Statistical Methods for Multiple Phenotypes and Gene-Set Association Analysis

by

Diptavo Dutta

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Seunggeun Lee, Chair
Professor Michael Boehnke
Professor Laura J. Scott
Professor Ananda Sen

Diptavo Dutta

diptavo@umich.edu

ORCID iD: 0000-0002-6634-9040

To Baba, Maa and Anwesha

ACKNOWLEDGEMENTS

I am extremely grateful to the Department of Biostatistics and the Center for Statistical Genetics for allowing me to be a part of this world-leading program and providing all the amenities, including a rich spectrum of real-world data and computing resources which has been essential in my dissertation.

I would like to express my deepest gratitude to my committee chair and advisor Seunggeun (Shawn) Lee for his support and guidance throughout my doctoral studies. His deep insights into the field of statistical genetics remain an inspiration and his mentorship has motivated me to pave my path as a researcher. He has been extremely accommodating and patient which helped me mature as an individual as well. I truly look up to him as a role model in my forthcoming years. I am deeply indebted to Laura J. Scott for her continued support throughout the five years first as research mentor and then as dissertation committee member. Every discussion and interaction with her has enriched me as a researcher and helped me to find the right balance in my research. Her dedication and enthusiasm has motivated me and will continue to do so throughout my career. I would also like to express my heartfelt thanks to Michael Boehnke. From the day I secured admission into this department he has been someone I learned a lot from and has provided many valuable insights as my dissertation committee member. I am extremely grateful to Ananda Sen for his constant encouragement, thoughtful advice and motivational guidance as a member of my dissertation committee.

Being from a different background, it had not been easy for me to merge into

applied research from the onset. This was facilitated by the rich collaborative environments in the department and my brief time as a research intern at AbbVie. I have been fortunate to have worked with Chad M. Brummett and Daniel Clauw as a research assistant in a project related to fibromyalgia. Stephanie Moser has also provided tremendous support in preparing and understanding the data. I'm also grateful to Lars Fritsche, Matthew Zawistowski and Joshua Weinstock for their help in analyzing of the Michigan Genomics Initiative data in context to multiple projects and also to Sarah Gagliano Taliun for her helpful comments on the SardiNIA data.

I am grateful to many people including Fan Zhang, Corbin Quick, Adrian Tan, Yeji Lee, Alan Kwong, Vincent Tan, Chris Lee, Andy Liu, Sai Dharmarajan, Pranav Yajnik, Zhangchen Zhao, Wenjian Bi and all the wonderful members of the Lee Lab and Center for Statistical Genetics. It was a really rewarding and enriching experience to have been a part of such intellectually stimulating groups of faculty and students. Special thanks to the ISI community in Ann Arbor including Sayantan Das, Sebanti Sengupta, Rounak Dey, Aritra Guha, Shrijita Bhattacharya, Debarghya Mukherjee, Moulinath Banerjee, Ananda Sen and Bhramar Mukherjee for all the fun hangouts, theater-carnivals, late-night parties and "*adda*" (for the lack of an appropriate translation). I am thankful to several of my friends from high school, St. Xavier's College and ISI including Aritro Pathak, Indranil Sahoo, Shreyan Ganguly, Abhijoy Saha, Arkopal Choudhury, Swagato Mukherjee, Arnab Kr. Pal, Sayan Das, Arindam Bhattacharya, Mainak Mitra, Arijit De, Charbak Das, Abhishek Poddar, Rituparno Guha, Songlap Saha, Kasturi Chatterjee, Ankur Lahiri, Rishika Sen, Kaustav Nandy, Dhritiman Gupta and many others for being a part of this journey in their own way.

I cannot begin to express my thanks to my family: Ma (Mukuta Dutta) and Baba (Sanjib Dutta) for their continuous encouragement and support throughout the last five years and showing faith in me whenever I doubted myself. They have been the

guiding stars in my life and although, Baba is no more with us, he and Ma were the ones who motivated me to pursue my doctoral studies and have inspired me in my moments of disappointment and anguish. Their work ethics and lessons of life have led me to achieve whatever I have in these years away from them. Finally I would like express my sincerest thanks to my fiancée and best friend Anwesha Bhattacharyya who has been a constant source of love, inspiration and motivation in my life for the past few years. Her being by my side has made my days in Ann Arbor that much more rewarding and full of life. Without her support, especially through a few turbulent months, it would have been a lot more difficult journey in graduate school.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

As association studies continue to advance, more efficient statistical methods are required to fully utilize existing data and to provide insight into genetic architecture of complex traits. Identifying association for a set of phenotypes or with respect to a set of variants can be particularly useful for understanding how biological networks might be affecting the patho-physiology of outcomes. In this dissertation, I attempt to develop computationally efficient statistical methods that facilitate insights into the mechanism of complex traits and understanding of their underlying biology.

In Chapter II, I propose a generalized framework for gene-based tests with multiple correlated phenotypes. In genetic association analysis, a joint test of multiple correlated phenotypes can increase power to identify sets of trait-associated variants within genes or regions of interest. Existing multi-phenotype tests for rare variants make specific assumptions about the patterns of association with underlying causal variants and the violation of these assumptions can reduce power to detect association. In this project we develop a general framework for testing pleiotropic effects of rare variants on multiple continuous phenotypes using multivariate kernel regression (Multi-SKAT). To increase power of detecting association across tests with different kernel matrices, we developed a fast and accurate approximation of the significance of the minimum observed p-value across tests. To account for related individuals, our framework uses random effects for the kinship matrix. Using simulated and exome-array data from the METSIM study, we show that Multi-SKAT can increase power over single-phenotype SKAT-O test and existing multiple phenotype tests, while maintaining type I error rate.

In Chapter III, I extend Multi-SKAT to a meta-analysis strategy, namely Meta-MultiSKAT, to combine results from several studies. Our method involves extracting score statistics and phenotype adjusted variant relationship matrix from individual studies which are then combined using a kernel that models the heterogeneity of effects between the studies. The proposed method accommodates situations where one or more phenotypes have not been observed in a particular study and studies have different correlation patterns among the phenotypes. With minor modifications our method can be used to test the combined effects of common and rare variants in a region, as well as incorporate functional information on individual variants. Meta-analysis of 4 white blood cell subtype traits from the MGI and SardiNIA studies show that Meta-MultiSKAT can identify associations which were not identified by existing methods and were not significant in individual studies.

In Chapter IV, I propose a subset based approach for gene-set (pathway) association analysis using variant level summary statistics. Existing gene-set association (GSA) methods can have low statistical power when only a small fraction of the genes is associated with the phenotype and interpreting results in terms of the underlying genetic mechanism can be challenging since they cannot identify possible active genes within the set. For this, we propose a maximum-type statistic that selects the subset of genes with maximal evidence of association as the driver-genes and evaluates its significance using efficient simulation techniques. Through the analysis of summary statistics from the UK Biobank data for 1201 phenotypes with 10679 gene-sets, we show that our method can be used to identify novel associations across a large number of phenotypes and gene-sets.

CHAPTER I

Introduction

In the past decade, biological datasets have increased in both size and scope due to the dramatic developments in high-throughput technologies. In particular, for quantitative human genetics, the availability of data on high-density single nucleotide polymorphisms (SNPs) through micro-arrays provided unprecedented opportunities, especially to design large-scale genome-wide association studies (GWAS). In a standard GWAS, the association between a genetic variant and the phenotype is evaluated typically through a regression model. These tests are carried out for millions of SNPs across the genome [Bush and Moore, 2012, Visscher et al., 2012]. The results can highlight the region(s) in the genome which might be significantly associated with the phenotype.

One of the earlier genetic association studies in 2005 had a relatively small sample size of 96 age-related macular degeneration cases and 50 controls [Haines et al., 2005, Edwards et al., 2005, Klein et al., 2005]. But its results led to increasing interest in this field and was followed by several other GWASs including that of Wellcome Trust Case Control Consortium (WTCCC) in 2007 [The Wellcome Trust Case Control Consortium, 2007]. Since then, the GWAS datasets have expanded at a fast pace identifying thousands of genetic variants associated with hundreds of different traits and continues to grow even now. For example, the current version of UK-Biobank

study has data on more than $500,000$ individuals for over $1,500$ phenotypes [Bycroft et al., 2018]. GWAS findings have now been reported for a numerous complex traits across several domains, including common diseases, complex traits, gene-expression and brain-imaging traits [Visscher et al., 2017]. As of May 30^{th}, 2019, GWAS catalog provides information on more than 120,000 reported associations across more than $2,500$ traits [Buniello et al., 2019].

Although array-based technologies has produced genotype data on millions of variants across thousands of samples within an affordable price range [Marchini and Howie, 2010, LaFramboise, 2009], the advent of cost-effective sequencing technology has further transformed the landscape of GWAS [Cirulli and Goldstein, 2010]. Current sequencing studies allows inferences to be based on the full set of variants for a given trait and fuels downstream analysis of their functional consequences. In particular, this allows researchers to evaulate the role of rare-variants in detail, which was not possible with the previous micro-array-based genotyping. However, the statistical methods for analysis of GWAS data are not well suited for detecting rare-variant associations. Especially in studies that do not have a huge sample size, these statistical methods are usually underpowered to detect rare-variant association. This has presented the research community with a need to develop novel statistical methods to analyze data generated by large-scale sequencing initiatives.

1.1 Rare variant association (RVA) studies

Despite the fact that large-scale meta-analysis and biobank-scale association analysis are being conducted, most of the genetic variants identified to date have low effect sizes and explain only a small proportion of the trait heritability. For example, a study on type 2 diabetes involving a sample of around 650,000 individuals identified 143 genome-wide significant loci but explained only about 21% of trait heritability [Xue et al., 2018]. Several explanation have been put forth for this problem of missing

heritability [Eichler et al., 2010, Zuk et al., 2012]. One of the possible explanations is that low-frequency ($1\% \leq \text{MAF} \leq 5\%$) and rare ($\text{MAF} \leq 1\%$) variants could explain additional disease risk or trait variability. Rare variants are known to play an important role in Mendelian disorders, rare forms of common diseases [Gibson, 2012] as well as complex diseases [Gudmundsson et al., 2012]. Additionally, evelutionary theory predicts that deleterious alleles are likely to be rare due to purifying selection [Kryukov et al., 2009]. This is confirmed by the fact that, loss-of-function variants, which prevent the generation of functional proteins, are especially rare [The 1000 Genomes Project Consortium, 2012, MacArthur et al., 2012]. Hence it is logical to think that rare or low frequency variants are involved in the genetic architecture of a given trait and hence incorporating them in the analysis could increase the fraction of the trait heritability that is explained by genetic variants.

Although, whole-genome or whole-exome sequencing have made it possible for researchers to analyze rare variants, empirical studies have shown that the standard statistical framework of GWAS is underpowered to detect associations with rare or low-frequency variants, unless the sample size or the effect size of the variant is very large. To boost power, region-based collapsing or binning approaches have become a standard for analyzing rare variants. In this approach, the variants are first grouped into biologically relevant regions (or genes) and then the association of joint effect of multiple rare variants in the region with the outcome is evaluated. Such groupings can increase the power to detect moderate to weak effects by the accumulating single-SNP effects within the region or gene and by lowering the multiple testing burden.

Several novel statistical methods have been developed to investigate the associations between rare variants within a region and a trait. For example, the burden test [Li and Leal, 2008] collapses variants in a gene or functional region into a single score and then tests the association between this collapsed score and the trait. The Sequence Kernel Association Test (SKAT) [Wu et al., 2011] employs a mixed effects

model and conducts a variance components test for the association. These methods and others have identified regions or genes harboring rare-variants associated to phenotypes like hypertriglyceridemia, type-2 diabetes and alzheimer's disease.

Inspite of the advancements made by RVA studies, the dramatic growth of genetic datasets continue to present the research community with several further interesting directions for research. One primary example of such problem, is that of detecting cross-phenotype (pleiotropy) effects.

1.2 Detecting pleiotropy

Pleiotropy occurs when one gene has an effect on multiple phenotypes [Sivakumaran et al., 2011, Li et al., 2014, Yang et al., 2015]. Twin and family studies have long provided evidence for genetic correlations among diseases (such as major depressive disorder and generalized anxiety disorder [Kendler et al., 1992] or rheumatoid arthritis and systemic lupus erythematosus [Criswell et al., 2005]), suggesting a role for pleiotropic genetic effects. In addition, the co-occurrence of multiple diseases in the same individual (for example, type 1 diabetes and autoimmune thyroid disease [Eaton et al., 2007]) point to shared genetic causes. The genetic loci identified through GWAS also show that variants can be associated with multiple and distinct phenotypes. There are numerous examples where the same variants show association with multiple traits (rs6983267 associated with prostate and colorectal cancer [Tomlinson et al., 2007, Thomas et al., 2008]); A recent evaluation of genome-wide-significant single-nucleotide polymorphisms (SNPs) listed in the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies found that 4.6% of SNPs and 16.9% of genes have cross phenotype effects [Sivakumaran et al., 2011, Yang et al., 2015]. Thus it remains important to identify pleiotropic effects for better understanding the biological mechanism underlying a complex trait.

Currently, most cross-phenotype association methods are designed to assess the effect of a single polymorphism at a time and have low power for detecting rare-variants with pleiotropic effects. Recently, several rare-variant cross-phenotype tests have been proposed in literature as well. For example, [Wang et al., 2015] proposed a gene-level test of pleiotropy that uses multivariate functional linear models (MFLM); [Broadaway et al., 2016] used a matrix-similarity based approach to test for cross phenotype effects of rare variants (GAMuT); [Wu and Pankow, 2016] developed a score based kernel association test for multiple traits, MSKAT, which has similar performance to GAMuT; and [Zhan et al., 2017] proposed DKAT, which uses the similarity-based approach as in GAMuT but is suitable for high-dimensional of phenotypes.

Despite these developments, existing methods have certain major limitations. Most methods were developed under a set of specific distributional assumptions regarding the effects of the variants on multiple phenotypes. Hence, they lose power if the actual underlying structures in the data are significantly different from them. Although there has been an recent attempt to make the association results more robust by combining analysis results across different models [Zhan et al., 2017], computational scalability of such methods for genome-wide applications is yet to be achieved. Further, very few of the methods can adjust for relatedness between individuals which is a common occurrence in many of the current association studies. Thus, to apply these methods, related individuals must be removed from the analysis to maintain type I error rate. This can cause a substantial loss in sample size and hence power, especially for datasets where related individuals are present.

In Chapters II and III, we address these problems and propose novel and computationally efficient methods to detect rare-variants with cross-phenotype effects.

1.3 Gene-set or Pathway association (GSA) analysis

Besides having rare and low frequency variants associated with a trait, one other possible explanation for missing heritability might be that the trait is highly polygenic with many weak to moderate associations with genetic variants which standard single-variant or single-gene analysis might be underpowered to detect [Manolio et al., 2009]. Indeed, with large number of genetic polymorphisms examined in GWAS and the massive amount of tests conducted, such real but weak associations are likely to be missed after multiple comparison adjustment [Liu et al., 2010]. Studies have found that the cumulative effect of a large number of weakly associated SNPs, most of which are not statistically significant on their own, can predict disease status or symptoms, for example in psychiatric conditions [Wray et al., 2014]. This highlights the role of such weak to moderate associations that are not identified by GWAS.

Gene set analysis (GSA), also termed pathway analysis [Cantor et al., 2010], has been suggested as a more powerful alternative in situations where standard GWAS is underpowered to detect weaker associations. In GSA, we group individual genes or variants according to biological or functional characteristics and then test the association of the trait with the gene-set for significance. This considerably reduces the number of tests that need to be performed and hence decreases the multiple testing burden.

GSA can provide a number of benefits when used as a tool for secondary analysis of a GWAS data set. Because of the polygenic nature of complex diseases, testing for association with sets of functionally related genes or variants can provide biological context for multiple genetic risk factors and can provide insights into disease mechanisms and possible treatment targets [Pers, 2016]. Further, by cumulating signals across numerous variants in the gene-set, GSA can increase the statistical power of detecting weaker signals that would otherwise be missed by GWAS [Fridley and Biernacka, 2011, Yu et al., 2009].

There is an extensive literature on numerous methods to perform GSA analysis and they have identified several novel gene-sets associated with complex traits like obesity [Locke et al., 2015], height [Allen et al., 2010] and bipolar disorder [Nurnberger et al., 2014], yielding important biological insights. However these methods have certain important limitations. For example, existing GSA methods have been known to have low power [Jia et al., 2011], especially when only a small fraction of the genes is associated with the phenotype. Further, several methods cannot maintain proper type-I error in presence of linkage disequilibrium between the variants or gene-gene correlations. Although p-values can be computed using permutations or simulations in such situations, but this can be both time and memory consuming [Moskvina et al., 2012, Holmans, 2010].

Another key challenge, which not many methods have tried to address, is the question of interpretation. Standard GSA analyses fail to identify the genes which individually or in congregation might be driving the association signal. Thus, although weak associations are identified through such GSA methods, it is difficult to target individual genes for follow-up analysis, especially for large gene-sets. A common practice is to recommend the top few genes with the lowest p-values as the active genes. However, the lack of any principled data-driven approach can lead to false results and difficulty in interpretation.

To address these issues, we introduce a powerful novel GSA method in Chapter IV, that can adaptively select the driver-set of genes within the set. Our method is computationally efficient and can be applied to large biobank-scale datasets.

1.4 Overview of the dissertation

The aim of this dissertation is to propose novel methodological solutions to the problems in pleiotropy and gene-set association analysis as stated above. We have also considered the fact that these methods should be computationally scalable to

be applied to current genomic datasets. The methods we have developed have been implemented in R-based softwares and made publicly available through public repositories and github. We have provided links to the softwares corresponding to each chapter at the end of each chapter correspondingly.

In Chapter II, we propose a computationally efficient and novel method to detect cross phenotype effects of rare variants, given individual level data in a study. Our method, Multi-SKAT is a general framework that extends variance components tests to a multivariate mixed-effects model framework. Mixed-effects models have been widely used for rare-variant association tests with a single phenotype such as SKAT [Wu et al., 2011] and SKAT-O [Lee et al., 2012b]. For multiple phenotypes, we use a phenotype kernel matrix (Σ_P) which models the relationship between the effect sizes of a particular variant in the region and the phenotypes. By using kernels to relate genetic variants to multiple continuous phenotypes, Multi-SKAT allows for flexible modeling of the genetic effects on the phenotypes. To avoid power loss due to model misspecification, we develop minimum p-value based omnibus tests that can aggregate results across different choices of kernels. It can be shown that several existing methods like GAMuT, MAAUSS and MF-KM are special cases of Multi-SKAT under particular choices of the phenotype kernel matrix. Through extensive simulations, we showed that Multi-SKAT can increase power over single-phenotype SKAT-O test and existing multiple phenotype tests while maintaining type-I error. We applied Multi-SKAT to exome-array data from the METSIM (Metabolic Syndrome in Men) study, where Multi-SKAT identified several novel rare-variant associations in addition to identifying those identified by standard methods.

In Chapter III, we extend Multi-SKAT to a meta-analysis strategy (Meta-MultiSKAT) to combine summary statistics from several studies. Meta-analysis of multiple studies, using association summary statistics, is a practical approach to increase power by increasing sample sizes [Panagiotou et al., 2013] which is especially important for

rare-variants. Various methods have been developed for meta-analysis of multiple phenotypes [Majumdar et al., 2018, Ray and Boehnke, 2018, Zhu et al., 2015], but most of them are single variant-based methods, which have low power to identify rare variant associations. To the best of our knowledge, Meta-MultiSKAT is the first method that focuses on rare-variants in this context. Our method involves sharing of score statistics and a phenotype adjusted variant relationship matrix from individual studies which are then combined using a kernel that models the heterogeneity of effects between the studies. To achieve robust power under different association models, we developed fast and accurate omnibus tests by combining different models of genetic effects and functional genomic annotation. Additionally, Meta-MultiSKAT accommodates situations where studies do not share exactly the same set of phenotypes or have differing correlation patterns among the phenotypes. Simulation studies confirm that Meta-MultiSKAT can maintain type-I error rate. Further simulations under different models of association show that Meta-MultiSKAT can substantially improve power of detection over single phenotype-based meta-analysis approaches. We demonstrate the utility and improved power of Meta-MultiSKAT in the meta-analyses of four white blood cell subtype traits from the Michigan Genomics Initiative (MGI) and SardiNIA studies.

In Chapter IV, we introduce Gene-Set association Using Sparse Signals (GAUSS), a method for GSA with summary statistics using a maximum-type statistic. GAUSS additionally selects the subset of genes with the maximal evidence of association as the driver-genes. The p-value for GAUSS, despite being simulation-based, can be efficiently calculated by using pre-computed matrices using a reference data. Simulations show that GAUSS can increase power over several existing methods while controlling type-I error under a variety of association models. Through the analysis of summary statistics from the UK Biobank data for 1201 phenotypes with 10679 gene-sets, we show that GAUSS can be used to identify novel associations across a

large number of phenotypes and gene-sets. Additionally, similarities or differences in genetic mechanism of phenotypes can be investigated by phenome-wide association analysis of a given gene-set, which has been unexplored to date.

Finally, in Chapter V, we discuss the implications of this dissertation and propose potential directions of future research.

CHAPTER II

Multi-SKAT: General framework to test for rare variant association with multiple phenotypes

2.1 Introduction

Since the advent of array genotyping technologies, genome-wide association studies (GWAS) have identified numerous genetic variants associated with complex traits. Despite these discoveries, GWAS loci explain only a small proportion of heritability for most traits. This may be due, in part, to the fact that these association studies are underpowered to identify associations with rare variants [Korte and Farlow, 2013]. To identify such rare variant associations, gene- or region-based multiple variant tests have been developed [Lee et al., 2014]. By jointly testing rare variants in a target gene or region, these methods can increase power over a single variant test and are now used as a standard approach in rare variant analysis.

Recent GWAS results have shown that many GWAS loci are associated with multiple traits [Solovieff et al., 2013], which is called pleiotropy or cross-phenotypic associations. Nearly 17% of variants in National Heart Lung and Blood Institute (NHLBI) GWAS categories are associated with multiple traits [Sivakumaran et al., 2011, Li et al., 2014, Yang et al., 2015]. For example, 44% of autoimmune risk single nucleotide polymorphisms (SNPs) have been estimated to be associated with two

or more autoimmune diseases [Cotsapas et al., 2011]. Detecting such pleiotropic effects is important to understand the underlying biological structure of complex traits. In addition, by leveraging cross-phenotype associations, the power to detect trait-associated variants can be increased.

Identifying the cross-phenotype effects requires a suitable joint or multivariate analysis framework that can incorporate information on the correlation of the phenotypes. Various methods have been proposed for multiple phenotype analysis in GWAS [Ferreira and Purcell, 2009, Huang et al., 2011, Zhou and Stephens, 2014a, Ried et al., 2012, Ray et al., 2016] primarily aimed at detecting cross-phenotype associations of common variants. Extending them, several groups have developed multiple phenotype tests for rare variants [Wang et al., 2015, Broadaway et al., 2016, Wu and Pankow, 2016, Lee et al., 2016, Sun et al., 2016, Maity et al., 2012, Yan et al., 2015, Zhan et al., 2017]. For example, [Wang et al., 2015] proposed a multivariate functional linear model (MFLM); [Broadaway et al., 2016] used a dual-kernel based distance-covariance approach to test for cross-phenotype effects of rare variants by comparing similarity in multivariate phenotypes to similarity in genetic variants (GAMuT); [Wu and Pankow, 2016] developed a score based sequence kernel association test for multiple traits, MSKAT, which has been shown to be similar in performance to GAMuT [Broadaway et al., 2016]; and [Zhan et al., 2017] proposed DKAT, which uses the dual kernel approach as in GAMuT but provides more robust performance when the dimension of phenotypes is high compared to the sample size.

Despite these developments, existing methods have certain limitations. Most methods were developed under specific assumptions regarding the effects of the variants on multiple phenotypes, and hence lose power if these assumptions are violated [Ray et al., 2016]. For example, if genetic effects are heterogeneous across multiple phenotypes, methods assuming homogeneous genetic effects can lose a substantial amount of power [Lee et al., 2016]. Although there has been a recent attempt to

combine analysis results from different models [Zhan et al., 2017], no computationally scalable methods have been developed to evaluate the significance of the combined results in genome-wide scale analysis. In addition, most existing methods and software cannot adjust for relatedness between individuals; thus, to apply these methods, related individuals must be removed from the analysis to maintain type I error rate. For example, in the METabolic Syndrome In Men (METSIM) study $\sim 15\%$ of individuals are estimated to be related up to the second degree. Thus to apply the existing methods on the data from METSIM study, we need to remove these individuals which will result in a lower sample size.

In this project, we develop Multi-SKAT, a general framework that extends the mixed effect model-based kernel association tests to a multivariate regression framework while accounting for family relatedness. Mixed effect models have been widely used for rare-variant association tests. Popular rare variant tests such as SKAT [Wu et al., 2011] and SKAT-O [Lee et al., 2012b] are based on mixed effect models. By using kernels to relate genetic variants to multiple continuous phenotypes, Multi-SKAT allows for flexible modeling of the genetic effects on the phenotypes. The idea of using kernels for genotypes and phenotypes was previously used by the dual kernel approaches such as GAMuT and DKAT. However, in contrast to these two similarity-based methods, Multi-SKAT is multivariate regression based and hence provides a natural way to adjust for covariates and also can account for sample relatedness by incorporating random effects for the kinship matrix. Many of the existing methods for multiple phenotype rare variant tests can be viewed as special cases of Multi-SKAT with particular choices of kernels. Furthermore, to avoid loss of power due to model misspecification, we develop computationally efficient omnibus tests, which allow for aggregation of tests over several kernels and provide fast p-value calculation.

In the next section, we present the multivariate mixed effect model and kernel

matrices. We particularly focus on the phenotype-kernel and describe omnibus procedures that can aggregate results across different choices of kernels and kinship adjustment. Following that, we describe the simulation experiments that demonstrate that Multi-SKAT tests have increased power to detect associations compared to existing methods like GAMuT, MSKAT and others under a wide range of association models. Finally we describe the results from applying Multi-SKAT on a set of nine amino acids measured on 8,545 Finnish men from the METSIM study, to detect the cross-phenotype effects of rare nonsynonymous and protein-truncating variants.

2.2 Material and Methods

2.2.1 Single-phenotype region-based tests

To describe the Multi-SKAT tests, we first present the existing model of the single-phenotype gene or region-based tests. Let $y_k = (y_{1k}, y_{2k}, \cdots, y_{nk})^T$ be an $n \times 1$ vector of the k^{th} phenotype over n individuals; X an $n \times q$ matrix of the q non-genetic covariates including the intercept; $G_j = (G_{1j}, \cdots, G_{nj})^T$ is an $n \times 1$ vector of the minor allele counts (0, 1, or 2) for a binary genetic variant j; and $G = [G_1, \cdots, G_m]$ is an $n \times m$ genotype matrix for m genetic variants in a target region. The regression model shown in equation (2.1) can relate m genetic variants to phenotype k,

$$y_k = X\alpha_k + G\beta_{.k} + \epsilon_k \tag{2.1}$$

where α_k is a $q \times 1$ vector of regression coefficients of q non-genetic covariates (can include top principal components to account for population structure), $\beta_{.k} = (\beta_{1k}, \cdots, \beta_{mk})^T$ is an $m \times 1$ vector of regression coefficients of the m genetic variants, and ϵ_k is an $n \times 1$ vector of non-systematic error term with each element following $N(0, \sigma_k^2)$. To test for $H_0 : \beta_{.k} = 0$, a variance component test under the mixed effects model have been proposed to increase power over the usual F-test [Wu et al., 2011].

The variance component test assumes that the regression coefficients, $\beta_{.k}$, are random variables and follow a centered distribution with covariance matrix $\tau^2\Sigma_G$, where Σ_G is an $m \times m$ matrix (for details on Σ_G see below). Under these assumptions, the test for $\beta_{.k} = 0$ is equivalent to testing $\tau = 0$. The score statistic for this test is

$$Q = (y_k - \hat{\mu}_k)^T G \Sigma_G G^T (y_k - \hat{\mu}_k) \tag{2.2}$$

where $\hat{\mu}_k = X\hat{\alpha}_k$ is the estimated mean of y_k under the null hypothesis of no association. The test statistic Q asymptotically follows a mixture of chi-squared distributions under the null hypothesis and p-values can be computed by inverting the characteristic function [Davies, 1980]. Although there are other methods to approximate the p-value using for a mixture of chi-square distributions, Davies' methods appears to work well in practice and is widely used in this context [Wu et al., 2011].

The kernel matrix Σ_G plays a critical role; it models the relationship among the effect sizes of the variants on the phenotypes. Any positive semidefinite matrix can be used for Σ_G providing a unified framework for the region-based tests. A frequent choice of Σ_G is a sandwich type matrix $\Sigma_G = WR_GW$, where $W = diag(w_1, .., w_m)$ is a diagonal weighting matrix for each variant, and R_G is a correlation matrix between the effect sizes of the variants. $R_G = I_{m\times m}$ implies uncorrelated effect sizes and corresponds to SKAT, and $R_G = 1_m {1_m}^T$ corresponds to the Burden test, where $I_{m\times m}$ is an $m \times m$ diagonal matrix and $1_m = (1, \cdots 1)^T$ is an $m \times 1$ vector with all elements being unity. Furthermore, a linear combination of these two matrices corresponds to $R_G = \rho 1_m {1_m}^T + (1-\rho) I_{m\times m}$, which is used for SKAT-O [Lee et al., 2012b].

2.2.2 Multiple-phenotype region-based tests

Extending the idea of using kernels, we build a model for multiple phenotypes. The multivariate linear model shown in equation (2.3) can relate genetic variants to K correlated phenotypes,

$$Y = XA + GB + E \tag{2.3}$$

where $Y = (y_1, \cdots, y_K)$ is an $n \times K$ phenotype matrix; A is a $q \times K$ matrix of coefficients of X; $B = (\beta_{ij})$ is an $m \times K$ matrix of coefficients where β_{ij} denotes the effect of the i^{th} variant on the j^{th} phenotype and E is an $n \times K$ matrix of non-systematic errors. Let $vec(\cdot)$ denote the matrix vectorization function, and then $vec(E)$ follows $N(0, I_n \otimes V)$, where V is a $K \times K$ covariance matrix and $\otimes$ represents the Kronecker product.

In addition to assuming that $\beta_{.k}$ follows a centered distribution with covariance $\tau^2\Sigma_G$, we further assume that $\beta_{i.} = (\beta_{i1}, \cdots, \beta_{iK})^T$, which is the vector of regression coefficients of variant i for K multiple phenotypes, follows a centered distribution with covariance $\tau^2\Sigma_P$, which implies that $vec(B)$ follows a centered distribution with covariance $\tau^2\Sigma_G \otimes \Sigma_P$. As before, the null hypothesis $H_0 : vec(B) = 0$ is equivalent to $\tau = 0$. The corresponding score test statistic is

$$Q = \{vec(Y) - vec(\widehat{\mu})\}^T \left\{(G\Sigma_G G^T) \otimes (\widehat{V}^{-1}\Sigma_P\widehat{V}^{-1})\right\} \{vec(Y) - vec(\widehat{\mu})\} \tag{2.4}$$

where $\widehat{\mu}$ and $\widehat{V}$ are the estimated mean and covariance of Y under the null hypothesis.

Σ_P plays a similar role as Σ_G but with respect to phenotypes. Σ_P represents a kernel in the phenotypes space and models the covariance between the effect sizes of a variant on each of the phenotypes. Any positive semidefinite matrix can be used as Σ_P.

The proposed approach provides a double-flexibility in modeling. Through the choice of structures for Σ_G and Σ_P, we can control the dependencies of genetic

effects. However, the kernel matrices Σ_G and Σ_P are nuisance parameters in the model and cannot be estiamted from the data. Additionally, similar to SKAT, the use of a sandwich type matrix WR_GW for Σ_G allows us to upweight rare variants by using $Beta(1, 25)$ weights as in [Wu et al., 2011]. Most of our hypotheses about the underlying genetic structure of a set of phenotypes can be modeled through varying structures of these two matrices.

2.2.3 Phenotype kernel structure Σ_P

The use of Σ_G has been extensively studied previously in literature [Wu et al., 2011, Lee et al., 2012b]. Here we propose several choices for Σ_P and study their effect from a modeling perspective.

Homogeneous (Hom) Kernel

It is possible that effect sizes of a variant on different phenotypes are homogeneous, in which case $\beta_{j1} = \cdots = \beta_{jK}$. Under this assumption,

$$\Sigma_{P,Hom} = 1_K 1_K^T \tag{2.5}$$

Under $\Sigma_{P,Hom}$, the effect sizes $\beta_{jk}, (k = 1, \cdots K)$ for a variant j are the same for all the phenotypes.

Heterogeneous (Het) Kernel

Effect sizes of a variant on different phenotypes can be heterogeneous in which $\beta_{j1} \neq \cdots \neq \beta_{jK}$. Under this assumption, we can construct

$$\Sigma_{P,Het} = I_{k\times k} \tag{2.6}$$

The $\Sigma_{P,Het}$ implies that the effect sizes $(\beta_{j1}, \cdots, \beta_{jK}^T)$ are uncorrelated among them-

selves. This also indicates that the correlation among the phenotypes is not affected by this particular region or gene.

Phenotype Covariance (PhC) Kernel

We may model Σ_P as proportional to the estimated residual covariance across the phenotypes as,

$$\Sigma_{P,PhC} = \widehat{V} \tag{2.7}$$

where $\widehat{V}$ is the estimated covariance matrix among the phenotypes. This model assumes that the covariance between the effect sizes is proportional to that between the residual phenotypes after adjusting for the non-genetic covariates.

Principal Component (PC) Kernel

Principal component analysis (PCA) is a popular tool for multivariate analysis. In multiple phenotype tests, PC-based approaches have been used to reduce the dimension in phenotypes [Aschard et al., 2014]. Here we show that PC-based approach can be included in our framework. Let $L = (L_1, \cdots, L_K)$ be the loading matrix with each column L_i produces the i^{th} PC score. In Appendix A.1, we show that using $\Sigma_{P,PC} = \widehat{V} L \widehat{V_P}^{-1} \widehat{V_P}^{-1} L^T \widehat{V}$ is equivalent to assuming heterogeneous effects with all PCs as phenotypes. Instead of using all the PC's, we can use selected PC's that represent the majority of cumulative variation in phenotypes. For example, we can jointly test the PC's that have cumulative variance of 90%. If the top t PC's have been chosen for analysis using ν% cumulative variance as cutoff, we can use

$$\Sigma_{P,PC-\nu} = \widehat{V} L_{sel} \widehat{V}_P^{-1} \widehat{V}_P^{-1} L_{sel}^T \widehat{V}$$

where $L_{sel} = [L_1, \cdots, L_t, 0, \cdots, 0]$ and 0 represents a vector of 0's of appropriate length.

Relationship with other Multiple-Phenotype rare variant tests

We have proposed a uniform framework of Multi-SKAT tests that depend on the kernels Σ_G and Σ_P. There are certain specific choices of these kernel matrices that correspond to other published methods.

- Using $\Sigma_{P,PhC}$ and $\Sigma_G = WI_mW^T$ is identical to the GAMuT[Broadaway et al., 2016] with the projection phenotype kernel and the MSKAT with the Q statistic [Wu and Pankow, 2016].
- Using $\Sigma_P = \widehat{V}^2$ and $\Sigma_G = WI_mW^T$ is identical to GAMuT[Broadaway et al., 2016] with the linear phenotype kernel and the MSKAT with the Q' statistic [Wu and Pankow, 2016].
- Using $\Sigma_{P,Hom}$ and $\Sigma_G = WI_mW^T$ is identical to hom-MAAUSS [Lee et al., 2016].
- Using $\Sigma_{P,Het}$ and $\Sigma_G = WI_mW^T$ is identical to het-MAAUSS [Lee et al., 2016] and MF-KM [Yan et al., 2015].

For the detailed proof, please see Appendix A.2.

2.2.4 Minimum p-value based omnibus tests (minP and minP$_{\text{com}}$)

The model and the corresponding test of association that we proposed through the Multi-SKAT test statistic (2.4) has two parameters, Σ_G and Σ_P, which are absent in the null model of no association. Since Multi-SKAT is a score test, Σ_G and Σ_P cannot be estimated from the data. One possible approach is to select Σ_G and Σ_P based on prior knowledge; however, if the selected Σ_G and Σ_P do not reflect underlying biology, the test may have substantially reduced power [Ray et al., 2016,

Lee et al., 2016]. In an attempt to achieve robust power, we aggregate results across different Σ_G and Σ_P using the minimum of p-values from different kernels.

Although this omnibus test approach has been used in rare variant tests and multiple phenotype analysis for combining multiple kernels from genotypes and phenotypes [Zhan et al., 2017, Wu et al., 2013, Urrutia et al., 2015, He et al., 2017], it is challenging to calculate the p-value, since the minimum p-value does not follow the uniform distribution. One possible approach is using permutation or perturbation to calculate the monte-carlo p-value [Urrutia et al., 2015, Zhan et al., 2017]; however, this approach is computationally too expensive to be used in genome-wide analysis. To address it, here we propose a fast copula based p-value calculation for Multi-SKAT, which needs only a small number of resampling steps to calculate the p-value.

Suppose p_h is the p-value for Q_h with given h^{th} Σ_G and Σ_P, $h = 1, \cdots, b$, and $T_P = (p_1, \cdots, p_b)^T$ is an $b \times 1$ vector of p-values of b such Multi-SKAT tests. The minimum p-value test statistic after the Bonferroni adjustment is $b \times p_{min}$, where p_{min} is the minimum of the b p-values. In the presence of positive correlation among the tests, this approach is conservative and hence might lack power of detection. Rather than using Bonferroni corrected p_{min}, more accurate results can be obtained if the joint null distribution or more specifically the correlation structure of T_P can be estimated and incorporated in the test. Here we adopt a resampling based approach to estimate this correlation structure. Note that our test statistic is equivalent to

$$Q = S^T \left\{ (G\Sigma_G G^T) \otimes (\widehat{V}^{-\frac{1}{2}} \Sigma_P \widehat{V}^{-\frac{1}{2}}) \right\} S, \tag{2.8}$$

where $S = (I_n \otimes \widehat{V}^{-\frac{1}{2}}) \{vec(Y) - vec(\widehat{\mu})\}$. Under the null hypothesis S approximately follows an uncorrelated multivariate normal distribution $N(0, I_{nK})$. Using this, we propose the following resampling algorithm

- Step 1. Generate nK samples from an $N(0,1)$ distribution, say S_R.
- Step 2. Calculate b different test statistics as $Q_R = S_R^T \left\{ (G\Sigma_G G^T) \otimes (\widehat{V}^{-\frac{1}{2}} \Sigma_P \widehat{V}^{-\frac{1}{2}}) \right\} S_R$ for all the choices of Σ_P and calculate p-values.
- Step 3. Repeat the previous steps independently for $R(= 1000)$ iterations, and calculate the correlation between the p-values of the tests from the R resampling p-values.

With the estimated null correlation structure, we use a Copula to approximate the joint distribution of T_P [Demarta and McNeil, 2005, He et al., 2017]. Copula is a statistical approach to construct joint multivariate distribution using marginal distribution of each variable and correlation structure. Since marginally each test statistic Q follows a mixture of chi-square distributions, which has a heavier tail than normal distribution, we propose to use a t-Copula to approximate the joint distribution, i.e, we assume the joint distribution of T_P to be multivariate t with the estimated correlation structure. The final p-value for association is then calculated from the distribution function of the assumed t-Copula.

When calculating the correlation across the p-values, Pearson's correlation coefficient can be unreliable since it depends on normality and homoscedasticity assumptions. To avoid such assumptions we recommend estimating the null-correlation matrix of the p-values through Kendall's tau (τ), which is a non-parametric approach based on concordance of ranks. The correlation matrix can be reliably estimated in a small number of iterations (≤ 1000).

The minimum p-value approach can be used to combine different Σ_P given Σ_G, or combine both Σ_P and Σ_G. For example two Σ_G's corresponding to SKAT (WI_mW) and Burden kernels ($W1_m1_m^TW$) and four Σ_P's ($\Sigma_{P,Hom}$, $\Sigma_{P,Het}$, $\Sigma_{P,PhC}$, $\Sigma_{P,PC-0.9}$) can be combined, which results in the omnibus test of these eight different tests. To differentiate the latter, we will call it minP_{com} which combines SKAT and Burden

type kernels of Σ_G.

2.2.5 Adjusting for relatedness

We formulated equation (2.3) and corresponding tests under the assumption of independent individuals. If individuals are related, this assumption is no longer valid, and the tests may have inflated type I error rate. Since our method is regression-based, we can relax the independence assumption by introducing a random effect term to account for the relatedness among individuals.

Let Φ be the kinship matrix of the individuals and V_g is a co-heritability matrix, denoting the shared heritability between the phenotypes. Extending the model presented in equation(2.3), we incorporate Φ and V_g as

$$Y = XA + GB + Z + E \tag{2.9}$$

where Z is an $n \times K$ matrix with $vec(Z)$ following $N(0, \Phi \otimes V_g)$. Z represents a matrix of random effects arising from shared genetic effects between individuals due to the relatedness. The remaining terms are the same as in equation (2.3). The corresponding score test statistic is

$$Q_{Kin} = S_{Kin}^T \widehat{V}_e^{-1/2} \left\{ (G\Sigma_G G^T) \otimes \Sigma_P \right\} \widehat{V}_e^{-1/2} S_{Kin} \tag{2.10}$$

where $S_{Kin} = \widehat{V}_e^{-1/2} \left\{ vec(Y) - vec(\widehat{\mu}) \right\}$ and $\widehat{V}_e = \Phi \otimes \widehat{V}_g + I_n \otimes \widehat{V}$ is the estimated covariance matrix of $vec(Y)$ under the null hypothesis. Similar to the previous versions for unrelated individuals, Q_{Kin} asymptotically follows a mixture of chi-square under the assumption of no association.

This approach depends on the estimation of the matrices Φ, V_g and V. The kinship matrix Φ can be estimated using the genome-wide genotype data [Manichaikul et al., 2010a]. Several of the published methods like LD-Score [Bulik-Sullivan et al.,

2015], PHENIX [Dahl et al., 2016] and GEMMA [Zhou and Stephens, 2014a, Zhou et al., 2013] can jointly estimate V_g and V. In our numerical analysis, we have used PHENIX, which implements a generalized restricted maximum likelihood estimation using individual level data to accurately estimate V_g and V. This is an efficient method to fit local maximum likelihood variance components in a multiple phenotype mixed model through an E-M algorithm.

Once the matrices Φ, V_g and V are estimated, we compute the asymptotic p-values for Q_{Kin} by using a mixture of chi-square distributions. The computation of Q_{Kin} requires large matrix multiplications, which can be time and memory consuming. To reduce computational burden, we employ several transformations. We perform an eigen-decomposition on the kinship matrix Φ as $\Phi = U\Lambda U^T$, where U is an orthogonal matrix of eigenvectors and Λ is a diagonal matrix of corresponding eigenvalues. We obtain the transformed phenotype matrix as $\widetilde{Y} = YU$, the transformed covariate matrix as $\widetilde{X} = XU$, the transformed random effects matrix $\widetilde{Z} = ZU$ and transformed residual error matrix $\widetilde{E} = EU$. Equation (2.9) can be transformed into

$$\widetilde{Y} = \widetilde{X}A + \widetilde{G}B + \widetilde{Z} + \widetilde{E}; \quad vec(\widetilde{Z}) \sim N(0, \Lambda \otimes V_g); \quad vec(\widetilde{E}) \sim N(0, I \otimes V) \quad (2.11)$$

All the properties of the tests developed from equation (2.3) are directly applicable to those from equation (2.11). Q_{Kin} can be computed from this transformed equation as,

$$Q_{Kin} = \widetilde{S}_{Kin}^T \widetilde{V}_e^{-1/2} \left\{ (\widetilde{G}\Sigma_G \widetilde{G}^T) \otimes \Sigma_P \right\} \widetilde{V}_e^{-1/2} \widetilde{S}_{Kin}, \quad (2.12)$$

where $\widetilde{S}_{Kin} = \widetilde{V}_e^{-1/2} \left\{ vec(\widetilde{Y}) - vec(\widetilde{\mu}) \right\}$, $\widetilde{\mu}$ is the estimated mean of $\widetilde{Y}$ under the null hypothesis and $\widetilde{V}_e = \Lambda \otimes \widehat{V}_g + I_n \otimes \widehat{V}$. Asymptotic p-values can be obtained from the corresponding mixture of chi-squares distribution. Further, omnibus strategies for the tests developed from equation (2.3) are applicable in this case with similar modifications. For example, the resampling algorithm for minimum p-value based

omnibus test can be implemented here as well by noting that $\widetilde{S}_{Kin}$ approximately follows an uncorrelated multivariate normal distribution.

2.3 Simulations

We carried out extensive simulation studies to evaluate the type I error and power of Multi-SKAT tests. For type I error simulations without related individuals and all power simulations, we generated 10,000 chromosomes over 1Mbp regions using a coalescent simulator with European demographic model [Schaffner et al., 2005]. The MAF spectrum of the simulated variants is shown in Appendix Figure A.6, showing that most of the variants are rare. We randomly selected a 3 kbps sub-region for each simulated dataset to test for associations. For the type I error simulations with related individuals, to have a realistic kinship structure, we used the METSIM study genotype data.

We generated phenotypes from the multivariate normal distribution as

$$y_i \sim MVN\{(\beta_1 G_1 + \cdots + \beta_m G_m)I, V\} \tag{2.13}$$

where $y_i = (y_{i1}, \cdots, y_{iK})^T$ is the outcome vector, G_j is the genotype of the j^{th} variant, and β_j is the corresponding effect size, and V is a covariance of the non-systematic error term. We use V to define level of covariance between the traits. I is a $k \times 1$ indicator vector, which has 1 when the corresponding phenotype is associated with the region and 0 otherwise. For example, if there are 5 phenotypes and the last three are associated with the region, $I = (0, 0, 1, 1, 1)^T$.

To evaluate whether Multi-SKAT can control type I error under realistic scenarios, we simulated a dataset with 9 phenotypes with a correlation structure identical to that of 9 amino acid phenotypes in the METSIM data (See Appendix Figure A.1). Phenotypes were generated using equation (2.13) with $\beta = 0$. Total $5,000,000$

datasets with $5,000$ individuals were generated to obtain the empirical type-I error rates at $\alpha = 10^{-4}, 10^{-5}$ and 2.5×10^{-6}, which are corresponding to candidate gene studies to test for 500 and 5000 genes and exome-wide studies to test for all $20,000$ protein coding genes, respectively.

Next, we evaluated type I error controls in the presence of related individuals. To have a realistic related structure we used the METSIM study genotype data. We generated a random subsample of 5000 individuals from the METSIM study individuals and generated null values for the 9 phenotypes from $MVN(0, V_e)$, where $V_e = \Phi_{5k} \otimes \widehat{V}_{g;5k} + I \otimes \widehat{V}_{5k}$, Φ_{5k} is the estimated kinship matrix of the 5000 selected individuals, $\widehat{V}_{g;5k}$ and $\widehat{V}_{5k}$ are estimated co-heritability and residual variance matrices respectively for these individuals as estimated using the MPMM function in the PHENIX R-package (version 1.0). For each set of 9 phenotypes, we performed the Multi-SKAT tests for a randomly selected 5000 genes in the METSIM data (For the details on the METSIM data, see next section). We carried out this procedure 1000 times and obtained $5,000,000$ p-values, and estimated type I error rate as proportions of p-values smaller than the given level α.

Our simulation studies focus on evaluating the power of the proposed tests when the number of phenotypes are 5 or 6. We chose the number of phenotypes to be relatively small since the METSIM data also has a small (9) number of phenotypes. We performed power simulations both in situations when there was no pleiotropy (i.e., only one of the phenotypes was associated with the causal variants) and also when there was pleiotropy. Under pleiotropy, since it is unlikely that all the phenotypes are associated with genotypes in the region, we varied the number of phenotypes associated. For each associated phenotype, 30% or 50% of the rare variants (MAF $< 1\%$) were randomly selected to be causal variants. We modeled the rarer variants to have stronger effect, as $|\beta_j| = c|log_{10}(MAF_j)|$. We used $c = 0.3$ which yields $|\beta_j| = 0.9$ for variants with MAF$= 10^{-3}$. Our choice of β yielded the average heritability of asso-

ciated phenotypes between 1% to 4%. We also considered situations that all causal variants were trait-increasing variants (i.e. positive β) or 20% of causal variants were trait-decreasing variants (i.e. negative β). Empirical power was estimated from 1000 independent datasets at exome-wide $\alpha = 2.5 \times 10^{-6}$.

In type I error and power simulations, we compared the following tests:

- Bonferroni adjusted minimum p-values from gene-based test (SKAT, Burden or SKAT-O) on each phenotype (minPhen)
- Multi-SKAT with $\Sigma_{P,Hom}$ (Hom)
- Multi-SKAT with $\Sigma_{P,Het}$ (Het)
- Multi-SKAT with $\Sigma_{P,PhC}$ (PhC)
- Multi-SKAT with $\Sigma_{P,PC-0.9}$ (PC-Sel)
- Minimum P-value of Hom, Het, PhC and PC-Sel using Copula (minP)
- Minimum P-value of Hom, Het, PhC and PC-Sel with Σ_G being SKAT and Burden, using Copula (minP_{com})

For the Multi-SKAT tests, we used two different Σ_G's corresponding to SKAT (i.e. $\Sigma_G = WW$) and Burden tests (i.e. $\Sigma_G = W1_m1_m^TW$). For the variant weighting matrix $W = diag(w_1, \cdots, w_m)$, we used $w_j = Beta(MAF_j, 1, 25)$ function to upweight rarer variants, as recommended by [Wu et al., 2011].

2.3.1 Computation Time

We estimated the computation time of Multi-SKAT tests and the existing methods. Using simulated datasets of 5000 related and unrelated individuals with 10 phenotypes and 20 genetic variants, we estimated the computation time of Multi-SKAT tests with and without kinship adjustments. To compare the computation

performance of Multi-SKAT tests with the existing methods (GAMuT and MSKAT), we generated datasets of unrelated individuals with five different sample sizes (N = 1000, 2000, 5000, 10000, 15000 and 20000) and four different number of variants ($m = 10, 20, 50, 100$). For each simulation setup, we generated 100 datasets and obtained the average value of the computation time.

2.3.2 Analysis of the METSIM study exomechip data

To investigate the cross-phenotype roles of low frequency and rare variants on amino acids, we analyzed data on 8545 participants of the METSIM study on whom 9 amino acids (Alanine, Leucine, Isoleucine, Glycine,Valine, Tyrosine, Phenylalanine, Glutamine, Histidine) were measured by proton nuclear magnetic resonance spectroscopy[Teslovich et al., 2018]. Individuals were genotyped on the Illumina ExomeChip and OmniExpress arrays and we included individuals that passed sample QC filters [Huyghe et al., 2013]. The kinship between the individuals was estimated via KING (version 2.0) [Manichaikul et al., 2010a]. We adjusted the amino acid levels for age, age^2 and BMI and inverse-normalized the residuals. The phenotype correlation matrix after covariate adjustment is shown in Appendix Figure A.1. Subsequently, we estimated the genetic heritability matrix and the residual covariance matrix using the MPMM function from PHENIX [Dahl et al., 2016] R package.

We included rare (MAF $< 1\%$) nonsynonymous and protein-truncating variants in our analysis. To avoid the effect of singletons or results purely driven by single-variant effect, we only considered the genes with a total rare minor allele count of at least 5 for genes that had at least 3 variants leaving 5207 genes for analysis. We set a stringent significance threshold at 9.6×10^{-6} corresponding to the Bonferroni adjustment for 5207 genes. Further, we also considered a less stringent threshold of 10^{-4}, corresponding to a candidate gene study of 500 genes, as suggestive to study the associations which were not significant but close to the threshold.

2.4 Results

2.4.1 Type I Error simulations

We estimated empirical type I error rates of the Multi-SKAT tests with and without related individuals. For unrelated individuals, we simulated $5,000$ individuals and 9 phenotypes based on the correlation structure for the amino acids phenotypes in the METSIM study data. For related individuals, we simulated $5,000$ individuals using the kinship matrix for randomly chosen METSIM individuals (see the Method section). We performed association tests and estimated type I error rate as the proportion of p-values less than the specified α levels. Type I error rates of the Multi-SKAT tests were well maintained at $\alpha = 10^{-4}, 10^{-5}$ and 2.5×10^{-6} for both unrelated and related individuals (Table 2.1), which correspond to candidate gene studies of 500 and 5000 genes and exome-wide studies to test for all $20,000$ protein coding genes, respectively. For example, at level $\alpha = 2.5 \times 10^{-6}$, the largest empirical type I error rate from any of the Multi-SKAT tests was 3.4×10^{-6}, which was within the 95% confidence interval (CI = $(1.6 \times 10^{-6}, 4 \times 10^{-6})$).

2.4.2 Power simulations

We compared the empirical power of the minPhen (Bonferroni adjusted minimum p-value for the phenotypes) and Multi-SKAT tests. For each simulation setting, we generated $1,000$ sequence datasets of $5,000$ unrelated individuals and for each test estimated empirical power as the proportion of p-values less than $\alpha = 2.5 \times 10^{-6}$, reflecting Bonferroni correction for testing $20,000$ independent genes. Since the Hom and Het tests are identical to hom-MAAUSS and het-MAAUSS, respectively, and using PhC is identical to both GAMuT (with projection phenotype kernel) and MSKAT, our power simulation studies effectively compare majority of the existing multiple phenotype tests.

In Figure 2.1, we show the results for 5 phenotypes with compound symmetric correlation structure with the correlation (ρ) being 0.3 or 0.7, where 30% of rare variants (MAF < 0.01) were positively associated with $1, 2$ or 3 phenotypes. Since it is unlikely that all the phenotypes are associated with the region, we restricted the number of associated phenotypes to at most 3. In most scenarios, PhC, PC-Sel and Het had greater power among the Multi-SKAT tests with fixed phenotype kernels (i.e. Hom, Het, PhC and PC-Sel) while minP, maintained high power as well. For example, when the correlation between the phenotypes was 0.3 (i.e. $\rho = 0.3$) and SKAT kernel was used for the genotype kernel Σ_G, if 3 phenotypes were associated with the region, minP and PhC were more powerful than the other tests. If the correlation between the phenotypes was $\rho = 0.7$ and Burden kernel was used for genotype kernel Σ_G, Het, PC-Sel and minP had higher power than the rest of the tests when 2 phenotypes were associated. It is noteworthy that Hom had the lowest power in all the scenarios of Figure 2.1.

Figure 2.2 demonstrates scenarios involving 6 phenotypes and clustered correlation structures where PhC was outperformed by other choices of the phenotype kernel Σ_P. When all three phenotype clusters had associated phenotypes and the correlation within the clusters was low ($\rho = 0.3$) (Figure 2.2, upper panel), Hom and minP tests outperformed PhC when the SKAT kernel was used. This may be because that the phenotype correlation structure did not reflect the genetic association pattern. When 2 small clusters had high within-cluster correlation ($\rho = 0.7$) and one large cluster had low within-cluster correlation ($\rho = 0.3$) (Figure 2.2, lower panel), Het and minP had higher power than PhC.

When 20% of causal variants were trait-decreasing variants (80% trait-increasing), the power of Multi-SKAT tests with Burden Σ_G was reduced (Appendix Figure A.2 and A.3). This is because the association signals were attenuated due to presence of both trait-increasing and trait-decreasing variants. Since SKAT is robust regardless

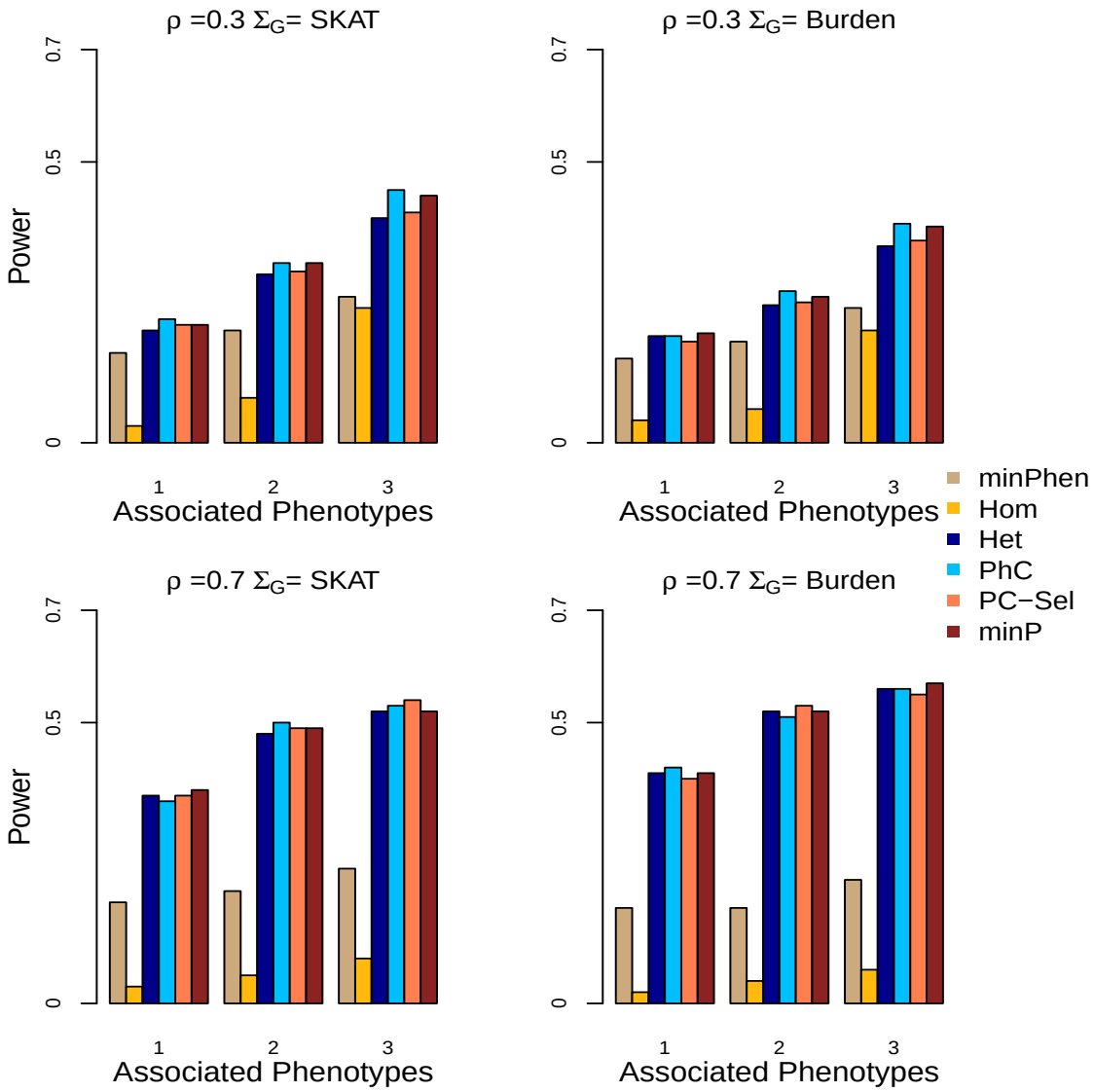


Figure 2.1: Power for Multi-SKAT tests when phenotypes have compound symmetric correlation structures. Empirical power for minPhen, Hom, Het, PhC, PC-Sel, minP plotted against the number of phenotypes associated with the gene of interest with a total of 5 phenotypes under consideration. Upper row shows the results for $\rho = 0.3$ and lower row for $\rho = 0.7$. Left column shows results with SKAT kernel Σ_G, and right columns shows results with Burden kernel. All the causal variants were trait-increasing variants.

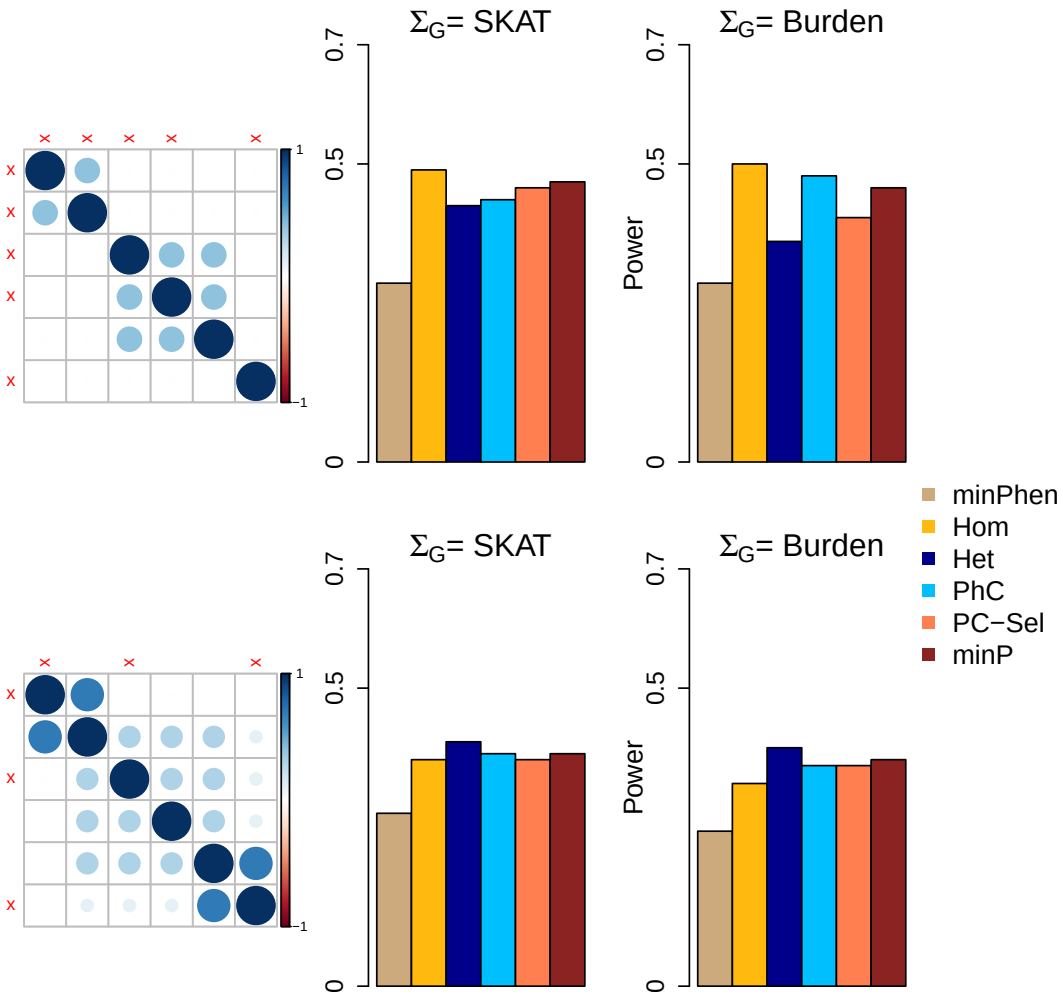


Figure 2.2: Power for Multi-SKAT tests when phenotypes have clustered correlation structures. Empirical powers for minPhen, Hom, Het, PhC, PC-Sel, minP are plotted under different levels of association with a total of 6 phenotypes and with clustered correlation structures. Middle column shows the empirical powers for different combinations of phenotypes associated with SKAT kernel Σ_G; the rightmost column shows the corresponding results with Burden kernel; left column shows the corresponding correlation matrices for the phenotypes. The associated phenotypes are indicated in red cross marks across the correlation matrices. All the causal variants were trait-increasing variants.

of the association direction, power with SKAT Σ_G was largely maintained. The relative performance of methods with different Σ_P given Σ_G was quantitatively similar to the results without trait-decreasing variants.

Further, we estimated power of minP_{com}, which combines tests across phenotype (Σ_P) and genotype Σ_G kernels. The power of minP_{com} was evaluated for the compound symmetric phenotype correlation structure presented in Figure 2.1 and was compared with the two minP tests of SKAT (minP-SKAT) and Burden (minP-Burden) Σ_G kernels. Figure 2.3 shows empirical power with and without trait-decreasing variants. When all genetic effect coefficients were positive (Figure 2.3, left panel) the performances of minP-SKAT and minP-Burden were similar for both the situations where the correlation between the phenotypes were low (i.e. $\rho = 0.3$) and high (i.e. $\rho = 0.7$). When 20% of genetic effect coefficients were negative (Figure 2.3, right panel), as expected, the power of minP-Burden was substantially decreased. Across all the situations, the power of minP_{com} was similar to the most powerful minP with fixed genotype kernel Σ_G. When 50% of variants were causal variants and all genetic effect coefficients were positive (Appendix Figure A.4, left panel), minP-Burden was more powerful than minP-SKAT, and minP_{com} had similar power than minP-Burden.

Overall, our simulation results show that the omnibus tests, especially minP_{com}, had robust power throughout all the simulation scenarios considered. When Σ_G and Σ_P were fixed, power depended on the model of association and the correlation structure of the phenotypes. Overall, the proposed Multi-SKAT tests generally outperformed the single phenotype test (minPhen), even when only one phenotype was associated with genetic variants.

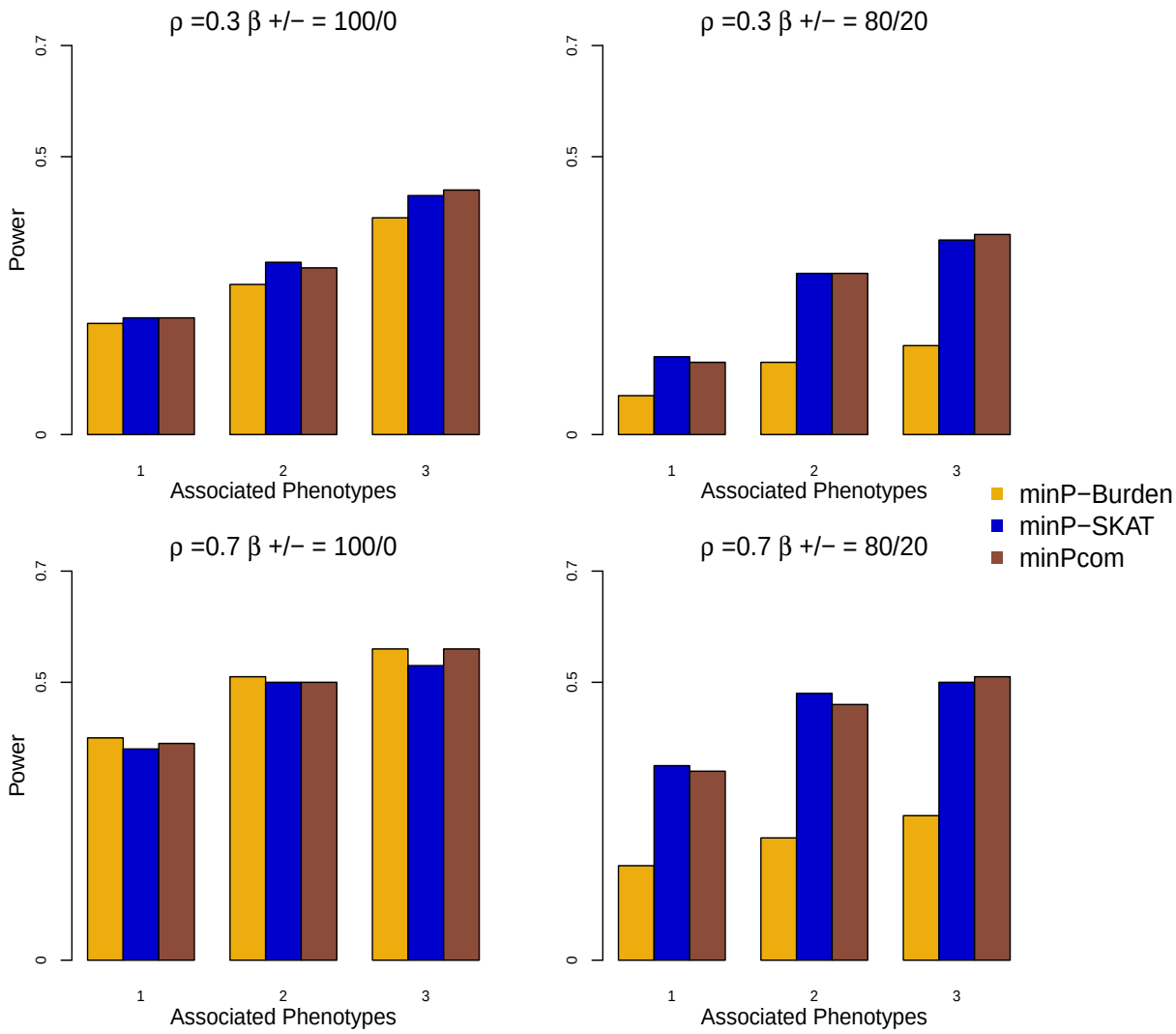


Figure 2.3: Power for Multi-SKAT by combining tests with Σ_P as Hom, Het, PhC, PC-Sel and Σ_G as SKAT and Burden when phenotypes have compound symmetric correlation structures. Empirical powers for minP-Burden, minP-SKAT and minP_{com} are plotted against the number of phenotypes associated with the gene of interest with a total of 5 phenotypes under consideration. Upper row shows the results for $\rho = 0.3$ and lower row for $\rho = 0.7$. Left column shows results when all the causal variants were trait-increasing variants, and right column shows results when 80%/20% of the causal variants were trait-increasing/trait-decreasing variants.

2.4.3 Application to the METSIM study exomechip data

Inborn errors of amino acid metabolism cause mild to severe symptoms including type 2 diabetes [Stankov et al., 2012, Wrtz et al., 2012, 2013] and liver diseases [Tajiri and Shimizu, 2013] among others. Amino acid levels are perturbed in certain disease states, e.g., glutamic and aspartic acid levels are reduced in Alzheimer disease brains [Allan Butterfield and Pocernich, 2003]; Isoleucine, glycine, alanine, phenylalanine, and threonine levels are increased in cerebo-spinal fluid (CSF) of individuals with motor neuron disease [de Belleroche et al., 2003]. To find rare variants associated with the 9 measured amino acid levels, we applied the Multi-SKAT tests to the METSIM study data [Teslovich et al., 2018]. The MAF spectrum of the genotyped variants is shown in Appendix Figure A.6, showing that most of the variants are rare variants. We estimated the relatedness between individuals by KING [Manichaikul et al., 2010a], and coheritability of the amino acid phenotypes and the corresponding residual variance using PHENIX [Dahl et al., 2016] (Appendix Figure A.1). Among the 8,545 METSIM participants with non-missing phenotypes and covariates, 1,332 individuals had a second degree or closer relationship with one or more of the METSIM participants. A total of $5,207$ genes with at least three rare variants were included in our analysis. The Bonferroni corrected significance threshold was $\alpha = 0.05/5207 = 9.6 \times 10^{-6}$. Further we used a less significant cutoff of $\alpha = 10^{-4}$ for a gene to be suggestive. After identifying associated genes, we carried out backward elimination procedure (Appendix A.3, Appendix Table A.2) to investigate which phenotypes are associated with the gene. This procedure iteratively removes phenotypes based on minP_{com} p-values.

QQ plots for the p-values obtained by minPhen and Multi-SKAT omnibus tests (minP and minP_{com}) are displayed in Figure 2.4. Due to the presence of several strong associations, for the ease of viewing, any p-value $< 10^{-12}$ was collapsed to 10^{-12}. The QQ plots are well calibrated with slight inflation in tail areas. The

genomic-control lambda (λ_{GC}) varied between 0.97 and 1.04, which indicates no inflation of test statistics. Table 2.2 shows genes with p-values less than 10^{-4} for minPhen or minP_{com}. Table 2.5 shows SKAT-O p-values for each of the gene - amino acid pairs.

Among the eight significant or suggestive genes displayed in Table 2.2, minP_{com} provides more significant p-values than minPhen for six genes: Glycine decarboxylase (*GLDC* [MIM: 238300]), Histidine ammonia-lyase (*HAL* [MIM: 609457]), Phenylalanine hydroxylase (*PAH* [MIM: 612349]), Dihydroorotate dehydrogenase (*DHODH* [MIM: 126064]), Mediator of RNA polymerase II transcription subunit 1 (*MED1* [MIM: 604311]), Serine/Threonine Kinase 33 (*STK33* [MIM: 607670]). Interestingly, *PAH* and *MED1* are significant by minP_{com}, but not significant by minPhen. *PAH* encodes an Phenylalanine hydroxylase, which catalyzes the hydroxylation of the aromatic side-chain of phenylalanine to generate tyrosine. *MED1* is involved in the regulated transcription of nearly all RNA polymerase II-dependent genes. This gene does not show any single phenotype association, but cross-phenotype analysis produced evidence of association. Using backward elimination we find that Phenylalanine and Tyrosine are the last two phenotypes to be eliminated (Appendix Table A.2). We have provided a detailed description of the function and clinical implications of the significant and suggestive genes in Appendix Table A.4.

Among other genes, *GLDC* has the smallest p-value. Variants in *GLDC* are known to cause glycine encephalopathy (MIM: 605899) [Hughes, 2009]. To investigate whether our results were supported by single phenotype associations, we applied SKAT-O to each of the 9 amino acid phenotypes. Univariate SKAT-O test with each of these phenotype reveals that this gene has a strong association with Glycine (p-value = 2.5×10^{-64}, Table 2.5). Among the variants genotyped in this gene, rs138640017 (MAF = 0.009) appears to drive the association (single variant p-value = 1.0×10^{-64}). Variants in *HAL* cause histidinemia (MIM: 235800) in hu-

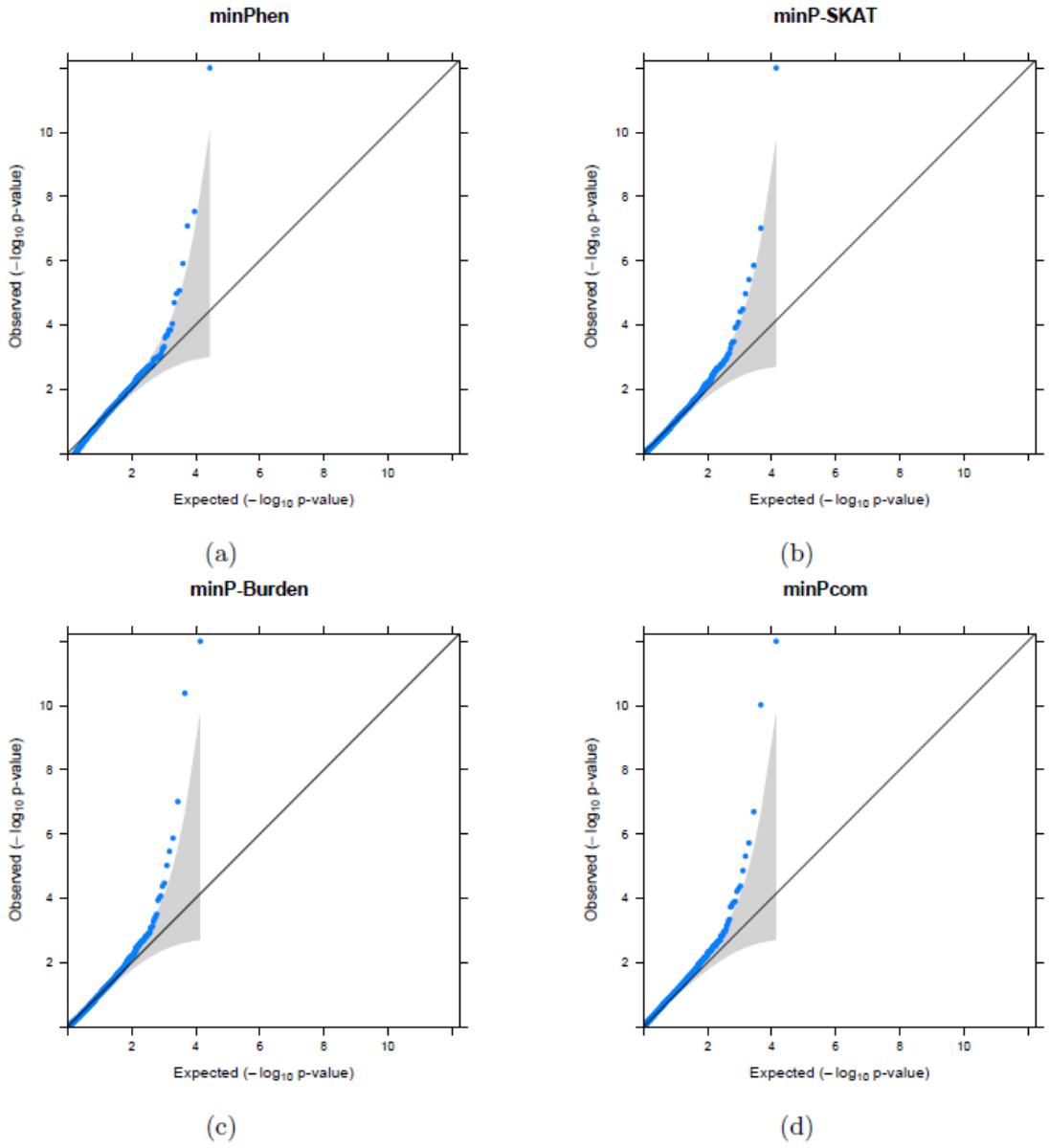


Figure 2.4: QQplot of the p-values of minPhen and Multi-SKAT omnibus tests for the METSIM data. For the ease of viewing, any associations with p-values $< 10^{-12}$ have been collapsed to 10^{-12}

man and mouse. This gene shows significant univariate association with Histidine (SKAT-O p-value = 3.2×10^{-8}, Table 2.5) which in turn is influenced by the association of rs141635447 (MAF = 0.005) with Histidine (single variant p-value = 3.7×10^{-13}). Similarly, variants in *DHODH*, which have been previously found to be associated with postaxial acrofacial dysostosis (MIM: 263750), have significant cross-phenotype association although the result us mostly driven by the association with Alanine (SKAT-O p-value =1.4×10^{-07}, Table 2.5) although, no single variant is significantly associated with Alanine. *ALDH1L1* catalyzes conversion of 10-formyltetrahydrofolate to tetrahydrofolate. Published results show that common variant rs1107366, 5kb upstream of *ALDH1L1*, is associated with Glycine-Seratinine ratio [Xie et al., 2013]. Down-regulation of *BCAT2* in mice causes elevated serum branched chain amino acid levels and features of maple syrup urine disease.

Table 2.3 shows p-values of Multi-SKAT kernel and minP with two genotype kernels (SKAT and Burden). Among phenotype kernels, PhC and Het generally produced the smallest p-values. We further applied Multi-SKAT tests without kinship adjustment on the whole METSIM study individuals. As expected, this produced inflation in QQ plots (Appendix Figure A.5) with λ_{GC} varying between 1.80 and 1.93. It is to be noted here, instead of using a Bonferroni threshold of 9.6×10^{-06}, using the exome-wide cut-off of 2.5×10^{-06} would not have changed the inference hugely. The gene *MED1* would not have remained significant in that case, although the association p-value is suggestive. The other genes would have remained significantly associated with the amino acid traits.

To directly compare our results with existing methods we applied GAMuT, DKAT and MSKAT to the METSIM dataset. Since these methods cannot be applied to related individuals, we eliminated 1332 individuals that were related up to second degree, leaving us 7213 individuals. Table 2.4 shows p-values of different methods on the eight significant or suggestive genes displayed in Table 2.2. Since

DKAT and GAMuT had nearly identical p-values when the same kernels were used, DKAT p-values were not shown in Table 2.4. For unrelated individuals, as expected, p-values produced by MSKAT with Q statistic, GAMuT with projection phenotype kernel and PhC (with SKAT Σ_G) were very similar, and minP_{com} provided similar or more significant p-values than PhC. Interestingly MSKAT with Q' statistic and GAMuT with linear phenotype kernel have less significant p-values than the other tests. We found that in 5 of the 8 genes in Table 2.4, using all individuals with kinship correction produced more significant PhC and minP_{com} p-values than using only unrelated individuals. Even, when we restricted our analysis to unrelated individuals, Multi-SKAT, identified more significant genes (using minP_{com}), compared to GAMuT or MSKAT. Further, we have listed the top 10 genes for each of PhC, GAMuT and MSKAT with unrelated individuals (Appendix Table A.4). Except for the genes in Table 2.4, no other genes were found to be significant or suggestive by any of the methods.

Overall, our METSIM amino acid data analysis suggests that the proposed method can be more powerful than the single phenotype tests as well as existing tests, while maintaining type I error rate even in the presence of the relatedness. It also shows that the omnibus tests (minP and minP_{com}) provides robust performance by effectively aggregating results of various kernels.

2.4.4 Computation Time

When Σ_P and Σ_G are given, p-values of Multi-SKAT are computed by the Davies method [Davies, 1980], which inverts the characteristic function of the mixture of chi-squares. On average, Multi-SKAT tests for a given Σ_P and Σ_G required less than 1 CPU sec (Intel Xeon 2.80 GHz) when applied to a dataset with 5000 independent individuals, 20 variants and 10 phenotypes (Appendix Table A.1). With the kinship adjustment for 5000 related individuals, computation time was increased to 3 CPU

sec. Since minP_{com} requires only a small number of resampling steps to estimate the correlation among tests, it is still scalable for genome-wide analysis. In the same dataset, minP_{com} required 4 and 10 CPU sec on average without and with the kinship adjustment, respectively. Further, Multi-SKAT given Σ_P and Σ_G, is computationally equivalent to MSKAT and takes less than 1 CPU-sec for up to 20,000 samples, with 20 variants (Appendix Figure A.7), while GAMuT takes considerably more time than these two. The performance of minP_{com} is similar to GAMuT for small and moderate sample sizes (7.5 and 7.1 CPU-secs respectively for 10,000 samples) and performs better than GAMuT for larger sample sizes (14.9 and 34.6 CPU-secs respectively for 20,000 samples). Computation time of all the methods were slightly increased when the number of variants were 100 (Appendix Figure A.7). Analyzing the METSIM dataset with minP_{com} required 10 hours when parallelized into 5 processes.

2.5 Discussion

In this chapter, we have introduced a general framework for rare variant tests for multiple phenotypes. As demonstrated, Multi-SKAT gains flexibility with regard to modeling the relationship between phenotypes and genotypes through the use of the kernels Σ_P and Σ_G. Many published methods, including GAMuT, MSKAT, MAAUSS and MF-KM, can be viewed as special cases of the Multi-SKAT test with corresponding values of Σ_P and Σ_G. This can potentially highlight the underlying assumptions of these methods and their relationships. In addition, by unifying existing methods to a common framework, our approach provides a way to combine different methods through the minimum p-value based omnibus test. Multi-SKAT can also adjust for sample relatedness. From simulation studies we have found that Multi-SKAT methods are scalable to genome-wide analysis and can outperform the single phenotype test and existing multiple phenotype tests. The METSIM data analysis demonstrates that the proposed methods perform well in practice.

It is natural to assume that different genes follow different models of association. For some genes, the effect of the variants on the phenotypes might be independent of each other, thus best detected by the Het phenotype kernel for Σ_P, while for others, the effects might be nearly the same and best detected by the Hom phenotype kernel. If the kernel structures Σ_G and Σ_P are chosen based on minor allele frequency, functional consequence or other prior knowledge about the structure of associations, but they do not reflect the underlying biology satisfactorily, the corresponding test of association may have substantially reduced power. The omnibus test, which uses the minimum p-value from the various choices of kernels, has been a useful approach under such situations in genetic association analysis [Lee et al., 2012b, Urrutia et al., 2015, Zhan et al., 2017]. By aggregating association results across several choices of kernels omnibus tests can produce robust results across a spectrum of different association models. We applied this omnibus test to Multi-SKAT and used a Copula to obtain p-values. As seen in simulation studies and real data analysis, our omnibus approaches (minP and minP_{com}) are scalable to genome-wide analysis and provide robust power regardless of underlying genetic models.

Multi-SKAT retains most of the desirable properties of SKAT. The asymptotic p-values of all the Multi-SKAT tests, other than minP and minP_{com}, can be analytically obtained via Davies' method. The p-value calculations for minP and minP_{com} depend on a resampling based approach and a reliable estimate can be obtained using a small number of resampling steps. Thus, computationally all the Multi-SKAT tests are scalable at the genome-wide level.

Additionally, Multi-SKAT can adjust for the relatedness among study individuals by accounting for their kinship matrix. This is an important aspect of our method because if we do not incorporate the between-sample-relatedness, the corresponding test of association might not be statistically valid. As shown in Appendix Figure A.5, in the presence of related individuals, not adjusting for relatedness can produce

inflated type I error rate. Since Multi-SKAT is a regression based approach, it can effectively incorporate the relatedness by including a random effect term for kinship.

Simulations and METSIM data analysis indicate that Multi-SKAT can have a greater power than alternative methods, like GAMuT and MSKAT, while controlling type I error rates. Majority of the the genes, with the exception of *ALDH1L1*, that were found to be significant by GAMuT and MSKAT, were again identified by Multi-SKAT. Additionally, Multi-SKAT identified additional genes to be significant (like *PAH, MED1* and *STK33*) that were neither identified by existing methods as well as single phenotype tests.

Although Multi-SKAT provides a general framework for gene-based multiple phenotype tests, it has certain important limitations. The current formulation is restricted to continuous phenotypes only. Hence Multi-SKAT cannot be used for phenotypes like disease status which are binary. In the future, using a generalized mixed effect model framework, we aim to extend Multi-SKAT to binary phenotypes. Further, the computation time for omnibus tests can be improved upon by recently developed cauchy-transformation techniques.

In summary, we have developed Multi-SKAT, a powerful multiple phenotype test for rare variants. The proposed method has robust power regardless of the underlying biology and can adjust for family relatedness. Multi-SKAT can be a scalable and practical solution to test for multiple phenotypes and will contribute to detecting rare variants with pleiotropic effects. We have implemented these methods in a publicly-available R-based software (see next section for URL).

2.6 Web Resources

MultiSKAT R-package: https://github.com/diptavo/MultiSKAT

GAMuT R-package: https://epstein-software.github.io/GAMuT

MSKAT R-package: https://github.com/baolinwu/MSKAT

PHENIX R-package: https://mathgen.stats.ox.ac.uk/genetics_software/phenix/phenix.html

Online Mendelian Inheritance in Man (OMIM): http://www.omim.org

Table 2.1: Empirical type I error rates of the Multi-SKAT tests. The number of phenotypes were nine and the correlation structure among the phenotypes were similar to that of the amino acid phenotypes in the METSIM study data. The sample size was 5000.

	Level	Σ_G	minPhen	Hom	Het	PhC	PC-Sel	minP	minP_{com}
Independent samples (without kinship adjustment)	2.5×10^{-06}	*SKAT*	2.4×10^{-06}	2.6×10^{-06}	2.6×10^{-06}	2.8×10^{-06}	2.6×10^{-06}	3.4×10^{-06}	2.8×10^{-06}
		Burden	2.4×10^{-06}	2.8×10^{-06}	2.6×10^{-06}	2.6×10^{-06}	2.6×10^{-06}	3.0×10^{-06}	
	10^{-05}	*SKAT*	9.6×10^{-06}	9.6×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.4×10^{-06}	1.2×10^{-05}
		Burden	9.4×10^{-06}	9.6×10^{-06}	9.8×10^{-06}	9.6×10^{-06}	9.6×10^{-06}	9.6×10^{-06}	
	10^{-04}	*SKAT*	9.6×10^{-05}	9.4×10^{-05}	9.6×10^{-05}	9.8×10^{-05}	9.8×10^{-05}	9.8×10^{-05}	1.1×10^{-04}
		Burden	9.8×10^{-05}	9.6×10^{-05}	9.4×10^{-05}	9.7×10^{-05}	9.6×10^{-05}	9.7×10^{-05}	
Related samples (with kinship adjustment)	2.5×10^{-06}	*SKAT*	2.2×10^{-06}	2.4×10^{-06}	2.4×10^{-06}	2.8×10^{-06}	2.6×10^{-06}	3.0×10^{-06}	2.6×10^{-06}
		Burden	2.2×10^{-06}	2.6×10^{-06}	2.4×10^{-06}	2.6×10^{-06}	2.4×10^{-06}	3.2×10^{-06}	
	10^{-05}	*SKAT*	9.6×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.8×10^{-06}	9.4×10^{-06}	1.4×10^{-05}
		Burden	9.7×10^{-06}	9.6×10^{-06}	9.4×10^{-06}	9.4×10^{-06}	9.4×10^{-06}	9.4×10^{-06}	
	10^{-04}	*SKAT*	9.4×10^{-05}	9.6×10^{-05}	9.7×10^{-05}	9.8×10^{-05}	9.4×10^{-05}	9.8×10^{-05}	1.2×10^{-04}
		Burden	9.5×10^{-05}	9.7×10^{-05}	9.6×10^{-05}	9.8×10^{-05}	9.8×10^{-05}	9.7×10^{-05}	

Table 2.2: Significant and suggestive genes associated with 9 amino acid phenotypes. Genes with p-values $< 10^{-4}$ by minPhen or any Multi-SKAT tests (minP-SKAT, minP-Burden and minP_{com}) were reported in this table. Multi-SKAT tests were applied with the kinship adjustment, and 5207 genes with at least three rare (MAF < 0.01) nonsynonymous and protein-truncating variants were used in this analysis. The total sample size was $N = 8545$. P-values smaller than the Bonferroni corrected significance $\alpha = 9.6 \times 10^{-6}$ were marked as bold. Smallest p-value for each gene among all the tests have been underlined. minPhen was calculated as the Bonferroni adjusted minimum SKAT-O p-value across each phenotype.

Gene	Chromosome	Rare SNPs	MAC	minPhen	minP-SKAT	minP-Burden	minP_{com}
GLDC	9	4	183	$\mathbf{2.2 \times 10^{-63}}$	$\underline{\mathbf{1.1 \times 10^{-72}}}$	$\mathbf{9.7 \times 10^{-64}}$	$\mathbf{2.3 \times 10^{-72}}$
HAL	3	6	42	$\mathbf{2.9 \times 10^{-08}}$	1.1×10^{-05}	$\underline{\mathbf{4.2 \times 10^{-11}}}$	$\mathbf{9.5 \times 10^{-11}}$
DHODH	16	5	90	$\mathbf{1.3 \times 10^{-06}}$	$\underline{\mathbf{9.7 \times 10^{-08}}}$	$\mathbf{7.7 \times 10^{-06}}$	$\mathbf{1.1 \times 10^{-07}}$
PAH	12	6	27	6.1×10^{-04}	$\mathbf{1.4 \times 10^{-06}}$	$\underline{\mathbf{1.3 \times 10^{-06}}}$	$\mathbf{1.9 \times 10^{-06}}$
MED1	17	3	147	6.2×10^{-01}	$\underline{\mathbf{3.9 \times 10^{-06}}}$	3.4×10^{-04}	$\mathbf{4.9 \times 10^{-06}}$
STK33	11	6	180	8.9×10^{-01}	$\underline{3.2 \times 10^{-05}}$	3.3×10^{-02}	4.2×10^{-05}
ALDH1L1	3	8	103	$\underline{\mathbf{8.4 \times 10^{-08}}}$	3.8×10^{-04}	4.2×10^{-05}	5.4×10^{-05}
BCAT2	19	3	133	$\underline{1.1 \times 10^{-05}}$	9.3×10^{-03}	3.4×10^{-05}	6.2×10^{-05}

Table 2.3: P-values for MultiSKAT tests (Hom, Het, PhC,PC-Sel, minP) with SKAT and Burden kernels for the genes reported in Table 2.2. Multi-SKAT tests were applied with the kinship adjustment, and 5207 genes with at least three rare (MAF < 0.01) nonsynonymous and protein-truncating variants were used in this analysis. The total sample size was $N = 8545$. P-values smaller than the Bonferroni corrected significance $\alpha = 9.6 \times 10^{-6}$ were marked as bold. For the upper part of the table, minPhen was calculated as the Bonferroni adjusted minimum SKAT p-value across each phenotype, while for the lower part it was calculated as the Bonferroni adjusted minimum Burden p-value across each phenotype.

	Gene					Multi-SKAT				
Σ_G	Name	Chromosome	Rare SNPs	MAC	minPhen	Hom	Het	PhC	PC-Sel	minP
SKAT	*GLDC*	9	4	183	$\mathbf{6.8 \times 10^{-63}}$	1.3×10^{-03}	$\mathbf{1.3 \times 10^{-17}}$	$\mathbf{2.9 \times 10^{-73}}$	$\mathbf{5.3 \times 10^{-59}}$	$\mathbf{1.1 \times 10^{-72}}$
	DHODH	16	5	90	$\mathbf{5.1 \times 10^{-08}}$	1.4×10^{-01}	3.2×10^{-03}	$\mathbf{3.1 \times 10^{-08}}$	$\mathbf{7.1 \times 10^{-06}}$	$\mathbf{9.7 \times 10^{-08}}$
	PAH	12	6	27	4.3×10^{-05}	7.9×10^{-02}	1.8×10^{-05}	$\mathbf{3.9 \times 10^{-07}}$	6.3×10^{-04}	$\mathbf{1.4 \times 10^{-06}}$
	MED1	17	3	147	8.7×10^{-02}	3.4×10^{-01}	$\mathbf{9.3 \times 10^{-07}}$	2.8×10^{-04}	1.9×10^{-04}	$\mathbf{3.9 \times 10^{-06}}$
	HAL	12	6	42	3.8×10^{-05}	5.5×10^{-01}	3.5×10^{-03}	$\mathbf{2.9 \times 10^{-06}}$	1.5×10^{-04}	1.1×10^{-05}
	STK33	11	6	180	4.0×10^{-01}	2.3×10^{-01}	$\mathbf{8.9 \times 10^{-06}}$	1.1×10^{-03}	5.7×10^{-03}	3.2×10^{-05}
	ALDH1L1	3	8	103	4.4×10^{-02}	1.1×10^{-01}	2.2×10^{-02}	1.6×10^{-04}	4.2×10^{-04}	3.8×10^{-04}
	BCAT2	19	3	133	9.7×10^{-04}	5.0×10^{-01}	8.1×10^{-02}	2.4×10^{-03}	3.5×10^{-03}	9.3×10^{-03}
Burden	*GLDC*	9	4	183	$\mathbf{1.4 \times 10^{-59}}$	7.5×10^{-04}	$\mathbf{1.5 \times 10^{-15}}$	$\mathbf{1.3 \times 10^{-64}}$	$\mathbf{2.1 \times 10^{-54}}$	$\mathbf{9.7 \times 10^{-64}}$
	HAL	12	6	42	$\mathbf{2.4 \times 10^{-08}}$	2.8×10^{-05}	4.6×10^{-01}	$\mathbf{4.2 \times 10^{-12}}$	7.9×10^{-03}	$\mathbf{4.2 \times 10^{-11}}$
	PAH	12	6	27	4.3×10^{-05}	2.0×10^{-01}	5.3×10^{-04}	$\mathbf{7.2 \times 10^{-06}}$	9.3×10^{-03}	$\mathbf{1.3 \times 10^{-06}}$
	DHODH	16	5	90	$\mathbf{1.3 \times 10^{-07}}$	5.3×10^{-02}	1.3×10^{-02}	$\mathbf{2.9 \times 10^{-06}}$	7.6×10^{-04}	$\mathbf{7.7 \times 10^{-06}}$
	BCAT2	19	3	133	$\mathbf{8.4 \times 10^{-06}}$	4.0×10^{-01}	1.7×10^{-02}	1.1×10^{-05}	6.1×10^{-03}	3.7×10^{-05}
	ALDH1L1	3	8	103	$\mathbf{8.3 \times 10^{-08}}$	1.7×10^{-02}	5.1×10^{-03}	$\mathbf{2.1 \times 10^{-06}}$	1.7×10^{-05}	4.2×10^{-05}
	MED1	17	3	147	9.4×10^{-01}	1.8×10^{-01}	8.7×10^{-05}	2.1×10^{-03}	7.1×10^{-01}	3.4×10^{-04}
	STK33	11	6	180	4.8×10^{-01}	7.1×10^{-01}	7.0×10^{-04}	3.8×10^{-02}	6.4×10^{-01}	6.3×10^{-03}

Table 2.4: P-values for genes reported in Table 2.2 using MSKAT, GAMuT and Multi-SKAT (PhC and minP_{com}). P-values in columns 2 to 7 were calculated with unrelated individuals ($N = 7213$), while those in columns 8 and 9 were calculated using all individuals ($N = 8545$). Both MSKAT (Q and Q' statistic) and GAMuT (Projection and Linear phenotype kernel) p-values were calculated with the linear weighted genotype kernel. DKAT p-values were nearly identical to those of GAMuT when the same kernels were used (data not shown). P-values smaller than the Bonferroni corrected significance $\alpha = 9.6 \times 10^{-6}$ were marked as bold.

	Unrelated individuals (N = 7213)						All individuals (N = 8545)	
Gene	MSKAT (Q)	MSKAT (Q')	GAMuT (Projection)	GAMuT (Linear)	PhC (Σ_G = SKAT)	minP_{com}	PhC (Σ_G = SKAT)	minP_{com}
GLDC	$\mathbf{8.9 \times 10^{-54}}$	$\mathbf{6.1 \times 10^{-15}}$	**0***	$\mathbf{6.2 \times 10^{-15}}$	$\mathbf{8.1 \times 10^{-54}}$	$\mathbf{1.3 \times 10^{-53}}$	$\mathbf{2.9 \times 10^{-73}}$	$\mathbf{2.3 \times 10^{-72}}$
HAL	9.5×10^{-05}	4.2×10^{-02}	9.6×10^{-05}	4.3×10^{-02}	9.5×10^{-05}	$\mathbf{2.9 \times 10^{-07}}$	$\mathbf{2.9 \times 10^{-06}}$	$\mathbf{9.5 \times 10^{-11}}$
DHODH	$\mathbf{2.1 \times 10^{-06}}$	2.2×10^{-03}	$\mathbf{2.4 \times 10^{-06}}$	2.2×10^{-03}	$\mathbf{1.9 \times 10^{-06}}$	$\mathbf{8.3 \times 10^{-06}}$	$\mathbf{3.1 \times 10^{-08}}$	$\mathbf{1.1 \times 10^{-07}}$
PAH	9.9×10^{-06}	1.3×10^{-02}	1.0×10^{-05}	1.3×10^{-02}	9.9×10^{-06}	5.1×10^{-05}	$\mathbf{3.9 \times 10^{-07}}$	$\mathbf{1.9 \times 10^{-06}}$
MED1	1.7×10^{-03}	2.7×10^{-01}	1.7×10^{-03}	2.7×10^{-01}	1.7×10^{-03}	2.8×10^{-05}	2.8×10^{-04}	$\mathbf{4.9 \times 10^{-06}}$
STK33	6.8×10^{-04}	6.9×10^{-01}	6.7×10^{-04}	6.9×10^{-01}	6.7×10^{-04}	$\mathbf{1.9 \times 10^{-06}}$	1.1×10^{-03}	4.2×10^{-05}
ALDH1L1	5.9×10^{-05}	3.1×10^{-02}	6.1×10^{-05}	3.0×10^{-02}	6.0×10^{-05}	$\mathbf{9.5 \times 10^{-06}}$	1.6×10^{-04}	5.4×10^{-05}
BCAT2	6.1×10^{-04}	1.5×10^{-02}	6.3×10^{-04}	1.5×10^{-02}	6.2×10^{-04}	3.7×10^{-05}	2.4×10^{-03}	6.2×10^{-05}

Table 2.5: Single phenotype SKAT-O with kinship adjustment test for the METSIM study data ($N = 8545$). P-values smaller than the Bonferroni corrected significance $\alpha = 9.6 \times 10^{-6}$ were marked as bold.

Gene	Ala	Gln	Gly	His	Ile	Leu	Phe	Tyr	Val
GLDC	2.9×10^{-01}	2.9×10^{-03}	$\mathbf{2.5 \times 10^{-64}}$	1.0×10^{-02}	1.5×10^{-01}	3.4×10^{-02}	8.5×10^{-01}	6.6×10^{-01}	4.7×10^{-03}
HAL	9.9×10^{-01}	1.1×10^{-01}	3.1×10^{-01}	$\mathbf{3.2 \times 10^{-09}}$	2.6×10^{-01}	2.5×10^{-01}	6.6×10^{-01}	1.2×10^{-01}	3.5×10^{-01}
DHODH	$\mathbf{1.4 \times 10^{-07}}$	9.9×10^{-01}	9.0×10^{-02}	3.6×10^{-01}	7.7×10^{-01}	1.7×10^{-01}	1.9×10^{-01}	3.0×10^{-01}	1.0×10^{-02}
PAH	7.9×10^{-01}	3.0×10^{-01}	2.8×10^{-01}	8.6×10^{-01}	4.5×10^{-01}	8.1×10^{-01}	6.8×10^{-05}	4.0×10^{-01}	9.9×10^{-01}
MED1	1.0×10^{-01}	5.1×10^{-01}	7.9×10^{-01}	5.9×10^{-01}	3.3×10^{-01}	1.4×10^{-01}	6.9×10^{-01}	6.8×10^{-02}	6.7×10^{-01}
STK33	8.4×10^{-01}	8.2×10^{-01}	3.5×10^{-01}	5.7×10^{-01}	6.6×10^{-01}	1.0×10^{-01}	9.9×10^{-01}	8.1×10^{-01}	4.1×10^{-01}
ALDH1L1	6.7×10^{-01}	7.6×10^{-01}	$\mathbf{9.3 \times 10^{-09}}$	7.3×10^{-01}	3.4×10^{-01}	3.2×10^{-01}	9.9×10^{-01}	5.9×10^{-01}	2.5×10^{-01}
BCAT2	1.4×10^{-01}	2.6×10^{-01}	8.2×10^{-01}	9.9×10^{-01}	1.0×10^{-01}	2.1×10^{-02}	5.4×10^{-01}	5.0×10^{-01}	$\mathbf{1.2 \times 10^{-06}}$

CHAPTER III

Meta-MultiSKAT: Multiple phenotype meta-analysis for region-based association test

3.1 Introduction

The advent of large scale genome-wide association studies (GWAS) has shown that many distinct phenotypes have substantial genetic correlation [Bulik-Sullivan et al., 2015] and many loci have pleiotropic effects [Cotsapas et al., 2011, Solovieff et al., 2013, Sivakumaran et al., 2011, Yang et al., 2015, Li et al., 2014].

To leverage the widespread pleiotropy, a statistical model to jointly test multiple phenotypes is beneficial. Although data on multiple related phenotypes are often collected in hospital or population based studies, association tests are usually performed with one phenotype at a time. Such methods that do not account for the correlation between phenotypes may lack power to detect cross-phenotype effects of associated loci [Ferreira and Purcell, 2009, Huang et al., 2011, Ray et al., 2016]. Alternatively, joint tests which aggregate association signals in multiple phenotypes can substantially improve power over single phenotype-based tests [Ferreira and Purcell, 2009, Ray et al., 2016, Ried et al., 2012, Zhou and Stephens, 2014b], although interpreting the results can prove difficult.

Meta-analysis of multiple studies, using association summary statistics, is a prac-

tical approach to increase power by increasing sample sizes [Panagiotou et al., 2013]. Meta-analysis is especially valuable for association analysis of variation on the lower end of the allele frequency spectrum, since detecting such associations often require large sample sizes. It seems logical to expect that meta-analyzing multiple phenotypes can further increase power of rare variant tests. Various methods have been developed for meta-analysis of multiple phenotypes [Majumdar et al., 2018, Ray and Boehnke, 2018, Zhu et al., 2015], but most of them are single variant-based methods, which have low power to identify rare variant associations. More powerful gene or region-based tests for multiple phenotypes have been developed for use within a single study [Broadaway et al., 2016, Lee et al., 2016, Wu and Pankow, 2016]. However, to the best of our knowledge, no work has been done to extend these methods to meta-analysis. This is partly because most of the methods are similarity-based non-parametric methods, which are difficult to extend to meta-analysis.

In the previous chapter we developed a regression-based method, Multiple phenotype sequence kernel association test (Multi-SKAT) [Dutta et al., 2019], that can aggregate signals across models with different kernels while correcting for sample relatedness, which only few methods have addressed. Through simulations and real-data analysis we showed that Multi-SKAT can have greater power than current methods under a wide range of association models while maintaining type-I error rate. In this project we develop Meta-MultiSKAT, a meta-analysis extension of Multi-SKAT, which uses summary statistics from individual studies to construct a test of of cross-phenotype associations. Meta-MultiSKAT models the relationship between effect sizes of different studies through a kernel matrix and performs a variance component test of association. Our method retains useful features of Multi-SKAT, including fast computation. Meta-MultiSKAT can incorporate various missing data scenarios, including situations where studies do not share exactly the same set of phenotypes, and test for only rare variants as well as for the combined effects of

both common and rare variants. The latter allows us to evaluate the overall effect of gene or region on multiple phenotypes. By using kinship adjusted score statistics, Meta-MultiSKAT can account for sample relatedness, an important feature to use in a study with widespread relatedness, such as the SardiNIA study [Sidore et al., 2015, Vacca et al., 2006]. To avoid loss of the power due to model misspecification, we have also developed a minimum p-value-based omnibus test that can aggregate results across different patterns of association. We evaluate the performance of our method through extensive type-I error and power simulations.

We applied Meta-MultiSKAT to meta-analyze four white blood cell (WBC) subtype traits from the Michigan Genomics Initiative (MGI) [Fritsche et al., 2018] study and the SardiNIA study. In addition to detecting the genes *PRG2* [MIM: 605601] and *RP11-872D17.8*, that had significant association signals with WBCs within one of the studies, Meta-MultiSKAT further identified two additionally associated genes (*IRF8* [MIM: 601565] and *CCL24* [MIM: 602495]) that did not have any significant signals in either of the studies but were identified as significant only as a result of meta-analysis.

3.2 Methods

Suppose we intend to conduct a meta-analysis with S studies each having K phenotypes. For the s^{th} study n_s subjects are genotyped in a region that has m_s variants. Let $y_{ks} = (y_{1ks}, y_{2ks}, \cdots, y_{n_sks})^T$ be the $n_s \times 1$ vector for the k^{th} phenotype on n_s individuals in the s^{th} study; $G_{js} = (g_{1js}, g_{2js}, \cdots, g_{n_sjs})^T$ is an $n_s \times 1$ vector for the minor allele counts (0, 1, or 2 variant alleles) for variant j and $G_s = (G_{1s}, \cdots G_{m_ss})$ is an $n_s \times m_s$ genotype matrix of the m_s genetic variants in the target gene or region. For a gene-based multiple phenotype test we consider the following regression model, similar to equation 2.3,

$$Y_s = X_s A_s + G_s B_s + E_s \tag{3.1}$$

where $Y_s = (y_{1s}, \cdots, y_{Ks})$ is an $n_s \times K$ phenotype matrix of n_s individuals and K phenotypes; $B_s = ((\beta_{jks}))$ is an $m_s \times K$ matrix where β_{jks} is the regression coefficient of phenotype k on G_{js}; A_s is a $q_s \times K$ matrix of regression coefficients for non-genetic covariates X_s; E_s is an $n_s \times K$ matrix of non-systematic error terms. The null hypothesis of no genetic association between variants in the region and the phenotypes is $H_0 : \beta_{jks} = 0$ for all j and k.

Let $L_s = G_s^T(Y_s - \widehat{\mu}_s)\widehat{V}_s^{-1}$ be the $m_s \times K$ score matrix for the s^{th} study where $\widehat{\mu}_s$ is an $n_s \times K$ matrix of the estimated mean of Y_s under the null hypothesis of no association and $\widehat{V}_s$ is the $K \times K$ estimated null residual covariance matrix among the K phenotypes in the s^{th} study. To test the null hypothesis of no association, we use a variance component test. Under the mixed effect model set-up, we assume that the vectorized form of matrix B_s represented as $vec(B_s)$ follows a distribution with mean 0 and variance $\tau^2\Sigma_G \otimes \Sigma_P$ (for details on Σ_G and Σ_P see below; also refer to Chapter II Methods sections), where $\otimes$ is a kronecker product. The null hypothesis of no genetic association can hence be written as $H_0 : \tau = 0$. The corresponding score statistic is

$$Q_s(\Sigma_P, \Sigma_G) = [vec(L_s)]^T(\Sigma_G \otimes \Sigma_P)[vec(L_s)] \tag{3.2}$$

Under the null hypothesis $vec(L_s)$ asymptotically follows a $N(0, \Phi_s)$ where Φ_s is the phenotype-adjusted variant relationship matrix

$$\Phi_s = (G_s^T G_s - G_s^T X_s (X_s^T X_s)^{-1} X_s^T G_s) \otimes (V_s^{-1})$$

$Q_s(\Sigma_P, \Sigma_G)$ asymptotically follows a mixture of chi-square distributions. The mixing parameters are the eigenvalues of $R\Phi_s R^T$ where $RR^T = \Sigma_G \otimes \Sigma_P$.

The kernel Σ_G represents the effect sizes of the variants to a phenotype. In general, Σ_G is assumed to be a sandwich matrix WR_GW where $W = diag(w_1, \cdots, w_{m_s})$

is a diagonal for the variant-weighting; $R_G = (1-\rho)I_{m_s \times m_s} + \rho J_{m_s} J_{m_s}^T$ is a compound symmetric correlation matrix with $I_{m_s \times m_s}$ being an identity matrix of order m_s and $J_{m_s} = (1, \cdots, 1)$ is an $m_s \times 1$ vector with all elements being 1. This model can cover a wide range of scenarios of the genetic effect distribution. For example, with one phenotype ($K = 1$), if $\rho = 1$ (i.e.$R_G = J_{m_s} J_{m_s}^T$), which assumes homogeneous effects of the variants on the phenotypes, the test reduces to a Burden test [Li and Leal, 2008]. Similarly if $\rho = 0$ (i.e. $(R_G = I_{m_s \times m_s})$, the test is equivalent to a SKAT test [Wu et al., 2011].

The kernel Σ_P represents the effect sizes of a variant on the phenotypes. For example, under the assumption that the genetic effects of a variant on each phenotype are independent, we can use $\Sigma_P = I_{K \times K}$ of a variant on the phenotypes. For example, under the assumption that the genetic effects of a variant on each phenotype are independent, we can use $\widehat{V}_s$ as Σ_P which results in the test equivalent to GAMuT [Broadaway et al., 2016], MSKAT [Wu and Pankow, 2016] and DKAT [Zhan et al., 2017].

3.2.1 Input summary statistics from each study for meta-analysis

Single-variant meta-analyses are conducted with single-variant summary statistics, such as the estimated effect sizes and their standard errors. For region based tests with a single phenotype, [Lee et al., 2013] showed that the score statistics of the variants, minor allele frequencies (MAFs) and the variant relationship matrix can be used as summary statistics for meta-analysis. With multiple phenotypes, the multivariate forms of these summary statistics from each study are needed. In particular from the s^{th} study, the score matrix L_s, the phenotype-adjusted variant relationship matrix Φ_s, the residual covariance structure of the phenotypes, $\widehat{V}_s$ and the MAFs of the variants in the region are needed for meta-analysis.

3.2.2 Meta-MultiSKAT: Meta-analysis of gene-based tests with multiple phenotypes

For simplicity, here we assume that all variants and phenotypes are observed in all S studies, so that $m = m_1 = \cdots = m_s$. We will relax this assumption later. Suppose summary statistics $(L_s, \Phi_s), s = 1, \cdots S$ is provided by S studies. We construct the meta-score-vector as $L_{meta} = (vec(L_1)^T, vec(L_2)^T, \cdots, vec(L_s)^T)^T$. The variance component test statistic for meta-analysis is

$$Q_{meta} = [vec(L_{meta})^T](\Sigma_S \otimes \Sigma_G \otimes \Sigma_P)[vec(L_{meta})] \tag{3.3}$$

Under the null hypothesis of no association, Q_{meta} follows a mixture of chi-square distributions and the corresponding p-value can be obtained by inverting the characteristic function (See Appendix B.1 for details on p-value calculation).

Here we have introduced another kernel Σ_S. Similarly as the other kernels, Σ_S models the heterogeneity between the effects of the contributing studies. In particular we will consider two special structures of Σ_S:

Homogeneous: $\Sigma_{S;Hom} = J_S J_S^T$ which assumes that across the S studies the effects of the variants on all the phenotypes are the same (homogeneous).

Heterogeneous: $\Sigma_{S;Hom} = I_S$ which assumes that across the S studies the effects of the variants on the phenotypes are uncorrelated or heterogeneous.

The test statistic in (3.3) assumes that the kernels Σ_G and Σ_P are the same across studies. This assumption is restrictive since different studies might be analyzed with different hypotheses, reflected in different Σ_G and Σ_P across studies. This can be resolved by modifying (3.3) as

$$Q_{meta} = [vec(\tilde{L}_{meta})^T](\Sigma_S \otimes I_m \otimes I_K)[vec(\tilde{L}_{meta})] \tag{3.4}$$

where $\tilde{L}_{meta} = (vec(\tilde{L}_1)^T, vec(\tilde{L}_2)^T, \cdots, vec(\tilde{L}_s)^T)^T$ and $\tilde{L}_s = \Sigma_{G;s}^{\frac{1}{2}} L_s \Sigma_{P;s}^{\frac{1}{2}}$ represents the kernelized scores incorporating study specific $\Sigma_{G;s}$ and $\Sigma_{P;s}$ for the s^{th} study (See Appendix B.2 for details).

3.2.2.1 Variant weighting scheme

In region-based analysis, [Wu et al., 2011] suggested a MAF- based weighting scheme. To upweight the rare-variants, they proposed to use Beta(1,25) weights. When the homogeneity across the studies are assumed (i.e. $\Sigma_S = \Sigma_{S;Hom}$), pooled MAFs across studies can be used to generate weights for variants. For $\Sigma_S = \Sigma_{S;Het}$, we use study specific weights obtained using MAFs of each study. Alternatively, functional scores, such as CADD [Kircher et al., 2014] and Eigen [Ionita-Laza et al., 2014] can be used to upweight functionally important variants. In addition to using the MAF-based weighting, we have also explored the use of CADD scores as weights for variants in the meta-analysis of MGI and SardiNIA datasets.

3.2.3 Combined effect of common and rare variants (Meta-MultiSKAT-Common-Rare)

The default setting for SKAT type tests (SKAT, MultiSKAT and Meta-MultiSKAT) is to use a MAF-based weighting scheme that up-weights the contribution of the rare variants and down-weights that of common variants. When there are common variants in the region associated with the phenotype, this weighting scheme can lead to a loss in power. Similar to [Ionita-Laza et al., 2013] we propose a test of the combined effects of common and rare variants on the phenotype. As in equation (3.3), the Meta-MultiSKAT test statistic is given by the quadratic form, $Q_{meta} = [vec(L_{meta})^T]K[vec(L_{meta})]$ where $K = (\Sigma_S \otimes \Sigma_G \otimes \Sigma_P)$. Given each study MAF, we compute the pooled MAFs for the variants in the region of interest and using a cut-off on that we partition the variants into common and rare. In practice,

cut-offs like 5% MAF or 1% MAF are commonly used. To explicitly separate the effects of common and rare variants, we construct the test statistic separately for common and rare variants, as

$$Q_{meta;common} = [vec(L_{meta;common})^T]K_{common}[vec(L_{meta;common})] \tag{3.5}$$

$$Q_{meta;rare} = [vec(L_{meta;rare})^T]K_{rare}[vec(L_{meta;rare})] \tag{3.6}$$

where $L_{meta;common}$ and K_{common} (alternatively $L_{meta;rare}$ and K_{rare}) are constructed using common variants (alternatively rare variants) only. The two matrices K_{common} and K_{rare} only differ in terms of the underlying Σ_G matrices and we can allow different weighting schemes for the Σ_G kernels corresponding to common and rare variants. In particular, here, we use Beta(0.5,0.5) weights for the common variants and Beta(1,25) weights for the rare variants.

The combined sum (Meta-MultiSKAT-Common-Rare) is then constructed as

$$Q_{meta;common-rare} = (1-\phi)Q_{meta;common} + \phi Q_{meta;rare} \tag{3.7}$$

with a given weight ϕ. A simple approach, as used in [Ionita-Laza et al., 2013], is to select ϕ such that the rare and common variants contribute equally to the test statistics, i.e.

$$\phi = \frac{SD(Q_{meta;rare})}{SD(Q_{meta;rare}) + SD(Q_{meta;common})}$$

. The asymptotic p-value of $Q_{meta;common-rare}$ can be calculated from a mixture of chi-squared distribution, similar to the previous discussion.

3.2.4 Kinship adjustment within studies

Individual studies might require adjustment for kinship if there are related individuals within the study. For instance, if study s has related individuals with kinship matrix Ψ, co-heritability matrix $V_{g;s}$ and the shared non-genetic effect matrix $V_{e;s}$ then we construct scores as

$$vec(L_s) = (G_s \otimes I_K)\tilde{V}_{t;s}^{-1}(vec(Y_s) - vec(\widehat{\mu}_s))$$

where $\tilde{V}_{t;s} = \Psi \otimes \widehat{V}_{g;s} + I \otimes \widehat{V}_{e;s}$ represents the estimated total covariance matrix for Y_s.

3.2.5 Discrepant phenotypes and genotypes across studies

The studies included in the meta-analysis may not have exactly the same set of variants genotyped (or sequenced). In particular, some variants may be observed in only a subset of studies. If variant j was not observed in study s, we set the $(j,k)^{th}$ element in $L_s(k = 1, 2, \cdots, K)$ and the corresponding elements in Φ_s to be zero, which implies that the studies with missing data do not contribute to the score statistic. This also corresponds to imputing the missing data with the respective mean under the null hypothesis of no association. Using the same framework, if phenotype k in study s was not collected, we set the $(j,k)^{th}$ element in $L_s(j = 1, 2, \cdots, m_s)$ and the corresponding elements in Φ_s to be zero. As above, this corresponds to the null hypothesis that the missing phenotype is not associated with the region of interest.

3.2.6 Minimum p-value-based omnibus tests: Meta-Hom, Meta-Het, and Meta-Com

The Meta-MultiSKAT model and tests have three parameters Σ_S , Σ_P and Σ_G that are absent in the null model. Since this is a score test, these parameters can-

not be estimated from the data. One possible solution is to select them based on a specific prior hypothesis about the underlying model of association, for example using $\Sigma_{S;Hom}$ or $\Sigma_{S;Het}$ for Σ_S. However if the selected values do not reflect the true model, then the corresponding test might have lower power [Lee et al., 2016, Ray et al., 2016]. To overcome such issues, minimum p-value-based omnibus tests have been proposed, which aggregate results across different values of the parameters to produce robust results [Dutta et al., 2019, Wu et al., 2013, He et al., 2017, Urrutia et al., 2015, Zhan et al., 2017]. Here we use the same strategy to formulate robust tests across different choices of Σ_S, Σ_P and Σ_G. We first calculate p-values from different choices of (Σ_S , Σ_P, Σ_G) and obtain the minimum of these p-values. Since the tests are correlated, using a Bonferroni correction can result in conservative type-I error and low power. Instead, we use a fast resampling approach to estimate the null correlation between the tests being aggregated and subsequently use a copula approach to estimate the p-value of the minimum p-value test statistic (See Appendix B.3 for details) [Demarta and McNeil, 2005, Dutta et al., 2019]. This approach has also been used previously to integrate information from multiple functional annotations [He et al., 2017]. Specifically we consider the following tests:

1. Meta-Hom: minimum p-value of Meta-MultiSKAT tests with Σ_S= $\Sigma_{S;Hom}$ across different choices of Σ_P and Σ_G. Specifically, we consider the following four different choices of (Σ_P, Σ_G):

   - $\Sigma_P = \widehat{V}_s$, Σ_G = SKAT
   - $\Sigma_P = \widehat{V}_s$, Σ_G = Burden
   - $\Sigma_P = I_K$, Σ_G = SKAT
   - $\Sigma_P = I_K$, Σ_G = Burden

   The minimum p-value across these four tests will be used as the test statistic to evaluate the associations.

2. Meta-Het: minimum p-value of Meta-MultiSKAT tests with $\Sigma_S = \Sigma_{S;Het}$ across different choices of Σ_P and Σ_G. We will use the same four sets of (Σ_P, Σ_G) as in Meta-Hom.

3. Meta-Com: combined test of Meta-Hom and Meta-Het. We use the minimum p-value of the tests used in Meta-Hom and Meta-Het as test statistics.

3.3 Simulation

We carried out extensive simulation studies to evaluate the type I error rate and power of Meta-MultiSKAT tests. For type-I error simulations and all power simulations, we generated 10,000 chromosomes over 1Mb-regions using a coalescent simulator with a European ancestry model [Schaffner et al., 2005]. We randomly selected a 3 kb sub-region for each simulated dataset to test for associations.

3.3.1 Simulation setting within individual study

In the s^{th} study, we generate K phenotypes according to the linear model:

$$y_i \sim MVN\{(\beta_1 G_1 + \beta_1 G_1 + \cdots + \beta_s G_s) I_s, V_{s;\rho}\}$$

where $V_{s;\rho}$ is the covariance of the non-systematic error term. We use $V_{s;\rho}$ to define level of residual covariance between the traits. The matrix $V_{s;\rho}$ is set to be compound symmetric throughout all the simulation settings with varying values of the correlation parameter ρ (low correlation $\rho = 0.3$; moderate correlation $\rho = 0.5$; high correlation $\rho = 0.7$). I_s is a $K \times 1$ indicator vector, which has 1 when the corresponding phenotype is associated with the region and 0 otherwise. Throughout our simulations we set $I_s = (1, 1, 1, 0, 0)^T$ meaning the first 3 phenotypes are associated with the region of interest in a particular study.

For estimating type-1 error rates we set $\beta_i = 0$ for all the variants in all the studies.

For power simulation, we used two different settings. In the first setting, to estimate the power of Meta-MultiSKAT as a rare-variant test, we set 30% of the rare variants (MAF $\leq 1\%$) to be causal. Next, to estimate the performance of Meta-MultiSKAT-Common-Rare as a test of combined effects of common and rare variants, we set 30% of all variants (common or rare) in the region to be causal. We modeled rare variants to have stronger association with the phenotypes than the common variants by setting $|\beta_j| = clog_{10}|MAF_j|$ with $c = 0.2$ for all the simulation scenarios. For both the settings, as mentioned earlier, the first three among the five phenotypes in each study were associated with the region of interest.

3.3.2 Simulation settings across studies

Throughout our simulations we have used settings which consist of three studies on European samples with five phenotypes of interest. The sample sizes for the studies were 2000, 2000 and 1000 respectively. To assess the performance of Meta-MultiSKAT under scenarios of missing data, we considered the following 3 scenarios:
Scenario A: all the individuals in each of the study have complete information on 5 correlated phenotypes
Scenario B: 10% samples (chosen completely at random) in the 3rd study have information on 4 phenotypes only. This means, 100 samples in study 3 have information on 4 phenotypes and misses information on 1 phenotype, while the rest 900 samples have information on all the 5 phenotypes. All the 2000 samples in study 1 and 2 have complete information on all the 5 phenotypes.
Scenario C: The 5th phenotype for study 3 is missing for all the samples. For study 1 and 2, all the 2000 samples have complete information on all the 5 phenotypes.
For these above scenarios, in addition to the Meta-MultiSKAT tests (Meta-Hom, Meta-Het and Meta-Com), we evaluated the following single phenotype-based approaches:

1. MinPhen-Het: Bonferroni-adjusted minimum p-value from the single phenotype region-based meta-analysis using Heterogeneous Meta-SKAT-O (Het-Meta-SKAT-O)

2. MinPhen-Hom: Bonferroni-adjusted minimum p-value from the single phenotype region-based meta-analysis using Homogeneous Meta-SKAT-O (Hom-Meta-SKAT-O)

3.4 Meta-analysis of white blood cell traits

To investigate the pleiotropic roles of low frequency and rare-variants on WBC subtypes, we analyzed data collected under the Michigan Genomics Initiative (MGI study) [Fritsche et al., 2018] Phase 2 (data-freeze on December 2017) and the SardiNIA [Sidore et al., 2015, Vacca et al., 2006] study. Data on four WBC subtypes percentages were included in the analysis: lymphocyte, monocyte, basophil and eosinophil. We excluded the data on percentage of neutrophils since it was highly correlated with lymphocytes (absolute value of correlation > 0.9 in both MGI and SardiNIA). European samples with at most two phenotypes missing were included in the analysis for each of the studies. In all, we included 11,049 and 5,899 samples from the MGI and the SardiNIA studies, respectively (Table 3.1). We annotated protein-coding variants and a region of 20kb ($\pm$ 10kb) around them to genes using Variant Effect Predictor [McLaren et al., 2010] software. Within each study, we included age, sex, and study specific top four principal components (PC) as fixed effect covariates in the analysis. In each study each of the four WBC subtypes were adjusted for the corresponding covariates and the residuals were quantile-normalized. Further, we estimated the kinship between the subjects in each study using KING [Manichaikul et al., 2010b] and estimated the co-heritability matrix of the phenotypes using PHENIX [Dahl et al., 2016]. The inverse normalized residuals were then

used in region-based multiple phenotype analysis (Multi-SKAT with kinship correction). The required summary statistics were calculated from the individual tests. We conducted three sets of analysis with the extracted summary statistics. First to test the rare-variant associations of the phenotypes, we used Meta-MultiSKAT tests (Meta-Het, Met-Hom and Meta-Com) to test groups of protein-coding variants with pooled MAF $\leq 1\%$. We only included the groups that had at least three variants and a total minor allele count of 5. We used a Beta(1,25) weighting scheme to upweight the effect of the rare variants. Next, to test the combined effect of common and rare variants, we used the Meta-MultiSKAT-Common-Rare versions of the above tests with groups of protein-coding variants without any MAF cutoff. This means both common (MAF $> 1\%$) and rare variants (MAF $\leq 1\%$) were present in the regions tested. For the rare variants we used Beta(1,25) weights and for the common variants we used Beta(0.5,0.5) weights (see Methods). Further, we annotated CADD scores for all the variants (common and rare) using ANNOVAR [Wang et al., 2010]. We used these scores as weights in the genotype kernel Σ_G and performed the above Meta-MultiSKAT tests.

3.5 Results

3.5.1 Type-I error

For type-I error simulations, we simulated 10^7 independent datasets with three studies each having five phenotypes with a compound symmetric null residual covariance structure having off-diagonal elements equal to 0.5, i.e. $V_{s;0.5}$. The MAF spectrum for the population allele frequencies shows that the majority of the simulated variants are rare (MAF $\leq 1\%$). We estimated the type-1 error rate as the proportion of p-values less than the specified α levels, with α set at 10^{-4}, 10^{-5} and 2.5×10^{-6}.

Type-I error rates were well maintained at all levels. For example, at $\alpha = 2.5 \times 10^{-6}$, the largest estimated type-I error rate for any of the Meta-MultiSKAT tests was 2.7×10^{-6}, which was well within the estimated 95% confidence interval (Table 3.2).

3.5.2 Power

We compared the empirical power of Meta-MultiSKAT tests with two possible existing approaches: minimum of the single phenotype MetaSKAT p-values (MinPhen-Hom and MinPhen-Het). For each simulation setting, we generated 1000 datasets and estimated the empirical power as the proportion of p-values less than 2.5×10^{-6}, reflecting the Bonferroni correction for testing 20,000 independent genes. In power simulations, the first scenario considered the case that each study has the same set of causal variants and all of them are trait-increasing. Meta-Hom and Meta-Com had the highest powers in all scenarios while the power for Meta-Het is lower (Figure 3.1). Also, there was a slight overall decrease in power from scenario A through scenario C. We expect this decrease in power since there is an increase of the amount of missing-ness in the scenarios A through C, though the power decrease is small (maximum relative decrease in empirical power $< 1\%$). Overall power of all the methods was higher when the correlation is high ($\rho = 0.7$).

Next, we considered a heterogeneous situation in which causal variants for each study were randomly selected so only small percentage of causal variants were shared among studies (Figure 3.2). As expected, Meta-Het and Meta-Com had high power among the tests being compared. Meta-Hom was underpowered compared to these tests, while MinPhen-Hom and MinPhen-Het had lower power than the rest. We then assumed that the causal variants for each study are chosen randomly within the region and 20% of the variants are trait-decreasing (80% are trait increasing) (Figure 3.3). Similar to the previous scenario, Meta-Het and Meta-Com had higher power than the rest of the tests. MinPhen-Hom and MinPhen-Het had lower power of

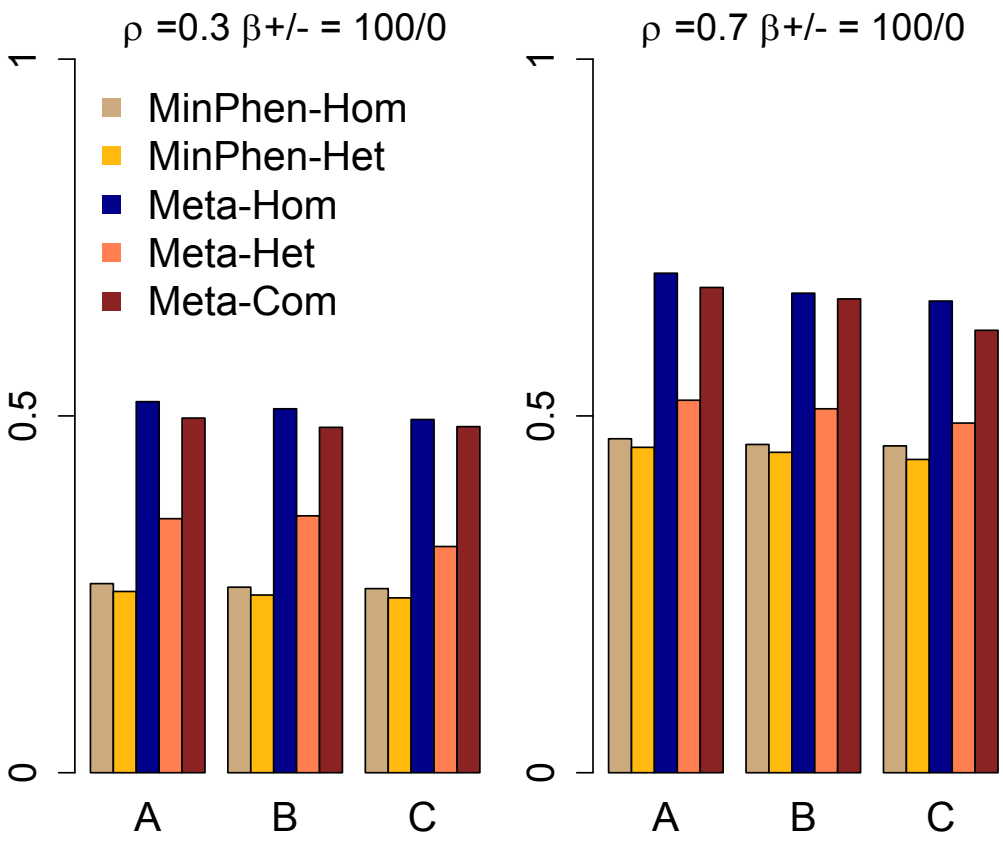


Figure 3.1: Power for Meta-MultiSKAT tests compared with the existing methods when the set of causal variants is the same across different studies and has the same direction of effect. Empirical power for Meta-Hom, Meta-Het and Meta-Com plotted for 3 different scenarios compared against MinPhen-Hom and MinPhen-Het (See Simulations for details). Left panel shows the results for low correlation ($\rho = 0.3$) among the phenotypes and right panel shows the results for high correlation ($\rho = 0.7$)

detecting association signals, and Meta-Hom consistently had the lowest power across all the settings. Next we considered a situation where the correlation structure among the phenotypes across studies varies. For the 1st and 2nd study the correlation among the 5 phenotypes is high ($\rho = 0.7$) while for the 3rd study, the correlation among the 5 phenotypes is moderate ($\rho = 0.5$). Similar to the previous cases, Meta-Het and Meta-Com maintained higher power than the rest of the tests (Figure 3.4). As before, Meta-Hom performed poorly when 20% of the causal variants are trait-decreasing. We further estimated type-1 error and power for the Meta-MultiSKAT-Common-Rare versions of the tests. The results are shown in Appendix Table B.2, Appendix Figure B.2 and B.3. Type-I error was well maintained at different levels and the patterns of estimated power remained the same.

Overall our simulations show that Meta-MultiSKAT tests can increase power over the existing single phenotype-based meta-analysis approaches, while controlling type-I error rates. In particular, Meta-Com maintains robust power across all the scenarios regardless of the underlying genetic model.

3.5.3 Meta-analysis of WBC subtype traits

White blood cells (WBCs) are major cellular components of the human immune system. They have been found to be associated with risk of cardiovascular disease [Kim et al., 2017] and cancer mortality [Erlinger et al., 2004] among others. Certain disease risk factors including high blood pressure, cigarette smoking, adiposity and increased levels of plasma inflammatory markers have been reported to be associated with elevated WBC counts [Hasegawa et al., 2002, Mu oz et al., 2012]. WBCs are classified into subtypes according to the functionality and morphology. Abundances (counts or percentage) of these WBC subtypes have been found to be important biomarkers for diseases including COPD [Kim et al., 2012] and rheumatoid arthritis [Salomon et al., 2017], and several GWAS have identified genetic variants associ-

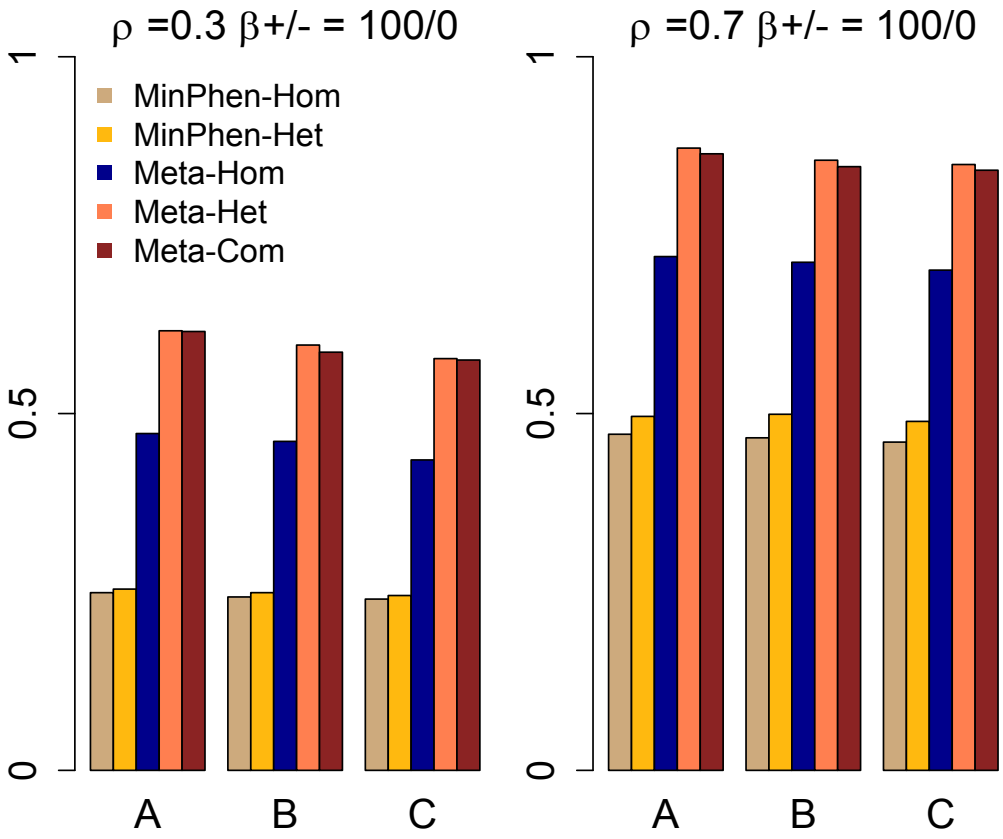


Figure 3.2: Power for Meta-MultiSKAT tests compared with the existing methods when the set of causal variants is randomly chosen for each study and has the same direction of effect. Empirical power for Meta-Hom, Meta-Het and Meta-Com plotted for 3 different scenarios compared against MinPhen-Hom and MinPhen-Het (See Simulations for details). Left panel shows the results for low correlation ($\rho = 0.3$) among the phenotypes and right panel shows the results for high correlation ($\rho = 0.7$).

ated with them [Astle et al., 2016, Crosslin et al., 2013, Kanai et al., 2018, Keller et al., 2014]. In this analysis, we used the abundance percentages (quantile normalized) of lymphocyte, monocyte, basophil and eosinophil as phenotypes. Correlations among the phenotypes for each study are shown in Appendix Figure B.4. Correlation estimates between the WBC subtype traits within MGI samples appear to be higher in magnitude as compared to samples in SardiNIA study. We applied Meta-MultiSKAT tests to the analysis of WBC subtypes from the MGI and SardiNIA studies (See Methods for details). In particular, we applied Meta-Het, Meta-Hom

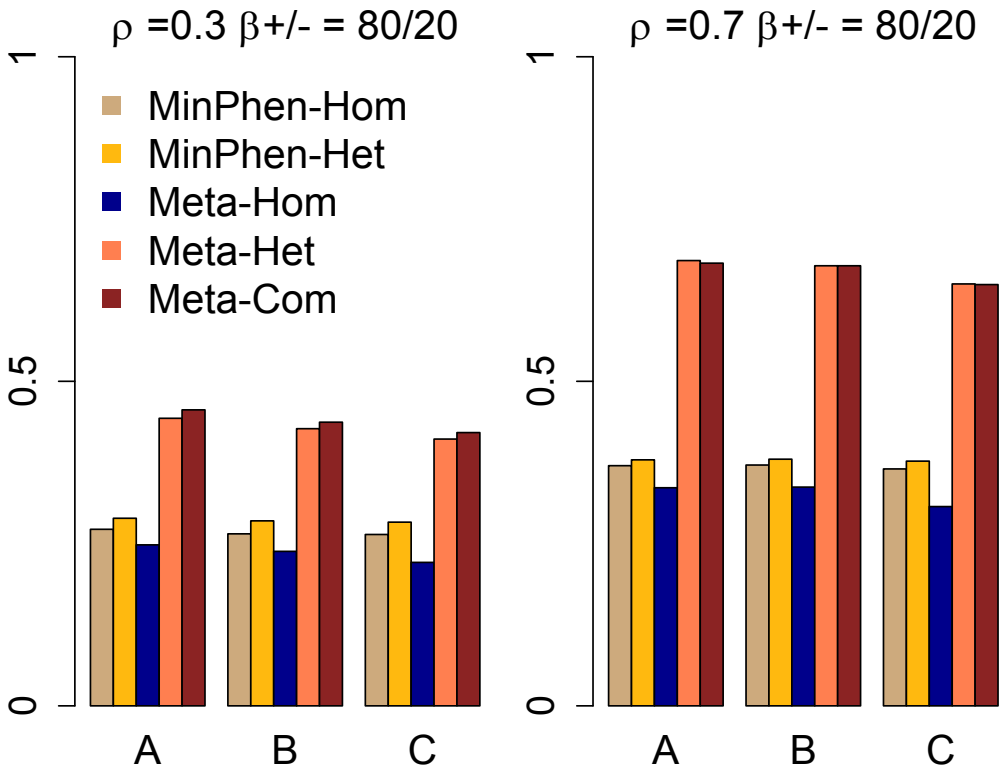


Figure 3.3: Power for Meta-MultiSKAT tests compared with the existing methods when the set of causal variants is randomly chosen for each study and 20% of the causal variants are traitdecreasing. Empirical power for Meta-Hom, Meta-Het and Meta-Com plotted for 3 different scenarios compared against MinPhen-Hom and MinPhen-Het (See Simulations for details). Left panel shows the results for low correlation ($\rho = 0.3$) among the phenotypes and right panel shows the results for high correlation ($\rho = 0.7$)

and Meta-Com tests along with MinPhen-Hom and MinPhen-Het. We also evaluated the single-phenotype tests and multiple-phenotype tests (Multi-SKAT) for each study (Appendix Table B.1).

3.5.3.1 Results for rare variants with MAF-based weighting

First, we used the MAF-based weighting scheme to upweight the rare variants as suggested by [Wu et al., 2011]. Using the variants with pooled MAF $\leq$1%, we used Beta(1,25) weights. Overall 5,109 genes with at least 3 variants and a total

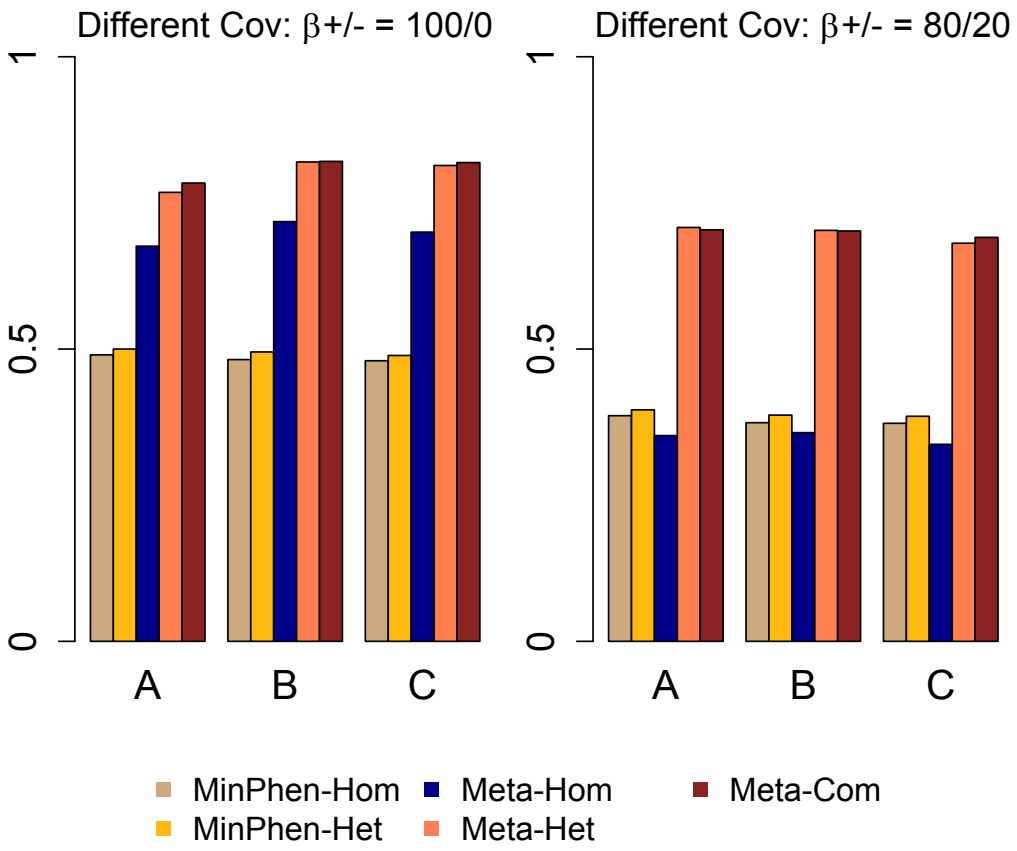


Figure 3.4: Power for Meta-MultiSKAT tests when the set of causal variants is randomly chosen for each study and the studies have different covariance structure across the phenotypes. Empirical power for Meta-Hom, Meta-Het and Meta-Com plotted for 3 different scenarios compared against MinPhen-Hom and MinPhen-Het (See Simulations for details). Left panel shows the results when all the causal variants are trait increasing; right panel shows the same when 20% of the causal variants are trait-decreasing

minor allele count > 5 were tested. This produces a Bonferroni cut-off of 9.8×10^{-6} (approximately 1×10^{-5}) The QQ-plots shown in Figure 3.5 corresponding to the Meta-MultiSKAT tests do not show any indication of inflation (genomic control varying from 0.998 to 1.003). Table 3.3 shows the genes that had p-values less than 1×10^{-5} for at least one of the tests. Two genes *PRG2* (p-value $= 5.9 \times 10^{-7}$) and *RP11-872D17.8* (p-value $= 1.7 \times 10^{-7}$) were identified as significant by Meta-MultiSKAT tests while the p-values for the existing tests did not reach significance.

PRG2 gene [MIM: 605601] encodes a protein, which is a major contributor to the crystalline core of the eosinophil granule. Multiple phenotype analysis (Multi-SKAT) shows evidence for a strong association in the SardiNIA study (p-value = 2.8×10^{-7}) whereas the p-value in the MGI study (p-value = 0.76) does not show evidence for association (Appendix Table B.1). This signal is driven by the association of the gene with eosinophils in SardiNIA (SKAT-O p-value = 3.7×10^{-7}; Appendix Table B.1). A low-frequency SNP at 11:57156106 (A/G; MAF 3% in SardiNIA), which is significantly associated with the eosinophil percentages (p-value = 9.3×10^{-12}). This variant is only observed for the individuals in SardiNIA study and was not observed in MGI. The signal for *RP11-872D17.8*, an adjacent gene to *PRG2*, is also driven by the same variant.

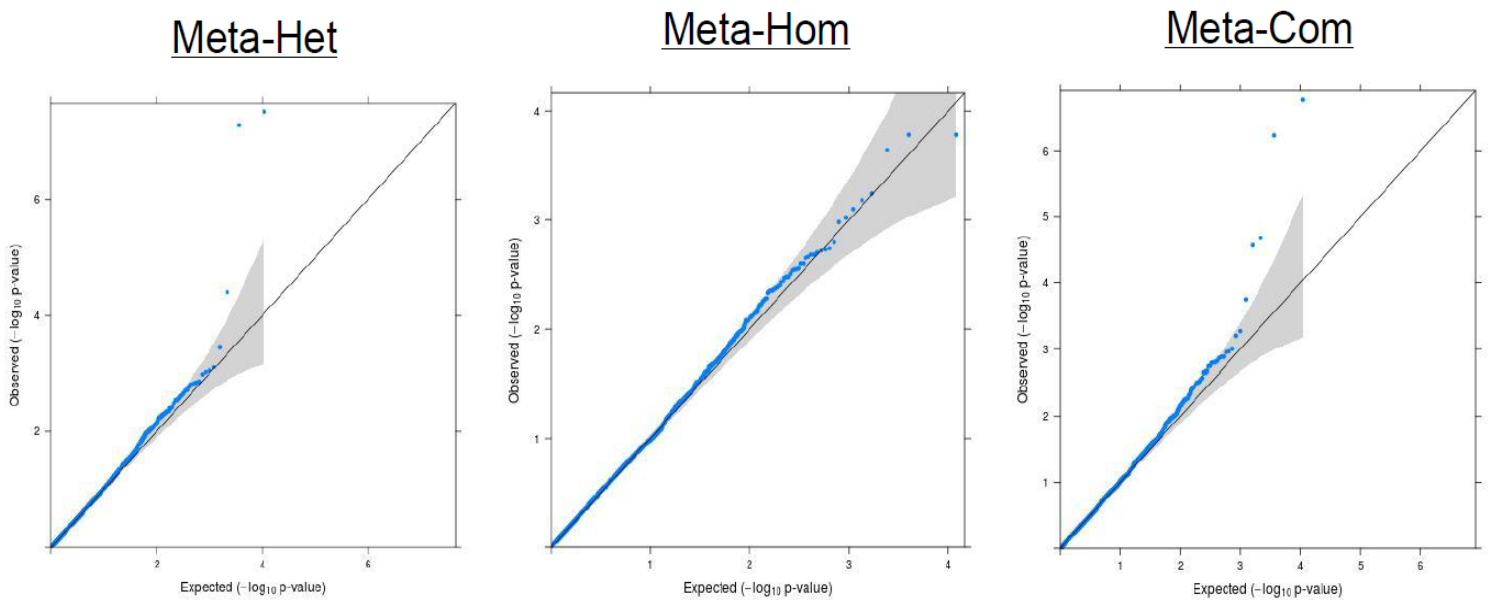


Figure 3.5: QQ plots for the Meta-MultiSKAT (Meta-Het, Meta-Hom and Meta-Com respectively) p-values obtained from MGI-SardiNIA meta-analysis.

3.5.3.2 Results with combined effects of common and rare variant

To illustrate how Meta-MultiSKAT can test the combined effect of common and rare variants, we used Meta-MultiSKAT-Common-Rare version of each test to analyze the WBC data from the MGI and SardiNIA studies. All the variants, common and rare, were used in this analysis. The same two genes, *PRG2* and *RP11-872D17.8*, had p-values less than 1×10^{-5}, but with different p-values along with *CCL24* and *IRF8* (Table 3.4). Among the genes that showed signal, Chemokine (C-C motif) ligand 24 gene (*CCL24* [MIM: 602495]) encodes a protein that interacts with chemokine receptor CCR3 to induce chemotaxis in eosinophils [White et al., 1997]. This chemokine is also strongly chemotactic for resting T lymphocytes and slightly chemotactic for neutrophils [Salcedo et al., 2002]. Multiple phenotype tests of *CCL24* did not reach significance in any of the individual studies (p-value in SardiNIA = 1.3×10^{-2}; p-value in MGI = 2.0×10^{-4};Appendix Table B.1). But Meta-Het (p-value = 4.8×10^{-7}) and Meta-Com (p-value = 8.1×10^{-7}) are significant indicating the utility of meta-analysis to identify this signal.

Interferon regulatory factor 8 (*IRF8* [MIM: 601565]) at 16q24.1 has been previously reported as associated with several WBC subtype traits. *IRF8* has been found to be associated with monocyte count [Sichien et al., 2016] and has also been identified as a multiple sclerosis susceptibility locus [De Jager et al., 2009]. Animal model studies showed that *IRF8* as a transcription factor plays an essential role in the regulation of lineage commitment during monocyte differentiation [Kurotaki et al., 2018, Yáñez et al., 2015]. [Astle et al., 2016] found several associations of *IRF8* with WBC subtype traits like Neutrophils (high correlation with Lymphocytes) and combinations (sum of neutrophil and basophil counts) in the UK Biobank. Meta-Hom (p-value=1.8×10^{-10}) and Meta-Com (p-value = 2.9×10^{-10}) show evidence for strong association. For *IRF8* (p-value = 2.9×10^{-10}), *CCL24* (p-value = 8.1×10^{-7}) and *PRG2* (p-value = 4.2×10^{-8}) the combined effect of common and rare variants

produces substantially more significant results compared to the MAF-based weighting with rare variants, without evidence of inflated false discoveries. In comparison, the p-values for *RP11-872D17.8* (p-value $= 1.7 \times 10^{-7}$) remain approximately of the same order of significance. The results from this analysis demonstrate that Meta-MultiSKAT can be applied as a region-based test for testing the combined effect of common and rare variants.

3.5.3.3 Results with CADD-score weighting with both common and rare variants

We reanalyzed the WBC data from the MGI and SardiNIA studies by weighting the variants with a functional score. Both common and rare variants were used in this analysis. We used CADD-scores as weights in $\Sigma_G = WR_GW$. The results are shown in Table 3.5. The same set of 4 genes identified using MAF-based weighting remained significant (p-value $< 1 \times 10^{-5}$), but with slightly different p-values. For *PRG2* (p-value $= 1.2 \times 10^{-8}$), *IRF8* (p-value $= 1.6 \times 10^{-7}$) and *CCL24* (p-value $= 1.1 \times 10^{-6}$), weighting by functional scores resulted in a slightly smaller p-value as compared to rare-variant test.

Further, Appendix Table B.1 lists the single phenotype and multiple phenotype p-values for these four genes in each study. It is to be noted that, no other gene had a p-value less than 1×10^{-5} in any of the single phenotype or multi-phenotype tests in each study. As a further illustration, we performed Meta-MultiSKAT tests by masking lymphocyte data in SardiNIA and treating that as a missing phenotype (See Appendix B.4 for details). The results (Appendix Table B.3) show that Meta-MultiSKAT has a robust power under such scenarios while controlling type-1 error.

3.5.4 Computation Time

We estimated the computation time of Meta-MultiSKAT tests using simulated datasets on 3 studies (as described in the Simulations section) with 5 phenotypes and 50 genetic variants. We set the number of perturbation iterations to 1000. On average, Meta-Hom and Meta-Het tests required approximately 8 CPU-seconds (Intel Xeon 2.80 GHz) and Meta-Com required 12 CPU-sec. Analyzing the MGI and SardiNIA datasets, using the extracted summary statistics from each study, required about thirty CPU-hours when parallelized to 10 processes.

3.6 Discussions

We propose a new method, Meta-MultiSKAT, which meta-analyzes region-based association of multiple phenotypes across studies. The model is based on study-specific summary statistics for the region and is flexible to accommodate a range of heterogeneity of genetic effects across studies. The simulation and the real data analysis results involving the summary statistics from MGI and SardiNIA demonstrate that Meta-MultiSKAT can substantially increase power compared to the existing tests and can identify additional association signals, while maintaining the desired type-I error rate. The method is implemented as an R-package.

We note that the test statistics, assuming homogeneous genetic effects, are essentially identical to joint analysis test statistics using all individual level data and accounting for study-specific covariate effects, resulting in nearly identical power using meta-analysis and joint analysis. Our power-simulations confirm this finding (Appendix Figure B.1).

For Meta-MultiSKAT tests with a given choice of Σ_S , Σ_P and Σ_G, asymptotic p-values can be calculated. For Meta-Hom and Meta-Het, we are aggregating four such Meta-MultiSKAT tests for a given choice of Σ_S , Σ_P and Σ_G. Although the

corresponding p-values depend on a resampling scheme, they still can be calculated using a small number of perturbations. Similarly, p-values for Meta-Com that aggregates 8 Meta-MultiSKAT tests with a given choice of Σ_S , Σ_P and Σ_G can also be calculated using a small number of resampling iterations. The reported computation times show that Meta-MultiSKAT tests are computationally manageable at a genome-wide level. In addition, Meta-MultiSKAT retains the desirable properties of Multi-SKAT. For instance, Multi-SKAT effectively incorporates kinship information through a regression framework, allowing the use of the whole sample rather than only unrelated individuals in a particular study. Meta-MultiSKAT can use the kinship adjusted summary statistics from the Multi-SKAT tests across several studies to produce a test of association, in which kinship information for each study has been incorporated. This integration allows for the use of all samples for each of the studies, further augmenting statistical power.

The asymptotic p-value calculations for Meta-MultiSKAT rely on the normality assumption of the score vectors. When at least one pair of the phenotypes is very strongly correlated (i.e. absolute correlation > 0.9), this assumption may be violated. Currently, we do not have a mechanism to adaptively select an active set of phenotypes which might produce the optimal association signal for a particular region. Hence, we recommend that the data be pre-pruned for correlation and such strongly correlated phenotypes be excluded before analysis. Currently, the framework of Meta-MultiSKAT is developed for continuous phenotypes. A direction of future research is to extend this framework for phenotypes that are a mixture of continuous and discrete types.

In summary, we have developed Meta-MultiSKAT, a meta-analysis method for testing rare-variant associations of multiple correlated phenotypes. Meta-MultiSKAT has robust power and can handle practical problems such as missing data and different covariance structures. The method provides a scalable and practical solution to

test multiple phenotypes jointly and thus can contribute to detecting regions in the genome with pleiotropic effects.

Table 3.1: Estimated Type-1 error rates for Meta-MultiSKAT tests

α	Meta-Hom	Meta-Het	Meta-Com
2.5×10^{-6}	2.3×10^{-6}	2.3×10^{-6}	2.7×10^{-6}
1×10^{-5}	1.1×10^{-5}	1.1×10^{-5}	1.2×10^{-5}
1×10^{-4}	1.2×10^{-4}	1.2×10^{-4}	1.3×10^{-4}

Table 3.2: Significant genes identified by Meta-MultiSKAT using rare variants. Genes/ regions identified by either of the Meta-MultiSKAT methods (Meta-Hom, Meta-Het or Meta-Com) or the existing approaches (MinPhen-Hom or MinPhen-Het). The p-values $< 10^5$ were marked in bold. Variants with pooled MAF $\leq$1% are included as rare.

Gene	Rare SNPs	Meta-Het	Meta-Hom	Meta-Com	MinPhen-Hom	MinPhen-Het
RP11-872D17.8	3	$\mathbf{5.3 \times 10^{-8}}$	2.1×10^{-4}	$\mathbf{1.7 \times 10^{-7}}$	2.9×10^{-5}	1.3×10^{-5}
PRG2	2	$\mathbf{3.1 \times 10^{-8}}$	6.1×10^{-4}	$\mathbf{5.9 \times 10^{-7}}$	6.3×10^{-5}	1.1×10^{-5}

Table 3.3: Significant genes identified by Meta-MultiSKAT-Common-Rare. Genes/ regions identified by either of the Meta-MultiSKAT-Common-Rare methods (Meta-Hom, Meta-Het or Meta-Com) or the existing approaches (MinPhen-Hom or MinPhen-Het). The p-values $< 10^5$ were marked in bold. Variants with pooled MAF $\leq$1% ($>$ 1%) are included as rare (common)

Gene	Common SNPs	Rare SNPs	Meta-Het	Meta-Hom	Meta-Com	MinPhen-Hom	MinPhen-Het
IRF8	28	3	1.3×10^{-4}	$\mathbf{1.8 \times 10^{-10}}$	$\mathbf{2.9 \times 10^{-10}}$	7.4×10^{-3}	4.9×10^{-4}
PRG2	5	2	$\mathbf{3.1 \times 10^{-8}}$	6.1×10^{-4}	$\mathbf{5.9 \times 10^{-7}}$	6.3×10^{-5}	1.1×10^{-5}
CCL24	12	1	$\mathbf{4.6 \times 10^{-7}}$	8.3×10^{-4}	$\mathbf{8.1 \times 10^{-7}}$	8.2×10^{-2}	2.2×10^{-4}
RP11-872D17.8	9	3	$\mathbf{5.3 \times 10^{-8}}$	2.1×10^{-4}	$\mathbf{1.7 \times 10^{-7}}$	2.9×10^{-5}	1.3×10^{-5}

Table 3.4: Significant genes identified by Meta-MultiSKAT using CADD weights. Genes/ regions identified by either of the Meta-MultiSKAT methods (Meta-Hom, Meta-Het or Meta-Com) with functional (CADD) score weights were used for the variants. The p-values $< 10^5$ were marked in bold. Variants with pooled MAF $\leq$1% ($>$ 1%) are included as rare (common)

Gene	Common SNPs	Rare SNPs	Meta-Het	Meta-Hom	Meta-Com	MinPhen-Hom	MinPhen-Het
IRF8	28	3	1.2×10^{-5}	$\mathbf{9.4 \times 10^{-8}}$	$\mathbf{1.6 \times 10^{-7}}$	7.4×10^{-3}	4.9×10^{-4}
PRG2	5	2	$\mathbf{7.4 \times 10^{-9}}$	3.6×10^{-4}	$\mathbf{1.2 \times 10^{-8}}$	6.3×10^{-5}	1.1×10^{-5}
CCL24	12	1	$\mathbf{8.2 \times 10^{-7}}$	5.7×10^{-4}	$\mathbf{1.1 \times 10^{-6}}$	8.2×10^{-2}	2.2×10^{-4}
RP11-872D17.8	9	3	$\mathbf{1.0 \times 10^{-6}}$	1.1×10^{-3}	$\mathbf{2.3 \times 10^{-6}}$	2.9×10^{-5}	1.3×10^{-5}

Table 3.5: Sample sizes for each phenotype in each study for MGI-SardiNIA meta-analysis

Phenotype	N (MGI)	N (SardiNIA)	N (Total)
Lymphocyte%	11049	5895	16944
Monocyte%	11049	5876	16925
Basophil%	11038	5866	16904
Eosinophil%	11037	5795	16830
Complete cases	11035	5735	16770
Total Samples	11049	5898	16947

CHAPTER IV

A powerful Gene-set analysis method to identify active genes with applications to Biobank-based association studies

4.1 Introduction

In the past few years, genome-wide association studies (GWAS) have identified thousands of genetic variants associated with hundreds of complex traits. However, the variants identified so far, individually or in combination, account for only a small proportion of the inherited component of disease risk for most of the complex diseases [Manolio et al., 2009]. A possible explanation is that due to the large number of genetic polymorphisms examined in GWAS and the massive amount of tests conducted, real but weak associations are likely to be missed after multiple comparison adjustment [Liu et al., 2010].

Gene-set analysis (GSA) have been suggested as potentially more powerful alternative to the single-variant or single-gene analysis performed in GWAS, specially in order to identify weak to moderate effects [Cantor et al., 2010]. In GSA, individual genes are aggregated to groups sharing certain biological or functional characteristics and association of the trait with the gene-set is tested for significance. This consid-

erably reduces the number of tests that need to be performed, and makes it possible to detect effects consisting of multiple weaker associations that would otherwise be missed [Fridley and Biernacka, 2011, Yu et al., 2009]. Additionally, the majority of complex phenotypes are manifested through a concerted activity of many variants. Thus in such cases a gene-set analysis can provide insight into the involvement of specific biological pathways, cellular functions or gene-sets defined using predefined biological functions in the genetic architecture of the phenotype [Pers, 2016].

In recent years, several novel statistical methods to perform GSA have been published [Lee et al., 2012a, Jia et al., 2011, Segre et al., 2010, O'Dushlaine et al., 2009, Holmans, 2010, Lips et al., 2015, Wang et al., 2007, Mooney et al., 2014, de Leeuw et al., 2015, Pan et al., 2015, Sun et al., 2019] and have successfully discovered gene-sets associated with numerous complex diseases [Locke et al., 2015, Allen et al., 2010, Nurnberger et al., 2014]. However there are several concerns regarding the properties and applicability of these methods. Existing GSA methods often demonstrate low power [Jia et al., 2011], especially in situations where only a few genes within the gene-set are associated with the trait [Sun et al., 2019]. Further, several methods depend on permutation or simulation-based p-values which can be computationally challenging and hence can reduce the applicability of the method [Holmans, 2010]. Additionally, in presence of linkage disequilibrium or gene-gene correlation, many existing methods cannot control the type-I error [Moskvina et al., 2012].

Another key challenge is the question of interpretability which only a few methods have tried to address. Although the existing GSA method produce a p-value for association between the gene-set and the trait, it remains important to understand the genes that are active within the gene-set. Usual set-based methods fail to identify such genes which individually or in congregation might be driving the association signal. Thus, the granularity of information on the active set of genes in the gene-set is lost which might be important in further downstream analysis and eventually

using the results as therapeutic targets.

In this project we describe a subset-based gene-set association method Gene-set analysis Association Using Sparse Signals (GAUSS) which aims to provide a solution to the issues mentioned above. In general, standard GSA aim to evaluate two types of null hypotheses, namely the competitive null hypothesis, i.e., the genes in a gene-set of interest are no more associated with the outcome than any other genes outside this set, and the self-contained null hypothesis, i.e., none of the genes in the gene-set is associated with the outcome. The GAUSS procedure focuses on the self-contained null hypothesis, as our main goal is to identify gene-sets associated with the trait of interest.

Our method identifies a subset within the gene-set which carries the maximum signal of association and evaluates its p-value through a fast-simulation approach. Additionally, this subset of genes within the gene-set, is the active set that appears to drive the association. Using several pre-computed correlation matrices from publicly available reference data-sets we show that the computational burden of our method can be substantially reduced and can be efficiently applied to large biobank-scale datasets. Furthermore, the test we propose can be conducted using publicly available GWAS summary level information only.

Using simulation studies we show that under a variety of association models, GAUSS can be more powerful than existing methods while maintaining the correct type-I error. We applied the method to evaluate the associations of $1,201$ phenotypes from UK Biobank [Bycroft et al., 2018] with $10,679$ gene-sets derived from the molecular signature database (MSigDB) [Liberzon et al., 2015], thus proving it feasible to be applied to such large-scale data and gaining newer insights into the genetic architecture of the phenotypes.

4.2 Methods

To conduct the GAUSS test we need p-values for the regions or genes present in the gene-set. Popular gene-based tests like SKAT [Wu et al., 2011], SKAT-Common-Rare [Ionita-Laza et al., 2013], SKAT-O [Lee et al., 2012b], prediXan [Gamazon et al., 2015] and others can be used for obtaining the p-values when individual level data are available. If GWAS summary statistics (effect size, standard error, p-value, minor allele frequency for each variant) are available, we can approximate the gene-based p-values using LD information from a suitable reference panel as outlined in 4.2.3.

Given the gene-based p-values, constructing the GAUSS test for a given gene-set or pathway consists of two major steps as follows:

- Step 1: Construct GAUSS test statistic for a given gene-set or pathway
- Step 2: Obtain the p-value for GAUSS test statistic using a suitable reference panel

4.2.1 Step 1: GAUSS test statistic

We start with the z-statistics for gene-based p-values for the J genes in the gene-set $\mathbb{G}$. In our data applications here, we have used publicly available GWAS summary statistics and a reference panel of Europeans from 1000 Genomes data [The 1000 Genomes Project Consortium, 2015] to estimate SKAT-Common-Rare p-values (4.2.3). Other gene-based tests which are not of the SKAT-family, like prediXan, can also be used to obtain gene-based p-values.

For any non-empty subset $B \subseteq \mathbb{G}$, we define $S(B)$ the association score for subset B as $S(B) = \frac{\sum_{b \in B} z_b}{\sqrt{|B|}}$ where $|B|$ is the number of genes in subset B. We define the GAUSS statistic for the gene-set $\mathbb{G}$ as the maximum score of any non-empty subset of $\mathbb{G}$.

$$GAUSS(\mathbb{G}) = \max_{B \subseteq \mathbb{G}} \frac{\sum_{b \in B} z_b}{\sqrt{|B|}} \tag{4.1}$$

where $|B|$ denotes the number of genes in the subset B. Such maximum-type statistics have previously been used in context of multiple phenotypes, meta-analysis [Bhattacharjee et al., 2012] and gene-environment interaction [Yu et al., 2018]. Additionally $B_{opt} = arg\max_{B\subseteq\mathbb{G}} \frac{\sum_{b\in B} z_b}{\sqrt{|B|}}$, the subset that has the maximum association score, is termed the active subset (AS)

The maximum is over all possible $2^J - 1$ subsets of $\mathbb{G}$, but the computational complexity can be greatly reduced by the following observations:

$$GAUSS(\mathbb{G}) = \max_{k\in\{1,\cdots J\}} \max_{B_k\subseteq\mathbb{G}} \frac{\sum_{b\in B_k} z_b}{\sqrt{|B_k|}}$$

where B_k denotes a subset of $\mathbb{G}$ with k elements. It is easy to see that

$$\max_{B_k\subseteq\mathbb{G}} \frac{\sum_{b\in B_k} z_b}{\sqrt{|B_k|}} = \frac{z_{(1)} + z_{(2)} + \cdots + z_{(k)}}{\sqrt{k}}$$

where $z_{(1)}, z_{(2)}, \cdots z_{(J)}$ are the z-statistics sorted in a decreasing order with $z_{(1)}$ as the maximum of the J z-statistic. We implement the following algorithm to obtain the GAUSS statistic as

- Order the z-statistics for the J genes as $z_{(1)}, z_{(2)}, \cdots z_{(J)}$ in a decreasing order, where $z_{(1)}$ denotes the maximum of the J z-statistic.
- Starting with $i = 1$, we compute $S_{(i)} = \frac{\sum_{l=1}^{i} z_{(l)}}{\sqrt{i}}$ for all $i = 1, \cdots, J$
- GAUSS test statistic is $\max_{i\in\{1,\cdots,J\}} S_{(i)}$

This algorithm has a computational complexity of $\mathcal{O}(JlogJ)$.

4.2.2 Step 2: Fast estimation of the p-value of GAUSS

We use fast two-step approach which uses normal-Copula to estimate p-values of GAUSS. We first estimate the null correlation structure ($\widehat{V}_{\mathbb{G}}$) among the z-statistics

$z_{(1)}, z_{(2)}, \cdots z_{(J)}$ in $\mathbb{G}$ through a small number of simulations using reference LD structure (See 4.2.4). Then we estimate the p-value of the GAUSS test statistic as follows

- Starting from $r = 1$, in the r^{th} step, generate a random $J \times 1$ vector v_r from the multivariate normal distribution $N(0, \widehat{V}_{\mathbb{G}})$
- Calculate the GAUSS statistic for v_r as GAUSS_r, using Step 1 in 4.2.1 above
- Repeat the above steps a large number of times, say R ($= 10^6$)
- P-value for observed GAUSS statistic (gs_{obs}) is caluclated as $\frac{\sum_{r=1}^{R} GAUSS_r > gs_{obs}}{R}$

Although it is a simulation-based method, the algorithm can be efficiently implemented since it only requires generating multivariate normal (MVN) random vectors. For example, generating 1 million MVN random vectors for a gene-set with 100 genes ($J = 100$) needs 2 CPU-seconds. We also implemented adaptive resampling scheme to perform small number of iterations if true p-value is large (> 0.001). Thus, if the true p-value is large the above algorithm estimates the it in less than 1 CPU-seconds and if the true p-values is small (< 0.001) the algorithm takes approximately 161 CPU seconds on an average to estimate it.

Our approach of simulating MVN random vectors is considerably faster than the existing approaches for simulation or permutation-based p-values. For example, in aSPUpath [Pan et al., 2015], which employs a permutation-based p-value, a new null trait vector (or score vector) is generated through permutation in each iteration. The test statistic is then calculated based on that null trait (or score) vector and this process is repeated for the specified number of iterations. This procedure has a substantially high computational burden since at each step the entire procedure of calculating a p-value starting from null traits (or scores) is being repeated. On the other hand, in our algorithm detailed above we have assumed that the z-statistics for the genes in the gene-set jointly follow a multivariate normal. Hence the simulations

can be carried out using the null distributions of the z-statistics rather than generating a null trait (or score) in each iteration which reduces the computation greatly. Additionally, since simulating MVN random vectors is considerably faster than generating permutation-based null traits (or scores), our algorithm has a significantly lower computational burden.

4.2.3 Estimating gene-based p-values from summary statistics: SKAT, Burden, SKAT-Common-Rare

Individual-level data on genotypes and phenotypes are often unavailable due to restrictions on data sharing, but summary statistics from such studies may be available. There are publicly accessible repositories of GWAS summary statistics which can be queried for numerous genetic variants as well as phenotypes [Buniello et al., 2019].

To complete Step 1 (4.2.1) above, we need the gene-based p-values for the gene-set $\mathbb{G}$. In this section, we demonstrate methods to estimate mixed model-based tests like SKAT, Burden and SKAT-Common-Rare when only GWAS summary statistics on individual variants are available. Thus, GAUSS can be applicable in a broad spectrum of scenarios with individual level GWAS data as well as summary statistics.

Let $y = (y_1, y_2, \cdots, y_n)^T$ be an $n \times 1$ vector of the phenotype over n individuals; X an $n \times q$ matrix of the q non-genetic covariates including the intercept; $G_j = (G_{1j}, \cdots, G_{nj})^T$ is an $n \times 1$ vector of the minor allele counts (0, 1, or 2) or dosages for a binary genetic variant j; and $G = [G_1, \cdots, G_m]$ is an $n \times m$ genotype matrix for m genetic variants in a target region. The regression model shown in equation (4.2) can relate m genetic variants to phenotype,

$$y = X\alpha + G\beta + \epsilon \tag{4.2}$$

where α is a $q \times 1$ vector of regression coefficients of q non-genetic covariates, $\beta = (\beta_1, \cdots, \beta_m)^T$ is an $m \times 1$ vector of regression coefficients of the m genetic variants, and ϵ is an $n \times 1$ vector of non-systematic error term that follows $N(0, \sigma^2 I_n)$. To test for $H_0 : \beta = 0$, under the random effects assumption $\beta_i \sim N(0, \tau^2)$. The SKAT test statistic is

$$Q = (y - \hat{\mu})^T GWWG^T (y - \hat{\mu}) \tag{4.3}$$

where $\hat{\mu} = X\hat{\alpha}$ is the estimated mean of y under the null hypothesis of no association and where $W = diag(w_1, .., w_m)$ is a diagonal weighting matrix. The test statistic Q asymptotically follows a mixture of chi-squared distributions under the null hypothesis and p-values can be computed by inverting the characteristic function. The mixing parameters are the eigenvalues of $WG^T P_0 GW$ where $P_0 = I_n - X(X^T X)^{-1} X^T$.

Equation (4.2) uses individual level data on the samples. However the test of association can be approximated by using summary level statistics on the m variants in the region [Lumley et al., 2018]. Given GWAS summary statistics (MAF_i, β_i, SE_i), the test statistic Q in (4.3) can be shown to be equal to

$$Q_{summary} = \sum_{i=1}^{p} 2p_i(1 - p_i) w_i^2 t_i^2 \tag{4.4}$$

where $t_i = \frac{\beta_i}{SE_i}$. Under the null hypothesis, Q follows a mixture of chi-squares and the mixing parameters are the eigen values of the matrix $WG^T P_0 GW$. Replacing P_0 by $\tilde{P}_0 = I - 11^T/n$, we can approximate the eigen values by that of the matrix $WG^T \tilde{P}_0 GW$. The matrix $G^T \tilde{P}_0 G$ is the LD-matrix of the p variants. We can estimate this matrix using a suitable publicly available reference panel.

To use (4.4) as a rare-variant test, [Wu et al., 2011] suggested using weights generated from a beta-distribution as $w_i = Beta(MAF_i, a_1, a_2)$, the beta distribution density function with prespecified parameters a_1 and a_2 evaluated at the sample minor-allele frequency MAF_i. [Wu et al., 2011] suggested using $a_1 = 1$ and $a_2 = 25$

upweights the contribution of rare-variants ($MAF < 1\%$), while putting lower non-zero weights for variants with $MAF \geq 1\%$.

To perform a test of the combined effect of common and rare variants in a region [Ionita-Laza et al., 2013] developed SKAT-Common-Rare, which modifies (4.4) by separating the contribution of common and rare variants. Given summary statistics as above, we construct the test statistic separately for common and rare variants as

$$Q_{summary;common} = \sum_{i=1}^{p} 2p_{i;common}(1 - p_{i;common})w_{i;common}^2 t_{i;common}^2 \tag{4.5}$$

$$Q_{summary;rare} = \sum_{i=1}^{p} 2p_{i;rare}(1 - p_{i;rare})w_{i;rare}^2 t_{i;rare}^2; \tag{4.6}$$

where $Q_{summary;common}$ ($Q_{summary;rare}$) is constructed using common (rare) variants only. The weights $w_{i;common}$ are generated using a Beta(0.5,0.5) distribution, which is equivalent to using inverse-variance weights, whereas the weights $w_{i;rare}$ uses a Beta(1,25) distribution, as mentioned earlier. SKAT-Common-Rare test is then constructed as

$$Q_{common-rare} = (1 - \phi)Q_{summary;common} + \phi Q_{summary;rare} \tag{4.7}$$

where $0 \geq \phi \geq 1$ is a weight. Here we set $\phi = \frac{SD(Q_{summary;rare})}{SD(Q_{summary;common} + SD(Q_{summary;rare}))}$ as suggested by [Ionita-Laza et al., 2013], which means $(1 - \phi)Q_{summary;common}$ and $\phi Q_{summary;rare}$ have the same variance. The asymptotic null distribution of $Q_{common-rare}$ is a mixture of chisquares and can be approximated using the LD matrices of common and rare variants.

In the data applications (4.3.1) and in the implementation of our method we have used data on Europeans in 1000 Genomes project as our reference panel to compute the LD between sets of variants. This data is widely used for estimation of LD and imputation in current association studies.

4.2.4 Reference data and the estimation of correlation structure $V_{\mathbb{G}}$

Starting from input GWAS summary statistic to obtaining the GAUSS p-value, we have used a reference panel in two contexts. Firstly, we use the reference panel to extract LD across variants in a gene or region. This LD information is used to construct the null distribution and evaluate the gene-based p-value. We have used emeraLD [Quick et al., 2019] for a fast extraction of LD from variant-call-format files. Second, we use the reference panel to estimate the null correlation matrix $V_{\mathbb{G}}$ between the z-statistics of a given gene-set $\mathbb{G}$. This is a pre-computed matrix that is used in calculating the simulation-based p-values. For this, we generate a null continuous phenotype from standard normal distribution, computed the gene-based p-values for the annotated genes using SKAT-Common-Rare and convert them to z-statistics. We repeat this procedure for 1000 iterations and $V_{\mathbb{G}}$ is calculated as the Pearson's correlation between the 1000 null z-statistic values. This greatly reduces the computational burden of the GAUSS test since we do not need to estimate $V_{\mathbb{G}}$ for every iteration or gene-set separately.

$V_{\mathbb{G}}$, the correlation between the p-values of the gene-set under null hypothesis (self-contained) of no association, can be estimated from a given reference panel. Here we use publicly available data on unrelated Europeans in the 1000 Genomes (Phase III) data as our reference panel to estimate $V_{\mathbb{G}}$.

4.3 Results

4.3.1 Simulation

We carried out extensive simulation studies to evaluate the type I error control and power of GAUSS. We selected two gene-sets from the gene-sets annotated by GO terms in MSigDB for our simulations. The gene-sets are sterol metabolic process (GO: 0016125) consisting of 123 genes and regulation of blood volume by renin

angiotensin (GO: 0002016) consisting of 11 genes. For a given gene-set we randomly set g_a genes to be active and within l^{th} active gene with t_l variants we set $v_{a;l}$ proportion of variants to have non-zero effects. Using genotypes of 5000 unrelated individuals from the UK Biobank we generate the phenotypes for individual i ($i = 1, \cdots, 5000$) according to the model

$$Y_i = \sum_{k=1}^{T} \beta_k G_{ik} + \epsilon_i \quad \epsilon_i \sim N(0,1) \tag{4.8}$$

where G_{ik} is the genotype of the i^{th} individual at k^{th} site and $T = \sum_{l=1}^{g_a} t_l v_{a;l}$ is the total number of variants with non-zero effects. The effect size of the k^{th} active variant with minor allele frequency MAF_k is generated as $\beta_k = c|log_{10}(MAF_k)|$ which upweights the effect of rare-variants. For type-I error simulation we used $c = 0$ while for power we set $c > 0$. We calculated the gene-based p-values for the genes belonging to the gene-sets that we considered and subsequently applied the GAUSS test. To compare the performance of GAUSS to several existing methods, we also applied MAGMA [de Leeuw et al., 2015], aSPUpath [Pan et al., 2015] and SKAT-Pathway (SKAT test of all the variants in the gene-set combined together) to the simulated data.

We evaluate the type-I error by simulating a phenotype independent of the genotypes and applying the above tests. This process is repeated 10^7 times. Type-I errors of GAUSS remains well calibrated at $\alpha = 1 \times 10^{-04}, 1 \times 10^{-05}$ and 5×10^{-06} (Table 4.1) for both the gene-sets under consideration.

Further, we evaluated power at a threshold of $\alpha = 5 \times 10^{-06}$ which represents the Bonferroni corrected threshold for testing association across $10,000$ independent gene-sets. For power simulations, with gene-set GO: 0016125, we first set $20(\sim 16\%)$ randomly chosen genes to be active and within each active gene 30% of the variants are set to have non-zero effects. The effect size of the k^{th} active variant with minor

allele frequency MAF_k is generated as $\beta_k = c|log_{10}(MAF_k)|$ such that rarer variants had larger effects. With varying magnitude of association determined by c, GAUSS maintains a high power to detect associations between the simulated phenotype and the gene-set. The power of GAUSS and MAGMA are quite similar (Figure 4.1) in most of the scenarios, while the power of SKAT-Pathway is consistently lower. GAUSS has a higher power than aSPUpath when the magnitude of the effects are weak or moderate ($c = 0.1, 0.12$), but for higher magnitude of effect sizes ($c = 0.2, 0.25$) the powers are similar.

When the signals are even sparser (Figure 4.1), i.e., 2 to 6 genes are active with $c = 0.1$, there is a noticeable power difference between GAUSS and MAGMA as well aSPUpath. The difference is maximum for the model with lowest number of active genes and is reduced as the number of active genes grows from 2 to 6. Thus, for gene-sets with many signals, weak or strong, GAUSS performs similar to MAGMA or aSPUpath, while the advantage of GAUSS is pronounced when only a few genes are active within the gene-set. This is becasue, GAUSS identifies the maximum attainable association signal within the gene-set while the other methods averages over all the signals within the gene-set. Thus, for them, the true signal can be overtaken by non-signals when the number of such non-signals is substantially higher than the number of signals. The power of SKAT-Pathway is lower than GAUSS throughout all the simulation scenarios.

We also report the sensitivity and specificty of GAUSS in identifying the active subset (AS) genes. Sensitivity and specificity are defined by the proportion of true active genes (genes having variants with non-zero effect sizes) correctly identified by GAUSS as AS genes and the proportion of true inactive genes (genes with all variants having zero effect size) that are not in AS genes, respectively. Both the quantities remain high ($> 75\%$) at different magnitudes of the effect size and for varying number of active genes (Figure C.1). High sensitivity implies that there is

a high overlap between the true active genes and the AS genes in GAUSS. High specificity implies that there is a high overlap between the true inactive genes and the genes in the gene-set that are not in AS genes. Overall, this means that the AS genes extracted by GAUSS approximates the actual active set of genes with high accuracy. We further evaluate the power to identify the exact set of active genes which is a more stringent criteria compared to sensitivity and specificity (Figure C.2). Under different magnitudes of effect size defined by different values of c, the empirical probability to identify the exact set of active genes through the AS genes, increases with the number of active genes as well the magnitude of effect size. For strong effects in 4 or more genes, estimated power to identify the exact set of active genes is more than 75% for both the gene-sets.

Simulation results highlight the utility of GAUSS compared to existing methods specially under the scenarios when only a few genes are active in the gene-set. Further by extracting AS genes, GAUSS can identify the set of such active genes with high probability and provides a natural way to interpret and utilize the findings.

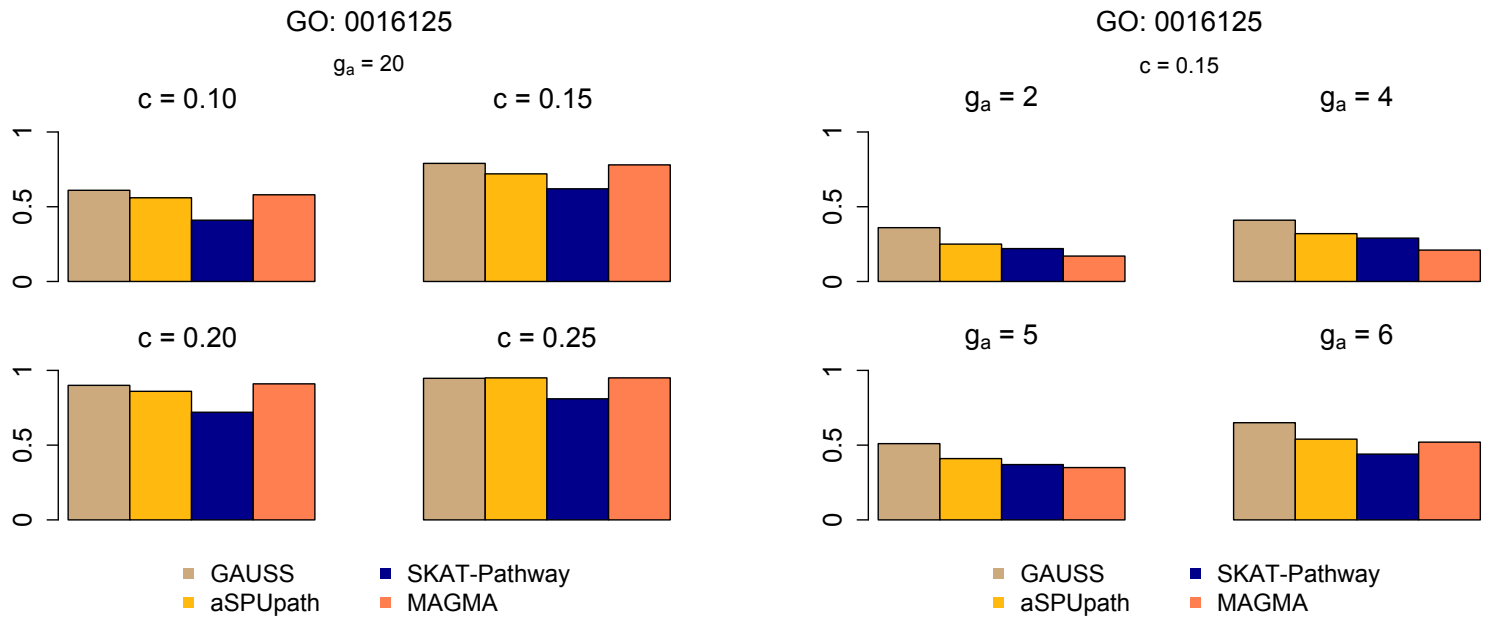


Figure 4.1: Power of GAUSS under different simulation settings using GO: 0016125 compared with that of aSPUpath, SKAT-Pathway and MAGMA. (a) Power of GAUSS when 20 genes are active and the variants with non-zero effect sizes have different magnitudes denoted by c. (b) Power of GAUSS with different number of active genes and the variants with non-zero effect sizes have a magnitude denoted by $c = 0.15$.

Association analysis in UK Biobank data

We next performed association analysis with GAUSS for 1201 phenotypes in UK Biobank data. We used two collections of gene-sets from MSigDB:

1. the curated gene-sets (C2) which contains gene-sets from KEGG, BioCarta and Reactome databases and also gene-sets representing expression signatures of genetic and chemical perturbations

2. Gene sets that contain genes annotated by the same gene ontology (GO) term (C5)

resulting in a total of 10,679 gene-sets. The Bonferroni corrected p-value threshold for testing association across these gene-sets is $4.68 \times 10^{-06} \approx 5 \times 10^{-06}$. For each pair of phenotypes and gene-set we computed the association test-statistic, corresponding p-value and the active subset (AS) of genes (if the gene-set is reported to be significant).

In our analysis, we used publicly available GWAS summary statistic for the phenotypes that were generated using SAIGE [Zhou et al., 2018]. The summary statistics files included results for markers directly genotyped or imputed by the Haplotype Reference Consortium (HRC) which produced approximately 28 million markers with MAC $\geq$ 20 and an imputation info score $\geq$ 0.3. We annotated a region of $\pm$1 kb around an exon to a gene using EPACTS. To do this, we used information from RefSeq gene database as a reference to indicate whether a particular variant resides in or near a gene and can potentially disrupt the corresponding protein sequence and its function.

From the summary statistics produced by SAIGE, we used effect size estimates (β), standard errors and minor allele frequencies (MAF) for the variants that were annotated to at least one gene. Using these statistics we constructed SKAT-Common-Rare test statistic for 18,216 genes and with LD information from a reference panel of 1000 Genomes data, approximated the corresponding p-values (See Methods). These

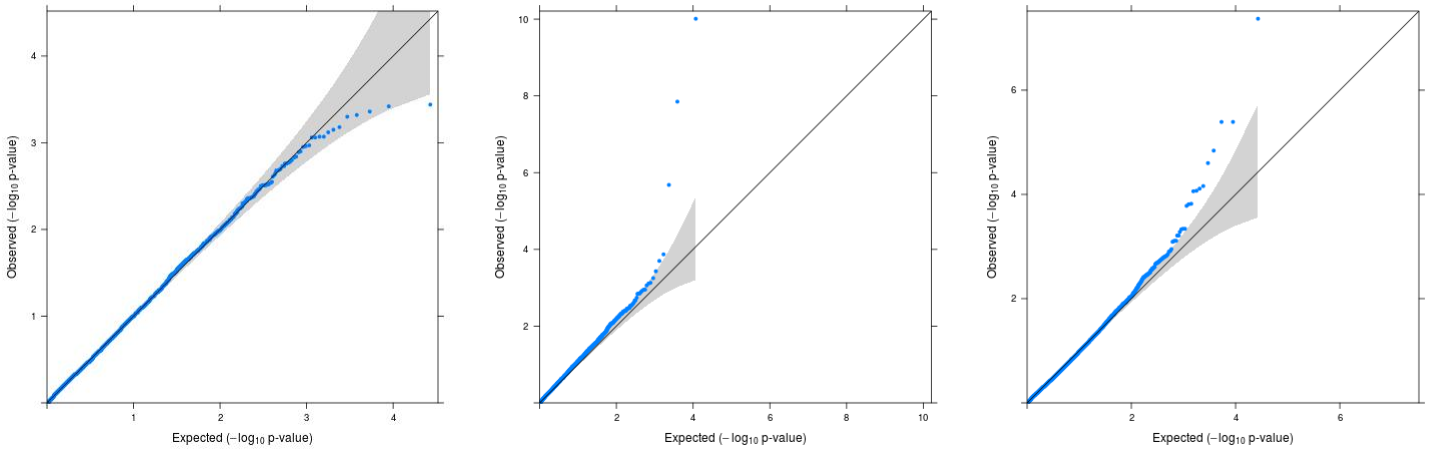
p-values were transformed to z-statistics and subsequently used in the gene-set analysis with GAUSS.

To illustrate the ultility of GAUSS in aggregating weak to moderate signals and also in improving interpretability through AS genes, association results for three exemplary phenotypes are shown here. The traits were: E.Coli infection (EC; PheCode: 041.4), Gastritis and duodenitis (GD; PheCode: 535) and Pernicious anemia (PA; PheCode: 281.11). Single variant GWAS for these traits has been presented in an online server (Pheweb; See URL). GD has five genome-wide significant loci while PA has one and EC has none. Figure 4.2 shows the QQ plots for approximated gene-based p-values for the 3 phenotypes. The p-values are well calibrated without any indication of inflation (λ_{GC} varies from 0.98 to 1.01). At an exome-wide cut-off of 2.5×10^{-06}, EC does not have any significantly associated genes; GD has 3 genes *HLA-DQA1* (p-value $= 9.8 \times 10^{-11}$), *HLA-DQB1* (p-value $= 1.4 \times 10^{-08}$) and *XXbac-BPG300A18.12* (p-value $= 2.1 \times 10^{-06}$) that are significantly associated; PA has one gene *PTPN22* (p-value $= 4.3 \times 10^{-08}$) that is significantly associated.

Gene-set association analysis with GAUSS are shown in Figure 4.4 and 4.3 and the significant gene-sets are detailed in Table 4.2. EC, which does not have any variant or any gene significantly associated with it, is associated with two gene-sets namely the gene-sets related to fatty acid catabolic process (GO: 0009062; p-value $< 1 \times 10^{-06}$) and fatty acid beta oxidation (GO: 0006635; p-value $= 2 \times 10^{-06}$). Although thorough gene-set analysis of EC has not been done before, but the antibacterial role of fatty acids has been reported in existing literature [Heipieper and Chiou, 2005, Ohya et al., 2000]. A set of 25 distinct genes (Table 4.2) are selected by GAUSS as the AS genes that are responsible for the association although none of them are marginally associated with EC.

GD is associated with 4 gene-sets in the Reactome database among those evaluated (Table 4.2). Although the gene-sets and the corresponding functions are biolog-

ically related, their role in GD is not easily identifiable. However, GAUSS selects a set of 10 genes to be the AS genes for the gene-sets, majority being from the different proteasome endopeptidase complex (PSM) subunits.



(a) (b) (c)

Figure 4.2: QQ plots for gene-based p-values of (a): E. Coli infection (EC), (b) Gastritis and duodenitis (GD) and (c) Pernicious anemia (PA)

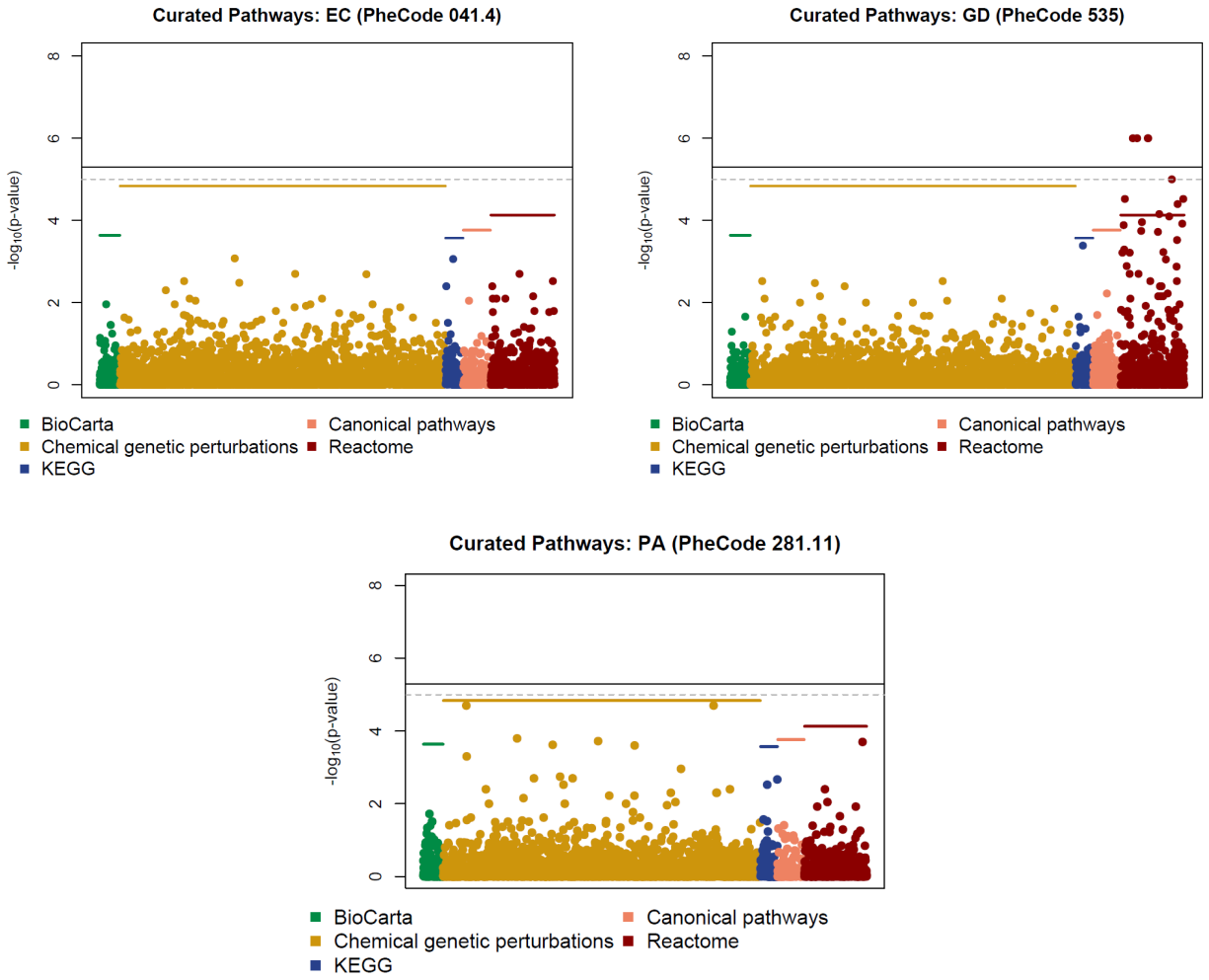


Figure 4.3: Significant gene-sets associated with (a) E. Coli infection (EC), (b) Gastritis and duodenitis (GD) and (c) Pernicious anemia (PA) among the curated gene-sets (C2). Colored horizontal lines denote the Bonferroni correction thresholds for corresponding groups. The horizontal solid black line denotes the signifiance threshold of 5×10^{-06}. The horizontal dashed line denotes a less stringent suggestive threshold of 1×10^{-05}

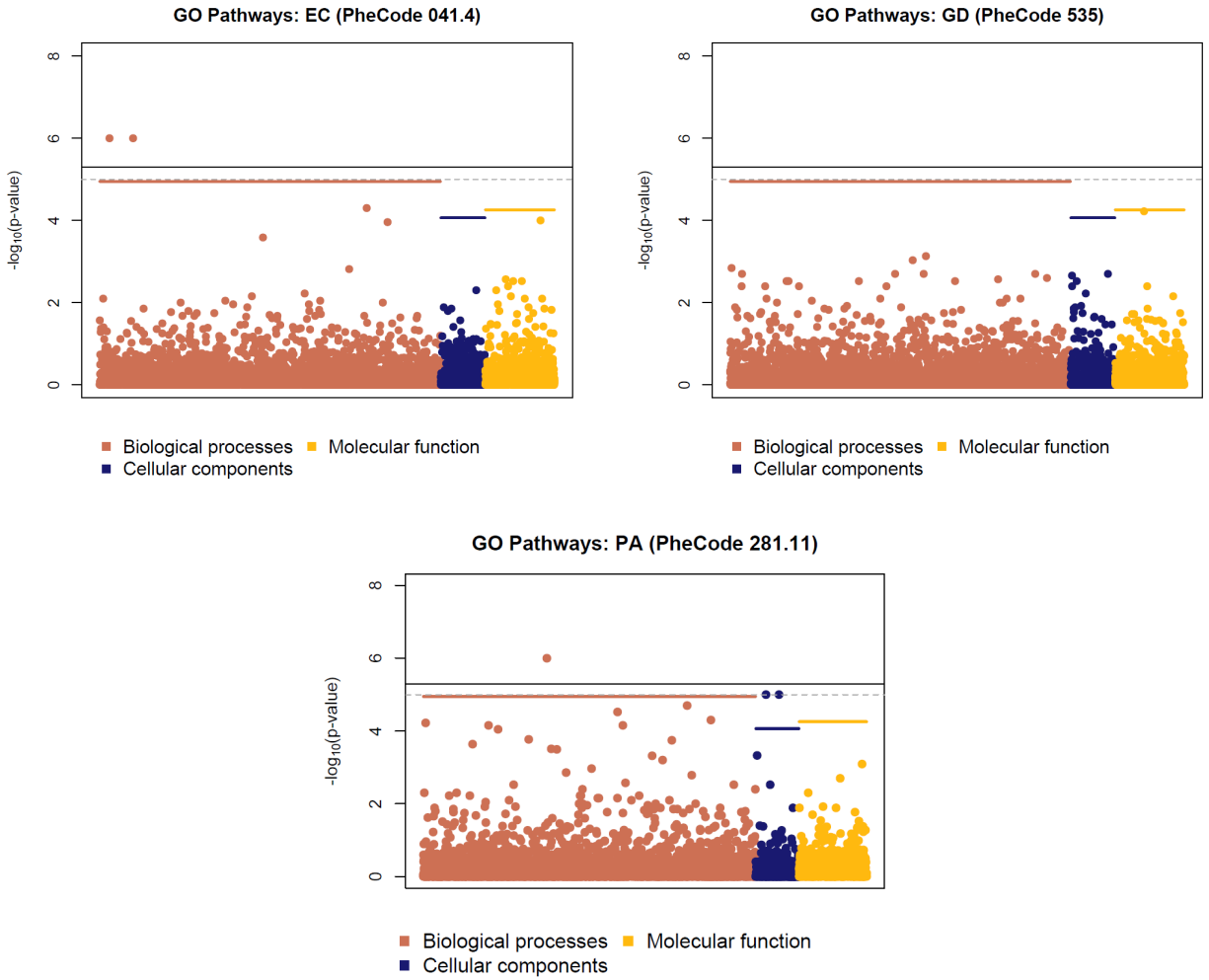


Figure 4.4: Significant gene-sets associated with (a) E. Coli infection (EC), (b) Gastritis and duodenitis (GD) and (c) Pernicious anemia (PA) among the GO gene-sets (C5). Colored horizontal lines denote the Bonferroni correction thresholds for corresponding groups. The horizontal solid black line denotes the signifiance threshold of 5×10^{-06}. The horizontal dashed line denotes a less stringent suggestive threshold of 1×10^{-05}.

In particular, the role of *PSMB8* in gastric cancer has been extensively reported in literature [Kwon et al., 2016]. Although, none of these genes are individually associated with GD, but they jointly drive the strong association signals of the identified gene-sets.

The set of genes involved in Cobalamin metabolic process (GO: 0009235) is significantly associated with PA (p-value $< 1 \times 10^{-06}$; Table 4.2). This result is expected since PA is a condition indicated by low levels of vitamin B12 (Cobalamin). Seven genes are selected as AS genes which include genes like *CUBN* and *TCN1*. The mutations in *CUBN* have been reported to encode the intrinsic factor-vitamin B12 receptor, cubilin [Aminoff et al., 1999] while mutations in *TCN1* result in severe cobalamin deficiency [Froese and Gravel, 2010]. None of the AS genes are significantly associated with PA, which means that GAUSS can successfully aggregate moderate to weak signals which are missed due to stringent exome-wide Bonferroni correction.

GAUSS can be applied to multitudes of phenotypes that are collected as a part of association studies, namely the phenotypes present in UK biobank. Using GAUSS for a given gene-set, we can investigate which phenotypes it is associated with and what are the corresponding AS genes. Such analysis can potentially be leveraged to gain mechanistic insights into the genetic relations between phenotypes and moderate to small pleiotropic effects of different regions or genes.

Association results for three exemplary gene-sets across the phenome of 1201 phenotypes in UK Biobank are shown here (Figure 4.5). The gene-sets are the ATP-binding cassette (ABC) transporters from KEGG database (ABC transporters; URL), the genes that are up-regulated in pyloric atrium with knockout of trefoil factor 2 (TFF2 targets) as reported in [Baus-Loncar et al., 2005] (URL) and genes constituting a supra-molecular assembly of fibrillar collagen complexes in the form of a long fiber (fibril) with transverse striations (GO: 0098643; URL).

Among the phenotypes associated with ABC transporters gene-set, transporter 2 (*TAP2*) is the most frequent gene in the AS genes selected by GAUSS (Table 4.3). This gene has previously been associated to several phenotypes including diastolic blood pressure [Warren et al., 2017], type-1 diabetes and autoimmune thyroid diseases [Tomer et al., 2015]. Our results show that the significant association of ABC transporters to disorders like psoriasis, celiac disease and type-1 diabetes are mainly driven by *TAP2* while those of gout, lipoid metabolism and cholelithiasis are driven mainly by members of ATP binding cassette subfamily G (*ABCG5* and *ABCG2*). Similarly for phenotypes associated with GO: 0098643, tenascin XB (*TNXB*) and genes from collagen alpha chain group (*COL11A2*, *COL27A1* etc.) drive the signals, especially with different forms of arthritis (Table 4.4). For TFF2 targets the length of the selected AS genes is usually more than one for different phenotypes but mostly comprising of protein tyrosine phosphatase, non-receptor type 22 (*PTPN22*) and members of proteasome subunit beta (*PSMB8* and *PSMB9*; Table 4.5).

The results highlight several important aspects of association results for pathway based analysis. First, GAUSS can detect and aggregate even weak to moderate association signals in a gene-set which might not be detected by standard genome-wide or exome-wide Bonferroni corrections. Second, a phenotype might be associated with several gene-sets but the signals might not be independent of each other, i.e., driven by the same AS genes (Table 4.2). Third, the phenome-wide association analysis of a given gene-set elucidates an aspect of gene-set analysis that has been unexplored until now. A particular gene-set may be associated to different phenotypes but the AS genes might be exactly the same (e.g. *TNXB* for Psoriasis, Type-1 and Type-2 diabetes in Table 4.5) or different (e.g. AS genes of monoarthritis and Osteoarthritis). This undelines the role that AS genes play in producing association signals and can highlight the underlying biological similarities or differences between phenotypes.

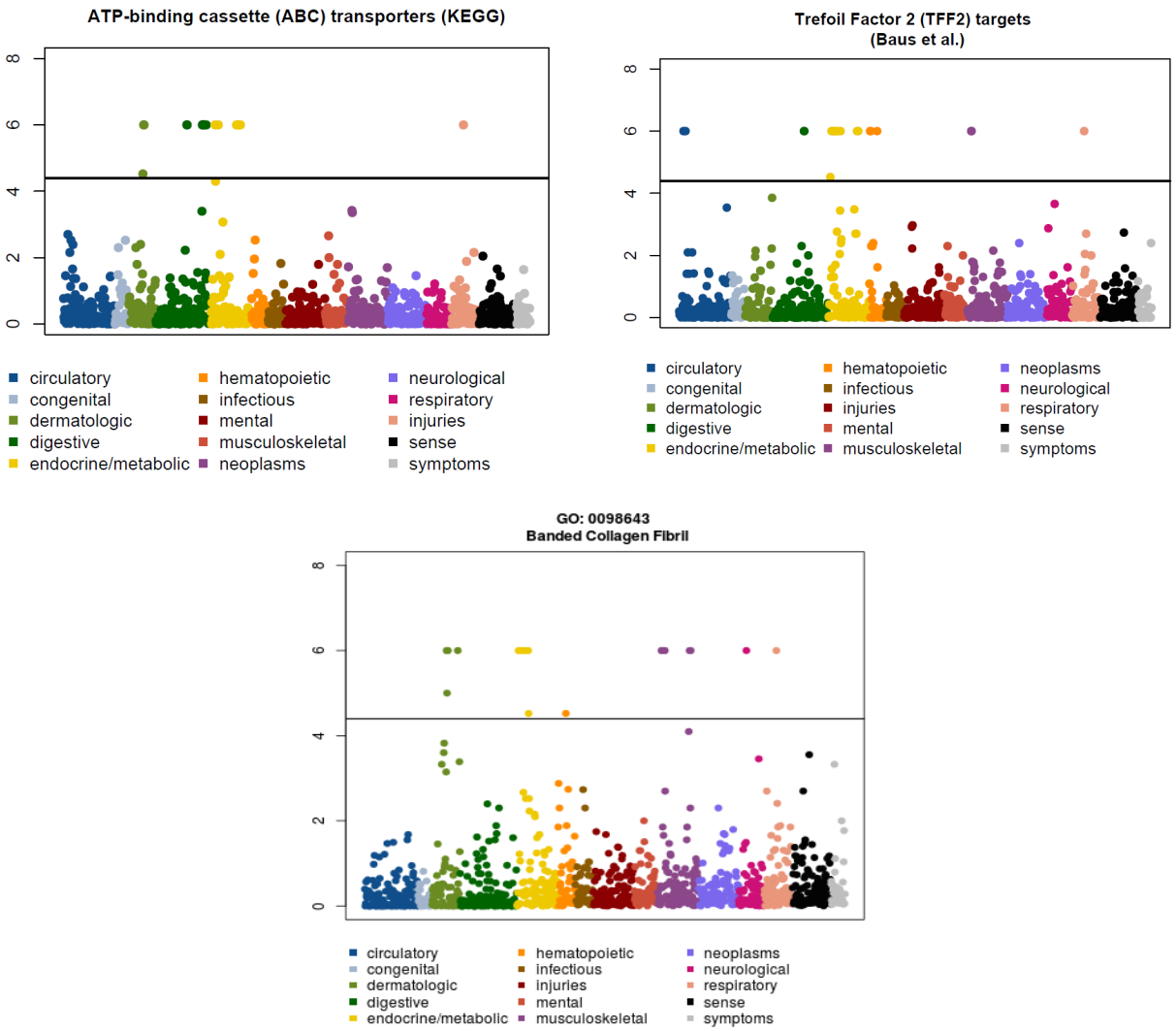


Figure 4.5: Phenotypes associated with (a) ABC transporters, (b) TFF2 targets and (c) GO: 0098643. P-values which were $< 1 \times 10^{-06}$ are collapsed to 1×10^{-06} for the ease of viewing. The horizontal black line denotes the Bonferroni cutoff of 4×10^{-05}.

4.4 Computation

One of the key features of GAUSS is the efficient computation which allows it to be applied to large genomic datasets. Although the p-value is based on a simulating multivariate normal random vectors, we have reduced the computational burden compared to other methods that also employ simulation by using pre-computed LD matrices within a gene and $\widehat{V}_{\mathbb{G}}$ between genes from a given reference panel (See Methods). We have employed selective-iteration resampling scheme to further reduce the computational cost. Figure C.3 shows the total run-time (in CPU-hours) of GAUSS, MAGMA and aSPUpath (we use aSPUspath, an implementation of aSPUpath that use summary statistics only) applied on pernicious anemia (PA) and type-2 diaebetes (T2D) from UK biobank as described above. Total run-time is calculated as the net time taken starting from the input of summary statistic till the p-values for the $10,679$ gene-sets are generated. GAUSS performs similar to MAGMA, while aSPUpath, which is also based on simulation p-values, is substantially slower than GAUSS.

Given summary statistics, the computation time to estimate the SKAT-Common-Rare p-values for EC, GD and PA were 5.5, 5.6 and 5.6 CPU-hours respectively. Subsequently, time taken to calculate the p-values for $10,679$ gene-sets for EC, GD and PA were 4.1, 4.0 and 4.7 CPU-hours respectively.

Run-time can be further reduced by computing the p-values for the gene-sets in parallel using a high-performance computing cluster. For example, when we parallelized the process into 11 chains, the clock-time was less than 2 hours.

4.5 Discussions

In this project we have presented GAUSS which introduces a maximum-type statistic to test the association between a gene-set and a phenotype. Similar to

several existing approaches like MAGMA, aSPUpath, GAUSS aims to cumulate weak to moderate association signals across a set of genes which might not have been detected due to stringent Bonferroni correction in standard single variant or region-based approaches. Given association z-statistics for the genes in the gene-set, GAUSS computes the maximum association score that can be achieved among any subset of the gene-set and computes a simulation-based p-value. Further, it also identifies the subset for which this maximum association score is obtained which is termed the active subset (AS) of genes.

The distinction between the AS genes and the rest of the gene-set highlights a key feature of GAUSS. To the best of our knowledge, there does not exist any other method to adaptively identify the subset that drives the signal. Most of the existing approaches usually suggest using the top few genes (genes with lowest p-values) in the gene-set. However, not tuning the choices according to a data driven approach might be misleading. For example, among the gene-sets presented ABC transporters have only one gene (*TNXB*) driving a association for at least 14 phenotypes. Although, it appears that only one gene-phenotype association drives the signal in these cases, such insights are eludicated by the extraction of AS genes. On the other hand a set of 25 genes drives the association of EC and GO:0009062. Hence, selecting the active set through a principled algorithm is helpful for interpreting the association signals and understanding the underlying mechanisms.

Computational scalability is another important aspect of GAUSS. Although GAUSS obtains simulation-based p-values, the computational cost is much lower than existing methods which employ direct resampling or permutations. This improvement is obtained since GAUSS uses copula to convert gene-based p-values to multivariate normal distribution and pre-computed correlation matrices. This allows GAUSS to be used for many phenotypes as well as many gene-sets.

Our UK Biobank analysis shows only a small percentage of genes in the pathway

are selected as active genes. Simulations show that GAUSS has substantial higher power than the existing methods in detecting associations in such scenarios. For example, MAGMA did not produce any significant results for PA(Figure C.4). When many of the genes in the gene-set are associated, the power of GAUSS was equivalent to that of the existing methods. Thus in most of the practical scenarios GAUSS has power better than or as good as the existing methods to detect association. Further, the type-I error for GAUSS remains calibrated at the desired level as well.

One of the limitations of GAUSS is that it only allows testing for the self-contained null hypothesis. Furthermore, the p-value being a simulation-based estimation can only provide estimates up to a level of accuracy determined by the number of iterations. We have explored a generalized pareto distribution based method to accurately estimate the very small p-values ($< 1 \times 10^{-06}$). But further research is needed in this respect.

The novel insights generated by GAUSS and its computational scalability make it a potentially attractive choice to perform gene-set analysis. We have made available the results from the analysis of UK Biobank data in a public repository and will continue to update that. We have also released an R-based package for our method.

4.6 URLS

Pheweb:`http://pheweb.sph.umich.edu/SAIGE-UKB/`

ABC transporters: `http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_ABC_TRANSPORTERS`

TFF2 targets: `http://software.broadinstitute.org/gsea/msigdb/cards/BAUS_TFF2_TARGETS_UP`

GO 0098643: `http://software.broadinstitute.org/gsea/msigdb/cards/GO_BANDED_COLLAGEN_FIBRIL`

emeraLD: `https://github.com/statgen/emeraLD`

Table 4.1: Estimated type-I error of GAUSS for gene-sets GO: 0016125 and GO: 0002016

α	GO: 0016125	GO: 0002016
1×10^{-04}	9.8×10^{-05}	9.7×10^{-05}
1×10^{-05}	9.9×10^{-06}	9.6×10^{-06}
5×10^{-06}	4.6×10^{-06}	4.2×10^{-06}

Table 4.2: Gene-sets associated with E. Coli infection (EC), Gastritis and duodenitis (GD) and Pernicious anemia (PA) corresponding p-values and the AS genes selected by GAUSS

Phenotype	Gene-Set	Genes	p-value	Active subset (AS) genes selected
EC	GO: Fatty acid catabolic process	73	$< 1 \times 10^{-06}$	*SLC27A2, CRAT, CPT1B, ACOX2, LPIN1, CPT1C, ETFB, SLC27A4, EHHADH, ACAA1, LEP, ABCD2, GCDH, HADH, MUT, BDH2, PLA2G15, PEX2, IVD, ACAAS, PEX13, ACAD8, ACADL, ECI1, ADIPOQ*
	GO: Fatty acid beta oxidation	51	2×10^{-06}	*SLC27A2, CRAT, CPT1B, ACOX2, CPT1C, ETFB, EHHADH, ACAA1, LEP, ABCD2, GCDH, HADH, BDH2, PEX2, IVD, ACAAS, ACAD8, ACADL, ECI1, ADIPOQ*
GD	Reactome: P53 independent G1/S DNA damage checkpoint	51	$< 1 \times 10^{-06}$	*PSMB2, PSMB9, PSMC5, CHEK1, PSMB8, PSMD9, PSMD2, RPS27A, PSMA6, PSMB7*
	Reactome: CDK mediated phosphorylation and removal of CDC6	48	$< 1 \times 10^{-06}$	*PSMB2, PSMB9, PSMC5, PSMB8, PSMD9, PSMD2, RPS27A, PSMA6, PSMB7*
	Reactome: Cyclin E associated events during G1/S transition	65	$< 1 \times 10^{-06}$	*PSMB2, PSMB9, PKMYT1, PSMC5, PSMB8, PSMD9, PSMD2, RPS27A, PSMA6, PSMB7*
	Reactome: P53 dependent G1 DNA damage response	57	$< 1 \times 10^{-06}$	*PSMB2, PSMB9, PSMC5, MDM2, PSMB8, PSMD9, PSMD2, RPS27A, PSMA6, PSMB7*
PA	GO: Cobalamin metabolic process	21	$< 1 \times 10^{-06}$	*CUBN, TCN1, ABCD4, GIF, CD320, MTRR, MMAA*

Table 4.3: Phenotypes associated with ABC transporters gene-set, corresponding p-values and the AS genes selected by GAUSS

Phenotype	Category	PheCode	p-value	Active subset (AS) genes selected
Psoriasis	dermatologic	696.4	$< 1 \times 10^{-06}$	*TAP2*
Psoriasis and related disorders	dermatologic	696	$< 1 \times 10^{-06}$	*TAP2*
Celiac disease	digestive	557.1	$< 1 \times 10^{-06}$	*TAP2*
Intestinal malabsorptions (non-celiac)	digestive	557	$< 1 \times 10^{-06}$	*TAP2*
Cholelithiasis with other cholecystitis	digestive	574.12	$< 1 \times 10^{-06}$	*ABCG5*
Cholelithiasis	digestive	574.1	$< 1 \times 10^{-06}$	*ABCG5*
Calculus of bile duct	digestive	574.2	$< 1 \times 10^{-06}$	*ABCG5*
Cholelithiasis without cholecystitis	digestive	574.3	$< 1 \times 10^{-06}$	*ABCG5,ABCC12,ABCA8,ABCB4*
Cholelithiasis and cholecystitis	digestive	574	$< 1 \times 10^{-06}$	*ABCG5*
Other biliary tract disease	digestive	575	$< 1 \times 10^{-06}$	*ABCG5*
Cholelithiasis and cholecystitis	digestive	574	$< 1 \times 10^{-06}$	*ABCG5*
Hypothyroidism NOS	endocrine/metabolic	244.4	$< 1 \times 10^{-06}$	*TAP2*
Type-1 diabetes	endocrine/metabolic	250.1	$< 1 \times 10^{-06}$	*TAP2*
Hypercholesterolemia	endocrine/metabolic	272.11	$< 1 \times 10^{-06}$	*ABCG5, TAP2, ABCC10, ABCA2, ABCA5, ABCA1, ABCA6, ABCC12, ABCC1, ABCA8, ABCB9*
Hyperlipidemia	endocrine/metabolic	272.1	$< 1 \times 10^{-06}$	*TAP2, ABCG5, ABCC10, ABCA6, ABCA2, ABCA5, ABCA1, ABCC1, ABCA8*
Disorders of lipoid metabolism	endocrine/metabolic	272	$< 1 \times 10^{-06}$	*ABCG2*
Gout	endocrine/metabolic	274.1	$< 1 \times 10^{-06}$	*ABCG2*
Asthma	respiratory	495	$< 1 \times 10^{-06}$	*TAP2*

Table 4.4: Phenotypes associated with GO:0098643, corresponding p-values and the AS genes selected by GAUSS

Phenotype	Category	PheCode	p-value	Active subset (AS) genes selected
Psoriasis vulgaris	dermatologic	696.41	$< 1 \times 10^{-06}$	*TNXB*
Psoriasis	dermatologic	696.4	$< 1 \times 10^{-06}$	*TNXB*
Psoriasis and related disorders	dermatologic	696	$< 1 \times 10^{-06}$	*TNXB*
Sicca syndrome	dermatologic	709.2	$< 1 \times 10^{-06}$	*TNXB*
Grave's disease	endocrine/metabolic	242.1	$< 1 \times 10^{-06}$	*TNXB*
Thyrotoxicosis with or without goiter	endocrine/metabolic	242	$< 1 \times 10^{-06}$	*TNXB*
Hypothyroidism NOS	endocrine/metabolic	244.4	$< 1 \times 10^{-06}$	*TNXB*
Type-1 diabetes with ketoacidosis	endocrine/metabolic	250.11	$< 1 \times 10^{-06}$	*TNXB*
Type-1 diabetes	endocrine/metabolic	250.1	$< 1 \times 10^{-06}$	*TNXB*
Type 2 diabetes with ophthalmic manifestations	endocrine/metabolic	250.23	$< 1 \times 10^{-06}$	*TNXB*
Diabetic retinopathy	endocrine/metabolic	250.7	$< 1 \times 10^{-06}$	*TNXB*
Rheumatoid arthritis	musculoskeletal	714.1	$< 1 \times 10^{-06}$	*TNXB*
Rheumatoid arthritis and other inflammatory polyarthropathies	musculoskeletal	714	$< 1 \times 10^{-06}$	*TNXB*
Unspecified monoarthritis	musculoskeletal	716.2	$< 1 \times 10^{-06}$	*COL11A2, COL2A1, TNXB, COL27A1, COL1A1*
Arthropathy NOS	musculoskeletal	716.9	$< 1 \times 10^{-06}$	*COL11A2, COL11A1, COL27A1, COL2A1, TNXB*
Other arthropathies	musculoskeletal	716	$< 1 \times 10^{-06}$	*COL11A2, COL11A1, COL2A1, COL27A1, TNXB*
Osteoarthritis; localized	musculoskeletal	740.1	$< 1 \times 10^{-06}$	*COL11A1, COL2A1, COL27A1, COL1A2*
Osteoarthritis	musculoskeletal	740	$< 1 \times 10^{-06}$	*COL11A1, COL1A2, COL2A1, COL27A1, COL5A1, TNXB*
Multiple sclerosis	neurological	335	$< 1 \times 10^{-06}$	*TNXB*
Asthma	respiratory	495	$< 1 \times 10^{-06}$	*TNXB,COL11A2*

Table 4.5: Phenotypes associated with TFF2 targets, corresponding p-values and the AS genes selected by GAUSS

Phenotype	Category	PheCode	p-value	Active subset (AS) genes selected
Essential hypertension	circulatory system	401.1	$< 1 \times 10^{-06}$	*PTPN22, SLC2A2, DPEP1, APOA4, PSMB8*
Hypertension	circulatory system	401	$< 1 \times 10^{-06}$	*PTPN22, SLC2A2, APOA4, LTB, PSMB8*
Celiac disease	digestive	557.1	$< 1 \times 10^{-06}$	*PSMB8, PSMB9*
Intestinal malabsorptions (non-celiac)	digestive	557	$< 1 \times 10^{-06}$	*PSMB8, PSMB9*
Thyrotoxicosis with or without goiter	endocrine/metabolic	242	$< 1 \times 10^{-06}$	*PSMB9, PSMB8, PTPN22*
Hypothyroidism NOS	endocrine/metabolic	244.4	$< 1 \times 10^{-06}$	*PTPN22, PSMB9, PSMB8, IRF7*
rule0pt8pt Hypothyroidism	endocrine/metabolic	244	$< 1 \times 10^{-06}$	*PTPN22*
Type-1 diabetes with ketoacidosis	endocrine/metabolic	250.11	$< 1 \times 10^{-06}$	*PTPN22, PSMB8, PSMB9*
Type-1 diabetes	endocrine/metabolic	250.1	$< 1 \times 10^{-06}$	*PSMB8, PTPN22, PSMB9*
Type 2 diabetes	endocrine/metabolic	250.2	$< 1 \times 10^{-06}$	*PSMB8, PTPN22, APOA4*
Gout	endocrine/metabolic	274.1	$< 1 \times 10^{-06}$	*ABCG2*
Gout and other crystal arthropathies	endocrine/metabolic	274	$< 1 \times 10^{-06}$	*ABCG2*
Iron deficiency anemias, unspecified or not due to blood loss	hematopoietic	280.1	$< 1 \times 10^{-06}$	*PSMB9, PSMB8, GLO1, IFIT1, BCAT2, TAP1, PTPN22, GGT1, APOA4*
Iron deficiency anemias	hematopoietic	280	$< 1 \times 10^{-06}$	*PSMB9, PSMB8, GLO1, BCAT2, IFIT1, TAP1, GGT1, PTPN22, APOA4*
Other anemias	hematopoietic	285	$< 1 \times 10^{-06}$	*PSMB9, PSMB8, PTPN22, TAP1, LTB*
Rheumatoid arthritis	musculoskeletal	714.1	$< 1 \times 10^{-06}$	*PTPN22, PSMB8, PSMB9*
Rheumatoid arthritis and other inflammatory polyarthropathies	musculoskeletal	714	$< 1 \times 10^{-06}$	*PTPN22, PSMB8, PSMB9, TAP1*
Asthma	respiratory	495	$< 1 \times 10^{-06}$	*PSMB8, PSMB9, PTPN22, TAP1*

CHAPTER V

Conclusion

The growth of biological datasets has presented us with enormous opportunities to gain insights into genetic architecture of complex traits and diseases. But, along with this, there are several challenges both in terms of statistical methodology as well as interpretation. In this dissertation, we have addressed a few such challenges regarding detecting pleiotropy in rare variants and identifying gene-set association with a phenotype.

5.1 Summary

The majority of the work on detecting pleiotropy has until now focused on common variants. Although recently several methods for rare variants have been published they have certain statistical and computational limitations. In Chapters II and III we have addressed these issues and developed novel and computationally efficient methods to identify rare-variants that have effect of multiple phenotypes.

In Chapter II, we provide a general framework, Multi-SKAT, to test the pleiotropic effects of rare variants, within a single study. Using kernel matrices we model the relation between multiple variants in the region and the set of phenotypes. One principle advantage of Multi-SKAT is that several existing methods can be expressed as special cases of it by specific choices of kernel matrices. This facilitates comparison

between the methods and highlights the modeling assumption for each. Through the use of fast and accurate omnibus tests, Multi-SKAT maintains robust power of detection across a range of association models. Further, it can effectively model between-sample relatedness which only a few methods have addressed. Hence, it can be applied to datasets that have substantial number of related individual without discarding samples or increasing type-I error. Multi-SKAT is computationally fast and can be applied to genome-wide datasets. We have demonstrated the performance of Multi-SKAT by using nine amino acid data from the METSIM study.

In Chapter III, we have extended Multi-SKAT to a meta-analysis framework, Meta-MultiSKAT. Meta-analysis is a powerful tool to jointly analyze genetic association results from multiple genome-wide association studies when individual level data are not available. Aggregating data across studies to increase effective sample size and power facilitates the discovery of trait-associated variants with modest effect sizes. Hence it is plausible that meta-analyzing multiple phenotypes can further increase power of rare variant tests. Our proposed method again uses a kernel matrix to model the heterogeneity between the effects of different studies and constructs a variance component test of association. Using data from Michigan Genomics Initiative and the SaridNIA study, we show that Meta-MultiSKAT can discover rare variants associated with white blood cell subtype traits and is more powerful than existing methods.

Next we addressed some challenges in gene-set analysis. In Chapter IV, we propose GAUSS, a subset-based powerful approach to gene-set analysis that facilitates interpretation. Although the p-value is evaluated through simulation approach, the method is computationally efficient since we use a suitable reference panel to precompute correlation matrices. One principal advantage of our method is that it can identify the set of core genes within the gene-set which contains the maximum association signal and hence appear to drive the signal. This has not been addressed

by any other method to our knowledge. Further, the computational scalability of GAUSS allows us to apply it to test the association between large numbers of gene-sets and phenotypes. Using GAUSS, we can evaluate phenome-wide associations for a given gene-set, which has not been performed to date. This can highlight genetic and mechanistic similarities or differences between different phenotypes.

5.2 Extensions and future work

Although our methods were developed focusing primarily on rare-variants in Chapters II and III, they can be used in a more general sense as region-based tests. With minor changes in the kernel structures, we can use Multi-SKAT as well as Meta-MultiSKAT for testing the combined effects of common and rare variants in a gene or region. This is particularly of interest when there might be one or more common variant in a region associated with the phenotype.

We depend on MAF based weighting of the individual variants in the region. However, in practice, it is not entirely straightforward to interpret such a weighting scheme. Researchers have suggested the use of functional scores which are used to predict the probability of a variant being deleterious or causing any alterations in protein structure. These scores, used as weights for variants, can yield a result that is more interpretable in terms of the biology.

The multiple phenotype based methods can be applied to variety of datasets including that of neuro-imaging phenotypes. Further, with minor modifications these can also be applied to longitudinal measurements of the same phenotype on an individual. For example, patients often have blood pressure or BMI measurements over time which can be leveraged to detect associations.

In spite of the increase in power by using multiple phenotypes, these multivariate tests give rise to the question of interpretability. We do not have any methods to adaptively select an active subset of phenotypes which might produce the optimal

association signal for a particular region. This requires further research and study.

In Chapter IV, we have used simulation-based approach to estimate the p-value of GAUSS. Thus, the p-value can only be estimated up to a fixed precision determined by the number of simulation iterations. Any analytical solution to this problem is worth exploring in the future.

5.3 Perspectives and conclusions

In this era, where rapid growth in technologies have radically transformed the landscape of GWAS, properly utilizing the scope of information is of critical importance. It is increasingly evident that several phenotypes could measure a disease in different dimensions and hence are likely to share the same genetic components. Empirical evidence of such shared genetics between phenotypes is abundant in literature and can be accurately estimated from summary statistics as well as individual level data. Thus, it follows that a multiple phenotypes test using a suitable multivariate framework can improve power to detect disease associated variants compared to single phenotype analysis. This higher power attained by using multivariate methods, such as Multi-SKAT and Meta-MultiSKAT might prove to be critical to detect disease variants in practice.

Further, understanding the mechanisms for complex traits or diseases is of paramount importance to translate the GWAS results into therapeutic targets. GSA have been used in this respect as a secondary follow up tool to GWAS. The additional power and insights gained by cumulating weak to moderate effect of variants not detected through GWAS can be pivotal in understanding the biology of complex diseases/traits.

In future, integrating multi-omics data such as epigenetic features, eQTLs, tissue-specific transcript expressions, chromatin conformation and others can improve our understanding of the functional and mechanistic roles of different variants. The ex-

pansion of GWAS and its integration with other efforts in understanding the molecular function of the human genome, will play a critical role in the study of gene coding and regulatory mechanisms and their contribution to complex diseases/traits. Understanding the mechanism by which genotype influences phenotype will ultimately lead to the identification of important targets for drug development and repositioning of known treatments. Continuing toward such targets will bring us closer to offering opportunities of innovative therapeutic strategies in precision as well as personalized medicine. We hope that topics elucidated in this dissertation will be one of the many starting steps in that direction.

APPENDICES

APPENDIX A

Appendix for Chapter II

A.1 Principal Component (PC) Kernel

Let L_i be the loading vector for the i^{th} PC, which produces the i^{th} PC score $P_i = YL_i$. In PCA-based analysis, PC scores are used as outcomes instead of original Y. Since the genetic information regarding the phenotypes may not be confined to the top few PCs Aschard et al. [2014], we first consider using all PCs. Let $P = (P_1, \cdots P_K)$. Since PCs are orthogonal, we assume genetic effects to multiple PCs are heterogeneous, which resulted in

$$Q = \{vec(P) - vec(\widehat{\mu}_P)\}^T \left\{ \left(G\Sigma_G G^T\right) \otimes \left(\widehat{V}_P^{-1}\widehat{V}_P^{-1}\right) \right\} \{vec(P) - vec(\widehat{\mu}_P)\} \quad \text{(A.1)}$$

where $\widehat{\mu}_P$ is the mean of P under the null hypothesis and $\widehat{V}_P$ is the estimated covariance matrix between the PC's. $\widehat{V}_P$ will be a diagonal matrix since PCs are orthogonal. Equation ((A.1)) can be written as

$$Q = \{vec(Y) - vec(\hat{\mu})\}^T \left\{ \left(G\Sigma_G G^T\right) \otimes \left(L\widehat{V}_P^{-1}\widehat{V}_P^{-1}L^T\right) \right\} \{vec(Y) - vec(\hat{\mu})\} \quad \text{(A.2)}$$

where $L = (L_1, \cdots, L_K)$ is a $K \times K$ PC loading matrix. Equation ((A.2)) shows that by using $\Sigma_{P,PC} = \widehat{V} L \widehat{V_P}^{-1} \widehat{V_P}^{-1} L^T \widehat{V}$, we can carry out PC-based tests. It is to be noted that the genetic effects of the PC's do not need to be assumed to be heterogeneous. Any kernel structure that is applicable to the test statistic in equation 2.4 can be applied here as well.

A.2 Relationship between Multi-SKAT and existing methods

For the ease of algebraic expressions, we will consider that all the K phenotypes have residual variance 1. For the general case of different residual variances, Σ_P should be replaced by $T_w^{-1} \Sigma_P T_w^{-1}$ where $T_w = diag(\sigma_1, \cdots, \sigma_K)$, σ_k being the residual standard error of k^{th} phenotype.

A.2.1 MSKAT

The Q statistic of MSKATWu and Pankow [2016] is given by

$$Q_{MSKAT} = vec(S_c)^T (WW \otimes \widehat{V}^{-1}) vec(S_c), \tag{A.3}$$

where $S_c = G^T(Y - \widehat{\mu})$ is a matrix of score statistics Wu and Pankow [2016]. Using row-vectorization properties

$$vec(S_c) = vec(G^T(Y - \widehat{\mu})) = (G^T \otimes I) vec(Y - \widehat{\mu}) = (G^T \otimes I) \{vec(Y) - vec(\widehat{\mu})\}$$

Then Q_{MSKAT} can be written as

$$\{vec(Y) - vec(\widehat{\mu})\}^T \left\{ (GWWG^T) \otimes \widehat{V}^{-1} \right\} \{vec(Y) - vec(\widehat{\mu})\},$$

which is the Multi-SKAT test statistics with $\Sigma_G = WW$ and $\Sigma_P = \widehat{V}$.

Further, the Q' of MSKAT is given by

$$Q'_{MSKAT} = vec(S_c)^T (WW \otimes I) vec(S_c). \tag{A.4}$$

Using the similar algebra as above, this can be written as

$$\{vec(Y) - vec(\widehat{\mu})\}^T \left\{(GWWG^T) \otimes I\right\} \{vec(Y) - vec(\widehat{\mu})\}$$

which is the Multi-SKAT test statistics with $\Sigma_G = WW$ and $\Sigma_P = \widehat{V}^2$.

A.2.2 GAMuT

Suppose $Y - \widehat{\mu} = Y_{adj} = HY$ and $G_{adj} = HG$ are covariate adjusted phenotype and genotype matrices where $H = I - X(X^TX)^{-1}X^T$. With the intercept in X, Y_{adj} and G_{adj} are mean centered. The covariate adjusted GAMuT test statistics is

$$Q_{GAMuT} = \frac{tr(P_c X_c)}{n}$$

where

$$P_c = \begin{cases} Y_{adj}(Y_{adj}^T Y_{adj})^{-1} Y_{adj}^T & \text{for projection phenotype kernel} \\ Y_{adj} Y_{adj}^T & \text{for linear phenotype kernel} \end{cases}$$

and $X_c = G_{adj} WW G_{adj}^T$. Using the fact that $Y_{adj}^T Y_{adj}/n = \widehat{V}$ is the estimate of variance after adjusting covariates and $G_{adj}^T Y_{adj} = G^T HY = G^T Y_{adj}$ (since H is a symmetric idempotent matrix), we show, for the projection kernel

$$
\begin{aligned}
tr(P_c X_c)/n &= tr(Y_{adj}\widehat{V}^{-1}Y_{adj}^T G_{adj} WWG_{adj}^T) \\
&= tr(\widehat{V}^{-\frac{1}{2}}Y_{adj}^T GWWG^T Y_{adj}\widehat{V}^{-\frac{1}{2}}) \\
&= vec(WG^T Y_{adj}\widehat{V}^{-\frac{1}{2}})^T vec(WG^T Y_{adj}\widehat{V}^{-\frac{1}{2}}) \\
&= \left\{(WG^T \otimes \widehat{V}^{-\frac{1}{2}})vec(Y_{adj})\right\}^T \left\{(WG^T \otimes \widehat{V}^{-\frac{1}{2}})vec(Y_{adj})\right\} \\
&= \{vec(Y) - vec(\widehat{\mu})\}^T (GWWG^T \otimes \widehat{V}^{-1}) \{vec(Y) - vec(\widehat{\mu})\}
\end{aligned}
$$

which is the same as the Multi-SKAT test statistic with $\Sigma_G = WW$ and $\Sigma_P = \widehat{V}$. Similarly for the linear kernel,

$$
\begin{aligned}
tr(P_c X_c)/n &= tr(Y_{adj}Y_{adj}^T G_{adj} WWG_{adj}^T) \\
&= \left\{(WG^T \otimes I)vec(Y_{adj})\right\}^T \left\{(WG^T \otimes I)vec(Y_{adj})\right\} \\
&= \{vec(Y) - vec(\widehat{\mu})\}^T (GWWG^T \otimes I) \{vec(Y) - vec(\widehat{\mu})\}
\end{aligned}
$$

which is the Multi-SKAT test statistic with $\Sigma_G = WW$ and $\Sigma_P = \widehat{V}^2$.

A.2.3 MAAUSS and MF-KM

There exists two different version of the MAAUSS tests. The homogeneous version of MAAUSS assumes that the effects of a variant on multiple phenotypes are identical and uses the following test statistic

$$
Q_{MAAUSS-HOM} = (vec(Y)-vec(\widehat{\mu}))^T (I_n \otimes \widehat{V}^{-1})(G \otimes I)(WW \otimes 1_m 1_m^T)(G^T \otimes I)(I_n \otimes \widehat{V}^{-1})(vec(Y)-vec(\widehat{\mu})) \tag{A.5}
$$

which is identical to the Multi-SKAT test statistic with $\Sigma_G = WW$ and $\Sigma_P = 1_m 1_m^T$. The heterogeneous version of MAAUSS assumes that the effects of a variant on

multiple phenotypes are independent, and uses the following test statistic

$$Q_{MAAUSS-HET} = (vec(Y)-vec(\widehat{\mu}))^T(I_n\otimes\widehat{V}^{-1})(G\otimes I)(WW\otimes I)(G^T\otimes I)(I_n\otimes\widehat{V}^{-1})(vec(Y)-vec(\widehat{\mu})) \tag{A.6}$$

which is identical to the Multi-SKAT test statistic with $\Sigma_G = WW$ and $\Sigma_P = I$. Note that the test statistic of MF-KM is exactly the same as $Q_{MAAUSS-HET}$.

A.3 Backward elimination procedure to identify associated phenotypes

After identifying the gene or region associated with multiple phenotypes, next question would be identifying truly associated phenotypes. Here we present a simple backward elimination algorithm to iteratively remove relatively less important phenotypes. A similar method has previously been applied to identify rare causal variants in an associated gene Ionita-Laza et al. [2014].

- Step 1. Start with a set of k phenotypes $Phen_{Current} = \{y_1, y_2, \cdots y_k\}$ and compute a Multi-SKAT test association p-value for the set $Phen_{Current}$ denoted by $p_{Current}$.

- Step 2. Remove each of the phenotypes one at a time from the set $Phen_{Current}$. The resulting set is $Phen_{-i} = \{y_1, y_2, \cdots y_{i-1}, y_{i+1} \cdots y_k\}$ for $i = 1, 2, \cdots, k$ and compute the corresponding p-values p_{-i} for that same Multi-SKAT test.

- Step 3. Remove the phenotype j that leads to the smallest p-value, i.e. $j = argmin\{p_{-1}, p_{-2}, \cdots, p_{-k}\}$. Update $Phen_{Current}$ to $Phen_{-j}$.

- Step 4. Continue removing phenotypes till only 1 phenotype is left.

Supplementary Table S2 shows the backward elimination results of 5 most significant and suggestive genes in the METSIM study data analysis as per the p-values

reported by minP_{com}. Although this procedure does not provide a set of phenotypes truly associated, it provides the relative importance of the phenotypes in driving association signals. For example, the minP_{com} p-value for *GLDC* was 2.3×10^{-72}. When each of the phenotypes were removed one at a time and the minP_{com} p-values were calculated on the remaining 8 phenotypes, we found that eliminating Isoleucine (Ile) actually improved the signal. The minP_{com} p-value of the set of 8 amino acids leaving out Ile was 2.8×10^{-73}. This indicates that Isoleucine has very minimal contribution to the association between the amino acids and *GLDC*. Subsequently, Valine was the next phenotype to be eliminated indicating that it has the next lowest contribution after Isoleucine. Carrying out this procedure further, we find that Glycine is the last phenotype to remain indicating that it is the strongest driver of the signal. This is in agreement to the single phenotype SKAT-O results (Table 5). Similary for genes *HAL*, *DHODH*, *PAH* and *MED1*, Histine, Alanine, Phenylalanine and Tyrosine were the most associated phenotypes, respectively. Interestingly for *PAH* and *MED1*, single phenotype p-values are not significant, which suggests that multiple phenotypes are associated with these genes.

Table A.1: Computation time of MultiSKAT. Computation time to analyze a dataset with 5000 individuals, 20 variants and 10 phenotypes. Analysis was done on a 2.80 GHz Intel Xeon CPU

	Method	CPU sec
Independent samples (Without kinship adjustment)	Multi-SKAT (given Σ_P and Σ_G)	0.014 secs
	minP	2.133 secs
	minP_{com}	3.971 secs
Related samples (With kinship adjustment)	Multi-SKAT (given Σ_P and Σ_G)	2.845 secs
	minP	6.961 secs
	minP_{com}	10.349 secs

Table A.2: Backward elimination results for the top 5 genes in Table 2.2. For a particular gene, each row indicates the phenotype eliminated and the p-value produced correspondingly. The last row indicates the remaining phenotype after backward elimination has been performed. This is the phenotype that drives the signal of association for the particular gene.

	GLDC		HAL		DHODH		PAH		MED1	
	Phenotype	p-values	Phenotype	p-values	Phenotype	p-values	Phenotype	p-values	Phenotype	p-values
	Ile	$\mathbf{2.8 \times 10^{-73}}$	Phe	$\mathbf{9.0 \times 10^{-12}}$	Gln	$\mathbf{3.5 \times 10^{-07}}$	Leu	$\mathbf{6.3 \times 10^{-09}}$	Gly	$\mathbf{9.3 \times 10^{-07}}$
	Val	$\mathbf{8.2 \times 10^{-74}}$	Ile	$\mathbf{3.2 \times 10^{-12}}$	His	$\mathbf{2.4 \times 10^{-08}}$	Ile	$\mathbf{3.2 \times 10^{-10}}$	Ala	$\mathbf{3.3 \times 10^{-06}}$
	Leu	$\mathbf{1.3 \times 10^{-74}}$	Leu	$\mathbf{8.6 \times 10^{-14}}$	Ile	$\mathbf{5.7 \times 10^{-09}}$	Ala	$\mathbf{2.5 \times 10^{-10}}$	His	$\mathbf{2.5 \times 10^{-06}}$
	Tyr	$\mathbf{3.3 \times 10^{-74}}$	Ala	$\mathbf{1.0 \times 10^{-14}}$	Phe	$\mathbf{2.3 \times 10^{-09}}$	His	$\mathbf{7.6 \times 10^{-11}}$	Gln	1.1×10^{-05}
	His	$\mathbf{4.8 \times 10^{-74}}$	Val	$\mathbf{2.6 \times 10^{-14}}$	Val	$\mathbf{2.9 \times 10^{-10}}$	Val	$\mathbf{2.6 \times 10^{-09}}$	Ile	2.3×10^{-05}
	Phe	$\mathbf{2.4 \times 10^{-76}}$	Gly	$\mathbf{2.7 \times 10^{-13}}$	Tyr	$\mathbf{1.2 \times 10^{-10}}$	Gln	$\mathbf{6.3 \times 10^{-07}}$	Phe	1.5×10^{-05}
	Ala	$\mathbf{3.2 \times 10^{-71}}$	Gln	$\mathbf{1.2 \times 10^{-11}}$	Gly	$\mathbf{6.7 \times 10^{-11}}$	Gly	$\mathbf{1.4 \times 10^{-06}}$	Val	1.0×10^{-03}
	Gln	$\mathbf{7.4 \times 10^{-64}}$	Tyr	$\mathbf{3.3 \times 10^{-09}}$	Leu	$\mathbf{1.5 \times 10^{-07}}$	Tyr	6.8×10^{-05}	Leu	4.9×10^{-02}
Remaining	Gly		His		Ala		Phe		Tyr	

Table A.3: Smallest 10 p-values and corresponding genes obtained by PhC($\Sigma_G = SKAT$), GAMuT (Projection and Linear kernel) and MSKAT (Q and Q' statistic). Each method produces the same set of top 10 genes, differing slightly by p-values. The tests were performed on unrelated individuals only (N = 7213).

PhC ($\Sigma_G = SKAT$)		GAMuT (Projection)		MSKAT (Q)		GAMuT (Linear)		MSKAT (Q')	
Gene	p-value	Gene	p-value	Gene	p-value	Gene	p-value	Gene	p-value
GLDC	$\mathbf{8.1 \times 10^{-54}}$	*GLDC*	$\mathbf{0}$	*GLDC*	$\mathbf{8.9 \times 10^{-54}}$	*GLDC*	$\mathbf{6.2 \times 10^{-15}}$	*GLDC*	$\mathbf{6.1 \times 10^{-15}}$
DHODH	$\mathbf{1.9 \times 10^{-06}}$	*DHODH*	$\mathbf{2.4 \times 10^{-06}}$	*DHODH*	$\mathbf{2.1 \times 10^{-06}}$	*METTL4*	3.0×10^{-05}	*METTL4*	3.0×10^{-05}
PAH	9.9×10^{-06}	*PAH*	1.0×10^{-05}	*PAH*	9.9×10^{-06}	*ASB10*	4.9×10^{-05}	*ASB10*	4.8×10^{-05}
ALDH1L1	6.0×10^{-05}	*DHODH*	6.1×10^{-05}	*ALDH1L1*	5.9×10^{-05}	*MEOX1*	6.4×10^{-05}	*MEOX1*	6.4×10^{-05}
HAL	9.5×10^{-05}	*HAL*	9.6×10^{-05}	*HAL*	9.5×10^{-05}	*PAH*	1.1×10^{-04}	*PAH*	1.3×10^{-04}
BCAT2	6.2×10^{-04}	*BCAT2*	6.3×10^{-04}	*BCAT2*	6.1×10^{-04}	*ABCC8*	2.5×10^{-04}	*ABCC8*	2.5×10^{-04}
STK33	6.7×10^{-04}	*STK33*	6.7×10^{-04}	*STK33*	6.8×10^{-04}	*OLFML2A*	2.8×10^{-04}	*OLFML2A*	2.8×10^{-04}
TBC1D4	1.7×10^{-04}	*TBC1D4*	1.6×10^{-04}	*TBC1D4*	1.6×10^{-04}	*OGG1*	3.9×10^{-04}	*OGG1*	4.0×10^{-04}
ABCC8	2.1×10^{-04}	*ABCC8*	2.3×10^{-04}	*ABCC8*	2.3×10^{-04}	*CPT1C*	4.7×10^{-04}	*CPT1C*	4.5×10^{-04}
MED1	1.7×10^{-03}	*MED1*	1.7×10^{-03}	*MED1*	1.7×10^{-03}	*DHODH*	2.2×10^{-03}	*DHODH*	2.2×10^{-03}

Table A.4: Functions and clinical implications for the significant and suggestive genes

Gene	Function	Clinical Implication / Associations
GLDC	catalyst in glycine cleavage system	glycine encephanlopathy, Autosomal recessive inheritance
HAL	catabolism of Histidine	Histidinemia, vitamin D measurement
DHODH	catalyzing pyrimidine de novo biosynthesis	postaxial acrofacial dysostosis, total cholesterol
PAH	iron containing enzyme	blood metabolite measurements
MED1	coactivator in the transcription of RNA polymerase II-dependent genes	asthma, inflammatory bowel disease
STK33	Serine/threonine protein kinase which phosphorylates VIME	BMI, small cell lung carcinoma
ALDH1L1	catalyzes the conversion of 10-formyltetrahydrofolate, NADP, and water to tetrahydrofolate, NADPH, and carbon dioxide	homocytosine, insulin sensitivity
BCAT2	catabolism of the branched chain amino acids leucine, isoleucine and valine	urinary metabolite, eye measurement, reticulocyte

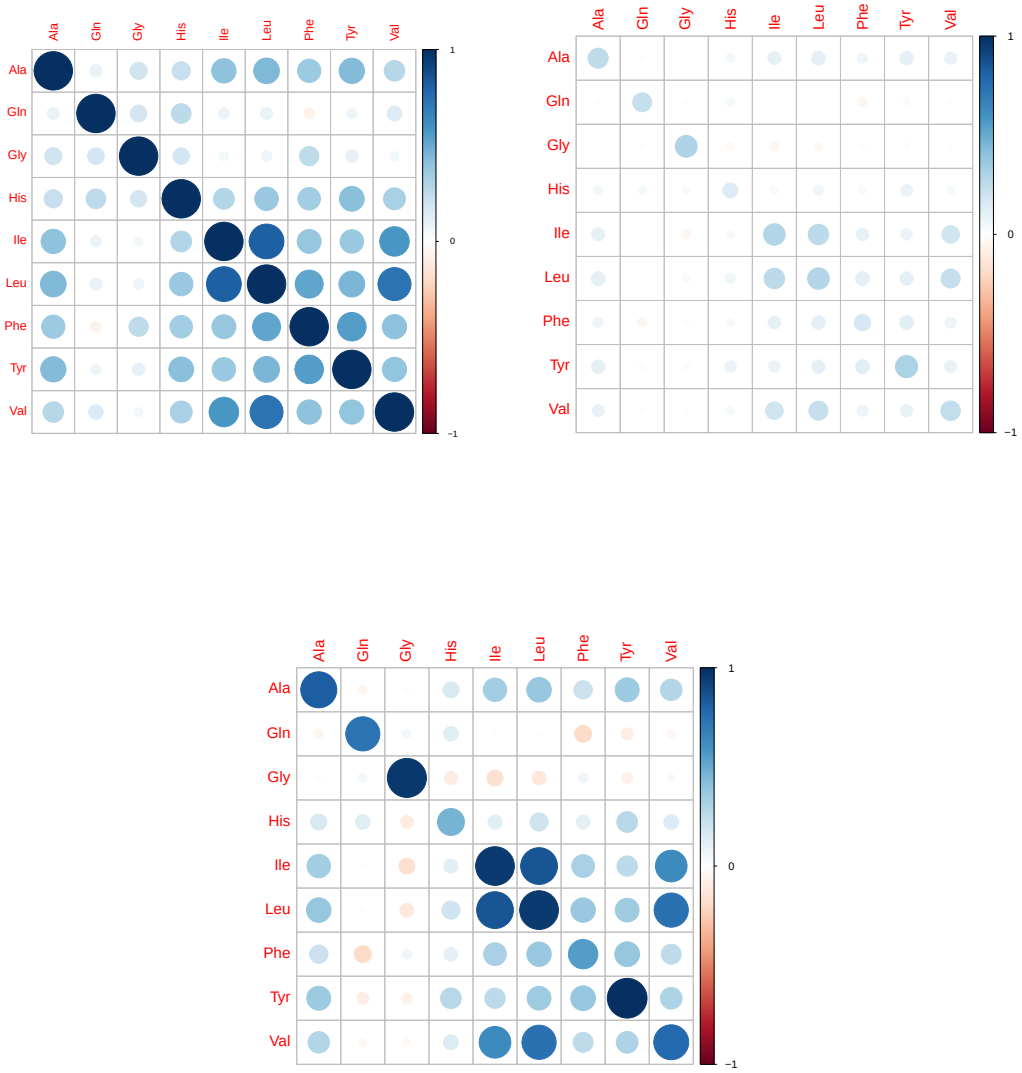


Figure A.1: Correlation and co-heritabilities of 9 amino acid phenotypes in METSIM. (a): Correlation matrix of the 9 amino acid phenotypes from METSIM study. (b): Co-heritability matrix of the same phenotypes as estimated from PHENIX.(c): Scaled co-heritability matrix: The elements in the matrix as shown in (b) were divided by the maximum diagonal element.
Ala: Alanine, Gln: Glutamine, Gly: Glycine, His: Histine, Ile: Isoleucine, Leu: Leucine, Phe: Phenylalanine, Tyr: Tyrosine, Val: Valine.

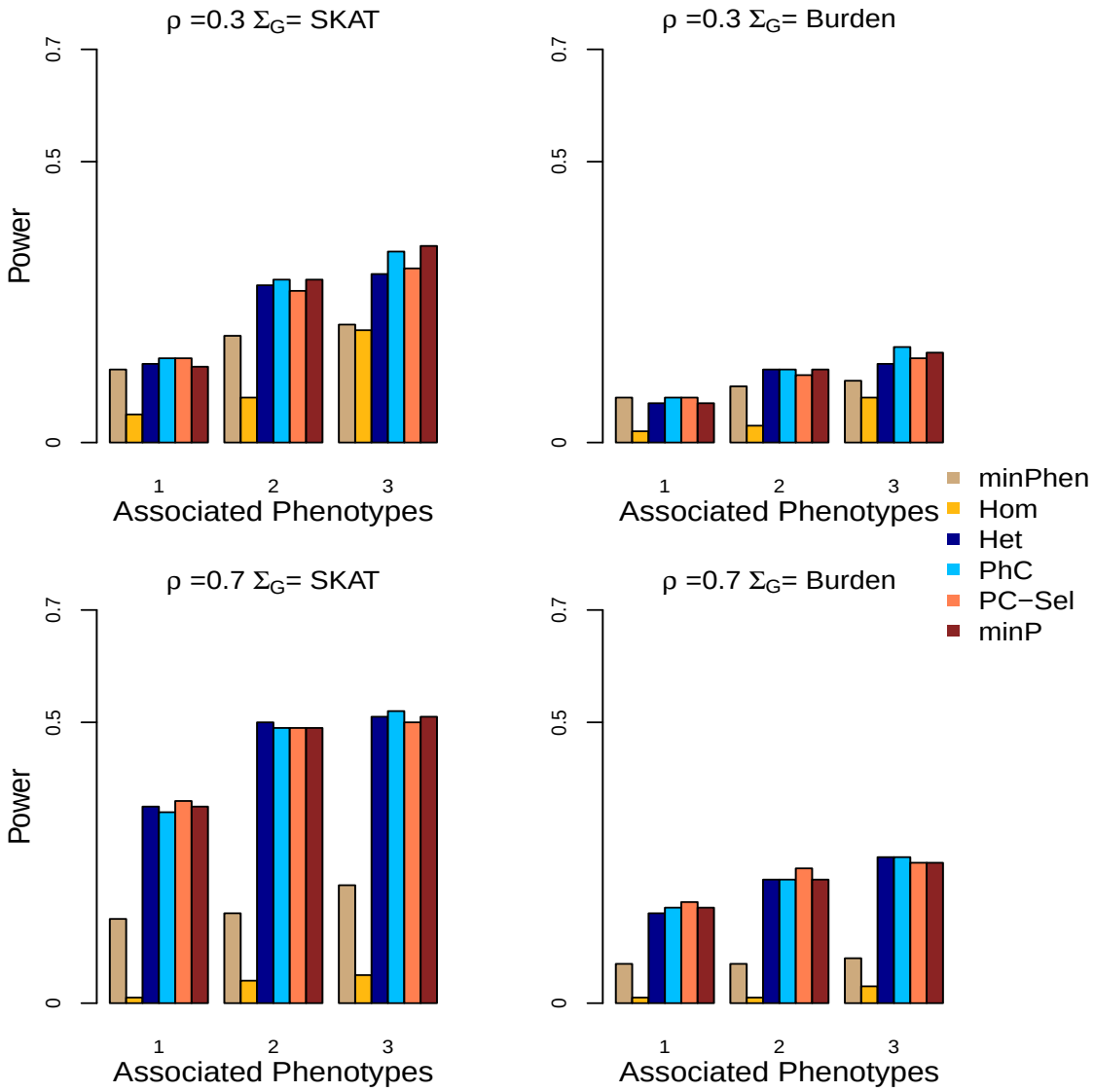


Figure A.2: Power for Multi-SKAT tests when phenotypes have compound symmetric correlation structures with a mixture of trait increasing and decreasing variants. Empirical power for minPhen, Hom, Het, PhC, PC-Sel, minP plotted against the number of phenotypes associated with the gene of interest with a total of 5 phenotypes under consideration. Upper row shows the results for $\rho = 0.3$ and lower row for $\rho = 0.7$. Left column shows results with SKAT kernel Σ_G, and right columns shows results with Burden kernel. 80%/20% of the causal variants were trait-increasing/trait-decreasing variants.

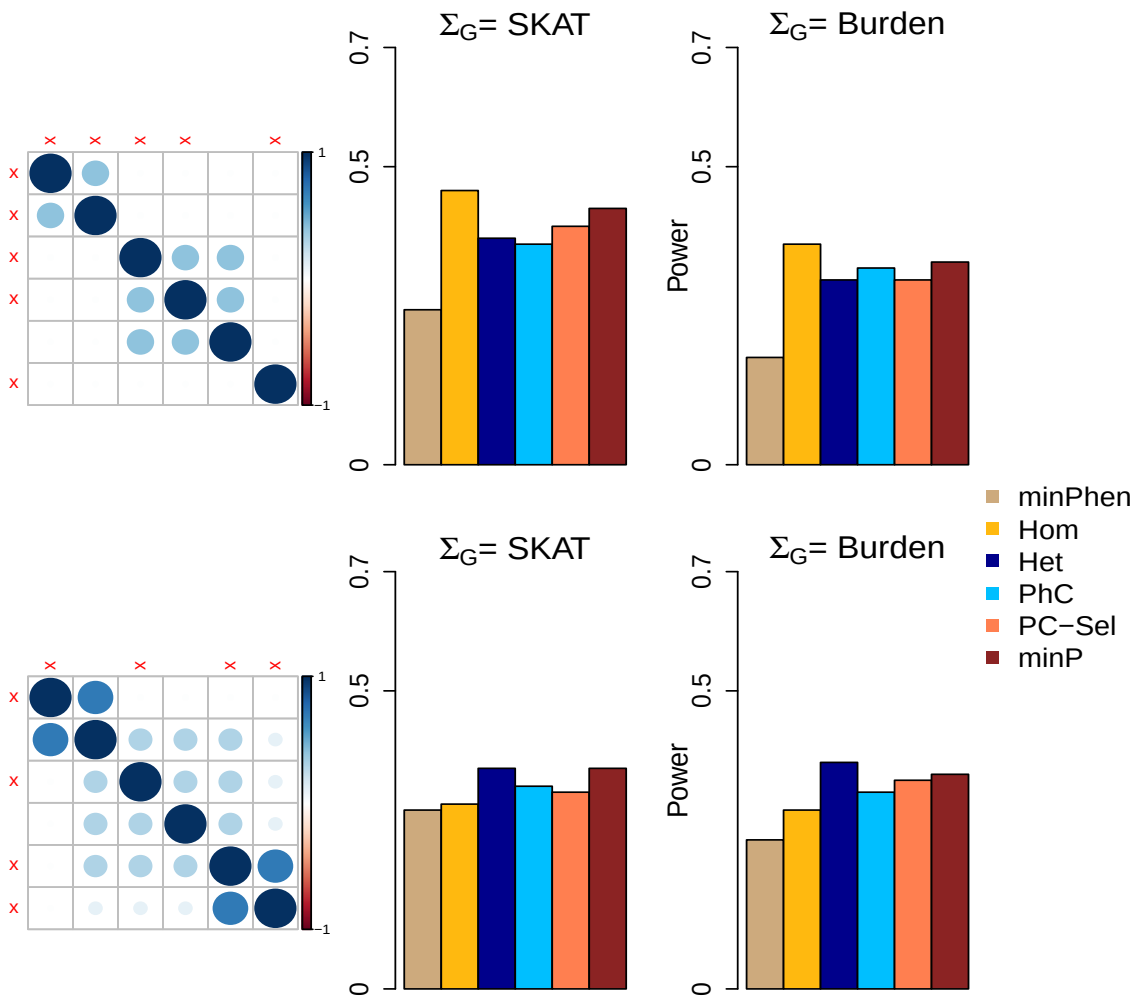


Figure A.3: Power for Multi-SKAT tests when phenotypes have clustered correlation structures with a mixture of trait increasing and decreasing variants. Empirical powers for minPhen, Hom, Het, PhC, PC-Sel, minP are plotted under different levels of association with a total of 6 phenotypes and with clustered correlation structures. Middle column shows the empirical powers for different combinations of phenotypes associated with SKAT kernel Σ_G; the rightmost column shows the corresponding results with Burden kernel; left column shows the corresponding correlation matrices for the phenotypes. The associated phenotypes are indicated in red cross marks across the correlation matrices. 80%/20% of the causal variants were trait-increasing/trait-decreasing variants.

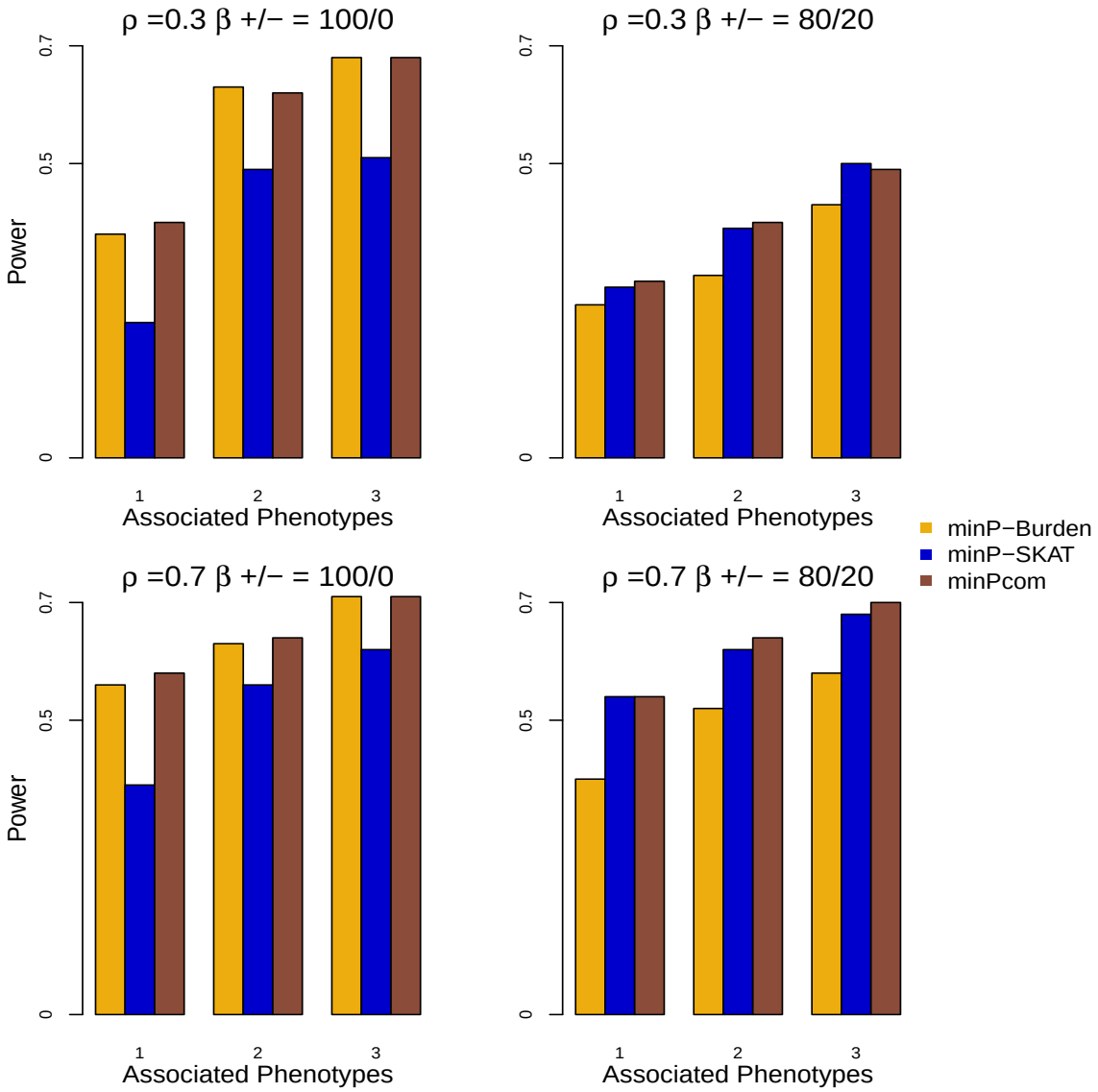


Figure A.4: Power for Multi-SKAT by combining tests with Σ_P as Hom, Het, PhC, PC-Sel and Σ_G as SKAT and Burden when phenotypes have compound symmetric correlation structures. Empirical powers for minP-Burden, minP-SKAT and minP_{com} are plotted against the number of phenotypes associated with the gene of interest with a total of 5 phenotypes under consideration and **50%** of the variants in the region are causal. Upper row shows the results for $\rho = 0.3$ and lower row for $\rho = 0.7$. Left column shows results when all the causal variants were trait-increasing variants, and right column shows results when 80%/20% of the causal variants were trait-increasing/trait-decreasing variants.

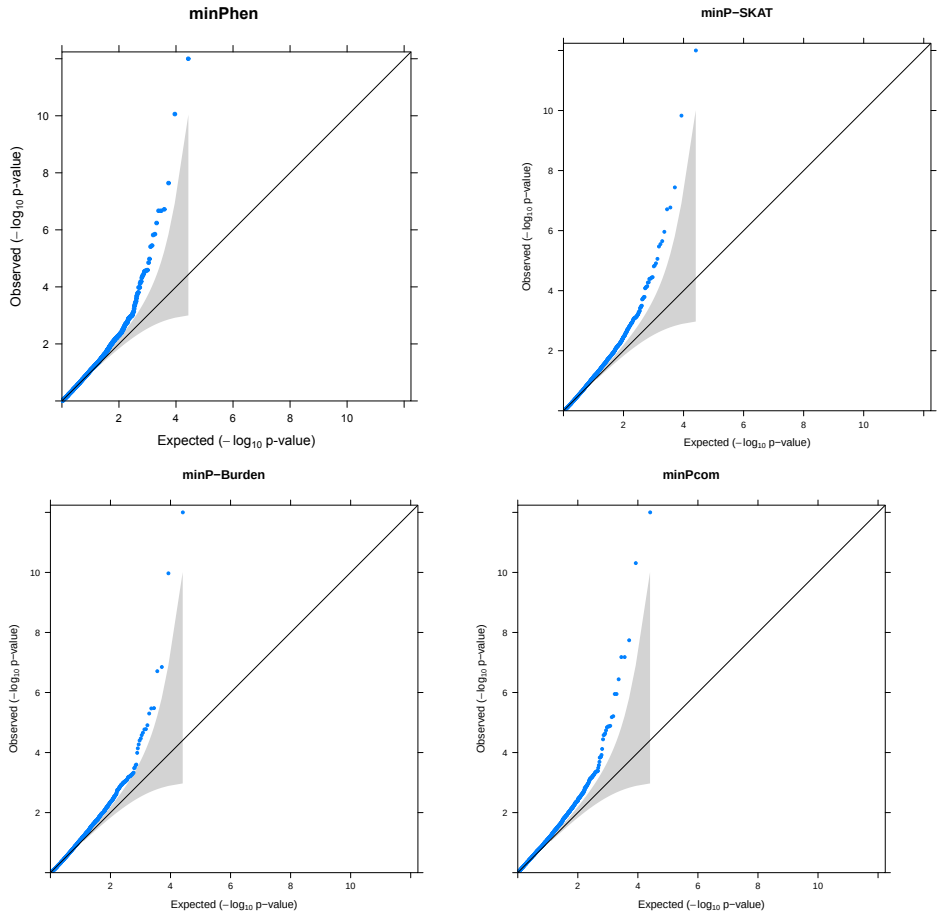


Figure A.5: QQplot of the p-values of Multi-SKAT omnibus tests without kinship adjustment for the METSIM data (N = 8545). For the ease of viewing, any associations with p-values $< 10^{-12}$ have been collapsed to 10^{-12}.

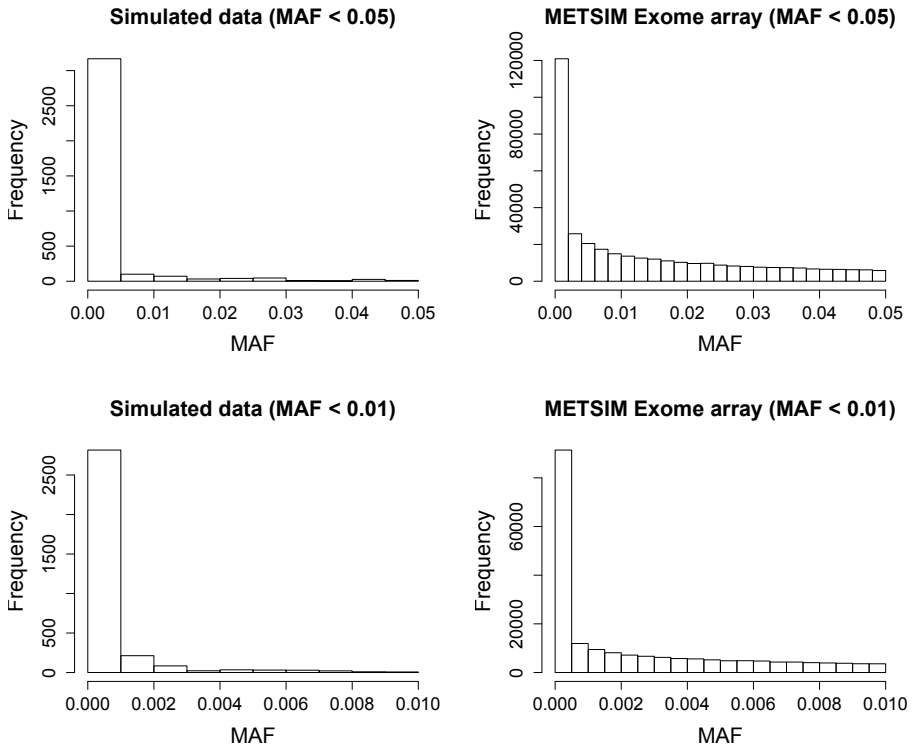


Figure A.6: Minor allele frequency (MAF) spectrum in simulations and METSIM data. Upper panel shows the MAFs for variants having MAF $< 5\%$. Lower panel zooms in into a region with variants having MAF $< 1\%$.

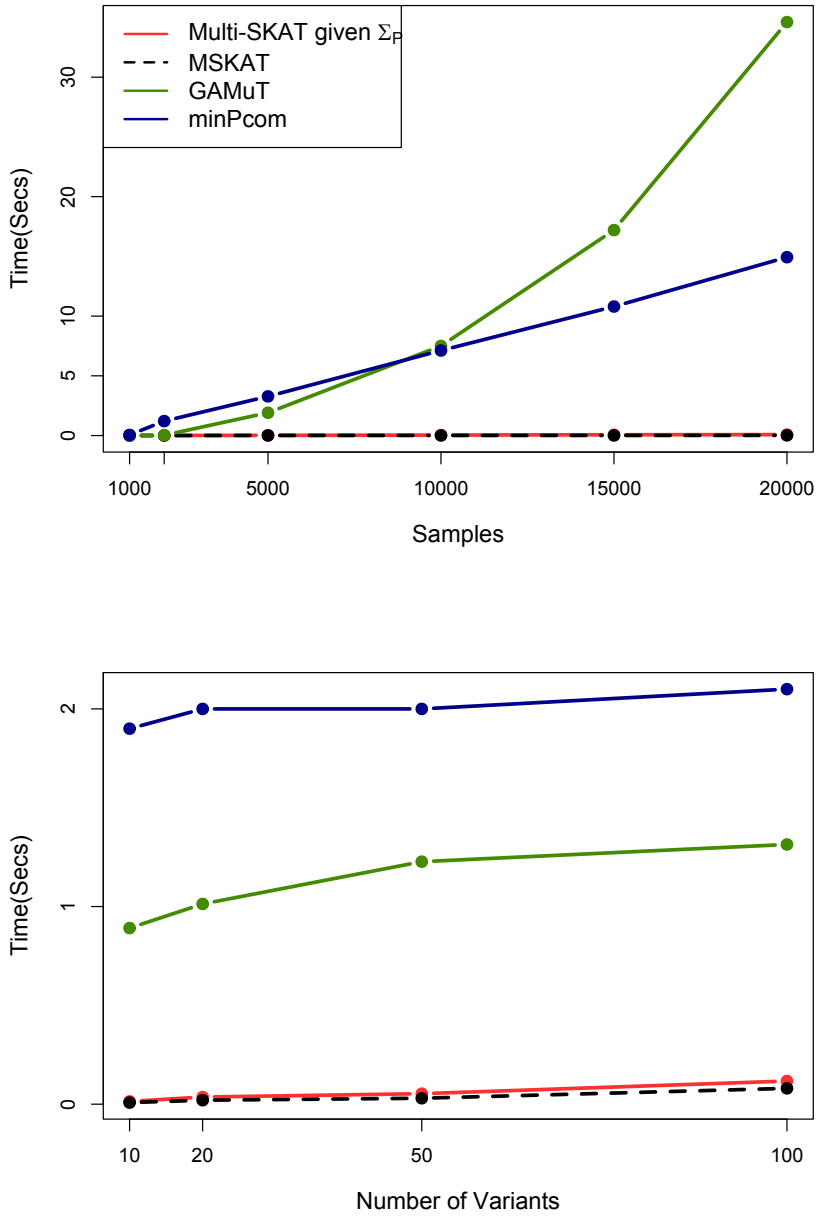


Figure A.7: Computation time of Multi-SKAT and existing methods with unrelated individuals and 10 phenotypes. (a) Estimated computation time for different sample sizes when the number of variant was 20. (b) Estimated computation time for different number of variants when the sample size was 5000. Each dot represents the average from 100 datasets.

APPENDIX B

Appendix for Chapter III

B.1 P-value for Meta-MultiSKAT tests

In (3.3), under the null hypotheses L_{meta} $N(0, \Phi_{meta})$ asymptotically where

$$\Phi_{meta} = \begin{bmatrix} \Phi_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Phi_S \end{bmatrix}$$

Hence under the null hypothesis, Q_{meta} in (3.3) follows a mixture of chi-squares. The mixing parameters of the distribution are eigenvalues of $\tilde{R}\Phi_{meta}\tilde{R}^T$ with $\Sigma_S \otimes \Sigma_G \otimes \Sigma_P = \tilde{R}\tilde{R}^T$. The p-value can be obtained by inverting the characteristic function of the null distribution.

B.2 Kernelized scores

It is to be noted that (3.3), assumes that Σ_G and Σ_P are the same for individual studies. This assumption is restrictive as different studies might be analysed with different hypotheses, reflected in different Σ_G and Σ_P across studies. We can relax

this assumption by using kernelized score matrix for each study in place of the score matrix. We construct the kernelized score statistic in each study as:

$$\tilde{L}_s = \Sigma_G^{\frac{1}{2}} L_s \Sigma_P^{\frac{1}{2}}$$

Under the null hypothesis, $vec(\widetilde{L_s}) \sim N(0, \tilde{\Phi}_s)$, where

$$\widetilde{\Phi}_s = (\Sigma_G^{\frac{1}{2}} G_s^T G_s \Sigma_G^{\frac{1}{2}} - \Sigma_G^{\frac{1}{2}} G_s^T X_s (X_s^T X_s)^{-1} X_s^T G_s \Sigma_G^{\frac{1}{2}}) \otimes (\Sigma_p^{\frac{1}{2}} V_s^{-1} \Sigma_P^{\frac{1}{2}})$$

is the kernelized phenotype adjusted variant relationship matrix. Given $(\widetilde{L}_s, \widetilde{\Phi}_s), s = 1, \cdots S$, we construct the kernelized meta-score-vector as $\widetilde{L}_{meta} = (vec(\widetilde{L}_1)^T, vec(\widetilde{L}_2)^T, \cdots, vec(\widetilde{L}_s)^T)^T$, which under the null hypothesis follows a normal distribution with mean 0 and variance-covariance matrix

$$\widetilde{\Phi}_{meta} = \begin{bmatrix} \widetilde{\Phi}_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \widetilde{\Phi}_S \end{bmatrix}$$

Then the test statistics can be constructed as in (3.4) with the null distribution being a mixture of chi-squares.

B.3 Resampling algorithm

In (3.3), $L_s \sim N(0, \Phi_s)$ and in particular $L_{meta} \sim N(0, \Phi_{meta})$. under the null hypothesis of no association. Suppose we have B Meta-MultiSKAT tests with corresponding kernel matrices Σ_S , Σ_P and Σ_G and p-values $T_P = (p_1, p_2, \cdots, p_B)$. Our test statistic is $p_{min} = \min(p_1, p_2, \cdots, p_B)$. First we adopt a resampling based approach to estimate this correlation structure as:

- Generate null observations $L_{s;null}$ from $N(0, \Phi_s)$ for $s = 1, 2, \cdots, S$ and con-

struct $L_{meta;null}$ as above

- Calculate Meta-MultiSKAT test statistic $Q_{meta;null}$ as in (3.3)(alternatively (3.4)) for each of the B combinations of Σ_S , Σ_P and Σ_G
- Calculate asymptotic null p-value
- Repeat the previous steps for R (= 500 or 1000) times and calculate the null correlation between the p-values

With the estimated null correlation structure, we use a t-Copula to approximate the joint distribution of T_P. The final p-value for p_{min} is then calculated from the distribution function of the assumed t-Copula. As the same way, the resampling approach can be used with the kernelized score $\widetilde{L}_{meta}$.

B.4 Illustration: Missing phenotype

To demonstrate the utility of Meta-MultiSKAT to handle missing phenotypes, we performed an analysis with all the 4 WBC phenotypes (lymphocyte, monocyte, basophil, eosinophill) in SardiNIA and only 3 (monocyte, basophil and eosinophil) in MGI. The models used for individual studies to extract the summary statistics remained the same.

We used Meta-MultiSKAT-Common-Rare tests in this analysis (Supplemental Table S2). All the variants, common and rare, were used in this analysis. The genes that were identified in the previous analysis were found to be significant or suggestive (p-value $< 10^5$) in this analysis as well, but with slightly differing p-values. As before, PRG2, RP11-872D17.8, IRF8 and CCL24 found to be significant using Meta-MultiSKAT methods.

Table B.1: Single phenotype and Multi-SKAT p-value for each of the 4 WBC subtypes in each of MGI and SardiNIA studies. For single-phenotype gene-based tests for rare-variants and CADD-score weighting, SKAT-O was used

		Lymphocyte		Monocyte		Basophil		Eosinophil		Joint (Multi-SKAT)	
Test	Gene	MGI	SardiNIA	MGI	SardiNIA	MGI	SardiNIA	MGI	SardiNIA	MGI	SardiNIA
Rare-variant test	*PRG2*	0.12	0.51	0.67	0.54	0.12	0.09	0.02	3.7×10^{-7}	0.48	8.8×10^{-7}
	RP11-872D17.8	0.23	0.77	0.58	0.91	0.16	0.11	0.01	5.9×10^{-7}	0.80	3.1×10^{-6}
Common-Rare tests	*PRG2*	0.39	0.54	0.73	0.61	0.12	0.07	0.08	7.8×10^{-7}	0.37	4.6×10^{-7}
	RP11-872D17.8	0.33	0.74	0.81	0.88	0.24	0.10	0.29	1.1×10^{-6}	0.80	9.1×10^{-6}
	IRF8	0.83	0.01	1.0×10^{-5}	3.9×10^{-5}	0.76	0.31	0.97	0.93	5.9×10^{-5}	3.5×10^{-4}
	CCL24	0.67	0.92	0.08	0.39	0.03	4.3×10^{-5}	9.1×10^{-5}	0.69	2.0×10^{-4}	1.2×10^{-3}
CADD-score weighting	*PRG2*	0.26	0.43	0.71	0.42	0.06	0.06	0.04	7.1×10^{-7}	0.17	2.9×10^{-8}
	RP11-872D17.8	0.41	0.59	0.63	0.81	0.11	0.08	0.10	8.9×10^{-6}	0.27	4.3×10^{-6}
	IRF8	0.61	0.24	1.7×10^{-5}	1.9×10^{-4}	0.63	0.29	0.88	0.81	1.1×10^{-5}	3.8×10^{-3}
	CCL24	0.48	0.76	0.12	0.16	0.07	1.0×10^{-4}	8.9×10^{-5}	0.53	1.3×10^{-4}	1.2×10^{-3}

Table B.2: Estimated Type-1 error rates for Meta-MultiSKAT-Common-Rare tests

α	Meta-Hom	Meta-Het	Meta-Com
1×10^{-5}	1.3×10^{-5}	1.3×10^{-5}	1.4×10^{-5}
1×10^{-4}	1.1×10^{-4}	1.1×10^{-4}	1.3×10^{-4}

Table B.3: Significant genes identified by Meta-MultiSKAT with missing phenotypes. Genes/regions identified by either of the Meta-MultiSKAT methods (Meta-Hom, Meta-Het or Meta-Com) in the example with missing phenotypes. The p-values $< 10^5$ were marked in bold. Variants with pooled MAF $\leq$1% ($>$ 1%) are included as rare (common)

Gene	Meta-Het	Meta-Hom	Meta-Com
IRF8	6.5×10^{-4}	$\mathbf{1.7 \times 10^{-6}}$	$\mathbf{2.8 \times 10^{-6}}$
PRG2	$\mathbf{2.4 \times 10^{-7}}$	7.6×10^{-5}	$\mathbf{4.1 \times 10^{-7}}$
CCL24	$\mathbf{4.8 \times 10^{-6}}$	8.2×10^{-3}	$\mathbf{8.9 \times 10^{-6}}$
RP11-872D17.8	$\mathbf{8.4 \times 10^{-7}}$	4.1×10^{-4}	$\mathbf{1.2 \times 10^{-6}}$

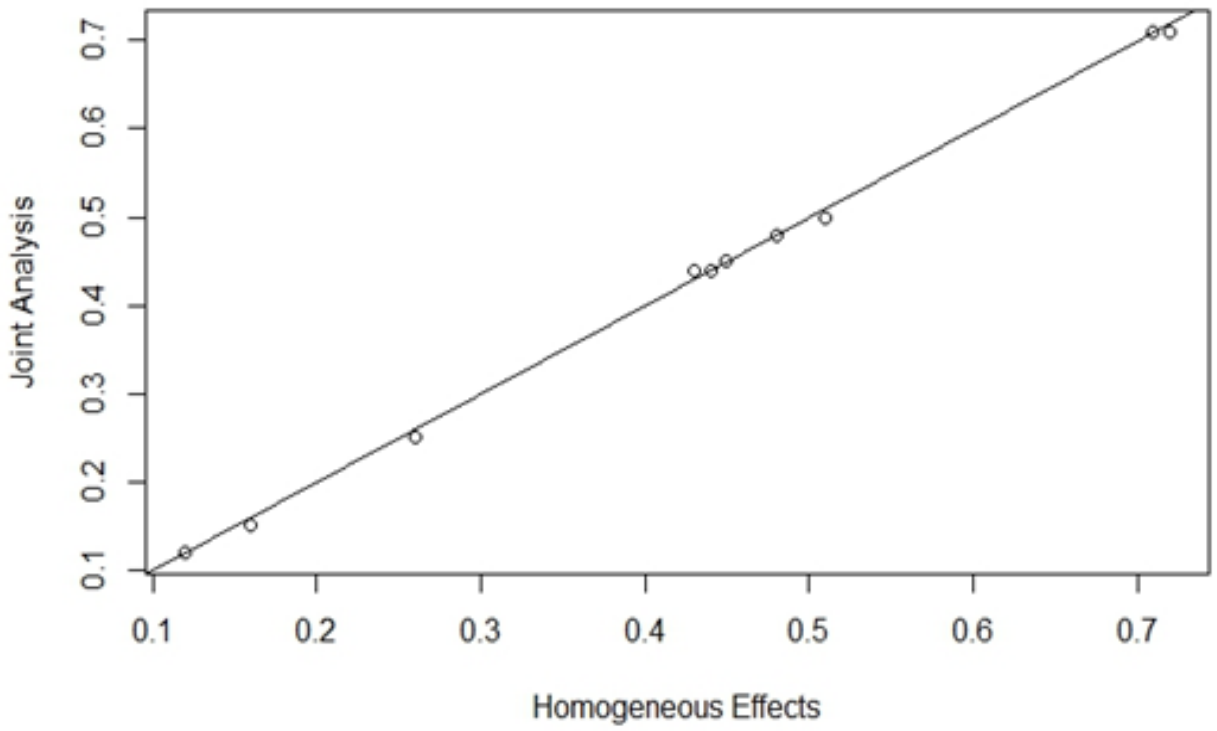


Figure B.1: Power comparison for Joint analysis and analysis with $\Sigma_s = \Sigma_{S;Hom}$

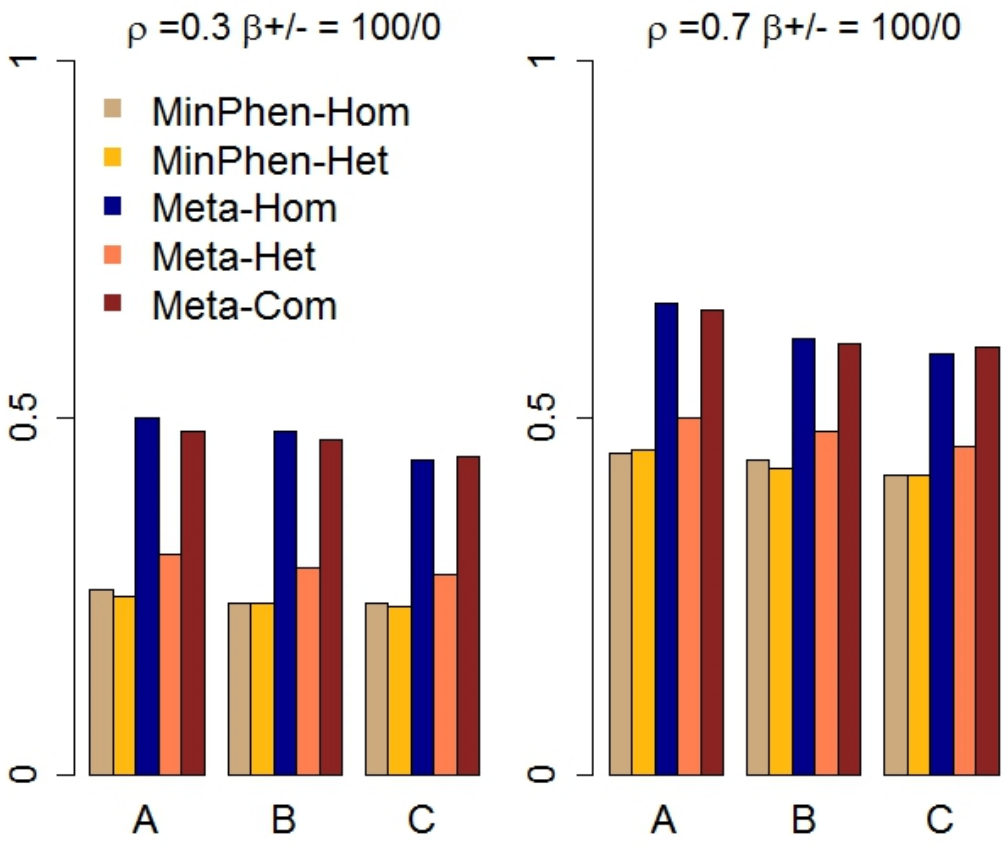


Figure B.2: Power for Meta-MultiSKAT-Common-Rare tests compared with the existing methods when the set of causal variants is the same across different studies and has the same direction of effect. Empirical power for Meta-Hom, Meta-Het and Meta-Com plotted for 3 different scenarios compared against MinPhen-Hom and MinPhen-Het (See Simulations for details). Left panel shows the results for low correlation ($\rho = 0.3$) among the phenotypes and right panel shows the results for high correlation ($\rho = 0.7$)

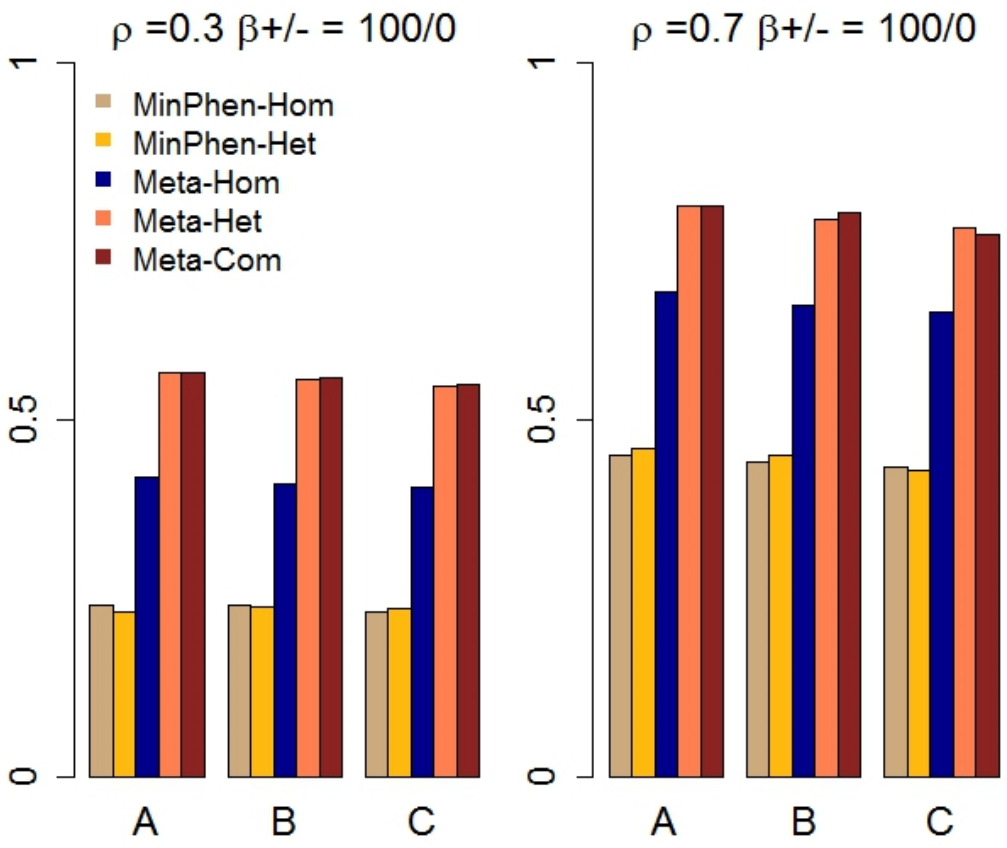


Figure B.3: Power for Meta-MultiSKAT-Common-Rare tests when the set of causal variants is randomly chosen for each study and has the same direction of effect. Empirical power for Meta-Hom, Meta-Het and Meta-Com plotted for 3 different scenarios compared against MinPhen-Hom and MinPhen-Het (See Simulations for details). Left panel shows the results for low correlation ($\rho = 0.3$) among the phenotypes and right panel shows the results for high correlation ($\rho = 0.7$)

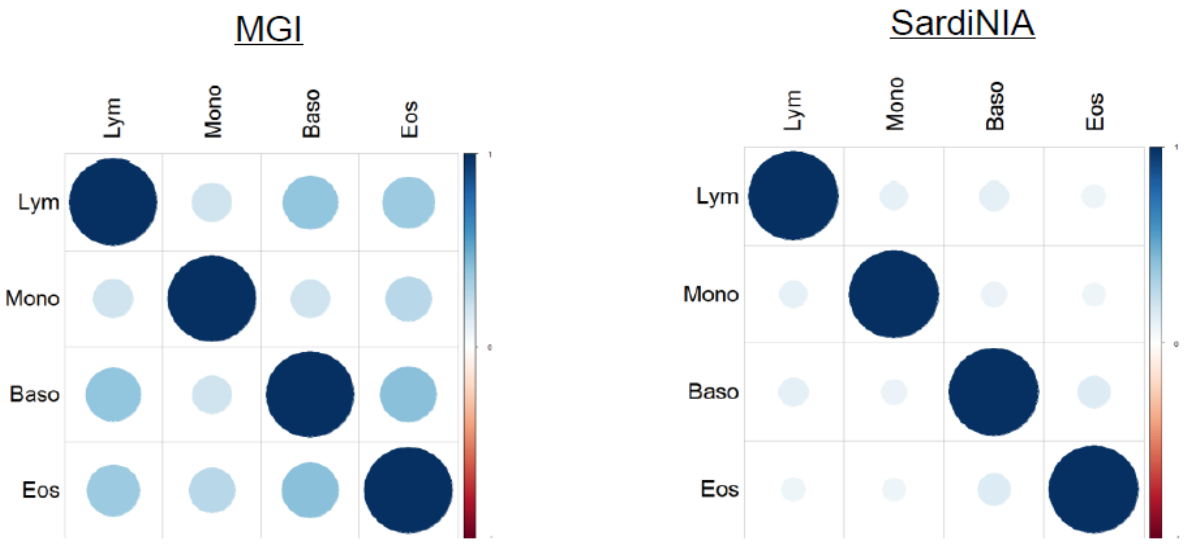


Figure B.4: Correlation structure of the WBC phenotypes in MGI and SardiNIA respectively. Lym: Lymphocytes; Mono: Monocytes; Baso: Basophils; Eos:Eosinophils

APPENDIX C

Appendix for Chapter IV

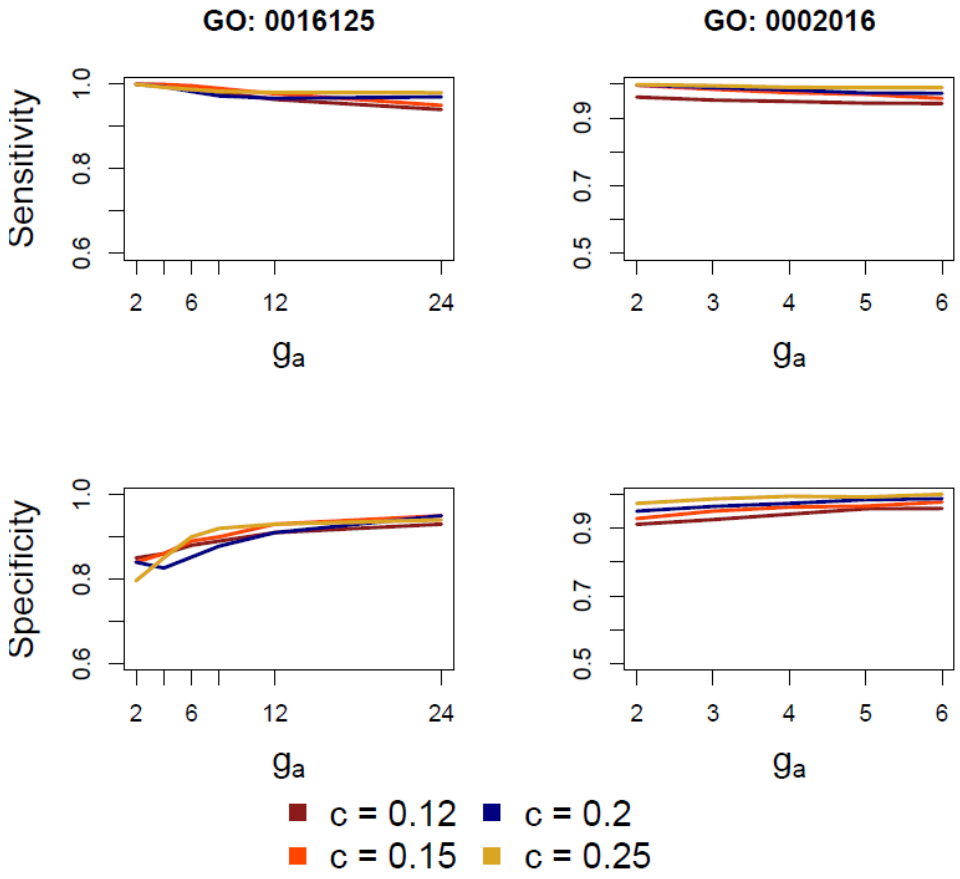


Figure C.1: Sensitivity and Specificity of GAUSS for GO: 0016125 and GO: 0002016 for different magnitudes of effects (c) and different number of active genes (g_a)

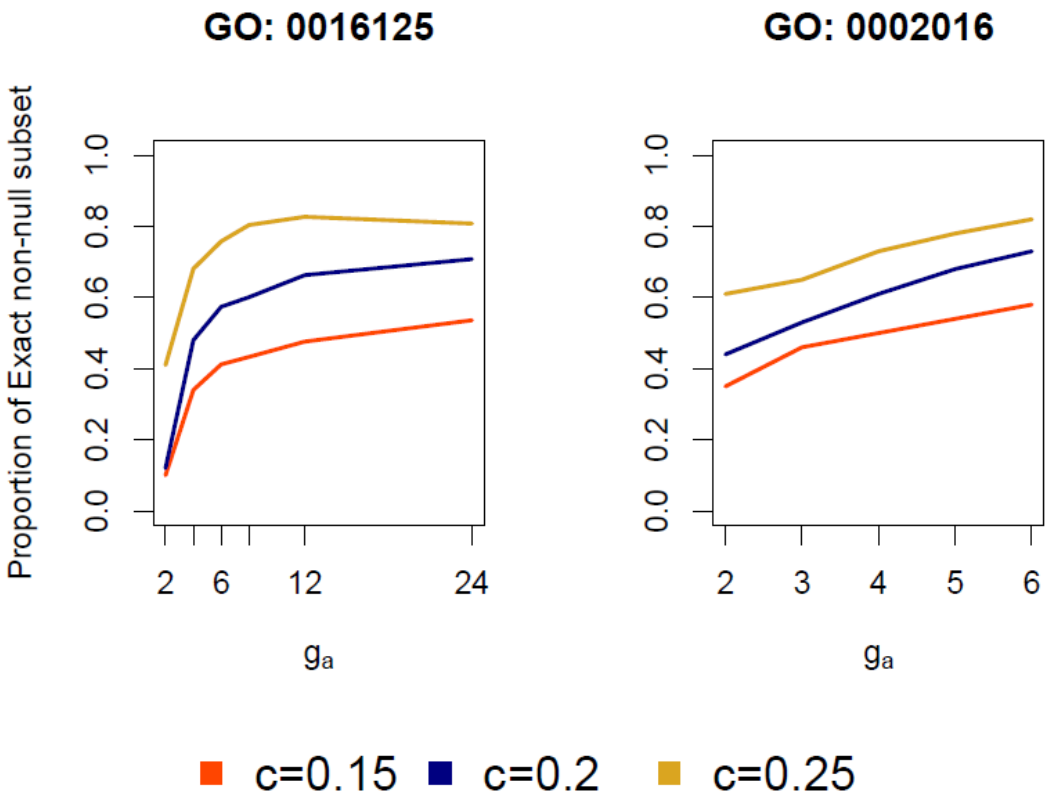


Figure C.2: Estimate of the probability of identifying the exact non-null subset by the active subset (AS) genes selected through GAUSS across different magnitudes of effects (c) and different number of active genes (g_a)

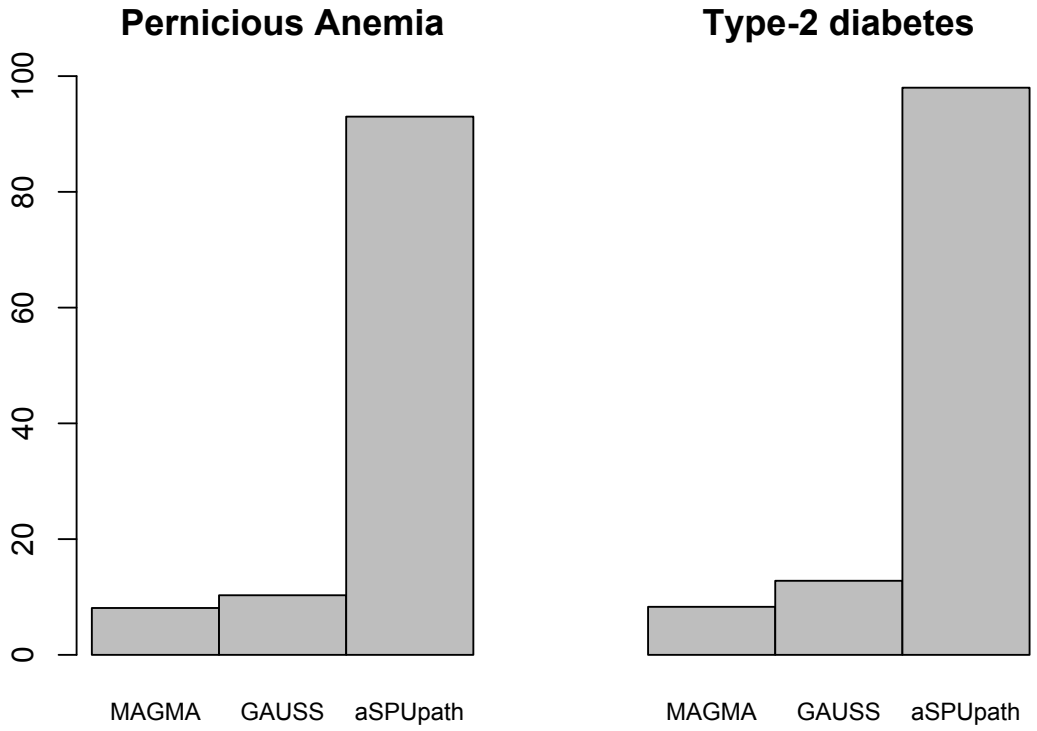


Figure C.3: Total run-time of GAUSS for Pernicious anemia and Type-2 diabetes in UK Biobank compared to that of MAGMA and aSPUpath. Total run-time is calculated as the net time taken starting from the input of summary statistic till the p-values for the $10,679$ gene-sets are generated

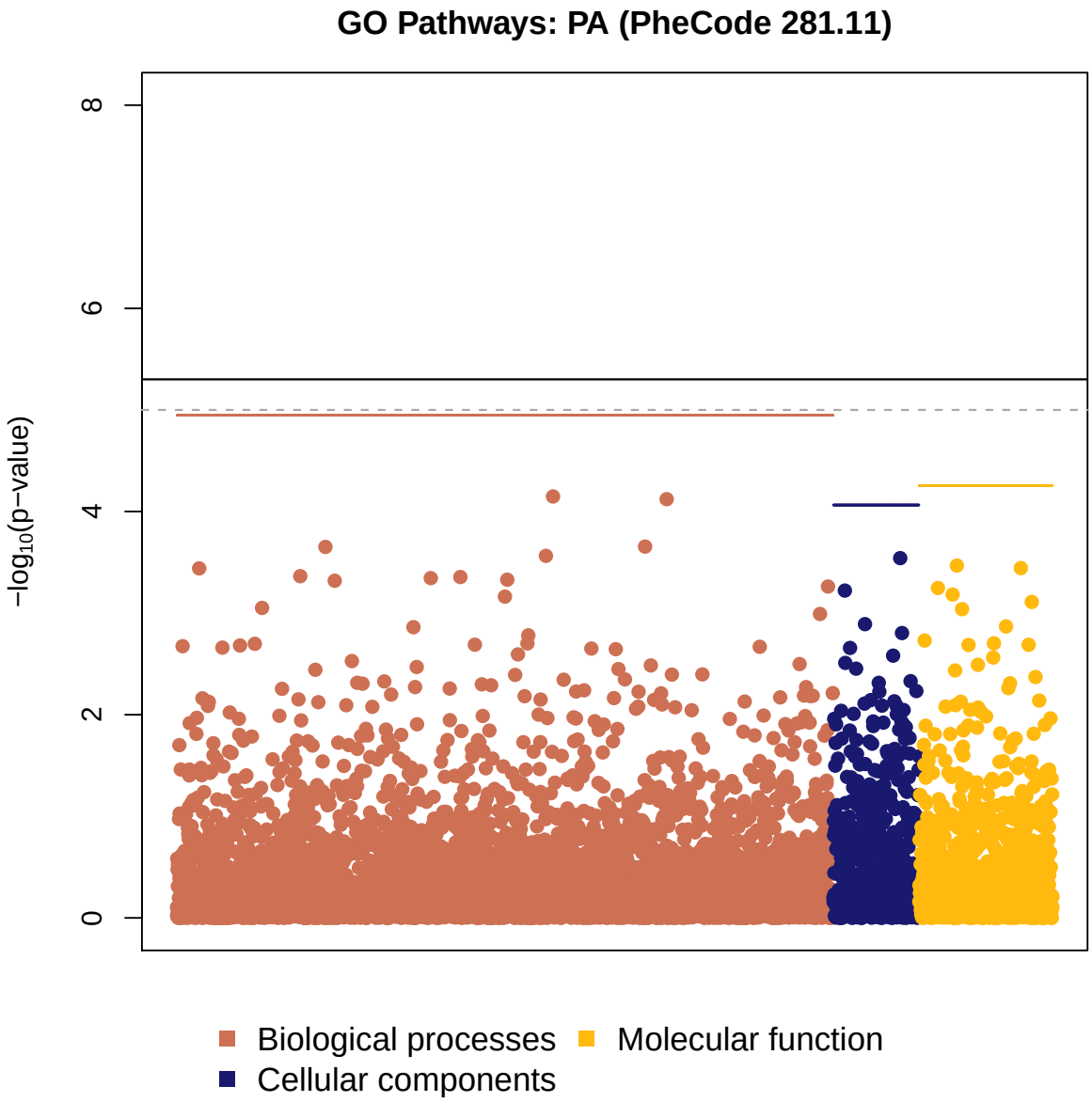


Figure C.4: P-values for association of Pernicious anemia (PA) with the GO gene-sets (C5) using MAGMA.

BIBLIOGRAPHY

BIBLIOGRAPHY

D. Allan Butterfield and C. B. Pocernich. The glutamatergic system and alzheimers disease. *CNS Drugs*, 17(9):641–652, 2003.

H. L. Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467:832–838, 2010.

M. Aminoff, J. E. Carter, R. B. Chadwick, C. Johnson, R. Gräsbeck, M. A. Abdelaal, H. Broch, L. B. Jenner, P. J. Verroust, S. K. Moestrup, A. de la Chapelle, and R. Krahe. Mutations in CUBN, encoding the intrinsic factor-vitamin B12 receptor, cubilin, cause hereditary megaloblastic anaemia 1. *Nature Genetics*, 21(3):309–313, mar 1999. ISSN 1061-4036. doi: 10.1038/6831. URL `http://www.nature.com/articles/ng0399{_}309`.

H. Aschard, B. J. Vilhjálmsson, N. Greliche, P.-E. Morange, D.-A. Trégouët, and P. Kraft. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *The American Journal of Human Genetics*, 94(5):662–676, 2014.

W. J. Astle, H. Elding, T. Jiang, D. Allen, D. Ruklisa, A. L. Mann, D. Mead, H. Bouman, F. Riveros-Mckay, M. A. Kostadima, J. J. Lambourne, S. Sivapalaratnam, K. Downes, K. Kundu, L. Bomba, K. Berentsen, J. R. Bradley, L. C. Daugherty, O. Delaneau, K. Freson, S. F. Garner, L. Grassi, J. Guerrero, M. Haimel, E. M. Janssen-Megens, A. Kaan, M. Kamat, B. Kim, A. Mandoli, J. Marchini, J. H. Martens, S. Meacham, K. Megy, J. O'Connell, R. Petersen, N. Sharifi, S. M. Sheard, J. R. Staley, S. Tuna, M. van der Ent, K. Walter, S.-Y. Wang, E. Wheeler, S. P. Wilder, V. Iotchkova, C. Moore, J. Sambrook, H. G. Stunnenberg, E. Di Angelantonio, S. Kaptoge, T. W. Kuijpers, E. Carrillo-de Santa-Pau, D. Juan, D. Rico, A. Valencia, L. Chen, B. Ge, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yang, R. Guigo, S. Beck, D. S. Paul, T. Pastinen, D. Bujold, G. Bourque, M. Frontini, J. Danesh, D. J. Roberts, W. H. Ouwehand, A. S. Butterworth, and N. Soranzo. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*, 167(5): 1415–1429.e19, nov 2016. ISSN 00928674. doi: 10.1016/j.cell.2016.10.042. URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867416314635http://www.ncbi.nlm.nih.gov/pubmed/27863252http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5300907`.

M. Baus-Loncar, J. Schmid, E. N. Lalani, I. Rosewell, R. A. Goodlad, G. W. Stamp, N. Blin, and T. Kayademir. Trefoil factor 2 (Tff2) deficiency in murine digestive tract influences the immune system. *Cellular Physiology and Biochemistry*, 2005. ISSN 10158987. doi: 10.1159/000087729.

S. Bhattacharjee, P. Rajaraman, K. B. Jacobs, W. A. Wheeler, B. S. Melin, et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Genetics*, 90(5):821–835, 2012. doi: 10.1016/j.ajhg.2012.03.015.

K. A. Broadaway, D. J. Cutler, R. Duncan, J. L. Moore, E. B. Ware, L. F. B. Min A. Jhun, W. Zhao, J. A. Smith, P. A. Peyser, S. L. Kardia, D. Ghosh, and M. P. Epstein. A statistical approach for testing cross-phenotype effects of rare variants. *American Journal of Human Genetics*, 98(3):525–540, 2016.

B. K. Bulik-Sullivan, P.-R. Loh, H. K. Finucane, S. Ripke, J. Yang, S. W. G. of the Psychiatric Genomics Consortium, N. Patterson, M. J. Daly, A. L. Price, and B. M. Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47:291–295, 2015.

A. Buniello, J. MacArthur, M. Cerezo, L. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acid Res.*, 47:D1005D1012, 2019.

W. S. Bush and J. H. Moore. Genome-wide association studies. *Plos Comput. Bio*, 8(12):e1002822, 2012.

C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. OConnell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, and J. Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562:203–209, 2018.

R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing gwas results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

E. Cirulli and D. Goldstein. Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Rev. Genetics*, 11:415–425, 2010.

C. Cotsapas, B. F. Voight, E. Rossin, K. Lage, B. M. Neale, C. Wallace, G. R. Abecasis, J. C. Barrett, T. Behrens, J. Cho, et al. Pervasive sharing of genetic effects in autoimmune disease. *PLoS genetics*, 7(8):e1002254, 2011.

L. A. Criswell, K. A. Pfeiffer, R. F. Lum, B. Gonzales, J. Novitzke, M. Kern, K. L. Moser, A. B. Begovich, V. E. Carlton, W. Li, A. Lee, W. Ortmann, T. Behrens, and P. K. Gregersen. Analysis of families in the multiple autoimmune disease

genetics consortium (madgc) collection: the ptpn22 620w allele associates with multiple autoimmune phenotypes. *The American Journal of Human Genetics*, 76(4):561–571, 2005.

D. R. Crosslin, A. McDavid, N. Weston, X. Zheng, E. Hart, M. de Andrade, I. J. Kullo, C. A. McCarty, K. F. Doheny, E. Pugh, A. Kho, M. G. Hayes, M. D. Ritchie, A. Saip, D. C. Crawford, P. K. Crane, K. Newton, D. S. Carrell, C. J. Gallego, M. A. Nalls, R. Li, D. B. Mirel, A. Crenshaw, D. J. Couper, T. Tanaka, F. J. van Rooij, M.-H. Chen, A. V. Smith, N. A. Zakai, Q. Yango, M. Garcia, Y. Liu, T. Lumley, A. R. Folsom, A. P. Reiner, J. F. Felix, A. Dehghan, J. G. Wilson, J. C. Bis, C. S. Fox, N. L. Glazer, L. A. Cupples, J. Coresh, G. Eiriksdottir, V. Gudnason, S. Bandinelli, T. M. Frayling, A. Chakravarti, C. M. van Duijn, D. Melzer, D. Levy, E. Boerwinkle, A. B. Singleton, D. G. Hernandez, D. L. Longo, J. C. Witteman, B. M. Psaty, L. Ferrucci, T. B. Harris, C. J. O'Donnell, S. K. Ganesh, E. B. Larson, C. S. Carlson, and G. P. Jarvik. Genetic variation associated with circulating monocyte count in the eMERGE Network. *Human Molecular Genetics*, 22(10):2119–2127, may 2013. ISSN 1460-2083. doi: 10.1093/hmg/ddt010. URL `https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddt010`.

A. Dahl, V. Iotchkova, A. Baud, A. Johansson, U. Gyllensten, N. Soranzo, R. Mott, A. Kranis, and J. Marchini. A multiple-phenotype imputation method for genetic studies. *Nature Genetics*, 48:466–472, 2016.

R. B. Davies. The distribution of a linear combination of x2 random variables. *Applied Statistics*, 29(3):323–333, 1980.

J. de Belleroche, A. Recordati, and F. Clifford Rose. Elevated levels of amino acids in the csf of motor neuron disease patients. *CNS Drugs*, 17(9):641–652, 2003.

P. L. De Jager, X. Jia, J. Wang, P. I. W. de Bakker, L. Ottoboni, N. T. Aggarwal, L. Piccio, S. Raychaudhuri, D. Tran, C. Aubin, R. Briskin, S. Romano, S. E. Baranzini, J. L. McCauley, M. A. Pericak-Vance, J. L. Haines, R. A. Gibson, Y. Naeglin, B. Uitdehaag, P. M. Matthews, L. Kappos, C. Polman, W. L. McArdle, D. P. Strachan, D. Evans, A. H. Cross, M. J. Daly, A. Compston, S. J. Sawcer, H. L. Weiner, S. L. Hauser, D. A. Hafler, and J. R. Oksenberg. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature Genetics*, 41(7):776–782, jul 2009. ISSN 1061-4036. doi: 10.1038/ng.401. URL `http://www.nature.com/doifinder/10.1038/ng.401`.

C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology*, 11(4):e1004219, apr 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004219. URL `https://dx.plos.org/10.1371/journal.pcbi.1004219`.

S. Demarta and A. J. McNeil. The t copula and related copulas. *International Statistical Review/Revue Internationale de Statistique*, pages 111–129, 2005.

D. Dutta, L. Scott, M. Boehnke, and S. Lee. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genetic Epidemiology*, 43(1):4–23, feb 2019. ISSN 07410395. doi: 10.1002/gepi.22156. URL `http://doi.wiley.com/10.1002/gepi.22156`.

W. W. Eaton, N. R. Rose, A. Kalaydjian, and P. B. Pedersen, M. G.and Mortensen. Epidemiology of autoimmune diseases in denmark. *J. Autoimmun.*, 29:1–9, 2007.

A. Edwards, R. Ritter, K. J. Abel, A. Manning, C. Panhuysen, and L. A. Farrer. Complement factor h polymorphism and age-related macular degeneration. *Science*, 308 (5720):421–424, 2005.

E. Eichler, J. Flint, G. Gibson, A. Kong, S. Leal, J. Moore, and J. Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Rev. Gen.*, 11:446–450, 2010.

T. P. Erlinger, P. Muntner, and K. J. Helzlsouer. WBC count and the risk of cancer mortality in a national sample of U.S. adults: results from the Second National Health and Nutrition Examination Survey mortality study. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 13 (6):1052–6, jun 2004. ISSN 1055-9965. URL `http://www.ncbi.nlm.nih.gov/pubmed/15184263`.

M. Ferreira and S. Purcell. A multivariate test of association. *Bioinformatics*, 25: 132–133, 2009.

B. L. Fridley and J. M. Biernacka. Gene set analysis of snp data: benefits, challenges and future directions. *European Journal of Human Genetics*, 19(8):837–843, 2011.

L. G. Fritsche, S. B. Gruber, Z. Wu, E. M. Schmidt, M. Zawistowski, S. E. Moser, V. M. Blanc, C. M. Brummett, S. Kheterpal, G. R. Abecasis, and B. Mukherjee. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *The American Journal of Human Genetics*, 102(6):1048–1061, jun 2018. ISSN 00029297. doi: 10.1016/j.ajhg.2018.04.001. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929718301332`.

D. S. Froese and R. A. Gravel. Genetic disorders of vitamin B12 metabolism: eight complementation groups eight genes. *Expert Reviews in Molecular Medicine*, 12: e37, nov 2010. ISSN 1462-3994. doi: 10.1017/S1462399410001651. URL `http://www.journals.cambridge.org/abstract{_}S1462399410001651`.

E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015. doi: 10.1038/ng.3367.

G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13:135–145, 2012.

J. Gudmundsson, P. Sulem, D. Gudbjartsson, G. Masson, B. Agnarsson, K. Benediktsdottir, A. Sigurdsson, O. Magnusson, S. Gudjonsson, and D. Magnusdottir. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genetics*, 44:1326–1329, 2012.

J. Haines, M. Hauser, S. Schmidt, W. K. Scott, L. M. Olson, P. Gallins, K. Spencer, S. Kwan, M. Noureddine, J. Gilbert, N. Schnetz-Boutaud, A. Agarwal, E. Postel, and M. Pericak-Vance. Complement factor h variant increases the risk of age-related macular degeneration. *Science*, 308 (5720):419–421, 2005.

T. Hasegawa, T. Negishi, and M. Deguchi. WBC Count, Atherosclerosis and Coronary Risk Factors. *Journal of Atherosclerosis and Thrombosis*, 9(5):219–223, 2002. ISSN 1880-3873. doi: 10.5551/jat.9.219. URL `http://joi.jlc.jst.go.jp/JST.JSTAGE/jat/9.219?from=CrossRef`.

Z. He, B. Xu, S. Lee, and I. Ionita-Laza. Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in metabochip data. *The American Journal of Human Genetics*, 101(3): 340–352, 2017.

H. J. Heipieper and R. Y.-Y. Chiou. Adaptation of Escherichia coli to Ethanol on the Level of Membrane Fatty Acid Composition. *Applied and Environmental Microbiology*, 71(6):3388–3388, jun 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.6.3388.2005. URL `http://aem.asm.org/cgi/doi/10.1128/AEM.71.6.3388.2005`.

P. Holmans. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv Genet.*, 72:141–179, 2010.

J. Huang, A. Johnson, and C. ODonnell. Prime: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. *Bioinformatics*, 27:1201–1206, 2011.

I. Hughes. Pediatric endocrinology and inborn errors of metabolism. *Journal of Paediatrics and Child Health*, 45(11):668–669, 2009.

J. R. Huyghe, A. U. Jackson, M. P. Fogarty, M. L. Buchkivoch, A. Stancakova, H. M. Stringham, X. Sim, L. Yang, C. Fuchsberger, H. Cederberg, and et al. Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nature Genetics*, 45:197–201, 2013.

I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, and X. Lin. Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, 92(6):841–853, 2013.

I. Ionita-Laza, M. Capanu, S. De Rubeis, K. McCallum, and J. D. Buxbaum. Identification of rare causal variants in sequence-based studies: methods and applications to vps13b, a gene involved in cohen syndrome and autism. *PLoS Genet*, 10(12): e1004729, 2014.

P. Jia, L. Wang, H. Y. Meltzer, and Z. Zhao. Pathway-based analysis of gwas datasets: effective but caution required. *Int J Neuropsychopharmacol*, 14:567–572, 2011.

M. Kanai, M. Akiyama, A. Takahashi, N. Matoba, Y. Momozawa, M. Ikeda, N. Iwata, S. Ikegawa, M. Hirata, K. Matsuda, M. Kubo, Y. Okada, and Y. Kamatani. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics*, 50(3):390–400, 2018. ISSN 15461718. doi: 10.1038/s41588-018-0047-6.

M. F. Keller, A. P. Reiner, Y. Okada, F. J. A. van Rooij, A. D. Johnson, M.-H. Chen, A. V. Smith, A. P. Morris, T. Tanaka, L. Ferrucci, A. B. Zonderman, G. Lettre, T. Harris, M. Garcia, S. Bandinelli, R. Qayyum, L. R. Yanek, D. M. Becker, L. C. Becker, C. Kooperberg, B. Keating, J. Reis, H. Tang, E. Boerwinkle, Y. Kamatani, K. Matsuda, N. Kamatani, Y. Nakamura, M. Kubo, S. Liu, A. Dehghan, J. F. Felix, A. Hofman, A. G. Uitterlinden, C. M. van Duijn, O. H. Franco, D. L. Longo, A. B. Singleton, B. M. Psaty, M. K. Evans, L. A. Cupples, J. I. Rotter, C. J. O'Donnell, A. Takahashi, J. G. Wilson, S. K. Ganesh, and M. A. Nalls. Trans-ethnic meta-analysis of white blood cell phenotypes. *Human Molecular Genetics*, 23(25):6944–6960, dec 2014. ISSN 0964-6906. doi: 10.1093/hmg/ddu401. URL `https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddu401`.

K. S. Kendler, M. C. Neale, R. C. Kessler, A. C. Heath, and L. J. Eaves. Major depression and generalized anxiety disorder. same genes, (partly) different environments? *Arch. Gen. Psychiatry*, 49:716–722, 1992.

D. K. Kim, M. H. Cho, C. P. Hersh, D. A. Lomas, B. E. Miller, X. Kong, P. Bakke, A. Gulsvik, A. Agustí, E. Wouters, B. Celli, H. Coxson, J. Vestbo, W. MacNee, J. C. Yates, S. Rennard, A. Litonjua, W. Qiu, T. H. Beaty, J. D. Crapo, J. H. Riley, R. Tal-Singer, and E. K. Silverman. Genome-Wide Association Analysis of Blood Biomarkers in Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory and Critical Care Medicine*, 186(12):1238–1247, dec 2012. ISSN 1073-449X. doi: 10.1164/rccm.201206-1013OC. URL `http://www.atsjournals.org/doi/abs/10.1164/rccm.201206-1013OC`.

J. H. Kim, S. Lim, K. S. Park, H. C. Jang, and S. H. Choi. Total and differential WBC counts are related with coronary artery atherosclerosis and increase the risk for cardiovascular disease in Koreans. *PLOS ONE*, 12(7):e0180332, jul 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0180332. URL `https://dx.plos.org/10.1371/journal.pone.0180332`.

M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, and J. Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, mar 2014. ISSN 1061-4036. doi: 10.1038/ng.2892. URL http://www.nature.com/articles/ng.2892.

R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. SanGiovanni, S. M. Mane, S. T. Mayne, M. B. Bracken, F. L. Ferris, J. Ott, B. C, and H. J. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308 (5720):385–389, 2005.

A. Korte and A. Farlow. The advantages and limitations of trait analysis with gwas: a review. *Plant Methods*, 9:9–29, 2013.

G. Kryukov, A. Shpunt, J. Stamatoyannopoulos, and S. Sunyaev. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Nat. Acad. Sci.*, 106:3871–3876, 2009.

D. Kurotaki, J. Nakabayashi, A. Nishiyama, H. Sasaki, W. Kawase, N. Kaneko, K. Ochiai, K. Igarashi, K. Ozato, Y. Suzuki, and T. Tamura. Transcription Factor IRF8 Governs Enhancer Landscape Dynamics in Mononuclear Phagocyte Progenitors. *Cell Reports*, 22(10):2628–2641, mar 2018. ISSN 22111247. doi: 10.1016/j.celrep.2018.02.048. URL https://linkinghub.elsevier.com/retrieve/pii/S2211124718302298.

C. H. Kwon, H. J. Park, Y. R. Choi, A. Kim, H. W. Kim, J. H. Choi, C. S. Hwang, S. J. Lee, C. I. Choi, T. Y. Jeon, D. H. Kim, G. H. Kim, and D. Y. Park. PSMB8 and PBK as potential gastric cancer subtype-specific biomarkers associated with prognosis. *Oncotarget*, 7(16), apr 2016. ISSN 1949-2553. doi: 10.18632/oncotarget.7411. URL http://www.oncotarget.com/fulltext/7411.

T. LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acid Res.*, 37(13):4181–4193, 2009.

P. H. Lee, C. O'Dushlaine, B. Thomas, and S. M. Purcell. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics*, 28(13):1797–1799, 2012a. ISSN 1460-2059. doi: 10.1093/bioinformatics/bts191. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts191.

S. Lee, M. C. Wu, and X. Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012b.

S. Lee, T. M. Teslovich, M. Boehnke, and X. Lin. General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies. *The American Journal of Human Genetics*, 93(1):42–53, jul 2013. ISSN 00029297. doi: 10.1016/j.ajhg.2013.05.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929713002231.

S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95 (1):5–23, 2014.

S. Lee, S. Won, Y. J. Kim, Y. Kim, B.-J. Kim, and T. Park. Rare variant association test with multiple phenotypes. *Genetic Epidemiology*, 2016.

B. Li and S. M. Leal. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics*, 83(3):311–321, sep 2008. ISSN 00029297. doi: 10.1016/j.ajhg.2008.06.024. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929708004084`.

L. Li, D. J. Ruau, C. J. Patel, S. C. Weber, R. Chen, N. P. Tatonetti, et al. Disease risk factors identified through shared genetic architecture and electronic medical records. *Sci. Trans. Med.*, 6:234ra57, 2014. doi: 10.1126/scitranslmed.3007191.

A. Liberzon, C. Birger, H. Thorvaldsdottir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The molecular signatures database (msigdb) hallmark gene set collection. *Cell Syst*, 1(6):417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

E. Lips, M. Kooyman, C. de Leeuw, and D. Posthuma. JAG: A Computational Tool to Evaluate the Role of Gene-Sets in Complex Traits. *Genes*, 6(2):238–251, may 2015. ISSN 2073-4425. doi: 10.3390/genes6020238. URL `http://www.mdpi.com/2073-4425/6/2/238`.

J. Z. Liu, A. F. Mcrae, D. R. Nyholt, S. E. Medland, N. R. Wray, K. M. Brown, et al. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139–145, 2010.

A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518:197–206, 2015.

T. Lumley, J. Brody, G. Peloso, A. Morrison, and K. Rice. Fastskat: Sequence kernel association tests for very large sets of markers. *Genetic Epidemiology*, 42(6):516–527, 2018. doi: 10.1002/gepi.22136.

D. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. Pickrell, S. Montgomery, and C. 1000Genomes. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335:823–828, 2012.

A. Maity, P. Sullivan, and J. Tzeng. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, 36:686–695, 2012.

A. Majumdar, T. Haldar, S. Bhattacharya, and J. S. Witte. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLOS*

Genetics, 14(2):e1007139, feb 2018. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007139. URL https://dx.plos.org/10.1371/journal.pgen.1007139.

A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010a.

A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W. M. Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010b. ISSN 13674803. doi: 10.1093/bioinformatics/btq559.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, , and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.

J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nat. Rev. Genetics*, 11(7):499–511, 2010.

W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–2070, aug 2010. ISSN 1460-2059. doi: 10.1093/bioinformatics/btq330. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq330.

M. A. Mooney, J. T. Nigg, S. K. McWeeney, and B. Wilmot. Functional and genomic context in pathway analysis of GWAS data. *Trends in Genetics*, 30(9):390–400, sep 2014. ISSN 01689525. doi: 10.1016/j.tig.2014.07.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0168952514001164.

V. Moskvina, K. M. Schmidt, A. Vedernikov, M. J. Owen, N. Craddock, P. Holmans, et al. Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis. *European Journal of Human Genetics*, 20:890–896, 2012.

A. Mu oz, Y. Peng, J. S. Chmiel, J. Margolick, J. Oishi, J. M. Samet, J. Sunyer, and L. Kingsley. Longitudinal Relation between Smoking and White Blood Cells. *American Journal of Epidemiology*, 144(8):734–741, 2012. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a008997.

J. I. Nurnberger, D. Koller, J. Jung, H. J. Edenberg, T. Foroud, I. Guella, et al. Identification of pathways for bipolar disorder: a meta-analysis. *JAMA Psychiatry*, 71:657–664, 2014.

C. O'Dushlaine, E. Kenny, E. A. Heron, R. Segurado, M. Gill, D. W. Morris, and A. Corvin. The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics*, 25(20):2762–2763, oct 2009. ISSN 1460-2059. doi: 10.1093/bioinformatics/btp448. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp448.

T. Ohya, T. Marubashi, and H. Ito. Significance of Fecal Volatile Fatty Acids in Shedding of Escherichia coli O157 from Calves: Experimental Infection and Preliminary Use of a Probiotic Product. *Journal of Veterinary Medical Science*, 62(11):1151–1155, 2000. ISSN 09167250. doi: 10.1292/jvms.62.1151. URL http://joi.jlc.jst.go.jp/JST.JSTAGE/jvms/62.1151?from=CrossRef.

W. Pan, I.-Y. Kwak, and P. Wei. A powerful pathway-based adaptive test for genetic association with common or rare variants. *The American Journal of Human Genetics*, 97(1):86–98, 2015. doi: 10.1016/j.ajhg.2015.05.018.

O. A. Panagiotou, C. J. Willer, J. N. Hirschhorn, and J. P. Ioannidis. The Power of Meta-Analysis in Genome-Wide Association Studies. *Annual Review of Genomics and Human Genetics*, 14(1):441–465, aug 2013. ISSN 1527-8204. doi: 10.1146/annurev-genom-091212-153520. URL http://www.annualreviews.org/doi/10.1146/annurev-genom-091212-153520.

T. H. Pers. Gene-set analysis for interpreting genetic studies. *Hum. Mol. Genet*, 25: R133–R140, 2016.

C. Quick, C. Fuchsberger, D. Taliun, G. Abecasis, M. Boehnke, and H. M. Kang. emerald: rapid linkage disequilibrium estimation with massive datasets. *Bioinformatics*, 35(1):164–166, 2019. doi: 10.1093/bioinformatics/bty547.

D. Ray and M. Boehnke. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genetic Epidemiology*, 42(2):134–145, mar 2018. ISSN 10982272. doi: 10.1002/gepi.22105. URL http://doi.wiley.com/10.1002/gepi.22105.

D. Ray, J. S. Pankow, and S. Basu. Usat: A unified score-based association test for multiple phenotype-genotype analysis. *Genetic epidemiology*, 40(1):20–34, 2016.

J. S. Ried, A. Doring, K. Oexle, C. Meisinger, J. Winkelmann, N. Klopp, T. Meitinger, A. Peters, K. Suhre, H. Wichmann, and C. Gieger. PSEA: Phenotype set enrichment analysis–a new method for analysis of multiple phenotypes. *Genetic Epidemiology*, 36:244–252, 2012.

T. Salcedo, Y. Li, B. L. Kreider, H. Li, V. P. Patel, K. Leung, D. Parmelee, R. Gentz, B. Nardelli, S. Gentz, V. Pippalla, G. Garotta, and R. Thotakura. Molecular and Functional Characterization of Two Novel Human C-C Chemokines as Inhibitors of Two Distinct Classes of Myeloid Progenitors. *The Journal of Experimental Medicine*, 185(7):1163–1172, 2002. ISSN 0022-1007. doi: 10.1084/jem.185.7.1163.

S. Salomon, C. Guignant, P. Morel, G. Flahaut, C. Brault, C. Gourguechon, P. Fardellone, J.-P. Marolleau, B. Gubler, and V. Goëb. Th17 and CD24hiCD27+ regulatory B lymphocytes are biomarkers of response to biologics in rheumatoid arthritis. *Arthritis Research & Therapy*, 19(1):33, dec 2017. ISSN 1478-6362. doi: 10.1186/s13075-017-1244-x. URL `http://arthritis-research.biomedcentral.com/articles/10.1186/s13075-017-1244-x`.

S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, 15(11):1576–1583, 2005.

A. V. Segre, L. Groop, V. K. Mootha, M. J. Daly, and D. Altshuler. Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits. *PLoS Genetics*, 6(8):e1001058, aug 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001058. URL `https://dx.plos.org/10.1371/journal.pgen.1001058`.

D. Sichien, C. L. Scott, L. Martens, M. Vanderkerken, S. Van Gassen, M. Plantinga, T. Joeris, S. De Prijck, L. Vanhoutte, M. Vanheerswynghels, G. Van Isterdael, W. Toussaint, F. B. Madeira, K. Vergote, W. W. Agace, B. E. Clausen, H. Hammad, M. Dalod, Y. Saeys, B. N. Lambrecht, and M. Guilliams. IRF8 Transcription Factor Controls Survival and Function of Terminally Differentiated Conventional and Plasmacytoid Dendritic Cells, Respectively. *Immunity*, 2016. ISSN 10974180. doi: 10.1016/j.immuni.2016.08.013.

C. Sidore, F. Busonero, A. Maschio, E. Porcu, S. Naitza, M. Zoledziewska, A. Mulas, G. Pistis, M. Steri, F. Danjou, A. Kwong, V. D. Ortega del Vecchyo, C. W. K. Chiang, J. Bragg-Gresham, M. Pitzalis, R. Nagaraja, B. Tarrier, C. Brennan, S. Uzzau, C. Fuchsberger, R. Atzeni, F. Reinier, R. Berutti, J. Huang, N. J. Timpson, D. Toniolo, P. Gasparini, G. Malerba, G. Dedoussis, E. Zeggini, N. Soranzo, C. Jones, R. Lyons, A. Angius, H. M. Kang, J. Novembre, S. Sanna, D. Schlessinger, F. Cucca, and G. R. Abecasis. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature Genetics*, 47(11):1272–1281, nov 2015. ISSN 1061-4036. doi: 10.1038/ng.3368. URL `http://www.nature.com/articles/ng.3368`.

S. Sivakumaran, F. Agakov, E. Theodoratou, J. G. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. F. Wilson, and H. Campbell. Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5):607–618, 2011.

N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature reviews. Genetics*, 14(7):483, 2013.

A. Stankov, M. Civelek, N. Saleem, P. Soininen, A. Kangas, J. Cederberg, H.and Paananen, J. Pihlajamki, L. Bonnycastle, M. Morken, et al. Hyperglycemia

and a common variant of gckr are associated with the levels of eight amino acids in 9,369 finnish men. *Diabetes*, 61:1895–1902, 2012.

J. Sun, K. Oualkacha, V. Forgetta, H.-F. Zheng, J. B. Richards, A. Ciampi, and C. M. Greenwood. A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. *European Journal of Human Genetics*, 24(9):1344–1351, 2016.

R. Sun, S. Hui, G. D. Bader, X. Lin, and P. Kraft. Powerful gene set analysis in gwas with the generalized berk-jones statistic. *Plos Genetics*, 15(3):e1007530, 2019.

K. Tajiri and Y. Shimizu. Branched-chain amino acids in liver diseases. *World Journal of Gastroentology*, 19:7620–7629, 2013.

T. M. Teslovich, , D. S. Kim, X. Yin, A. Stancakova, A. U. Jackson, M. Weilscher, A. Naj, and et al. Identification of seven novel loci associated with amino acid levels using single-variant and gene-based tests in 8545 finnish men from the metsim study. *Human Molecular Genetics*, 27:1664–1674, 2018.

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. doi: 10.1038/nature15393.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447 (7145):661–678, 2007.

G. Thomas, K. B. Jacobs, M. Yeager, P. Kraft, S. Wacholder, N. Orr, K. Yu, N. Chatterjee, R. Welch, A. Hutchinson, et al. Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics*, 40(3):310–315, 2008.

Y. Tomer, L. M. Dolan, G. Kahaly, J. Divers, R. B. D'Agostino, G. Imperatore, D. Dabelea, S. Marcovina, M. H. Black, C. Pihoker, A. Hasham, S. S. Hammerstad, et al. Genome wide identification of new genes and pathways in patients with both autoimmune thyroiditis and type 1 diabetes. *Journal of Autoimmunity*, 2015. ISSN 10959157. doi: 10.1016/j.jaut.2015.03.006.

I. Tomlinson, E. Webb, L. Carvajal-Carmona, P. Broderick, Z. Kemp, S. Spain, S. Penegar, I. Chandler, M. Gorman, W. Wood, et al. A genome-wide association scan of tag snps identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics*, 39(8):984–988, 2007.

E. Urrutia, S. Lee, A. Maity, N. Zhao, J. Shen, Y. Li, and M. Wu. Rare variant testing across methods and thresholds using the multi-kernel sequence kernel association test (MK-SKAT). *Stat Interface*, 8(4):495–505, 2015.

L. Vacca, A. Scuteri, C. Mameli, M. Dei, P. Loi, G. R. Abecasis, A. Terracciano, S. S. Najjar, A. Sharov, M. Masala, G. Pilia, P. Costa, A. B. Zonderman, A. Cao, G. Albai, N. Olla, T. Nedorezov, S. Lai, M. Deiana, W.-M. Chen, G. Usala, M. Orrú, M. Lai, D. Schlessinger, and E. Lakatta. Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLoS Genetics*, 2(8):e132, 2006. ISSN 1553-7390. doi: 10.1371/journal.pgen.0020132.

P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101:5–22, 2017.

K. Wang, M. Li, and M. Bucan. Pathway-Based Approaches for Analysis of Genomewide Association Studies. *The American Journal of Human Genetics*, 81(6):1278–1283, dec 2007. ISSN 00029297. doi: 10.1086/522374. URL `https://linkinghub.elsevier.com/retrieve/pii/S0002929707637756`.

K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16): e164–e164, sep 2010. ISSN 0305-1048. doi: 10.1093/nar/gkq603. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq603`.

Y. Wang, A. Liu, J. Mills, M. Boehnke, A. Wilson, J. Bailey-Wilson, M. Xiong, C. Wu, and R. Fan. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genetic Epidemiology*, 39:259–275, 2015.

H. R. Warren, E. Evangelou, C. P. Cabrera, H. Gao, M. Ren, B. Mifsud, I. Ntalla, P. Surendran, C. Liu, J. P. Cook, A. T. Kraja, et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature Genetics*, 2017. ISSN 15461718. doi: 10.1038/ng.3768.

J. R. White, C. Imburgia, E. Dul, E. Appelbaum, K. O'Donnell, D. J. O'Shannessy, M. Brawner, J. Fornwald, J. Adamou, N. A. Elshourbagy, K. Kaiser, J. J. Foley, D. B. Schmidt, K. Johanson, C. Macphee, K. Moores, D. McNulty, G. F. Scott, R. P. Schleimer, and H. M. Sarau. Cloning and functional characterization of a novel human CC chemokine that binds to the CCR3 receptor and activates human eosinophils. *Journal of Leukocyte Biology*, 62(5):667–675, nov 1997. ISSN 07415400. doi: 10.1002/jlb.62.5.667. URL `http://doi.wiley.com/10.1002/jlb.62.5.667`.

N. R. Wray, S. H. Lee, D. Mehta, A. A. Vinkhuyzen, F. Dudbridge, and C. M. Middeldorp. Research review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol Pyschiatry*, 55(10):1068–1087, 2014.

P. Wrtz, V.-P. Mkinen, P. Soininen, A. Kangas, T. Tukiainen, J. Kettunen, M. Savolainen, T. Tammelin, J. Viikari, T. Rnnemaa, et al. Metabolic signatures of insulin resistance in 7,098 young adults. *Diabetes*, 61:1372–1380, 2012.

P. Wrtz, P. Soininen, A. Kangas, T. Rnnemaa, T. Lehtimki, M. Khnen, J. Viikari, O. Raitakari, and M. Ala-Korpela. Branched-chain and aromatic amino acids are predictors of insulin resistance in young adults. *Diabetes Care*, 36:648–655, 2013.

B. Wu and J. Pankow. Sequence kernel association test of multiple continuous phenotypes. *Genetic Epidemiology*, 40(2):91–100, 2016.

M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Americal Journal of Human Genetics*, 89:82–93, 2011.

M. C. Wu, A. Maity, S. Lee, E. M. Simmons, et al. Kernel machine snp-set testing under multiple candidate kernels. *Genetic Epidemiology*, 37(3):267–275, 2013.

W. Xie, A. R. Wood, V. Lyssenko, M. N. Weedon, J. W. Knowles, S. Alkayyali, T. L. Assimes, T. Qaurtermous, F. Abbasi, J. Paananen, and et al. Genetic variants associated with glycine metabolism and their role in insulin sensitivity and type 2 diabetes. *Diabetes*, 62:2141–2150, 2013.

A. Xue, Y. Wu, Z. Zhu, F. Zhang, K. E. Kemper, Z. Zheng, L. Yengo, L. R. Lloyd-Jones, J. Sidorenko, Y. Wu, eQTLGen Consortium, A. F. McRae, P. M. Visscher, J. Zeng, and J. Yang. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications*, 9, 2018.

Q. Yan, D. E. Weeks, J. C. Celedón, H. K. Tiwari, B. Li, X. Wang, W.-Y. Lin, X.-Y. Lou, G. Gao, W. Chen, et al. Associating multivariate quantitative phenotypes with genetic variants in family samples with a novel kernel machine regression method. *Genetics*, 201(4):1329–1339, 2015.

A. Yáñez, M. Y. Ng, N. Hassanzadeh-Kiabi, and H. S. Goodridge. IRF8 acts in lineage-committed rather than oligopotent progenitors to control neutrophil vs monocyte production. *Blood*, 125(9):1452–1459, 2015. ISSN 15280020. doi: 10.1182/blood-2014-09-600833.

C. Yang, C. Li, Q. Wang, D. Chung, and H. Zhao. Implications of pleiotropy: challenges and opportunities for mining big data in biomedicine. *Frontiers in Genetics*, 6:229, 2015. ISSN 1664-8021. doi: 10.3389/fgene.2015.00229. URL `https://www.frontiersin.org/article/10.3389/fgene.2015.00229`.

K. Yu, Q. Li, A. W. Bergen, R. M. Pfeiffer, P. S. Rosenberg, N. Caporaso, et al. Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology*, 33:700–709, 2009.

Y. Yu, L. Xia, S. Lee, X. Zhou, H. M. Stringham, et al. Subset-based analysis using gene-environment interactions for discovery of genetic associations across multiple studies or phenotypes. *BiorXiv*, 2018. doi: https://doi.org/10.1101/326777.

X. Zhan, N. Zhao, A. Plantinga, T. Thornton, K. Conneely, M. Epstein, and M. Wu. Powerful genetic association analysis for common or rare variants with high-dimensional structured traits. *Genetics*, 206(4):1779–1790, 2017.

W. Zhou, J. B. Nielsen, L. G. Fritsche, R. Dey, M. E. Gabrielsen, B. N. Wolford, J. LeFaive, P. VandeHaar, S. A. Gagliano, A. Gifford, L. A. Bastarache, W.-Q. Wei, J. C. Denny, M. Lin, K. Hveem, H. M. Kang, G. R. Abecasis, C. J. Willer, and S. Lee. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9):1335–1341, sep 2018. ISSN 1061-4036. doi: 10.1038/s41588-018-0184-y. URL `http://www.nature.com/articles/s41588-018-0184-y`.

X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11:407–409, 2014a.

X. Zhou and M. Stephens. Efficient Algorithms for Multivariate Linear Mixed Models in Genome-wide Association Studies. *Nat Genet*, 11(4):407–409, 2014b. ISSN 1546-1718. doi: 10.1038/ng.2310.

X. Zhou, P. Carbonetto, and M. Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, 9(2):e1003264, 2013.

X. Zhu, T. Feng, B. O. Tayo, J. Liang, J. H. Young, N. Franceschini, J. A. Smith, L. R. Yanek, Y. V. Sun, T. L. Edwards, W. Chen, M. Nalls, E. Fox, M. Sale, E. Bottinger, C. Rotimi, Y. Liu, B. McKnight, K. Liu, D. K. Arnett, A. Chakravati, R. S. Cooper, S. Redline, and D. Levy. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *American Journal of Human Genetics*, 96(1):21–36, 2015. ISSN 15376605. doi: 10.1016/j.ajhg.2014.11.011.

O. Zuk, E. Hechter, S. S.R., and E. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Nat. Acad. Sci.*, 109:1193–1198, 2012.