal outcomes collected for a questionnaire or survey [126; 127; 130]. GAMBITS offers an alternative to the standard analytic approach, which requires collapsing multivariate questionnaire data into a single cumulative outcome. The approach allows for modeling phenotypic heterogeneity, in which a genetic risk factor only affects a subcategory within the questionnaire. An example of this is if a gene increased risk of sleep issues in patients with depression, but did not impact feelings of self-dislike or feeling guilty. In standard cumulative approaches, including KMR and linear regression, phenotypic heterogeneity can dilute the association between gene and trait, making the association extremely difficult to detect. For example, in our simulations, we found that the cumulative approaches had almost no power to detect an effect when over half of the questions were associated with the causal SNP. In contrast, GAMBITS maintains considerable power under these assumptions. We did observe a drop in power using GAMBITS when all (or nearly all) of the questions were associated the causal gene. While further investigation into this phenomenon is warranted, it has been observed in other multivariate approaches, including multivariate analysis of variance (MANOVA) [185].

We applied GAMBITS to the GTP dataset to test for associations between the BDI questionnaire and 15,175 genes. While no genetic association met genome-wide significance, the applied results demonstrate appropriate statistical size after accounting for important covariates, including ancestry, gender, and age.

GAMBITS derives analytic *P*-values based on Davies' exact method, thereby improving computational efficiency and permitting application on a genome-wide scale. Like the popular KMR framework for univariate analysis, our approach allows for inclusion of prior information, such as biological plausibility of the SNPs under study. We provide R software implementing the approach on our website (see Web Resources) which can be run through PLINK, if desired.

The framework used by GAMBITS is amenable to several promising extensions that we will explore in future work. First, although the results presented here are focused on analysis of common SNPs, the approach is readily extendible to rare variant analysis. Second, mediation analysis would allow investigators to tease apart which underlying factors are directly associated with the gene of interest. Although existing mediation analyses are not intended to handle high-dimensional traits, we intend to develop a KDC-based procedure that will be appropriate for multivariate questionnaire data. Third, GAMBITS framework should allow for inclusion of non-ordinal measures, as well as ordinal measures, within the same test. This would allow important disease status and continuous measures to be included with ordinal questionnaire responses.

ACKNOWLEDGEMENTS

This work was supported by NIH grants HG007508, HL086694, HL119443, MH071537, and AR060893. For purposes of disclosing duality of interest, Michael Epstein is a consultant for Amnion Laboratories.

WEB RESOURCES

Epstein Software: http://www.genetics.emory.edu/labs/epstein/software

OMIM: http://www.omim.org

Table 1: Top GTP Results

Gene Name	Chromosome	Number of SNPs	Projection Matrix	Linear Kernel	KMR	Linear Regression
DUSP5P	1	13	5.5E-02	5.4E-05	4.1E-05	1.9E-04
C5orf45	5	6	5.9E-01	1.4E-04	5.8E-05	3.7E-02
NUP214	9	9	1.6E-01	5.0E-03	4.0E-03	7.9E-05
SLC6A13	12	47	8.8E-02	3.8E-02	5.3E-02	8.8E-05

We identified 4 genes in the GTP dataset with *P*-values of suggestive significance ($P<1x10^{-4}$) using GAMBITS, KMR, or linear regression using P_{ACT} to adjust for multiple testing of SNPs within each gene. For the genes we provide gene name, chromosomal location of gene, number of SNPs (MAF > 5%) found in each gene in the GTP dataset, and *P*-values for the four approaches.

FIGURE LEGENDS

Figure 1: Correlation among the 21 BDI question responses in the GTP dataset, after adjusting for covariates. Dark green indicates correlation of 0.6 while white indicates correlation of 0.2. Correlation among all questions was positive. All correlations are significant (Pearson's product-moment correlation *P*-value < $1x10^{-15}$).

Figure 2: Pairwise LD (R^2) heatmap and MAF for all SNPs reported in HapMap ±2kb of the *STAT3* gene. MAF is plotted below, with genotyped SNPs denoted by the 'x' on the bottom of the MAF plot.

Figure 3: Frequency of scores for each of the 21 BDI questions in the GTP dataset. The answers to each question are scored from 0 (no symptoms) to 3 (severe symptoms).

Figure 4: The QQ plots applying GAMBITS, KMR, and linear regression to 10,000 simulated null data sets assuming a sample size of 1000 (top row) and 2500 (bottom row). For each simulation, 21 ordinal questionnaires were generated. For KMR and linear regression, the 21 questions were summed together to yield a single cumulative score.

Figure 5: The QQ plots of 10,000 simulated null datasets assuming a sample size of 1000 with a confounding variable. Questionnaire responses are independent of genotypes, but both responses and genotypes are associated with a continuous covariate. Left shows QQ plots without adjustment for confounding, while right shows QQ plots after adjustment for confounding by residualization.

Figure 6: Power for GAMBITS (red), KMR (blue), and linear regression (green) is plotted as a function of causal SNP. Left plot assumes the causal SNP is associated with 18 of the 21 BDI questions. Middle plot assumes 12 of 21 questions are associated with causal SNP. Right plot assumes only 6 of 21 questions are associated with the causal SNP.

Figure 7: Results of the GTP analysis. First column shows Manhattan and QQ plots for GAMBITS using a projection matrix for phenotypes. Second column shows Manhattan and QQ plots for GAMBITS using a linear kernel for phenotypes. Third column shows Manhattan and QQ plots for KMR, and fourth column shows Manhattan and QQ plots for linear regression, using P_{ACT} to correct for multiple testing. Horizontal blue line indicates suggestive significance threshold. Horizontal red line indicates study-wide significance.

Figure 1: Correlation of GTP BDI Scores by Question

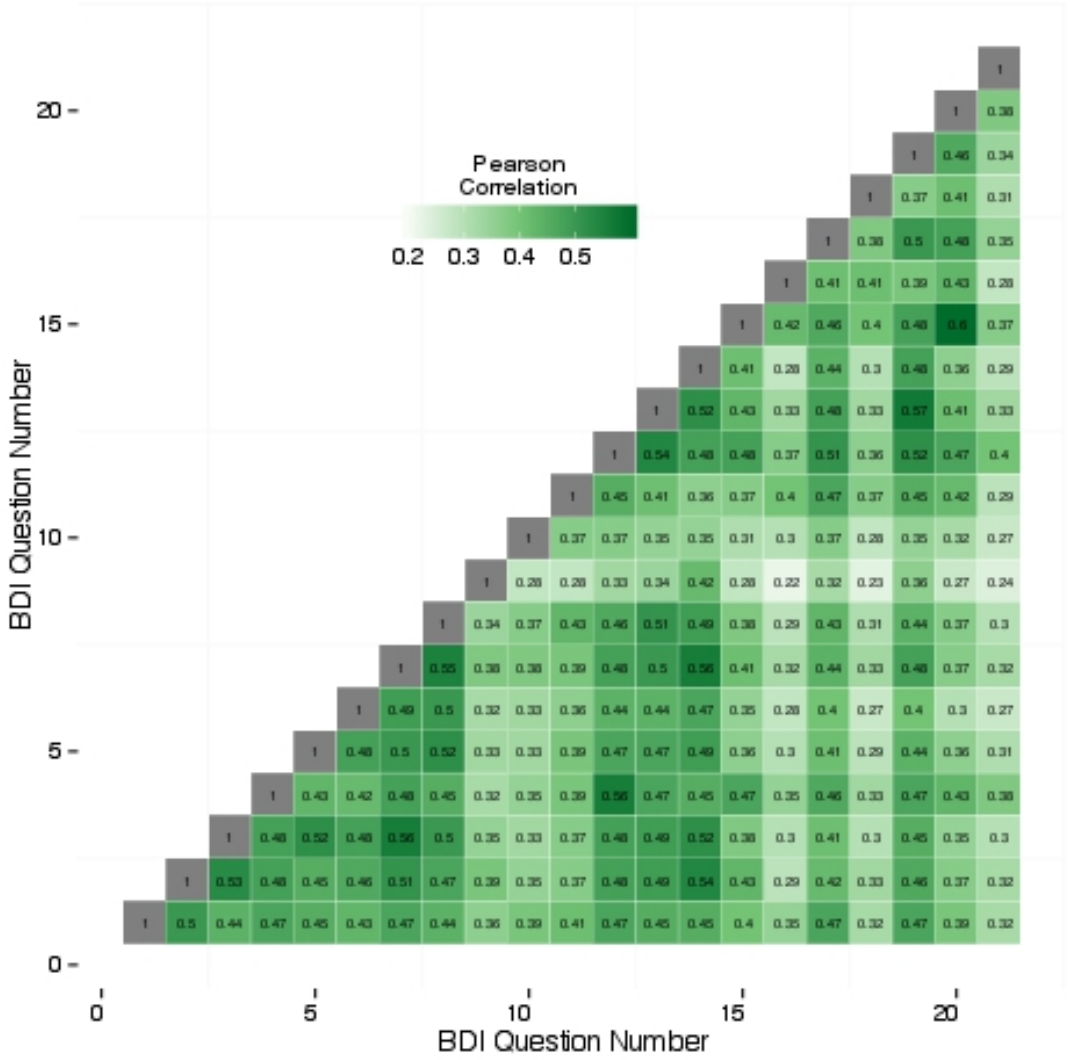


Figure 2: Pairwise LD and MAF of *STAT3*

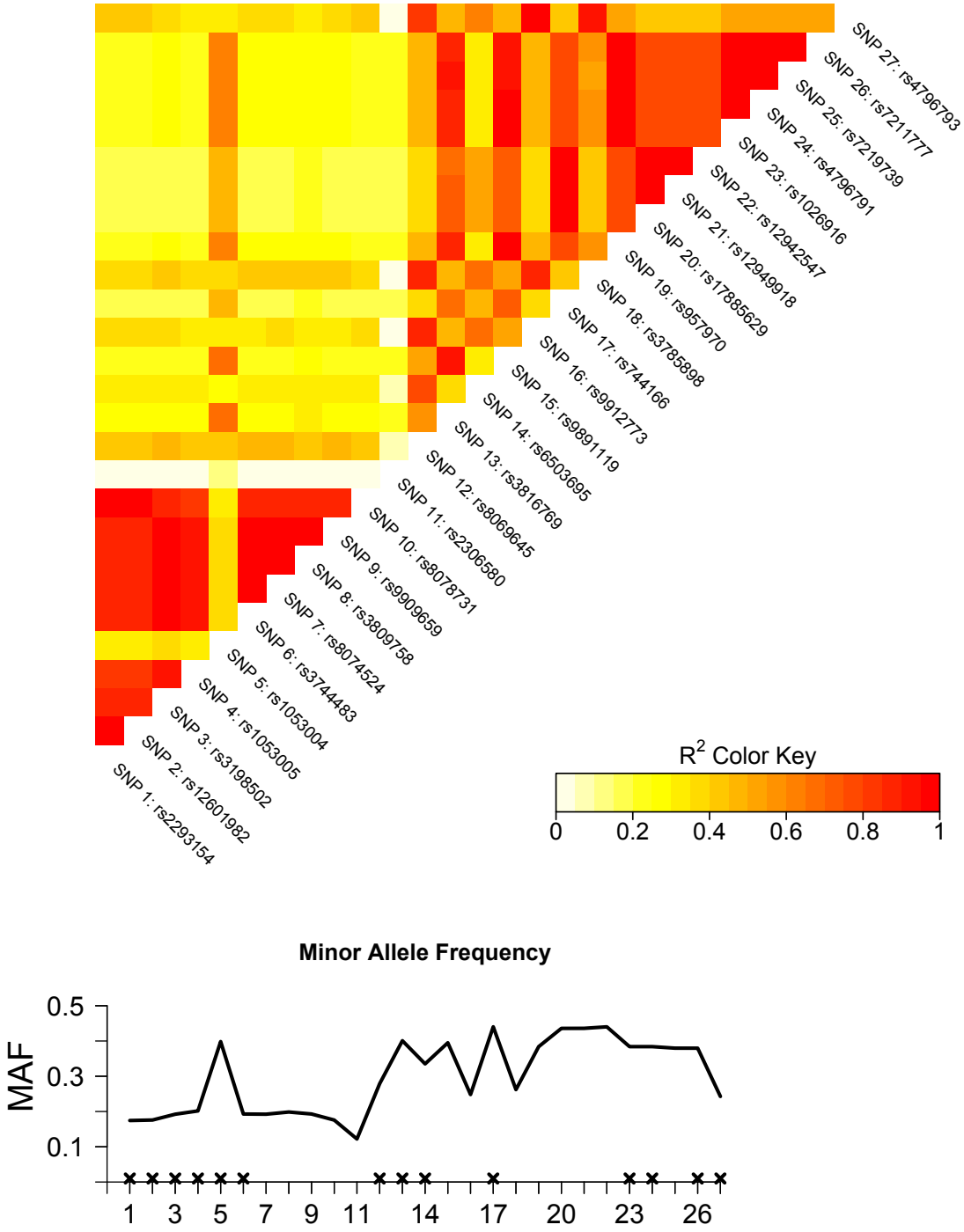


Figure 3: Frequency of BDI Scores

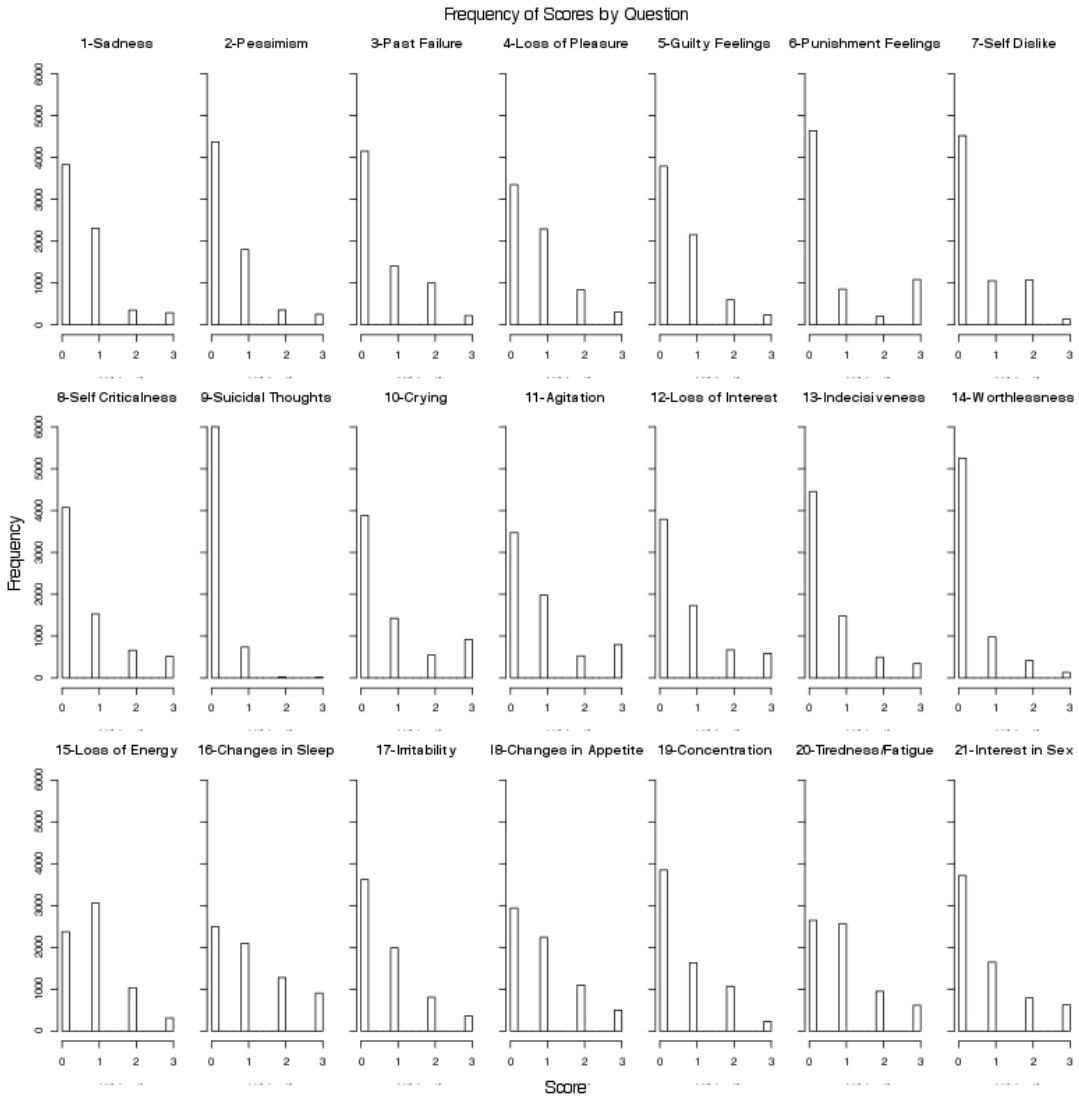


Figure 4: QQ Plots for GAMBITS, KMR, and Linear Regression

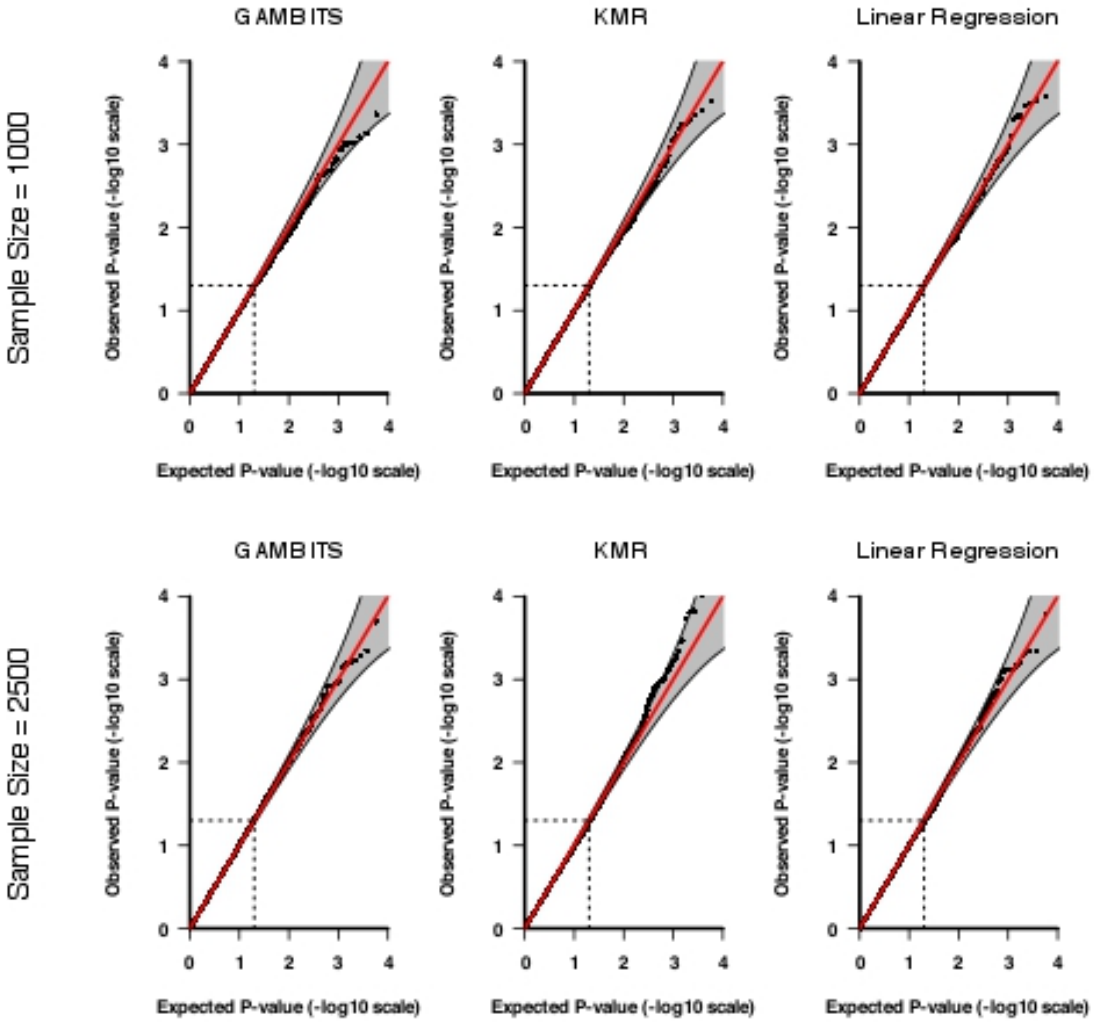


Figure 5: QQ Plots with Confounder

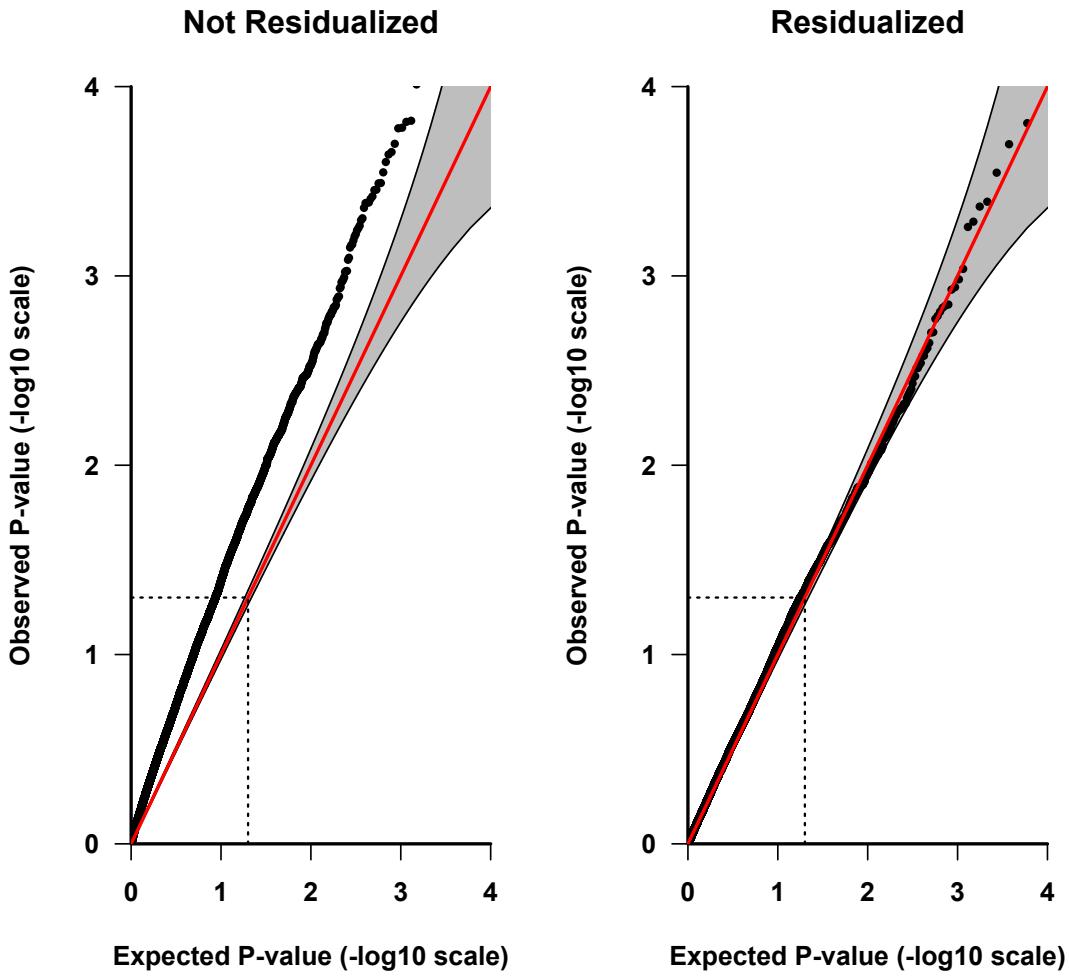


Figure 6: Power to Detect Genetic Effects

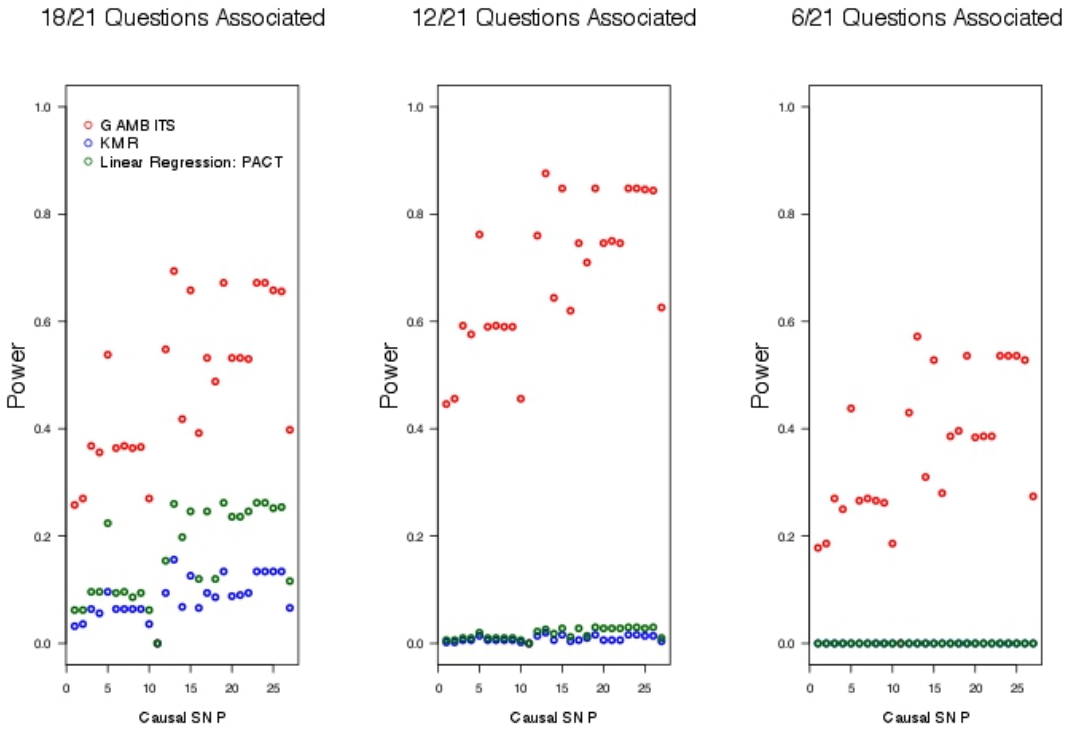
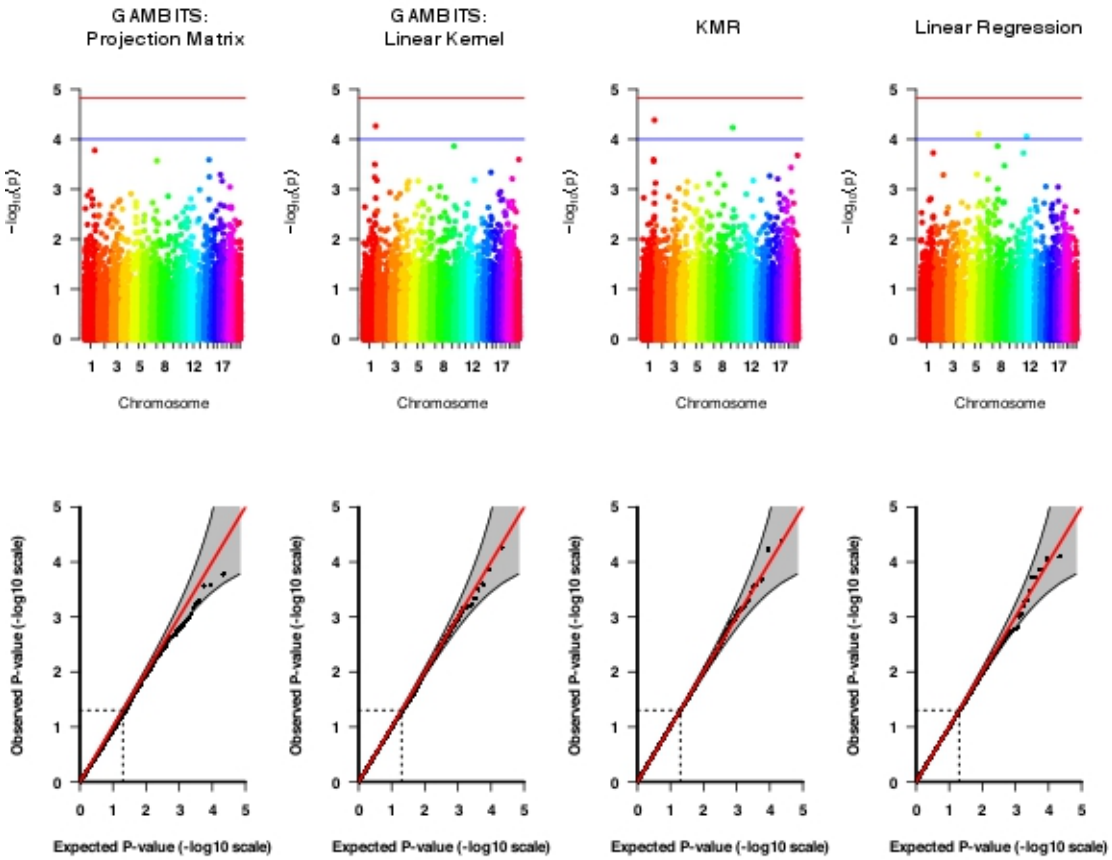


Figure 7: Results of GTP Analyses



CHAPTER 5:
Discussion

DISCUSSION

The challenges in detecting genetic variants associated with human diseases and traits are daunting. However, we can improve our ability to detect such variants by creating novel methods calibrated to detect the types of causal variation predicted by population-genetics theory. For example, the infinitesimal model predicts that causal alleles with large effect sizes will be kept at low frequency in the population, while causal variants of intermediate frequencies will have very small effect sizes [23-25]. Empirical evidence supports this expectation [19-21]. In order to find genetic associations with complex traits, we must then develop analytic approaches designed to find rare variants of potentially large effect and/or common variants of small effect size. Advancements in sequencing and genotyping technologies are now allowing researchers to generate array or sequence data for sample sizes that will be sufficient to detect these types of causal variants. However, novel statistical tests are necessary to better query these emerging data types.

Evidence and theory indicate that any one causal variant will explain very little of trait heritability. Under this model of genetic architecture, attempting to test for an effect of a single variant is inadvisable. If we instead analyze the effects of multiple genetic variants simultaneously, such as testing for an effect across all variants within a gene or pathway, we potentially increase power [97; 98; 106; 131]. This method of aggregated testing is already the norm for rare variant analyses, since ability to detect any effect of a single rare genetic variant would require either prohibitively large sample sizes or an implausibly large effect size [131; 186-188]. The assumption behind the gene-level approach is that while causal sequence variation

may be individually rare, as a class, affected individuals may harbor rare genetic variants that are at the same gene. While gene-level analysis is less frequently implemented in GWAS, common SNP analysis has also been shown to benefit from such an approach [97; 98; 189]. Examining sets of SNPs rather than each variant independently will greatly reduce the need for multiple hypothesis testing. In addition, since multiple typed markers are likely to be in linkage disequilibrium (LD) with the causal SNP, joint consideration of those variants will capture the effect of a true causal variant more effectively than independent SNP testing. Moreover, both common and rare variant analyses will benefit from the implicit consideration of within-gene epistatic interactions in the association tests.

In this dissertation, we make use of kernel approaches (both kernel-machine regression and kernel distance-covariance techniques) to develop powerful gene-based statistical tests. Kernel machine techniques allow researchers embrace the infinitesimal view of complex traits. Testing for a combined effect of multiple genetic variants simultaneously implicitly acknowledges the assumption that any single variant is unlikely to have substantial effect. In addition, KMR and KDC allow explicit inclusion of *a priori* expectations of genetic effect size for each variant. For example, one can preferentially up-weight contributions of very rare variants in the kernel function. KMR has been shown to be a powerful framework for testing main genetic effects of both common and rare variants [97-99; 105; 106; 131].

In Chapter 2, I introduced a gene-based approach for association mapping of common trait variation using multiple SNPs to consider joint tests of gene- and gene-environment interaction. I contrasted the approach to traditional linear

regression approaches that model either main genetic effects or joint effects of gene and gene-environment interaction. Our approach performed particularly well relative to the traditional tests when the underlying causal model involved a blend of both interaction and main genetic effects. In Chapter 3, I introduced GAMuT, a new statistical method that tests for association of multiple rare variants with multiple phenotypes. Under simulation models, GAMuT provides markedly greater power over univariate SKAT and multivariate MFLM statistical approaches. In Chapter 4, I introduced GAMBITS, a gene-based association test to be used in analyses of multivariate ordinal outcomes collected in questionnaires and surveys. Typically, this type of data is collapsed into a single cumulative score; I therefore compared the power of GAMBITS against the power of traditional cumulative approaches across a range of simulation models.

Although kernel-based approaches are powerful, there are two drawbacks to using this framework for genetic analysis. First, there is a loss of information by using a gene-level test rather than variant-level test. Researchers using a gene-level test cannot identify which variants within the gene are associated with the phenotype(s), and which are neutral with respect to the outcome(s). However, true findings are more likely to be replicable, since genic function is highly consistent while individual variants may differ in frequency and LD structure across different populations [177; 190]. Second, because the effects of multiple variants are modeled simultaneously, kernel machine tests cannot measure size of effect. Since the framework only allows testing and not estimation, results from KDC and KMR tests cannot be used in risk prediction.

There is also need for future work. Each of the previous Chapters focused on either common or rare variants, not both. However, each of the methods is readily extendible to both rare and common variant analysis. In addition, KMR and KDC approaches have shown some promise in other high-dimensional data such as neuroimaging [191], gene expression [67], and microbiome [192]. Extending the approaches introduced in this dissertation to other fields driven by high-dimensional data might prove useful in elucidating etiology of human traits. The GxE method could be extended to simultaneously model multiple environmental exposures, which would be useful in cases where several environmental measurements might be expected to correlate with a true latent exposure that is interacting with genotype to influence outcome. The GAMuT and GAMBITS methods would benefit from mediation analysis. A mediation analysis could be performed after GAMuT or GAMBITS, to determine if the associations discovered by either test were biological (a causal locus directly affecting more than one trait or question in a questionnaire) or mediation (a causal locus affecting only one trait, which in turn affects another trait). By performing a mediation analysis, researchers could determine which underlying phenotypes or questions within a questionnaire were directly associated with a gene of interest. Existing mediation analyses are not intended to handle high-dimensional traits; we propose the creation of KDC procedures to identify whether an observed association is mediated by a different set of phenotypes or questions. Additionally, GAMuT and GAMBITS currently assume unrelated subjects; however, the work by Jiang et al. [172] provides a framework to extend our KDC approaches to allow for case-parent trio studies. While the Jiang method uses a KMR approach and

is therefore only appropriate for univariate phenotype analyses, an analogous approach, using GAMuT, should allow for high-dimensional phenotype data. Finally, a meta-analysis approach introduced by Lee et al. [174], which is designed to combine results of multiple KMR-based studies, should be readily extendible to KDC results, such as those obtained via GAMuT or GAMBITS. Such an extension would allow combining association results from multiple KDC studies through a meta-analysis. We will explore all these ideas in future work.

REFERENCES

1. Collins, F.S., and Guttmacher, A.E. (2001). Genetics moves into the medical mainstream. JAMA : the journal of the American Medical Association 286, 2322-2324.
2. Barton, N.H., and Turelli, M. (1989). Evolutionary quantitative genetics: how little do we know? Annual review of genetics 23, 337-370.
3. Lander, E.S. (1996). The new genomics: global views of biology. Science 274, 536-539.
4. Gillespie, J.H., and Turelli, M. (1989). Genotype-environment interactions and the maintenance of polygenic variation. Genetics 121, 129-138.
5. Turelli, M. (1984). Heritable genetic variation via mutation-selection balance: Lerch's zeta meets the abdominal bristle. Theoretical population biology 25, 138-193.
6. Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Soceity of Edinburgh 52, 399-433.
7. Lande, R. (2007). The maintenance of genetic variability by mutation in a polygenic character with linked loci. Genetical research 89, 373-387.
8. Gillespie, J.H. (1994). The Causes of Molecular Evolution.(Oxford University Press).
9. Long, A.D., and Langley, C.H. (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome research 9, 720-731.
10. Marjoram, P., Zubair, A., and Nuzhdin, S.V. (2014). Post-GWAS: where next? More samples, more SNPs or more biology? Heredity 112, 79-88.
11. Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187, 367-383.
12. Collins, F.S., Guyer, M.S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. Science 278, 1580-1581.
13. Chakravarti, A. (1999). Population genetics--making sense out of sequence. Nature genetics 21, 56-60.
14. Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A., Goddard, M.E., Witte, J.S., et al. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nature genetics 45, 984-994.
15. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nature genetics 43, 519-525.
16. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature genetics 42, 565-569.

17. Hindorff LA, M.J.E.B.I., Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: http://www.genome.gov/gwastudies. Accessed July 28 2015.
18. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research 42, D1001-1006.
19. Gibson, G. (2011). Rare and common variants: twenty arguments. Nature reviews Genetics 13, 135-145.
20. Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature 456, 18-21.
21. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.
22. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature genetics 46, 1173-1186.
23. Nalls, M.A., Pankratz, N., Lill, C.M., Do, C.B., Hernandez, D.G., Saad, M., DeStefano, A.L., Kara, E., Bras, J., Sharma, M., et al. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nature genetics 46, 989-993.
24. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature 518, 197-206.
25. Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T.L., Thompson, J.R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B.A., et al. (2013). Large-scale association analysis identifies new risk loci for coronary artery disease. Nature genetics 45, 25-33.
26. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics 11, 446-450.
27. Falconer, D., and Mackay, T. (1996). Introduction to Quantitative Genetics.(Harlow, Essex, UK: Longmans Green).
28. Zammit, S., Owen, M.J., and Lewis, G. (2010). Misconceptions about gene-environment interactions in psychiatry. Evidence-based mental health 13, 65-68.
29. Cordell, H.J. (2009). Detecting gene-gene interactions that underlie human diseases. Nature reviews Genetics 10, 392-404.
30. Kraft, P., Yen, Y.C., Stram, D.O., Morrison, J., and Gauderman, W.J. (2007). Exploiting gene-environment interaction to detect genetic associations. Human heredity 63, 111-119.
31. Bearden, C.E., Reus, V.I., and Freimer, N.B. (2004). Why genetic investigation of psychiatric disorders is so difficult. Current opinion in genetics & development 14, 280-286.

32. Craddock, N., Kendler, K., Neale, M., Nurnberger, J., Purcell, S., Rietschel, M., Perlis, R., Santangelo, S.L., Schulze, T.G., Smoller, J.W., et al. (2009). Dissecting the phenotype in genome-wide association studies of psychiatric illness. The British journal of psychiatry : the journal of mental science 195, 97-99.
33. American Psychiatric Association (2000). Diagnostic and statistical manual of mental disorders, Fourth Edition, Text Revision: DSM-IV-TR. Washington, DC: American Psychiatric Publishing, Inc.
34. Beck, A.T., Steer, R.A., Ball, R., and Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. Journal of personality assessment 67, 588-597.
35. Goldstein, D.B., Allen, A., Keebler, J., Margulies, E.H., Petrou, S., Petrovski, S., and Sunyaev, S. (2013). Sequencing studies in human genetics: design and interpretation. Nature reviews Genetics 14, 460-470.
36. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? American journal of human genetics 69, 124-137.
37. Simonson, M.A., Wills, A.G., Keller, M.C., and McQueen, M.B. (2011). Recent methods for polygenic analysis of genome-wide data implicate an important effect of common variants on cardiovascular disease risk. BMC medical genetics 12, 146.
38. Morris, A.P., Voight, B.F., Teslovich, T.M., Ferreira, T., Segre, A.V., Steinthorsdottir, V., Strawbridge, R.J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics 44, 981-990.
39. Beecham, A.H., Patsopoulos, N.A., Xifara, D.K., Davis, M.F., Kemppinen, A., Cotsapas, C., Shah, T.S., Spencer, C., Booth, D., Goris, A., et al. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nature genetics 45, 1353-1360.
40. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., and Wray, N.R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nature genetics 44, 247-250.
41. Auer, P.L., and Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. Genome medicine 7, 16.
42. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Current opinion in genetics & development 19, 212-219.
43. Roth, E.M., McKenney, J.M., Hanotin, C., Asset, G., and Stein, E.A. (2012). Atorvastatin with or without an antibody to PCSK9 in primary hypercholesterolemia. The New England journal of medicine 367, 1891-1900.
44. Stein, E.A., Mellis, S., Yancopoulos, G.D., Stahl, N., Logan, D., Smith, W.B., Lisbon, E., Gutierrez, M., Webb, C., Wu, R., et al. (2012). Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. The New England journal of medicine 366, 1108-1118.
45. Kondrashov, A.S., and Crow, J.F. (1993). A molecular approach to estimating the human deleterious mutation rate. Human mutation 2, 229-234.

46. Haldane, J.B.S. (1935). The rate of spontaneous mutation of a human gene. Journal of Genetics 31, 317-326.
47. Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297-304.
48. Williams, G.C. (1957). Pleiotropy, natural selection, and the evolution of senescence. Evolution; international journal of organic evolution 11, 398-411.
49. Barton, N.H. (1990). Pleiotropic models of quantitative variation. Genetics 124, 773-782.
50. Waxman, D., and Peck, J.R. (1998). Pleiotropy and the preservation of perfection. Science 279, 1210-1213.
51. Hodgkin, J. (1998). Seven types of pleiotropy. The International journal of developmental biology 42, 501-505.
52. Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J.G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J.F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. American journal of human genetics 89, 607-618.
53. He, X., and Zhang, J. (2006). Toward a molecular understanding of pleiotropy. Genetics 173, 1885-1891.
54. Blau, N., van Spronsen, F.J., and Levy, H.L. (2010). Phenylketonuria. Lancet 376, 1417-1427.
55. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet 381, 1371-1379.
56. Cotsapas, C., Voight, B.F., Rossin, E., Lage, K., Neale, B.M., Wallace, C., Abecasis, G.R., Barrett, J.C., Behrens, T., Cho, J., et al. (2011). Pervasive sharing of genetic effects in autoimmune disease. PLoS genetics 7, e1002254.
57. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119-124.
58. Kessler, R.C., Chiu, W.T., Demler, O., Merikangas, K.R., and Walters, E.E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Archives of general psychiatry 62, 617-627.
59. Long, A.N., and Dagogo-Jack, S. (2011). Comorbidities of diabetes and hypertension: mechanisms and approach to target organ protection. J Clin Hypertens (Greenwich) 13, 244-251.
60. Solomon, D.H., Goodson, N.J., Katz, J.N., Weinblatt, M.E., Avorn, J., Setoguchi, S., Canning, C., and Schneeweiss, S. (2006). Patterns of cardiovascular risk in rheumatoid arthritis. Annals of the rheumatic diseases 65, 1608-1612.
61. Galesloot, T.E., van Steen, K., Kiemeney, L.A., Janss, L.L., and Vermeulen, S.H. (2014). A comparison of multivariate genome-wide association methods. PloS one 9, e95923.
62. Allison, D.B., Thiel, B., St Jean, P., Elston, R.C., Infante, M.C., and Schork, N.J. (1998). Multiple phenotype modeling in gene-mapping studies of

quantitative traits: power advantages. American journal of human genetics 63, 1190-1201.
63. Ferreira, M.A., and Purcell, S.M. (2009). A multivariate test of association. Bioinformatics 25, 132-133.
64. Huang, J., Johnson, A.D., and O'Donnell, C.J. (2011). PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. Bioinformatics 27, 1201-1206.
65. O'Reilly, P.F., Hoggart, C.J., Pomyen, Y., Calboli, F.C., Elliott, P., Jarvelin, M.R., and Coin, L.J. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PloS one 7, e34861.
66. Ried, J.S., Doring, A., Oexle, K., Meisinger, C., Winkelmann, J., Klopp, N., Meitinger, T., Peters, A., Suhre, K., Wichmann, H.E., et al. (2012). PSEA: Phenotype Set Enrichment Analysis--a new method for analysis of multiple phenotypes. Genetic epidemiology 36, 244-252.
67. Maity, A., Sullivan, P.F., and Tzeng, J.Y. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. Genetic epidemiology 36, 686-695.
68. Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature methods 11, 407-409.
69. Wang, Y., Liu, A., Mills, J.L., Boehnke, M., Wilson, A.F., Bailey-Wilson, J.E., Xiong, M., Wu, C.O., and Fan, R. (2015). Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genetic epidemiology 39, 259-275.
70. Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Schizophrenia Psychiatric Genome-Wide Association Study, C., International Schizophrenia, C., Molecular Genetics of Schizophrenia, C., Sullivan, P.F., Goddard, M.E., Keller, M.C., et al. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nature genetics 44, 247-250.
71. International Schizophrenia, C., Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748-752.
72. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. American journal of human genetics 88, 294-305.
73. Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nature genetics 46, 881-885.
74. Ioannidis, J.P. (2007). Non-replication and inconsistency in the genome-wide association setting. Human heredity 64, 203-213.
75. Shi, G., Gu, C.C., Kraja, A.T., Arnett, D.K., Myers, R.H., Pankow, J.S., Hunt, S.C., and Rao, D.C. (2009). Genetic effect on blood pressure is modulated by age: the Hypertension Genetic Epidemiology Network Study. Hypertension 53, 35-41.

76. Province, M.A., Tishler, P., and Rao, D.C. (1989). Repeated-measures model for the investigation of temporal trends using longitudinal family studies: application to systolic blood pressure. Genet Epidemiol 6, 333-347.
77. Simino, J., Shi, G., Bis, J.C., Chasman, D.I., Ehret, G.B., Gu, X., Guo, X., Hwang, S.J., Sijbrands, E., Smith, A.V., et al. (2014). Gene-age interactions in blood pressure regulation: a large-scale investigation with the CHARGE, Global BPgen, and ICBP Consortia. American journal of human genetics 95, 24-38.
78. Caspi, A., Sugden, K., Moffitt, T.E., Taylor, A., Craig, I.W., Harrington, H., McClay, J., Mill, J., Martin, J., Braithwaite, A., et al. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 301, 386-389.
79. Wilhelm, K., Mitchell, P.B., Niven, H., Finch, A., Wedgwood, L., Scimone, A., Blair, I.P., Parker, G., and Schofield, P.R. (2006). Life events, first depression onset and the serotonin transporter gene. The British journal of psychiatry : the journal of mental science 188, 210-215.
80. Caspi, A., McClay, J., Moffitt, T.E., Mill, J., Martin, J., Craig, I.W., Taylor, A., and Poulton, R. (2002). Role of genotype in the cycle of violence in maltreated children. Science 297, 851-854.
81. Uher, R. (2014). Gene-environment interactions in severe mental illness. Frontiers in psychiatry 5, 48.
82. Duncan, L.E., and Keller, M.C. (2011). A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. The American journal of psychiatry 168, 1041-1049.
83. Ober, C., and Vercelli, D. (2011). Gene-environment interactions in human disease: nuisance or opportunity? Trends in genetics : TIG 27, 107-115.
84. Jarvik, G.P., Goode, E.L., Austin, M.A., Auwerx, J., Deeb, S., Schellenberg, G.D., and Reed, T. (1997). Evidence that the apolipoprotein E-genotype effects on lipid levels can change with age in males: a longitudinal analysis. American journal of human genetics 61, 171-181.
85. Shaffer, J.R., Carlson, J.C., Stanley, B.O., Feingold, E., Cooper, M., Vanyukov, M.M., Maher, B.S., Slayton, R.L., Willing, M.C., Reis, S.E., et al. (2015). Effects of enamel matrix genes on dental caries are moderated by fluoride exposures. Human genetics 134, 159-167.
86. Gauderman, W.J., Zhang, P., Morrison, J.L., and Lewinger, J.P. (2013). Finding novel genes by testing G x E interactions in a genome-wide association study. Genetic epidemiology 37, 603-613.
87. Xu, X., Shi, G., and Nehorai, A. (2013). Meta-regression of gene-environment interaction in genome-wide association studies. IEEE transactions on nanobioscience 12, 354-362.
88. Kim, Y.K., Kim, Y., Hwang, M.Y., Shimokawa, K., Won, S., Kato, N., Tabara, Y., Yokota, M., Han, B.G., Lee, J.H., et al. (2014). Identification of a genetic variant at 2q12.1 associated with blood pressure in East-Asians by genome-wide scan including gene-environment interactions. BMC medical genetics 15, 65.
89. Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene

and gene-environment interactions. American journal of human genetics 79, 1002-1016.
90. Wang, H., Liu, Y., Tan, W., Zhang, Y., Zhao, N., Jiang, Y., Lin, C., Hao, B., Zhao, D., Qian, J., et al. (2008). Association of the variable number of tandem repeats polymorphism in the promoter region of the SMYD3 gene with risk of esophageal squamous cell carcinoma in relation to tobacco smoking. Cancer science 99, 787-791.
91. Liao, S.Y., Lin, X., and Christiani, D.C. (2013). Gene-environment interaction effects on lung function- a genome-wide association study within the Framingham heart study. Environmental health : a global access science source 12, 101.
92. Bradley, B., Davis, T.A., Wingo, A.P., Mercer, K.B., and Ressler, K.J. (2013). Family environment and adult resilience: contributions of positive parenting and the oxytocin receptor gene. European journal of psychotraumatology 4.
93. Binder, E.B., Bradley, R.G., Liu, W., Epstein, M.P., Deveau, T.C., Mercer, K.B., Tang, Y., Gillespie, C.F., Heim, C.M., Nemeroff, C.B., et al. (2008). Association of FKBP5 polymorphisms and childhood abuse with risk of posttraumatic stress disorder symptoms in adults. JAMA : the journal of the American Medical Association 299, 1291-1305.
94. Munafo, M.R., Zammit, S., and Flint, J. (2014). Practitioner Review: A critical perspective on gene-environment interaction models - what impact should they have on clinical perceptions and practice? Journal of child psychology and psychiatry, and allied disciplines.
95. Sullivan, P.F. (2007). Spurious genetic associations. Biological psychiatry 61, 1121-1126.
96. Hunter, D.J., and Kraft, P. (2007). Drinking from the fire hose--statistical issues in genomewide association studies. The New England journal of medicine 357, 436-439.
97. Kwee, L.C., Liu, D., Lin, X., Ghosh, D., and Epstein, M.P. (2008). A powerful and flexible multilocus association test for quantitative traits. American journal of human genetics 82, 386-397.
98. Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. American journal of human genetics 86, 929-942.
99. Maity, A., and Lin, X. (2011). Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. Biometrics 67, 1271-1284.
100. Wang, W.Y., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. Nature reviews Genetics 6, 109-118.
101. International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.
102. Tzeng, J.Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M.I., Sale, M.M., Worrall, B.B., Hsu, F.C., Thomas, D.C., and Sullivan, P.F. (2011). Studying gene and gene-environment effects of uncommon and common variants on

continuous traits: a marker-set approach using gene-trait similarity regression. American journal of human genetics 89, 277-288.
103. Conneely, K.N., and Boehnke, M. (2007). So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. American journal of human genetics 81, 1158-1168.
104. Buzkova, P., Lumley, T., and Rice, K. (2011). Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. Annals of human genetics 75, 36-45.
105. Schaid, D.J. (2010). Genomic Similarity and Kernel Methods II: Methods for Genomic Information. Human heredity 70, 132-140.
106. Schifano, E.D., Epstein, M.P., Bielak, L.F., Jhun, M.A., Kardia, S.L., Peyser, P.A., and Lin, X. (2012). SNP Set Association Analysis for Familial Data. Genetic epidemiology.
107. Breiman, L. (1995). Better subset regression using the nonnegative garrote. Technometrics 37, 373-384.
108. Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 63, 1079-1088.
109. Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. Biometrics 2, 110-114.
110. Davies, R.B. (1980). Algorithm AS 155: the distribution of a linear combination of 2 random variables. Journal of the Royal Statistical Society Series C Applied Statistics 29, 323-333.
111. Spencer, C.C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS genetics 5, e1000477.
112. Flint, J., Shifman, S., Munafo, M., and Mott, R. (2008). Genetic variants in major depression. Novartis Foundation symposium 289, 23-32; discussion 33-42, 87-93.
113. Agrawal, A., Nelson, E.C., Littlefield, A.K., Bucholz, K.K., Degenhardt, L., Henders, A.K., Madden, P.A., Martin, N.G., Montgomery, G.W., Pergadia, M.L., et al. (2012). Cannabinoid receptor genotype moderation of the effects of childhood physical abuse on anhedonia and depression. Archives of general psychiatry 69, 732-740.
114. Barrero, F.J., Ampuero, I., Morales, B., Vives, F., de Dios Luna Del Castillo, J., Hoenicka, J., and Garcia Yebenes, J. (2005). Depression in Parkinson's disease is related to a genetic polymorphism of the cannabinoid receptor gene (CNR1). The pharmacogenomics journal 5, 135-141.
115. Monteleone, P., Bifulco, M., Maina, G., Tortorella, A., Gazzerro, P., Proto, M.C., Di Filippo, C., Monteleone, F., Canestrelli, B., Buonerba, G., et al. (2010). Investigation of CNR1 and FAAH endocannabinoid gene polymorphisms in bipolar disorder and major depression. Pharmacological research : the official journal of the Italian Pharmacological Society 61, 400-404.
116. Vinod, K.Y., and Hungund, B.L. (2006). Role of the endocannabinoid system in depression and suicide. Trends in pharmacological sciences 27, 539-545.

117. Castelli, M.P., Fadda, P., Casu, A., Spano, M.S., Casti, A., Fratta, W., and Fattore, L. (2013). Male and Female Rats Differ in Brain Cannabinoid CB1 Receptor Density and Function and in Behavioural Traits Predisposing To Drug Addiction: Effect of Ovarian Hormones. Current pharmaceutical design.
118. Riebe, C.J., Hill, M.N., Lee, T.T., Hillard, C.J., and Gorzalka, B.B. (2010). Estrogenic regulation of limbic cannabinoid receptor binding. Psychoneuroendocrinology 35, 1265-1269.
119. Bradley, R.G., Binder, E.B., Epstein, M.P., Tang, Y., Nair, H.P., Liu, W., Gillespie, C.F., Berg, T., Evces, M., Newport, D.J., et al. (2008). Influence of child abuse on adult depression: moderation by the corticotropin-releasing hormone receptor gene. Archives of general psychiatry 65, 190-200.
120. Ressler, K.J., Mercer, K.B., Bradley, B., Jovanovic, T., Mahan, A., Kerley, K., Norrholm, S.D., Kilaru, V., Smith, A.K., Myers, A.J., et al. (2011). Post-traumatic stress disorder is associated with PACAP and the PAC1 receptor. Nature 470, 492-497.
121. Lin, X., Lee, S., Christiani, D.C., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics 14, 667-681.
122. Chavali, S., Barrenas, F., Kanduri, K., and Benson, M. (2010). Network properties of human disease genes with pleiotropic effects. BMC systems biology 4, 78.
123. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nature reviews Genetics 14, 483-495.
124. Lees, C.W., Barrett, J.C., Parkes, M., and Satsangi, J. (2011). New IBD genetics: common pathways with other diseases. Gut 60, 1739-1753.
125. Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M.M., Hysi, P.G., Wollstein, A., Lao, O., de Bruijne, M., Ikram, M.A., et al. (2012). A genome-wide association study identifies five loci influencing facial morphology in Europeans. PLoS genetics 8, e1002932.
126. Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., and Smola, A.J. (2008). A kernel statistical test of independence. Advances in Neural Information Processing Systems, pp 585-592.
127. Hua, W.Y., and Ghosh, D. (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. Biometrics.
128. Szekely, G.J., and Rizzo, M.L. (2009). Brownian Distance Covariance. Ann Appl Stat 3, 1236-1265.
129. Szekely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007). Measuring and testing dependence by correlation of distances. Ann Stat 35, 2769-2794.
130. Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. arXivpreprintarXiv:12023775.
131. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 89, 82-93.

132. Daniels, P.R., Kardia, S.L., Hanis, C.L., Brown, C.A., Hutchinson, R., Boerwinkle, E., and Turner, S.T. (2004). Familial aggregation of hypertension treatment and control in the Genetic Epidemiology Network of Arteriopathy (GENOA) study. The American journal of medicine 116, 676-681.
133. Zapala, M.A., and Schork, N.J. (2012). Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. Frontiers in genetics 3, 190.
134. Wessel, J., and Schork, N.J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. American journal of human genetics 79, 792-806.
135. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature methods 7, 248-249.
136. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4, 1073-1081.
137. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 46, 310-315.
138. Duchesne, P., and Lafaye De Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. Computational Statistics and Data Analysis 54, 858-862.
139. Schork, N.J., Wessel, J., and Malo, N. (2008). DNA sequence-based phenotypic association analysis. Advances in genetics 60, 195-217.
140. Wu, M.C., Maity, A., Lee, S., Simmons, E.M., Harmon, Q.E., Lin, X., Engel, S.M., Molldrem, J.J., and Armistead, P.M. (2013). Kernel machine SNP-set testing under multiple candidate kernels. Genetic epidemiology 37, 267-275.
141. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38, 904-909.
142. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature genetics 42, 348-354.
143. Schaffner, S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome research 15, 1576-1583.
144. Gao, X., Becker, L.C., Becker, D.M., Starmer, J.D., and Province, M.A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. Genetic epidemiology 34, 100-105.
145. Vattikuti, S., Guo, J., and Chow, C.C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS genetics 8, e1002637.
146. Zarkesh, M., Daneshpour, M.S., Faam, B., Fallah, M.S., Hosseinzadeh, N., Guity, K., Hosseinpanah, F., Momenan, A.A., and Azizi, F. (2012). Heritability of the

metabolic syndrome and its components in the Tehran Lipid and Glucose Study (TLGS). Genetics research 94, 331-337.
147. Hottenga, J.J., Boomsma, D.I., Kupper, N., Posthuma, D., Snieder, H., Willemsen, G., and de Geus, E.J. (2005). Heritability and stability of resting blood pressure. Twin research and human genetics : the official journal of the International Society for Twin Studies 8, 499-508.
148. Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., Hwang, S.J., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 478, 103-109.
149. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707-713.
150. Zabaneh, D., and Balding, D.J. (2010). A genome-wide association study of the metabolic syndrome in Indian Asian men. PloS one 5, e11961.
151. Kraja, A.T., Vaidya, D., Pankow, J.S., Goodarzi, M.O., Assimes, T.L., Kullo, I.J., Sovio, U., Mathias, R.A., Sun, Y.V., Franceschini, N., et al. (2011). A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium. Diabetes 60, 1329-1339.
152. Avery, C.L., He, Q., North, K.E., Ambite, J.L., Boerwinkle, E., Fornage, M., Hindorff, L.A., Kooperberg, C., Meigs, J.B., Pankow, J.S., et al. (2011). A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. PLoS genetics 7, e1002322.
153. Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U., Lamina, C., et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nature genetics 41, 25-34.
154. Herbert, A., Gerry, N.P., McQueen, M.B., Heid, I.M., Pfeufer, A., Illig, T., Wichmann, H.E., Meitinger, T., Hunter, D., Hu, F.B., et al. (2006). A common genetic variant is associated with adult and childhood obesity. Science 312, 279-283.
155. Fall, T., and Ingelsson, E. (2014). Genome-wide association studies of obesity and metabolic syndrome. Molecular and cellular endocrinology 382, 740-757.
156. Lange, L.A., Lange, E.M., Bielak, L.F., Langefeld, C.D., Kardia, S.L., Royston, P., Turner, S.T., Sheedy, P.F., 2nd, Boerwinkle, E., and Peyser, P.A. (2002). Autosomal genome-wide scan for coronary artery calcification loci in sibships at high risk for hypertension. Arteriosclerosis, thrombosis, and vascular biology 22, 418-423.
157. Epstein, M.P., Duren, W.L., and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. American journal of human genetics 67, 1219-1231.
158. Moskvina, V., and Schmidt, K.M. (2008). On multiple-testing correction in genome-wide association studies. Genetic epidemiology 32, 567-573.

159. Tregouet, D.A., Barbaux, S., Escolano, S., Tahri, N., Golmard, J.L., Tiret, L., and Cambien, F. (2002). Specific haplotypes of the P-selectin gene are associated with myocardial infarction. Human molecular genetics 11, 2015-2023.
160. Herrmann, S.M., Ricard, S., Nicaud, V., Mallet, C., Evans, A., Ruidavets, J.B., Arveiler, D., Luc, G., and Cambien, F. (1998). The P-selectin gene is highly polymorphic: reduced frequency of the Pro715 allele carriers in patients with myocardial infarction. Human molecular genetics 7, 1277-1284.
161. Zee, R.Y., Cook, N.R., Cheng, S., Reynolds, R., Erlich, H.A., Lindpaintner, K., and Ridker, P.M. (2004). Polymorphism in the P-selectin and interleukin-4 genes as determinants of stroke: a population-based, prospective genetic analysis. Human molecular genetics 13, 389-396.
162. Davi, G., Romano, M., Mezzetti, A., Procopio, A., Iacobelli, S., Antidormi, T., Bucciarelli, T., Alessandrini, P., Cuccurullo, F., and Bittolo Bon, G. (1998). Increased levels of soluble P-selectin in hypercholesterolemic patients. Circulation 97, 953-957.
163. Ikeda, H., Takajo, Y., Ichiki, K., Ueno, T., Maki, S., Noda, T., Sugi, K., and Imaizumi, T. (1995). Increased soluble form of P-selectin in patients with unstable angina. Circulation 92, 1693-1696.
164. Koyama, H., Maeno, T., Fukumoto, S., Shoji, T., Yamane, T., Yokoyama, H., Emoto, M., Shoji, T., Tahara, H., Inaba, M., et al. (2003). Platelet P-selectin expression is associated with atherosclerotic wall thickness in carotid artery in humans. Circulation 108, 524-529.
165. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Magi, R., et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature genetics 42, 937-948.
166. Visscher, P.M., Yang, J., and Goddard, M.E. (2010). A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). Twin research and human genetics : the official journal of the International Society for Twin Studies 13, 517-524.
167. Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S.E., Liewald, D., Ke, X., Le Hellard, S., Christoforou, A., Luciano, M., et al. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular psychiatry 16, 996-1005.
168. Lee, S.H., Harold, D., Nyholt, D.R., Goddard, M.E., Zondervan, K.T., Williams, J., Montgomery, G.W., Wray, N.R., and Visscher, P.M. (2013). Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Human molecular genetics 22, 832-841.
169. Plomin, R., Haworth, C.M., Meaburn, E.L., Price, T.S., and Davis, O.S. (2013). Common DNA markers can account for more than half of the genetic influence on cognitive abilities. Psychological science 24, 562-568.
170. Desachy, G., Croen, L.A., Torres, A.R., Kharrazi, M., Delorenze, G.N., Windham, G.C., Yoshida, C.K., and Weiss, L.A. (2015). Increased female autosomal burden of rare copy number variants in human populations and in autism families. Molecular psychiatry 20, 170-175.

171. Krumm, N., Turner, T.N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B.P., Stessman, H.A., He, Z.X., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nature genetics 47, 582-588.
172. Jiang, Y., Conneely, K.N., and Epstein, M.P. (2014). Flexible and robust methods for rare-variant testing of quantitative traits in trios and nuclear families. Genetic epidemiology 38, 542-551.
173. Abecasis, G.R., Cardon, L.R., and Cookson, W.O. (2000). A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66, 279-292.
174. Lee, S., Teslovich, T.M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. American journal of human genetics 93, 42-53.
175. van der Sluis, S., Verhage, M., Posthuma, D., and Dolan, C.V. (2010). Phenotypic complexity, measurement bias, and poor phenotypic resolution contribute to the missing heritability problem in genetic association studies. PloS one 5, e13929.
176. van der Sluis, S., Posthuma, D., and Dolan, C.V. (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS genetics 9, e1003235.
177. Van der Sluis, S., Dolan, C.V., Li, J., Song, Y., Sham, P., Posthuma, D., and Li, M.X. (2015). MGAS: a powerful tool for multivariate gene-based genome-wide association analysis. Bioinformatics 31, 1007-1015.
178. Minica, C.C., Boomsma, D.I., van der Sluis, S., and Dolan, C.V. (2010). Genetic association in multivariate phenotypic data: power in five models. Twin research and human genetics : the official journal of the International Society for Twin Studies 13, 525-543.
179. Medland, S.E., and Neale, M.C. (2010). An integrated phenomic approach to multivariate allelic association. European journal of human genetics : EJHG 18, 233-239.
180. He, L.N., Liu, Y.J., Xiao, P., Zhang, L., Guo, Y., Yang, T.L., Zhao, L.J., Drees, B., Hamilton, J., Deng, H.Y., et al. (2008). Genomewide linkage scan for combined obesity phenotypes using principal component analysis. Annals of human genetics 72, 319-326.
181. Zhang, F., Guo, X., Wu, S., Han, J., Liu, Y., Shen, H., and Deng, H.W. (2012). Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses. PloS one 7, e53320.
182. Aschard, H., Vilhjalmsson, B.J., Greliche, N., Morange, P.E., Tregouet, D.A., and Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. American journal of human genetics 94, 662-676.
183. Basu, S., Zhang, Y., Ray, D., Miller, M.B., Iacono, W.G., and McGue, M. (2013). A rapid gene-based genome-wide association test with multivariate traits. Human heredity 76, 53-63.
184. Broadaway, K.A., Cutler, D.J., Duncan, R., Moore, J.L., Ware, E.B., Jhun, M.A., Bielak, L.F., Zhao, W., Smith, J.A., Peyser, P.A., Kardia, S.L.R., Epstein, M.P.

(2015). A statistical approach for testing pleiotropic effects of rare variants. Manuscript submitted for publication, under revision.
185. Cole, D.A., Maxwell, S.E., Arvey, R., Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among dependent variables. Psychological bulletin 115, 465-474.
186. Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M.A., Gaulton, K.J., Albers, P.K., McVean, G., Boehnke, M., Altshuler, D., et al. (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS genetics 11, e1005165.
187. Bansal, V., Libiger, O., Torkamani, A., and Schork, N.J. (2011). An application and empirical comparison of statistical analysis methods for associating rare variants to a complex phenotype. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 76-87.
188. Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. Genetic epidemiology 35, 606-619.
189. Fridley, B.L., and Biernacka, J.M. (2011). Gene set analysis of SNP data: benefits, challenges, and future directions. European journal of human genetics : EJHG 19, 837-843.
190. Huang, H., Chanda, P., Alonso, A., Bader, J.S., and Arking, D.E. (2011). Gene-based tests of association. PLoS genetics 7, e1002177.
191. Ge, T., Nichols, T.E., Ghosh, D., Mormino, E.C., Smoller, J.W., and Sabuncu, M.R. (2015). A kernel machine method for detecting effects of interaction between multidimensional variable sets: an imaging genetics application. NeuroImage 109, 505-514.
192. Zhao, N., Chen, J., Carroll, I.M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J.J., Ringel, Y., Li, H., and Wu, M.C. (2015). Testing in Microbiome-Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. American journal of human genetics 96, 797-807.