Generalization of kernel machine methods for association testing of multi-omics data

Amarise Little

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Michael Wu, Chair
Timothy Thornton
Ali Shojaie

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Generalization of kernel machine methods for association testing of multi-omics data

Amarise Little

Chair of the Supervisory Committee:
Michael Wu
Department of Biostatistics

Over the past couple of decades, genome-wide association studies (GWASs) have successfully identified thousands of loci associated with complex traits and diseases in humans. Despite the immense success of these statistical tools, post-GWAS, we are often left underwhelmed by findings that are difficult to interpret or fail to to lead to causal mechanisms and deeper understanding of trait etiology. Studies utilizing omics, including transcriptomics, proteomics, metabolomics, etc, are gaining popularity, and, used in conjunction with genomics, may aid in providing insight into complex trait etiology and disease pathogenesis. To fully harness the availability of multi-omics data types, we propose to jointly evaluate, at the gene or pathway level, the cumulative effect of all data types simultaneously. We perform these analyses using the kernel machine regression (KMR) testing framework. Within this context, we propose three projects. For project one, we extend an existing KMR testing method to accommodate joint association testing of two data types with a trait of interest in correlated samples. For project two, we generalize existing KMR testing methods to allow for joint association testing of as many data types as desired against a trait of interest in correlated samples. Finally in project three, we propose a pseudo-permutation approach to association testing of an omics data type with a trait in correlated samples for studies with small sample sizes. These statistical tools facilitate analysis of complex multi-omics studies that are applicable to a broad range of studies with correlated samples, including family-based studies with

extensive relatedness and studies in ancestrally diverse populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Completion of this dissertation is not my achievement alone – it was a community effort. To me, this dissertation represents the culmination of my lifetime's journey of learning, and several folks impacted that journey along the way. First, I acknowledge the individuals serving on my committee – Ali Shojaie, Alex Reiner, and Sara Lindström – as guiding and supportive forces throughout my entire time at the University of Washington. I've learned so much from each of you, and I am incredibly grateful. I also acknowledge Ni Zhao for regularly meeting with me for the the past year and for being incredibly generous with your time and knowledge. Ni's efforts greatly elevated my dissertation.

To Tim Thornton, my advisor throughout my entire time in graduate school, thank you for everything! It has been one of my life's greatest pleasures to be able to be advised by a Black faculty member during my PhD program. Being a Black woman in this field can feel isolating and odd sometimes, but I'm glad I had Tim in my corner this whole time. Tim also put me in front of so many great opportunities that expanded my skills and breadth outside of research alone, which has made my professional path much more clear.

I also acknowledge my chair, Mike Wu. It has been a such a pleasure getting to know Mike this past year. I appreciate Mike's approach to research and advising. He always know when to be kind and supportive and when to push me, all in the spirit of me reaching my full potential. Mike changed my graduate experience for the better while also changing my outlook on engaging in scientific research. Thank you for everything!

To all my friends throughout the years, thank you for your support and constant maintenance of our friendships while I've been busy these past several years. I would especially like to thank those friends I made at graduate school. The bond we created via studying

together has forged some incredible friendships, and I hope we stay connected for life.

To my family, thank you for your unwavering support, love, and belief in me throughout the years, especially my parents and big sister who coached me through some of my toughest moments. You all inspired and encouraged me all throughout my journey, and I wouldn't have the confidence to finish without your transformative love.

Finally, I acknowledge my dearly departed little sister, Jordan Little. As I submit this dissertation on the day that would have been her 25^{th} birthday, I can't help but remember the last occasion I saw her in person. Jordan made the trip from Florida to California to see me graduate from college. I distinctly remember no one being more proud of me (not even our parents!) than her. Jordan, I feel you with me today. During some of my hardest and most doubt-filled moments throughout the PhD program, I was only able to push through if I thought about how you would have been proud to see me finish this, too. It was a great honor to be your big sister, and I miss you everyday.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2140004. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

DEDICATION

I dedicate this dissertation to Jordan Little, my little sister, who greatly influenced me during her too-short time with us. I miss you, and your memory got me through.

Chapter 1
INTRODUCTION

1.1 Motivation

Genes, transcripts, proteins, metabolites, and other macro/micro molecules work in concert to perform complex cellular processes. Classical analysis of omics data focuses primarily on individual data types (including transcriptomic, genetic, epigenetic, etc.). Although these single data type studies have been vastly successful in identifying individual features (transcripts, SNPs, epigenetic marks) associated with traits and phenotypes [16, 45, 41], integration of multi-omics data sets can help in unraveling the underlying mechanisms at multiple omics levels and also improve power to identify associations [41, 43, 38]. Thus, large scale multi-omics studies can be instrumental in overcoming many biological, medical, and public health challenges.

Despite the immense potential of emerging large-scale multi-omics studies, such as the NHLBI's Trans-Omics for Precision Medicine (TOPMed) initiative, there are a dearth of statistical methods available for the analysis of multi-omics data, and it is often unknown which methods are optimal under different settings. We focus particularly on the problem of gene discovery, wherein we are interested in understanding whether individual genes (or pathways) are associated with a complex trait. To fully harness the availability of multi-omics data types, we propose to jointly evaluate, at the gene or pathway level, the cumulative effect of all data types simultaneously. We perform these analyses using the kernel machine regression (KMR) testing framework. Following the general strategy of KMR testing, we embed individual data types into separate kernel matrices to capture specific characteristics of data types. Within this context, we propose three projects. The proposed methods in this

dissertation result in a powerful suite of statistical tools for facilitating analysis of complex multi-omics studies that are applicable to a broad range of studies with correlated samples, including family-based studies with extensive relatedness and studies in ancestrally diverse populations.

1.2 Association testing with multi-omics data

Joint analysis of multi-omics data is not a new statistical problem. Yet, despite the considerable interest and notable progress in this area, the best way to integrate multiple data types continues to be challenging. First, we need to address the usual challenges associated with high-dimensional omics data, including large numbers of features, modest effect size, complex (nonlinear or interactive) effects, and stringent type I error control levels. Second, we need to accommodate the complexities of the study design, e.g. family-based designs or longitudinal designs. Third, we need to accommodate the characteristics of individual data types including structure intrinsic to the data (LD for SNPs, phylogeny for microbiome data, etc.). Finally, the biggest challenge lies in identifying the specific analytic objective and scientific question to be addressed. With regard to the scientific question, a wide range of scientific objectives could be of interest in multi-omic studies including studies focused on understanding relationships among data types, discovering disease subtypes, prediction modeling, and many others. Although any multi-omics analysis approach needs to accommodate the first three challenges, addressing the last challenge is the ultimate goal and requires a clearly defined objective.

In this dissertation, we focus on the problem of gene discovery and identification of genomic features associated with a complex trait while harnessing multiple data types. The methods developed are applicable to general studies, allowing for correlated samples. Operationally, we operate under the KMR framework. Under this framework, a group of omic features (e.g. SNPs in a gene, expression of genes in a pathway, etc.) are embedded within a kernel function. The kernel function measures pairwise similarity between individuals based on the set of omic features. One can then construct the kernel matrix, $\mathbf{K}$, of all pairwise

similarities between individuals in the study. Then, one can assess the global association between the group of features by exploiting the connection between KMR and mixed models to construct a score statistic. Intuitively, the approach compares similarity between subjects based on the specific omic features (as measured through the kernel function) to similarity between subjects based on the outcome, while adjusting for covariates and accounting for dependence of outcomes. An analytic p-value can be calculated to evaluate the global null hypothesis.

A key feature of kernel methods is that the kernel functions can be tailored to individual data types, thereby capturing important structure in the data and potentially complex relationships. Kernels tailored for gene expression, metabolites, common and rare genetic variants, microbiome composition, and other omics features have been developed ref [6, 32, 36]. Further, as a multi-feature approach it allows for aggregation of small to modest effects and further reduces the multiple testing burden. Given these features, KMR methods are a natural strategy for multi-omics analysis.

Previous methods using KMR approaches for multi-omics analysis have been proposed. One such method comes from Zhao et al. (2018) who integrate methylation and genotype data for genome-wide association analysis. This gene-based method interrogates association between a gene and a general trait of interest by building a score test statistic based on a composite kernel matrix and covariate adjusted trait values. The composite kernel matrix is built from two different kernel matrices: one for methylation data and one for genotype data. This choice allows for each data type to be appropriately incorporated into the model, in a way that takes into account the nature of the data type. A perturbation approach and a kernel PCA approach to p-value calculation are presented [59]. This method fails to take into account potential correlation amongst observations, and it is limited to interrogating association for only two data types.

The Omnibus-Fisher method is a KMR approach that interrogates association between a quantitative or binary trait of interest and methylation, expression, and genotype data. Limiting focus to a gene of interest, a score test statistic is built for each data type. These

test statistics embed methylation, expression, or genotype data via a linear kernel function, while also capturing the covariate adjusted traits. Finally, these three test statistics are used to generate three different p-values using asymptotic results, which are then combined using a modified Fisher's test [55]. This method also fails to account for potential correlation amongst outcomes, is limited to three data types, and is restricted to modelling the omics data types linearly.

Finally, Huang et al. (2014) interrogates associations between a binary trait of interest and genotypes and gene expression. Looking at a single gene, this method constructs a test statistic that incorporates covariate adjusted traits and a composite kernel matrix that gives equal weight to a genotype kernel matrix, an expression kernel matrix, and a kernel matrix capturing genotype/expression interaction. All three kernel matrices are built using linear kernel functions. P-values are calculated using asymptotic results. The authors also propose an omnibus test with perturbation to address the case in which not all three data types influence the trait of interest [20]. This method fails to account for potential correlation amongst outcomes, is limited to two data types, s restricted to modelling the omics data types linearly.

Despite the numerous capabilities of these existing strategies, they are limited in terms of the number of data types, applicability to different study designs, or reliance on particular data assumptions. Thus, we seek to overcome the limitations of existing methods by proposing hypothesis tests using the KMR testing framework that (1) flexibly relate each data type to the trait of interest; (2) are applicable to general study designs, allowing for correlated samples; and (3) allow for testing using an arbitrary number of data types.

1.3 Dissertation Overview

In Chapter 2, we present the first project, in which we propose an extension of existing KMR testing methods to accommodate joint association testing of two data types with a single phenotype in correlated samples. We extend the existing composite kernel machine regression model as in [59] to integrate two multi-omics data types while accommodating general

correlation structure amongst outcomes. Here, we focus on scientific questions that aim to interrogate the association between a functional grouping (such as a gene or a pathway) and a quantitative trait of interest. We utilize a KMR model to integrate and model the two multi-omics data types, as they may relate to the trait, and perform a global test of association. We demonstrate the advantage of using this approach over single data type association tests via simulation. Finally, we applied this method to a large, multi-ethnic TOPMed sample that includes related subjects, to investigate how predicted gene expression and rare genetic variation may be related to two platelet traits. Due to the KMR framework, our methods allow for the integration of high-dimensional omics data with small, nonlinear, and interactive effects, and accommodation of general study designs.

In Chapter 3, we present the second project, where we extend work presented in Chapter 2 to allow for joint association testing of as many data types as desired to elucidate the multi-faceted effect of several omics data types on a trait of interest. . Again, we extend existing KMR models to integrate an arbitrary number of multi-omics data types, while accommodating for general correlation structure amongst outcomes. We perform a global test of association of multi-omics and a quantitative trait of interest in a functional grouping (such as a gene or a pathway). We demonstrate the advantage of using this approach over single data type association tests via simulation. Finally, we applied this method to a multi-ethnic TOPMed sample that includes related subjects, to assess association between platelet count and common genetic variation, methylation, and gene expression.

We present the third project in Chapter 4 – a pseudo-permutation method for valid, yet powerful, association testing of a quantitative trait under the KMR framework for a study with small sample size. As sequencing technologies emerge and gain popularity, samples for new technologies tend to be small. For example, microbiome studies typically have small sample sizes. The methods presented in Chapters 2 and 3 are most appropriate for large-scale studies, as they rely on asymptotic results. Use of these methods in small samples may result in loss of statistical power to detect signals. Thus, in Chapter 4 we provide a pseudo-permutation approach to p-value calculation for the association between a quanti-

tative trait and omics data under the KMR framework, while accounting for dependence of traits. One may want to use the p-value calculation presented here in conjunction with methods presented in Chapters 2 and 3 for studies with small sample sizes.

In Chapter 5, we conclude with summary of the dissertation and ideas for future work.

Chapter 2

GENERAL KERNEL MACHINE METHODS FOR MULTI-OMICS INTEGRATION AND GENOME-WIDE ASSOCIATION TESTING OF TWO DATA TYPES WITH RELATED INDIVIDUALS

2.1 Introduction

Over the last decade, genome-wide association studies (GWAS) have been successful in uncovering links between genetic variation and a wide range of diseases and complex traits in humans [37, 49, 11, 40, 7]. However, DNA sequence alone does not entirely explain the variation in complex traits. GWAS are a single-faceted approach to understanding the inherently multi-faceted nature of diseases and phenotypes which result from concerted biological cascades involving genes, transcripts, proteins, metabolites, and other macro/micro molecules performing complex cellular processes. For example, alternative splicing may generate more than one million proteins from just 25,000 protein-coding genes [50]. Other microbiome and epigenetic modifiers (e.g. DNA methylation) can further influence the expression of genes and proteins, leading to incredible variation from human to human [5, 54]. Consequently, there has been considerable interest in the integrative analysis of GWAS in conjunction with other omics to aid in elucidating the relationships between genes and complex traits and gaining a fuller understanding of disease and trait etiology [47, 43, 44, 48].

Unfortunately, despite the considerable interest and progress, the best way to integrate multiple data types remains unclear due to a number of critical challenges. First, we must address the usual challenges associated with high-dimensional omics data, including large numbers of features, small to modest effect sizes, complex (nonlinear or interactive) effects, and stringent type I error control levels. Second, we need to accommodate the characteristics

of individual data types including structure intrinsic to the data (LD for SNPs, phylogeny for microbiome data, etc.). Finally, a central challenge is the need to accommodate the complexities of the study design, e.g. family-based designs or longitudinal designs. This last problem is particularly pervasive and important as genetic studies are commonly conducted with related individuals where relatedness is known or cryptic [3, 8, 24, 19]. Failure to account for relatedness is known to lead to seriously inflated false positive rates, yet few integrative approaches directly allow for relatedness.

In this paper, we focus on the problem of gene mapping and identification of genomic features associated with a complex trait while harnessing multiple data types and accommodating relatedness amongst subjects. Specifically, we propose to operate at the gene or pathway level, as this represents a common and natural unit of analysis for many different data types. Within this context, integrating multi-omics data to perform association analysis helps improve statistical power by aggregating signals across multiple data types [20, 59, 55]. Gene level analysis further allows aggregation of effects within a data type (e.g. across the multiple SNPs within the gene) and also reducing multiple testing burden. We then propose a new test that assesses the joint effects of genetic variants as well as other omics data that is tailored to the individual data types and that further accommodates relatedness.

Operationally, we propose to utilize the kernel machine (KM) regression framework and develop a gene or pathway level test for the cumulative effect of multiple SNPs (e.g. within a gene/pathway) as well as one or more other -omics features (e.g. CpGs in a gene or a single gene expression level) [27, 26]. In particular, we utilize the multi-omics kernel machine model of [59], but we further provide allowance for relatedness. Under this framework, the effects of the SNPs and the omics features on the outcomes are embedded within devices called kernels which are measures of similarity based on the respective data types. A key feature of kernel methods is that the kernels can be tailored to individual data types, thereby capturing important structure in the data and potentially complex relationships. Kernels tailored for gene expression, metabolites, common and rare genetic variants, microbiome composition, and other omics features have been developed. Then, one can assess the global association

between the group of features by exploiting the connection between kernel regression and mixed models to construct a score statistic. Intuitively, the approach compares similarity between subjects based on the specific omic features (as measured through the kernel) to similarity between subjects based on the outcome, while adjusting for covariates and further incorporating random effects for relatedness.

The major contribution of this work is the development of a powerful integrated test that accommodates general correlation. We find that integrating multiple data types can lead to improved power to implicate genes related to complex traits, particularly when both omics data types have effects on the outcome. We further find that appropriately accounting for relatedness and family structure is critical to protecting type I error. In addition, we apply our method to a large multi-ethnic Trans-Omics for Precision Medicine (TOPMed) initiative [46] sample to investigate genes associated with two common quantitative clinical measures: platelet count ($n = 60,409$) and mean platelet volume ($n = 23,373$). We find that the integrative approach to association testing identified signals at genes *TNFAIP*, *1TRIM58*, and *ITGA2B* for the MPV trait and at genes *BCC4*, *TTC31*, and *ITGA2B* for the PLT trait. These signals were not identified using single data types alone.

For the remainder of this chapter, we first introduce the kernel regression framework for multi-omics data integration while accommodating relatedness amongst study subjects. We then discuss three approaches for p-value calculation for our association test. In simulation, we demonstrate poor type I error control of the test if we assume independence of outcomes and the utility of these three approaches for hypothesis testing. Finally, we present the results of applying one of the hypothesis testing approaches to a TOPMed sample.

2.2 Methods

2.2.1 Notation

Suppose our study has n potentially related individuals with $q-1$ covariates measured and two omics data types measured on q_1 features for the first data type and q_2 features measured

on the second data type. Then, for the i^{th} individual, $\boldsymbol{z}_i^{(1)} = (z_{i,1}^{(1)}, \ldots, z_{i,q_1}^{(1)})$ is the vector of feature values for the first omics data type (e.g. reference allele counts of all q_1 SNPs in a gene), $\boldsymbol{z}_i^{(2)} = (z_{i,1}^{(2)}, \ldots, z_{i,q_2}^{(2)})$ is the vector of values for the second omics data type (e.g. methylation values of all q_2 CpG sites in a gene), $\boldsymbol{X}_i = (1, x_{i,1}, \ldots, x_{i,q-1})$ is the vector of covariate measures (e.g. sex, age, population structure PCs, etc.), and y_i is an outcome value.

Then, $\boldsymbol{Z}_1$ is a $n \times q_1$ matrix of all features of all individuals of the first data type, $\boldsymbol{Z}_2$ is a $n \times q_2$ matrix of the second omics data type, $\boldsymbol{X}$ is a $n \times q$ covariate matrix, and $\boldsymbol{y}$ is a $n-$length vector of outcome values. Finally, denote $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ as the $n \times (q_1 + q_2)$ matrix containing all omics features of interest.

2.2.2 Model

We relate the features to the outcome via the kernel machine regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + f(\boldsymbol{Z}) + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$. We design $\boldsymbol{\Sigma}$ to take into account genetic relatedness amongst individuals in the study. In particular, let $\boldsymbol{\Sigma} = \sigma_a^2\boldsymbol{\Phi} + \sigma_e^2\boldsymbol{I}$. Here, σ_a^2 is the variance component attributed to the genetic relatedness matrix (GRM) and σ_e^2 is the variance component attributed to residual error. Finally, $\boldsymbol{\Phi}$ is a $n \times n$ GRM estimated from SNP data of the study participants [56, 14].

Function f relates the both omics data types to the outcome. Under the KMR framework, in (2.1), $f(\boldsymbol{Z})$ is an unknown, centered, smooth function assumed to lie in a space spanned by positive definite kernel functions $k(\cdot, \cdot)$. The kernel function gives rise to a $n \times n$ kernel matrix $\boldsymbol{K}$ which has $(j,k)^{\text{th}}$ entry $k(\boldsymbol{z}_j, \boldsymbol{z}_k)$, a measurement of the similarity in features between individuals j and k.

We are interested in testing if there is a joint effect of $\boldsymbol{Z}$ on $\boldsymbol{y}$ via hypothesis testing. We

aim to test null hypothesis

$$H_0 : f(\boldsymbol{Z}) = 0$$

against an alternative hypothesis of

$$H_1 : f(\boldsymbol{Z}) \neq 0.$$

To carry out the hypothesis test we exploit the connection between KMR and linear mixed models [27, 10]. Consider the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{f}$ is a $n \times 1$ vector of random effects with distribution $N(\boldsymbol{0}, \tau\boldsymbol{K})$, for some non-negative constant τ. Under this formulation, the equivalent hypotheses for our hypothesis test are

$$H_0 : \tau = 0, \text{ and } H_1 : \tau > 0. \tag{2.3}$$

This formulation of the model gives rise to a variance component score test statistic:

$$Q := (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}),$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ are estimated under the null model. This test statistic, Q, reduces to the familiar SKAT test statistic for continuous outcomes when $\boldsymbol{\Sigma} = \sigma_e^2\boldsymbol{I}$, indicating no relatedness amongst subjects [27, 23, 51, 52, 59]. The famSKAT test statistic [10] is a specific case of this test statistic, when $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{W}\boldsymbol{G}'$, for genotype matrix $\boldsymbol{G}$ and genetic weight matrix $\boldsymbol{W}$.

Intuitively, the test statistic Q captures how similar a pair of individuals in the study are in outcome and omics data. We more easily see this if we rewrite Q as

$$Q = \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\right).$$

The component $\boldsymbol{K}_Y = \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}$ can be thought of as a $n \times n$ similarity matrix of covariance re-weighted residuals or a similarity matrix of outcomes after covariate adjustment. As previously defined, $\boldsymbol{K}$ is a $n \times n$ similarity matrix of omics data. Due to a trace property, Q is the sum of all of the entries of the element-wise product of $\boldsymbol{K}_Y$ and $\boldsymbol{K}$. Thus, if a pair of individuals is similar in both outcome and omics at the same time, they contribute to a larger test statistic, ultimately leading us to reject the null hypothesis when there is enough evidence of concordance.

Asymptotically, Q follows a mixture of chi-square distribution, with mixture probabilities based on the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1/2}$, where $\boldsymbol{P}_0 = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'$. This distribution can be approximated using exact methods like Davies' method [15] or via moment matching methods [28], allowing for analytic p-value calculation.

2.2.3 Composite kernel matrix for combined omics with fixed weights

Following [59] we propose to construct a composite kernel matrix $\boldsymbol{K}$ for omics data as follows:

$$\boldsymbol{K}(\omega) = \omega k_1(\boldsymbol{Z}_1, \boldsymbol{Z}_1) + (1-\omega)k_2(\boldsymbol{Z}_2, \boldsymbol{Z}_2) \equiv \omega\boldsymbol{K}_1 + (1-\omega)\boldsymbol{K}_2$$

where k_1 and k_2 are kernel functions specific to their respective data type and $\omega \in [0, 1]$ is a weighting parameter. Thus, in (2.2), we specify $\boldsymbol{f}$ to have mean zero and covariance matrix $\tau(\omega\boldsymbol{K}_1 + (1-\omega)\boldsymbol{K}_2)$. In the model, ω controls how much emphasis we place on one data type over the other. Ideally, we choose ω to reflect the true effect model, however, if ω is chosen incorrectly, our test will suffer a loss in power, but the test remains valid. If ω is determined or fixed without regard to the y_i's in the present study, then one can construct the score statistic Q and treat $\boldsymbol{K}(\omega)$ as a fixed kernel matrix, which allows for analytic p-value calculation for the hypothesis test, as previously described.

2.2.4 Composite kernel matrix for combined omics when weight is unknown

In practice, however, ω is an unknown quantity that depends on the data. Thus, we consider a few different strategies for p-value calculation in the subsequent sections. In either strategy, we first proceed by standardizing each data type's kernel matrix.

Various omics data types tend to be on different scales, thus it is possible that $\boldsymbol{K}_1$ or $\boldsymbol{K}_2$ may dominate the other in the construction of the composite kernel matrix. If ω is known or fixed, as in the previous section, then this weight should take into account the inherent discrepancy of scale and order of magnitudes present in the individual data type kernel matrices. However, when the weight, ω, is unknown we first proceed by standardizing kernel matrices, such that one data type does not dominate information conveyed in the composite kernel matrix.

Let η_j be the standard error of $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}_j\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$. Then the quadratics $\frac{\eta_2}{\eta_1+\eta_2}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}_1\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ and $\frac{\eta_1}{\eta_1+\eta_2}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}_2\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ have the same standard error. Thus, we proceed with composite kernel matrix construction by using $\frac{\eta_2}{\eta_1+\eta_2}\boldsymbol{K}_1$ in place of $\boldsymbol{K}_1$ and $\frac{\eta_1}{\eta_1+\eta_2}\boldsymbol{K}_2$ in place of $\boldsymbol{K}_2$. Now, the composite kernel matrix is

$$\boldsymbol{K}(\omega) \;\; = \omega\tfrac{\eta_2}{\eta_1+\eta_2}\boldsymbol{K}_1 + (1-\omega)\tfrac{\eta_1}{\eta_1+\eta_2}\boldsymbol{K}_2,$$

and the test statistic is

$$\begin{aligned} Q(\omega) \;\; &= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}(\omega)\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \omega(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\tfrac{\eta_2}{\eta_1+\eta_2}\boldsymbol{K}_1\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &\quad +(1-\omega)(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\tfrac{\eta_1}{\eta_1+\eta_2}\boldsymbol{K}_2\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &\equiv \omega Q_1 + (1-\omega)Q_2. \end{aligned} \tag{2.4}$$

Perturbation hypothesis testing

In this section, instead of fixing ω, we proceed by considering multiple different choices of ω, $\{\omega_1, \ldots, \omega_L\}$, giving rise to L candidate composite kernel matrices. For example, we may

want to consider a grid of ω ranging from 0 to 1: $\{0, 0.5, 1\}$. In this case, our candidate composite kernel matrices comprise of (1) only the kernel matrix of data type one, (2) both kernel matrices, equally weighted, and (3) only the kernel matrix of data type two, respectively. In considering multiple candidate kernel matrices, we have a better chance that at least one of the choices of ω_j corresponds to a statistical test with high power. For each of the L candidate composite kernel matrices, we obtain its corresponding test statistic, $Q(\omega_j)$ and p-value, p_j. Set $p_0 = \min_{1 \leq j \leq L} p_j$.

Directly using p_0 as the p-value of the hypothesis test leads to inflated type I error. To account for using the minimum p-value amongst candidate composite kernel matrices, we proceed by conducting an intermediate hypothesis test: we test the null hypothesis that p_0 is a typical minimum p-value amongst these L composite kernel matrices when the null hypothesis in (2.3) is true. We use p_0 as a test statistic for this intermediate hypothesis test, and we approximate the null distribution of the minimum p-value amongst multiple candidate kernel matrices using an adapted version of the perturbation approach [53]. The p-value obtained from the perturbation approach is the final p-value of our hypothesis test in (2.3).

In short, the perturbation approach proceeds by noticing that when the null hypothesis is true the test statistic $Q(\omega_j)$ consists of two normal distributed random variables sandwiching a symmetric matrix:

$$\begin{aligned} Q(\omega_j) &= \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)' \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \\ &= \widehat{\boldsymbol{\epsilon}}' \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\epsilon}} \\ &= \boldsymbol{\epsilon}' \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\epsilon} \\ &\equiv \boldsymbol{w}' \widehat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{w}. \end{aligned} \tag{2.5}$$

Denote $\boldsymbol{w} \equiv \widehat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{\epsilon}$. Asymptotically, the random variable $\boldsymbol{w}$ is normally distributed and its value does not change with varying ω_j, however, the matrix $\widehat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1/2}$ does change as the value of ω_j varies. Therefore, perturbation proceeds by sampling new

standard normal vectors, rotating them appropriately, and constructing new test statistics for each value of ω_j. For each "perturbed" test statistic, we compute its corresponding p-value. Then we select the minimum p-value of the L p-values. We repeat this process many times to get an empirical distribution of the lowest p-value under the null hypothesis. The final p-value of the hypothesis test is the proportion of perturbed p-values less than or equal to p_0, the p-value from the original dataset. Additional details of the perturbation approach are in Supplement section B.

Kernel PCA

The aforementioned perturbation procedure is potentially more powerful than assuming a fixed ω due to considering multiple options of a composite kernel matrix. However, it tends to be time consuming due to its Monte-Carlo generation of many correlated random normal variables. In this section, we consider another procedure that also considers multiple options of a composite kernel matrix but allows for analytic p-value calculation.

When we consider multiple test statistics generated from a grid of ω values, the resulting L test statistics are correlated with one another. The correlation between these test statistics complicates direct, analytic computation of a final p-value. Moreover, there may be correlation of the two data types, leading to a test statistic that takes the form of the sum of two correlated quantities. For example, if we test with data types methylation and gene expression, we expect these data types to be correlated, as methylation may directly affect what genes are transcribed and ultimately expressed. We address this correlation between summands, Q_1 and Q_2, of the test statistic in (2.4) and amongst candidate test statistics by using kernel principal component analysis (kPCA) and basis projection. We decorrelate the summands of the test statistic in order to analytically calculate a p-value for significance.

The model with a composite kernel matrix, $\boldsymbol{K}(\omega)$, is equivalent to

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} + f_1(\boldsymbol{Z}_1) + f_2(\boldsymbol{Z}_2), \tag{2.6}$$

where f_1 and f_2 are functions in the space generated by kernel functions k_1 and k_2, respectively. We use kernel PCA to linearize functions f_1 and f_2. First, we eigendecompose the data kernel matrices, such that $\boldsymbol{K}_1 = \frac{\eta_2}{\eta_1+\eta_2}\boldsymbol{U}_1\boldsymbol{\Lambda}_1\boldsymbol{U}_1'$ and $\boldsymbol{K}_2 = \frac{\eta_1}{\eta_1+\eta_2}\boldsymbol{U}_2\boldsymbol{\Lambda}_2\boldsymbol{U}_2'$. Let $\boldsymbol{W}_1 = \boldsymbol{U}_1\boldsymbol{\Lambda}_1^{1/2}$, and $\boldsymbol{W}_2$ is defined similarly.

Now we express the model in (2.6) linearly:

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}_1\boldsymbol{\beta}_1 + \boldsymbol{W}_2\boldsymbol{\beta}_2 \tag{2.7}$$

In this linear form, we can more easily manipulate functions f_1 and f_2. In particular, we project $\boldsymbol{W}_2$ onto the space orthogonal to $\boldsymbol{W}_1$ and construct the following model

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}_1\boldsymbol{\beta}_1^* + \boldsymbol{W}_2^*\boldsymbol{\beta}_2^*, \tag{2.8}$$

where $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are regression coefficients for this transformed model and

$$\boldsymbol{W}_2^* = \left(\boldsymbol{I} - \boldsymbol{W}_1^*\left(\boldsymbol{W}_1^{*\prime}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}_1^*\right)^{-1}\boldsymbol{W}_1^{*\prime}\widehat{\boldsymbol{\Sigma}}^{-1}\right)\boldsymbol{W}_2.$$

Here, $\boldsymbol{W}_1^* = \boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{W}_1$. Then, $\boldsymbol{W}_2^*$ is a vector of the residuals that result from linear regression of features of data type two on covariate adjusted features of data type one, while accounting for relatedness.

We re-construct kernel matrix $\boldsymbol{K}_2^* = \boldsymbol{W}_2^*\boldsymbol{W}_2^{*\prime}$. Since, $\boldsymbol{W}_2^*$ lies in a subspace orthogonal to the column space of $\boldsymbol{W}_1$, the hypothesis test is transformed. Now we assume that $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are random effects with mean $\boldsymbol{0}$ and variance $\tau\omega\boldsymbol{I}$ and $\tau(1-\omega)\boldsymbol{I}$, respectively. Thus, in model (2.8), the effect due to data type one is modeled to have variance $\tau\omega\boldsymbol{K}_1$ and the effect due to data type two is modeled to have variance $\tau(1-\omega)\boldsymbol{K}_2^*$. Thus, the null hypothesis of $\tau = 0$ holds true when there is no effect from the first data type, and there is no effect from the second data type adjusted for data type one.

Under this transformed null hypothesis, we use test statistic $Q^*(\omega)$:

$$\begin{aligned} Q^*(\omega) &= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}^*(\omega)\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \omega(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}_1\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (1-\omega)(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}_2^*\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \omega Q_1 + (1-\omega)Q_2^* \end{aligned}$$

Now, Q_1 is independent of Q_2^*, and each summand follows a mixture of chi-square distribution. Details on calculating the final p-value for the hypothesis test are in Supplement section C.

2.2.5 Simulation Studies

Simulations when data types are uncorrelated

First, we assessed performance via simulation using uncorrelated data types. We conduct simulations integrating simulated genotype and methylation data, and we conduct simulations integrating simulated genotype and univariate continuous data. We simulate data for sample sizes of $n = \{1000, 2500, 5000\}$.

Genotype data were simulated using cosi2 [42]. We simulated $2n$ haplotypes for a 1Mb region to mimic the linkage disequilibrium (LD) pattern, local recombination rate, and the coalescent population history of European, African, and East Asian populations. Specifically, we designated three populations to be similar to three 1000 Genomes populations: Yoruba in Ibadan, Nigeria (YRI) of size $14,474$; Utah residents with Northern and Western European ancestry (CEU) of size $338,000$; and Han Chinese in Beijing, China (CHB) and Japanese in Tokyo, Japan (JPT) of size $454,000$ [4]. In the coalescent simulator, demographic events mimic those of these three populations. The simulated sample consists of 30% YRI haplotypes, 40% CEU haplotypes, and 30% CHB/JPT haplotypes. Of the $2n$ haplotypes, 30% were designated as a genetically unrelated subset of the population. We randomly created diploids with the remaining 70% of the haplotypes to give rise to generation 1. We randomly paired all diploids of generation 1 - denoting one in the pair female and the other male. We

then randomly sampled haplotypes from each individual in the pair to give rise to 2 diploid offspring for each pair, again, denoting one offspring female and the other male. This is generation 2. We aggregated sibling pairs from generation 2 into groups of 5 sibling pairs. We paired the female from pair 1 with the male from pair 5, the female from pair 2 with the male from pair 1, the female from pair 3 with the male from pair 2, etc. Then we randomly sample haplotypes from each pair to create 2 diploid offspring, giving rise to generation 3. Ultimately, a sample of size n is comprised of

1. all $0.3n$ unrelated diploids,

2. a random subset of $0.1n$ diploids from generation 1,

3. a random subset of $0.2n$ diploids from generation 2, and

4. a random subset of $0.4n$ diploids from generation 3.

This mimics a study in which a modest proportion of subjects are closely related. These simulated samples feature 10%, 10% and 14% of the sample being close relatives (3rd degree or more closely related) for the samples of size 1000, 2500, and 5000, respectively.

Using variants from the 1Mb region with minor allele frequency of at least 0.1%, we estimated a GRM, $\mathbf{\Phi}$, using the GCTA method from the SNPRelate R package [61]. We defined a gene as a 5kb region, giving rise to 200 genes. We only considered variants within the gene with minor allele frequency of at least 1%. Genes contained between 7 and 36 variants.

Methylation data was randomly sampled as multivariate normal with mean 0 and covariance matrix $\mathbf{\Sigma}_M$ for each simulation. The covariance matrix $\mathbf{\Sigma}_M$ was estimated from publicly available methylation data for 21 CpG sites within gene *RB1* [2].

Finally, univariate continuous data were randomly sampled as univariate standard normal.

For simulations integrating genotype and methylation data, we performed $50,000$ simulations to assess type I error at various significance thresholds. Each simulated gene was used 50 times across the simulations, whereas a unique set of methylation values were produced for each simulation. We embedded genotype and methylation data into kernel matrices using a linear kernel function. Quantitative traits were generated as follows:

$$\boldsymbol{y} = \boldsymbol{X} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{\epsilon} \sim N\left(\boldsymbol{0}, \frac{3}{7}\boldsymbol{\Phi} + \boldsymbol{I}\right)$.

We evaluated type I error of 5 approaches: (1) the kernel PCA approach assuming independence of outcomes, (2) the perturbation approach, (3) the kernel PCA approach, (4) genotype kernel matrix only, and (5) methylation kernel matrix only. We use a grid of $\omega = \{0, 0.25, 0.5, 0.75, 1\}$ for approaches 1-3. Note, the methylation kernel matrix test and the genotype kernel matrix test corresponds to using a fixed weight of 0 and 1, respectively, and for these tests, we calculated p-values as described in section 2.2.3. We calculated type I error as the proportion of simulated data sets whose hypothesis test results in a p-value less than the specified significance threshold.

For simulations integrating genotype and univariate continuous data, simulations to assess type I error were performed similarly as previously described.

For simulations integrating genotype and methylation data, we performed $1,000$ simulations to evaluate power under various settings. For a gene, quantitative traits were generated as follows:

$$\boldsymbol{y} = \boldsymbol{X} + \beta_g \sum_{j \in J_1} \boldsymbol{G}_j + \beta_m \sum_{j \in J_2} \boldsymbol{M}_j + \boldsymbol{\epsilon}, \tag{2.9}$$

where $\boldsymbol{G}_j$ is a column vector of the genotypes for the j^{th} variant in the gene, and $\boldsymbol{M}_j$ is a column vector of methylation values for the j^{th} CpG site. Denote q_1 the number of columns of the genotype matrix. If the gene contains fewer than 10 variants, then J_1 is one randomly selected column index of the genotype matrix, and if the gene contains 10 or more variants,

then J_1 is a set of $\lfloor \frac{q_1}{10} \rfloor$ randomly selected column indices of the genotype matrix. Finally, $J_2 = \{10, 20\}$.

For simulations integrating genotype and univariate continuous data, we performed $1,000$ simulations to evaluate power under various settings. For a gene, quantitative traits were generated as follows:

$$\boldsymbol{y} = \boldsymbol{X} + \beta_g \sum_{j \in J_1} \boldsymbol{G}_j + \beta_c \boldsymbol{C} + \boldsymbol{\epsilon}, \tag{2.10}$$

where $\boldsymbol{C}$ is a vector of simulated continuous values.

We evaluated power of 4 approaches: (1) the perturbation approach, (2) the kernel PCA approach, (3) genotype kernel matrix only, and (4) methylation kernel matrix only. We calculated power as the proportion of simulated data sets whose hypothesis test results in a p-value less than 0.05.

Simulations when data types are correlated

Next, we assessed performance via simulation using correlated data types. Again, we conduct simulations integrating simulated genotype and methylation data, and we conduct simulations integrating simulated genotype and univariate continuous data. We simulate data for sample sizes of $n = \{1000, 2500, 5000\}$.

Genes were defined the same as in section 2.2.5.

For each gene, we simulated 21 CpG sites. For each of the 21 CpG sites, four randomly selected variants from the gene give rise the methylation value:

$$\boldsymbol{M}_j = 3.5 \sum_{b \in B} \boldsymbol{G}_b + \boldsymbol{\upsilon},$$

where B is a set of 4 randomly selected indices of the columns of the genotype matrix, and $\boldsymbol{\upsilon} \sim N(\boldsymbol{0}, \boldsymbol{I})$. When generated in this manner, methylation and genotype values are highly correlated.

For each gene, we simulated a univariate continuous variable:

$$\boldsymbol{C} = 3.5 \sum_{b \in B} \boldsymbol{G}_b + \boldsymbol{v},$$

where B is a set of 4 randomly selected indices of the columns of the genotype matrix. Again, this creates strong correlation between the univariate continuous variable and the gene.

We carried out the simulation procedure the same way as in the previous section.

2.2.6 Data Application

We applied this approach to conduct genome-wide gene-based association analysis of two platelet traits using data from a multi-ethnic population. We here utilized whole genome sequencing (WGS) from NHLBI's Trans-Omics for Precision Medicine (TOPMed) initiative and gene transcript expression predicted for whole blood [33].

We fit the same null models as in [25] on $60,409$ subjects for platelet count (PLT) and $23,373$ subjects for mean platelet volume (MPV). Gene-based tests were performed using one predicted gene expression per gene transcript and all rare variants within the gene boundaries. Variant sets consisted of variants with minor allele frequency less than 1%, and flat weights were applied to all variants. We applied a linear kernel function to variant data and predicted gene expression data and used the computationally faster kernel PCA approach to obtain p-values for each gene. We determined the genome-wide significance threshold was using Bonferroni correction.

We used three gene-based filters for variant inclusion as in [25]. Described briefly, coding filters 1, 2, and 3 decrease in inclusion stringency and incorporate high-confidence loss of function variants, missense variants, and protein altering or synonymous variants. Expression only and genotype only tests used fixed weights, ω, of 0 and 1, respectively.

2.3 Results

2.3.1 Simulation Results

Type I error results are presented in Table 2.1 for uncorrelated and correlated data types simulations when integrating simulated genotype and methylation data types, and type I error results are presented in Table 2.2 for uncorrelated and correlated data type simulations when integrating simulated genotype and univariate continuous data types. Across the board tests that assume independence of the outcomes is poorly calibrated – the empirical type I error tends to be several fold the nominal level. Type I error is more poorly controlled as sample size increases. This indicates that if we were to incorrectly assume independence of outcomes while performing the hypothesis test, we run the risk of most significant hits being false positives. The magnitude of the discrepancy between the empirical type I error and the nominal level is more drastic as the significance threshold decreases. For example, when genotypes and methylation are uncorrelated and the sample size is $5,000$, the empirical type I error is about 15 times, 111 times, and 465 times the nominal level of 0.05, 0.005, and 0.001, respectively. This is concerning as the significance threshold of genome-wide gene-based tests is typically on the order of magnitude of 10^{-5} or 10^{-6}. We expect even poorer type I error control at these smaller significance thresholds.

On the other hand, we observe appropriate type I error control at all three significance thresholds of the perturbation, kernel PCA, and single data type only tests. The kernel PCA test tends to be slightly more conservative than the perturbation test, both when data types are simulated uncorrelated and correlated. However, both tests remain valid at all significance thresholds we examined.

Correlated Data Types	n	α	Independent	Perturbation	Kernel PCA	Genotype Only	Methylation Only
No	1,000	0.05	0.288	0.048	0.046	0.049	0.049
		0.005	0.1280	0.0046	0.0043	0.0048	0.0049
		0.001	0.0776	0.0007	0.0010	0.0007	0.0008
	2,500	0.05	0.512	0.050	0.048	0.052	0.051
		0.005	0.3090	0.0050	0.0048	0.0051	0.0045
		0.001	0.2250	0.0010	0.0010	0.0009	0.0011
	5,000	0.05	0.737	0.048	0.046	0.048	0.049
		0.005	0.5570	0.0050	0.0043	0.0048	0.0048
		0.001	0.4650	0.0009	0.0013	0.0010	0.0009
Yes	1,000	0.05	0.287	0.049	0.045	0.048	0.049
		0.005	0.1300	0.0046	0.0043	0.0049	0.0043
		0.001	0.0799	0.0008	0.0010	0.0009	0.0009
	2,500	0.05	0.514	0.050	0.047	0.050	0.050
		0.005	0.3100	0.0042	0.0044	0.0044	0.0045
		0.001	0.2260	0.0007	0.0011	0.0006	0.0008
	5,000	0.05	0.735	0.049	0.046	0.049	0.049
		0.005	0.5560	0.0048	0.0044	0.0050	0.0052
		0.001	0.4630	0.0012	0.0010	0.0012	0.0013

Table 2.1: Empirical type 1 error results when genotype and methylation are simulated uncorrelated and correlated. The independent test is kernel PCA assuming independence of outcomes. Perturbation and kernel PCA are integrative approaches that jointly test for genotype and methylation effect. Genotype only and methylation only approaches solely use a genotype and methylation kernel matrix for testing, respectively. Empirical type I error is calculated as the proportion of $50,000$ simulated data sets whose hypothesis test results in a p-value less than the specified significance threshold, α.

Correlated Data Types	n	α	Independent	Perturbation	Kernel PCA	Genotype Only	Univariate Only
No	1,000	0.05	0.290	0.048	0.048	0.049	0.047
		0.005	0.1279	0.0046	0.0051	0.0044	0.0046
		0.001	0.0787	0.0010	0.0008	0.0008	0.0010
	2,500	0.05	0.511	0.050	0.047	0.051	0.047
		0.005	0.3051	0.0056	0.0048	0.0053	0.0051
		0.001	0.2224	0.0009	0.0008	0.0011	0.0009
	5,000	0.05	0.738	0.050	0.048	0.050	0.047
		0.005	0.5611	0.0054	0.0050	0.0049	0.0045
		0.001	0.4672	0.0013	0.0009	0.0012	0.0011
Yes	1,000	0.05	0.290	0.050	0.047	0.049	0.049
		0.005	0.1270	0.0045	0.0051	0.0048	0.0050
		0.001	0.0763	0.0008	0.0009	0.0007	0.0008
	2,500	0.05	0.517	0.050	0.048	0.049	0.047
		0.005	0.3132	0.0050	0.0049	0.0051	0.0044
		0.001	0.2289	0.0008	0.0008	0.0008	0.0009
	5,000	0.05	0.734	0.049	0.049	0.049	0.048
		0.005	0.5555	0.0046	0.0043	0.0045	0.0047
		0.001	0.4647	0.0010	0.0009	0.0008	0.0010

Table 2.2: Empirical type 1 error results when genotype and a univariate continuous data are simulated uncorrelated and correlated. The independent test is kernel PCA assuming independence of outcomes. Perturbation and kernel PCA are integrative approaches that jointly test for an effect from genotype and a univariate continuous variable. Genotype only and univariate only approaches solely use a genotype and a continuous variable kernel matrix for testing, respectively. Empirical type I error is calculated as the proportion of $50,000$ simulated data sets whose hypothesis test results in a p-value less than the specified significance threshold, α.

Power results of the perturbation, kernel PCA, and single data type tests are in Table 2.3

for uncorrelated and correlated data types simulations when integrating simulated genotype and methylation data types, and power results are in Table 2.4 for uncorrelated and correlated data type simulations when integrating simulated genotype and univariate continuous data types. When data types are uncorrelated and there is only one data type effect, the most powerful approach is to test with only that data type. The joint tests suffer a slight penalty in power due to incorporating the irrelevant information conveyed by the other data type. This loss in power is less pronounced when data types are correlated, especially for smaller sample sizes. On the other hand, when there is truly an effect due to both data types, we benefit in using a joint test over testing with one data type at a time.

In Table 2.3, we see that when data types are simulated uncorrelated and the true effect is due to genotype only, the type I error for the methylation only test is around the nominal level, as expected. Similarly, when data types are simulated uncorrelated and there is only a methylation effect, the type I error for the genotype only test is around the significance threshold of 0.05. We see a similar phenomenon in Table 2.4. However, when data types are correlated, we note inflated type I error of the genotype only test when there is only an effect due to the other data type. This is due to the high correlation between between the data types - the other data types are directly calculated as a weighted sum of genotype values, thus a genotype effect is induced. We see a similar phenomenon when there is only an effect due to methylation or the univariate variable.

Generally, power increases as sample size increases, and power increases as effect size increases when there is an effect due to both data types. The perturbation approach is slightly more powerful than kernel PCA both when data types are correlated and uncorrelated. However, the kernel PCA approach can be computationally faster than perturbation, and perturbation may be infeasible when sample sizes are large.

Correlated Data Types	n	β_g	β_m	Perturbation	Kernel PCA	Genotype Only	Methylation Only
No	1,000	0.5	0	0.557	0.517	0.591	0.051
		0	0.05	0.147	0.150	0.042	0.181
		0.25	0.025	0.333	0.339	0.332	0.105
		0.5	0.05	0.578	0.598	0.552	0.208
	2,500	0.5	0	0.554	0.530	0.579	0.049
		0	0.05	0.360	0.362	0.056	0.427
		0.25	0.025	0.414	0.404	0.410	0.123
		0.5	0.05	0.727	0.713	0.600	0.447
	5,000	0.5	0	0.699	0.659	0.718	0.056
		0	0.05	0.715	0.689	0.063	0.794
		0.25	0.025	0.594	0.582	0.544	0.252
		0.5	0.05	0.910	0.907	0.702	0.784
Yes	1,000	0.5	0	0.527	0.531	0.527	0.506
		0	0.05	0.914	0.874	0.881	0.921
		0.25	0.025	0.776	0.686	0.744	0.788
		0.5	0.05	0.969	0.926	0.953	0.975
	2,500	0.5	0	0.603	0.537	0.612	0.561
		0	0.05	0.950	0.926	0.915	0.955
		0.25	0.025	0.840	0.750	0.766	0.846
		0.5	0.05	0.985	0.972	0.955	0.987
	5,000	0.5	0	0.672	0.671	0.682	0.638
		0	0.05	0.984	0.983	0.960	0.987
		0.25	0.025	0.930	0.883	0.884	0.928
		0.5	0.05	0.997	0.996	0.992	0.995

Table 2.3: Empirical power results when genotype and methylation are simulated uncorrelated and correlated. Perturbation and kernel PCA are integrative approaches that jointly test for genotype and methylation effect. Genotype only and methylation only approaches solely use a genotype and methylation kernel matrix for testing, respectively. Empirical power is calculated as the proportion of $1,000$ simulated data sets whose hypothesis test results in a p-value less than 0.05.

Correlated Data Types	n	β_g	β_c	Perturbation	Kernel PCA	Genotype Only	Univariate Only
No	1,000	0.5	0	0.536	0.557	0.570	0.043
		0	0.05	0.210	0.178	0.051	0.273
		0.25	0.025	0.354	0.324	0.357	0.113
		0.5	0.05	0.647	0.631	0.595	0.278
	2,500	0.5	0	0.562	0.534	0.607	0.048
		0	0.05	0.511	0.468	0.060	0.601
		0.25	0.025	0.452	0.457	0.421	0.208
		0.5	0.05	0.783	0.782	0.597	0.600
	5,000	0.5	0	0.698	0.647	0.731	0.047
		0	0.05	0.845	0.854	0.054	0.895
		0.25	0.025	0.628	0.612	0.510	0.357
		0.5	0.05	0.958	0.952	0.710	0.899
Yes	1,000	0.5	0	0.534	0.522	0.551	0.382
		0	0.05	0.835	0.642	0.618	0.864
		0.25	0.025	0.618	0.531	0.550	0.582
		0.5	0.05	0.911	0.824	0.813	0.873
	2,500	0.5	0	0.588	0.530	0.588	0.365
		0	0.05	0.957	0.823	0.662	0.972
		0.25	0.025	0.722	0.622	0.565	0.681
		0.5	0.05	0.985	0.932	0.884	0.956
	5,000	0.5	0	0.699	0.663	0.711	0.443
		0	0.05	0.994	0.969	0.769	0.996
		0.25	0.025	0.869	0.783	0.727	0.838
		0.5	0.05	0.998	0.989	0.945	0.978

Table 2.4: Empirical power results when genotype and univariate are simulated uncorrelated and correlated. Perturbation and kernel PCA are integrative approaches that jointly test for an effect from genotype and a univariate continuous variable. Genotype only and univariate only approaches solely use a genotype and a univariate continuous variable kernel matrix for testing, respectively. Empirical power is calculated as the proportion of $1,000$ simulated data sets whose hypothesis test results in a p-value less than 0.05.

2.3.2 Data Application

For all analyses, we employed a Bonferroni corrected significance threshold for determining significance of association at a gene. Namely, we defined the threshold as 0.05 divided by the number of genes tested. For the PLT trait, we tested $11,610$, $11,606$, and $11,599$ genes using filters 1, 2, and 3, respectively. Therefore, we used significance thresholds of 4.31×10^{-6}, 4.31×10^{-6}, and 4.32×10^{-6}, respectively. For the MPV trait, we tested $11,446$, $11,505$, and $11,573$ genes using filters 1, 2, and 3, respectively. Therefore, we used significance thresholds of 4.37×10^{-6}, 4.35×10^{-6}, and 4.31×10^{-6}, respectively.

Joint association analysis using predicted transcript expression and rare variants identified several gene transcripts associated with MPV and PLT. Figure 2.1 presents significant hits. Across filters, significant signals overlapped. For MPV, using filter 1 we identified 23 signals, using filter 2 we identified 23 signals, and using filter 3 we identified 24 signals. Across all 3 filters, we identified 26 distinct signals. For PLT, using filter 1 we identified 37 signals, using filter 2 we identified 35 signals, and using filter 3 we identified 38 signals. Across all 3 filters, we identified 42 distinct signals.

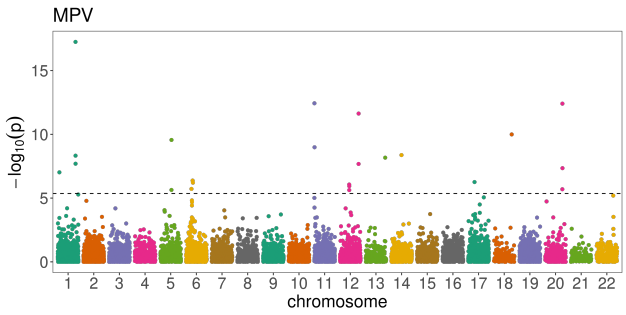
We identified several signals using the joint test that the single data type tests did not identify (Figure 2.2). Across all filters, the joint test for MPV identified three signals that tests with a single data type did not; these transcripts come from genes *TNFAIP*, *1TRIM58*, and *ITGA2B* (Table 2.5). The associations at *TNFAIP*, *1TRIM58*, and *ITGA2B* are driven by predicted gene expression, but we also observe a modest association with rare variation. The signals due to expression were not strong enough to attain genome-wide significance on their own, but when we integrated additional information about the rare variants in the gene, the aggregate signal of both data types is strong enough to attain genome-wide significance.

The joint test for PLT identified three signals that tests with a single data failed to identify. These signals fall in genes *ABCC4*, *TTC31*, and *ITGA2B* (Table 2.5). The association tests at *TTC31* and *ITGA2B* have nearly genome-wide significant p-values for predicted gene expression along with weak association with rare variation. The signal at

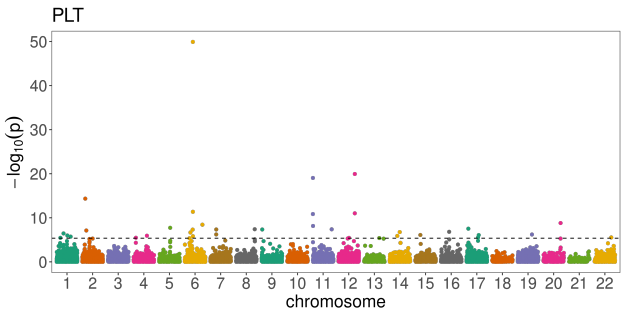
ABCC4 has a small p-value for both expression and rare variants, but neither test attains genome-wide significance. However, integrating these data types in the joint test results in enough combined signal to attain genome-wide significance.

Trait	Filter	Gene	Transcript	Lowest p-value ω	Joint	Expression Only	Genotype Only
MPV	1	*TNFAIP*	ENSG00000109079.9	0.5	4.37E-6	5.21E-6	3.69E-3
	3	*1TRIM58*	ENSG00000162722.8	0.5	2.97E-6	7.64E-6	1.61E-2
	3	*ITGA2B*	ENSG00000005961.17	0.5	3.27E-6	5.19E-5	1.50E-3
PLT	1	*ABCC4*	ENSG00000125257.13	0.5	4.00E-6	4.16E-4	1.01E-4
	3	*TTC31*	ENSG00000115282.19	0.5	1.75E-6	5.47E-6	1.94E-2
	3	*ITGA2B*	ENSG00000005961.17	0.5	1.35E-6	2.15E-5	1.30E-3

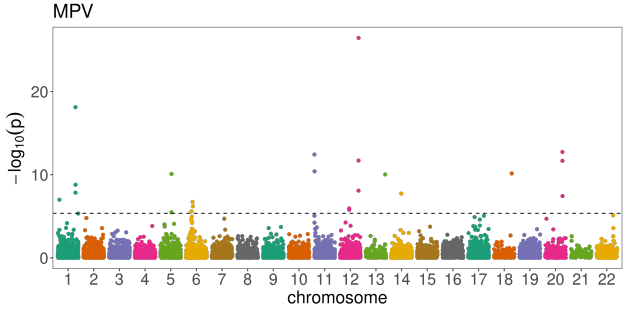
Table 2.5: Signals identified by the joint test that were not detected by individual data type tests



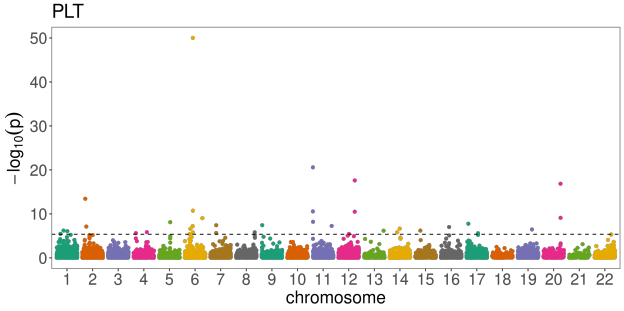
(a) Filter 1. Significance threshold of 4.37×10^{-6} based on $11,446$ genes tested.



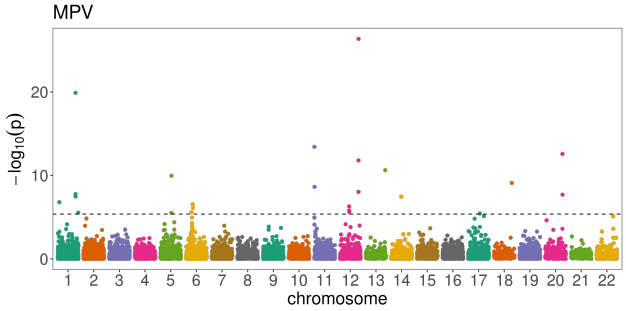
(b) Filter 1. Significance threshold of 4.31×10^{-6} based on $11,610$ genes tested.



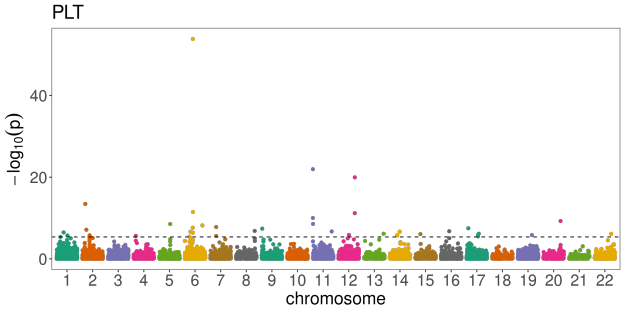
(c) Filter 2. Significance threshold of 4.35×10^{-6} based on $11,505$ genes tested.



(d) Filter 2. Significance threshold of 4.31×10^{-6} based on $11,606$ genes tested

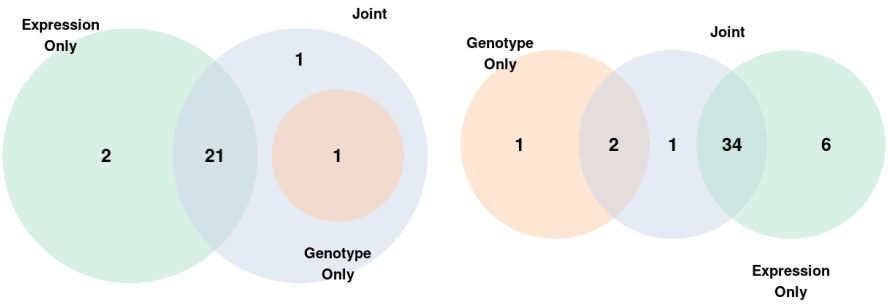


(e) Filter 3. Significance threshold of 4.31×10^{-6} based on $11,573$ genes tested



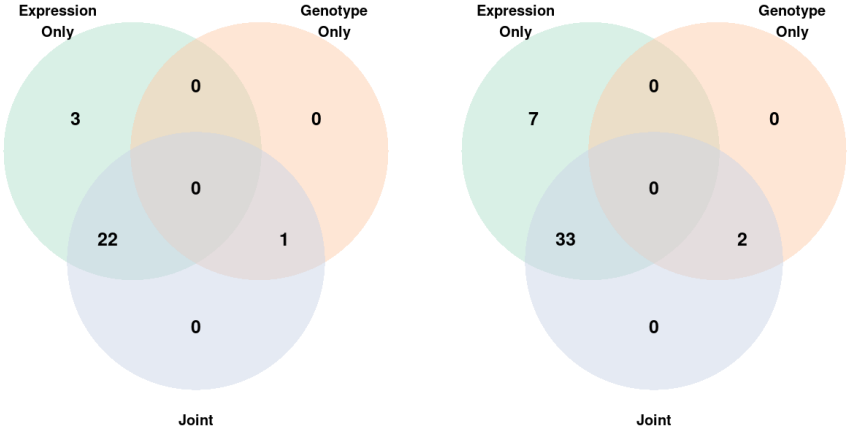
(f) Filter 3. Significance threshold of 4.32×10^{-6} based on $11,599$ genes tested

Figure 2.1: Manhattan plots from genome-wide gene-based association analysis integrating rare variant and predicted gene expression.



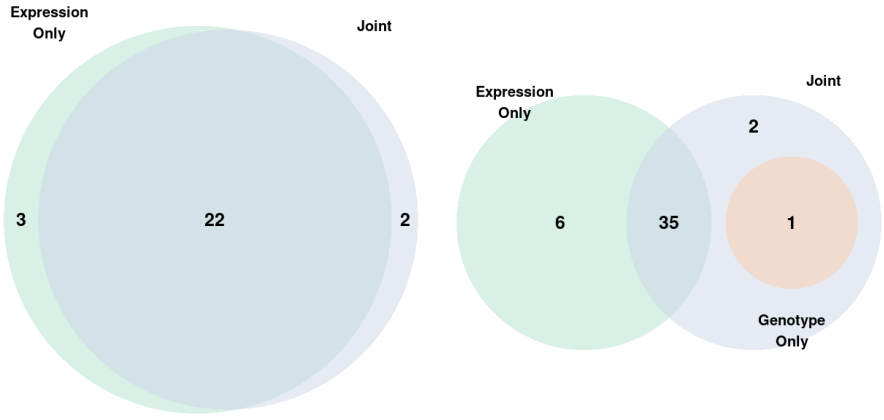
(a) Filter 1; MPV

(b) Filter 1; PLT



(c) Filter 2; MPV

(d) Filter 2; PLT



(e) Filter 3; MPV

(f) Filter 3; PLT

Figure 2.2: Venn diagram of number of genome-wide significant gene transcripts using three different tests: predicted gene expression only in green, rare variants (genotype) only in pink, and the joint test in blue.

2.4 Discussion

Our proposed approach integrates two omics data types while taking into account the architecture of the specific data types, including high dimensional data and omics that may interact with the trait in a non-linear fashion. Using this method, we may be able to identify additional factors that are associated with a quantitative trait of interest by combining small to modest effects to amplify signals that are relevant to the underlying trait etiology. We also get closer to bridging the gap in missing heritability by contributing a method that can be used on correlated measures and multi-ethnic samples. We demonstrated the utility of this approach on a large, multi-ethnic sample, including related individuals. Hopefully, identifying missing heritability contributes to disease prevention or treatment by learning about the function of identified variants or identification of the true causal facets implicated in disease pathology.

Depending on the scientific goal, it may not be appropriate to use this test in lieu of association analysis with a single data type, such as GWAS. This approach is well-suited for scientific questions that interrogate about a functional grouping of multi-omics, such as those features within or near a gene. One must take care that the features included in the test are consistent with the underlying biology. For example, we may not want to include both eQTLs along with expression, as these may be redundant pieces of information, artificially inflating the signal observed, which could potentially lead to false positives. On the other hand, including irrelevant information weakens the power of the hypothesis test, so it is best that the features included in the functional groupings are scientifically motivated. We see that this is the case, as several signals in the TOPMed study were identified via TWAS alone, but not via the joint test. Including rare variants that may not have been associated with platelet traits led to the joint test failing to attain significance in these particular cases. Therefore, one may want to conduct association analysis with single data types to answer slightly different, relevant scientific questions.

This approach is not limited to gene-based association analysis. One may be interested

in integrating multiple facets of a pathway, for example, to see if there exists association between a pathway and a trait of interest. We also are not limited to integrating data types that are implicated in genetic architecture. For example, one may want to integrate environmental effects alongside genetic effects. Embedding data types via a kernel function easily allows us to accommodate any type of high-dimensional exposures.

Another choice that may be guided by the underlying biology is the direction of projection when using kernel PCA. If one suspects that one data type has more of an effect than the other, then one would want to preserve the data type with the bigger suspected effect. While the KM regression approach offers flexibility in the choice of kernel function, as with any other association analysis, we are limited in that the underlying interaction between each data type and the trait is unknown. We benefit here compared to other methods in that we have the option of embedding data in a non-linear fashion. One could consider multiple different kernel functions to embed data in a kernel matrix, carry out the association analyses for all different kernel functions considered, and apply a multiple testing correction or use an omnibus test approach. Moreover, we may be computationally limited in the choice of kernel function when the sample size is very large. Specifically, if we use a linear kernel function we avoid computations that involve decomposition of $n \times n$ matrices. For this reason, since our sample sizes were relatively large for the computational task, we applied linear kernel functions to the TOPMed data, albeit this may not be the best choice of embedding these data. However, this approach is still a good choice to jointly assess association of two data types, while taking the correlation of observations into account.

Chapter 3

GENERAL KERNEL MACHINE METHODS FOR INTEGRATION OF TWO OR MORE OMICS AND GENOME-WIDE ASSOCIATION TESTING WITH RELATED INDIVIDUALS

3.1 Introduction

Classical analysis of omics data (including transcriptomic, genetic, epigenetic, etc.) focused primarily on individual data types. Although these single-data-type studies have been vastly successful in identifying individual features (transcripts, SNPs, epigenetic marks, etc.) associated with complex traits and phenotypes, each data-type represents inquiry and examination of just a single facet of the inherent multifaceted nature of disease. To achieve a full understanding of complex outcomes requires integration and multi-omics profiling approaches that can simultaneously yield a more holistic view of the biological systems as well as improve power to detect true effects since cellular and physiologic phenomena occur through a concerted biological cascade involving multiple omics. To these ends, large-scale multi-omic studies, such as the TOPMed project, are underway and promise comprehensive achievement of many biological, medical, and public health challenges.

Despite the immense potential of the emerging large-scale multi-omics studies, best approaches for statistical analysis of such studies remains unclear, particularly in the setting in which there is correlation in subjects due to relatedness or other clustering effects. We focus particularly on the problem of gene discovery wherein we are interested in understanding whether individual genes (or pathways) are associated with a complex trait. We focus on genes and pathways because data integration requires common units of analyses across data types and the gene represents a fundamental unit that can be characterized by many different

data types. For example, SNPs within and near the gene can simultaneously impact genetic regulation of expression as well as eventual protein sequence; epigenetic marks may increase or decrease expression; individual transcripts can represent the dynamic output of the gene; and protein represents the final, but often poorly measured, molecule for catalyzing chemical reactions. Similarly, genes (and the features associated with a gene) can be aggregated into pathways and functional groupings. Thus, to fully harness the availability of multiple omics data types, we propose to jointly evaluate, at the gene or pathway level, the cumulative effect of all the data types simultaneously while further accommodating correlation among the outcomes.

Operationally, we will utilize the kernel machine regression framework. Kernel machine testing (KMT), was originally proposed for gene expression analysis but has now become popular for genetic association analysis of common or rare variants. The approach has been extended to accommodate CNVs, methylation, microbiome, metabolomic, and scRNA data, among others. KM testing for a single data type proceeds by comparing pairwise similarity in the outcome to pairwise similarity in the features (e.g. SNPs in a gene). The data in the test appear purely through their embedding into pairwise similarities which is calculated via a kernel function. Importantly, the kernel function can be tailored to individual data types to capture key features of the data, e.g. epistatic relationships for SNP data or phylogenetic information for microbiome data. Thus, the framework can naturally and seamlessly accommodate the complex personalities of individual data types. For a given kernel, an analytic p-value can be calculated for the association between the individual data type and the outcome.

Under the KMT framework, we first identify groups of related features within and across data types, e.g. the SNPs, CpGs, transcripts associated with a single gene. We then propose to embed the features from each individual data type into separate data-type specific kernels. This allows for capture of data-type specific characteristics. For example, weighted linear kernels are popular for rare variants while other kernels have been tailored to other data types. Then, to assess the joint effect of all of the data types on the outcome, we

will construct a composite kernel as the weighted average/sum of the individual data type kernels. Since the weighted combination of kernels is just another kernel, for a fixed set of weights, we can analytically calculate a p-value for the association. To incorporate correlation among outcomes, we include additional random effects which capture effects that may arise through relatedness or other clustering effects. However, a central challenge lies in the choice of weights. The optimal weights depends on the true state of nature which is unknown. Therefore, we propose to use a coarse grid search. Specifically, we consider a range of different potential weights, construct a composite kernel based on each set of candidate weights, test for the association using the composite kernel while taking into account correlation, and then aggregating the p-values from the different choices of composite kernels. The major advantage of our approach is that we borrow information across data types in assessing the association. We emphasize that the combination is at the kernel level such that we are allowing differential contributions of the individual data types.

To aggregate the results across different sets of candidate weights and candidate composite kernels, we will combine the p-values using a novel truncated Cauchy-Combination test (TCCT), which is a variation on the commonly used Cauchy-Combination test (CCT). Specifically, we note that aggregation of some large p-values (e.g. from poor choice of weights) will dilute the power. Thus, we propose to truncate and remove large p-values when aggregating the individual p-values. In the same spirit of the CCT, the truncation of the TCCT does not affect the lower tail behavior of the p-values ensuring continued type I error protection while improving power in the presence of some large p-values.

Joint testing for the effect of multiple data types is not an entirely new problem. Our work is based on related models that have previously been used to integrate methylation and genotype data for genome-wide association analysis. As with out approach, this gene-based method interrogates the association between a gene and a general trait of interest by building a score test statistic based on a composite kernel matrix and covariate adjusted trait values. However, this method fails to take into account correlation amongst observations and is limited to interrogating association for only two data types [59]. While we have previously

developed a novel approach in the previous chapter that accommodates correlation among observations, the approach still is restricted to only two data types. Another approach, the Omnibus-Fisher, is somewhat reminiscent of our proposal in that it uses a modified Fisher's method to combine separate p-values of association testing of a quantitative trait and a set of SNPs, methylation markers, and RNA sequencing. The association testing for each data type is done via kernel machine regression, and afterwards, the three individual p-values are combined to generate a single p-value for a gene is produced. Despite the similarity to our approach in terms of combining p-values, a key conceptual difference is that it essentially relies on analyzing the individual data types separately and only combines the results at the end. This implicitly makes the unrealistic assumption that the contribution of each data type is the same and does not accommodate the correlation among data types. Further, the existing use of the approach does not directly account for relatedness amongst subjects in the study, though modifications can be made.

The remainder of this chapter is organized as follows. In the next section, we describe our model and proposed testing framework, including the construction of composite kernels and the truncated Cauchy Combination Test. We further describe the simulation scenarios and introduce the TOPMed data application. In the Result section, we then present the simulation results as well as the application of our approach to examine associations between genes and platelet counts in TOPMed. We conclude with a brief discussion.

3.2 Methods

For simplicity, we focus on testing a single gene (or pathway) with the understanding that the proposed approach can then be applied to testing all genes across the genome with appropriate multiple testing adjustments.

3.2.1 Kernel Model and Global Testing Framework

We assume a study in which there are n potentially related individuals. For each individual, we have a quantitative phenotype, y_i, for $i \in \{1, \ldots, n\}$ as well as L different data types

collected. For the i^{th} individual, we can then relate the data types to a quantitative outcome through the additive linear kernel model:

$$y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \sum_{\ell=1}^{L} f_\ell\left(\boldsymbol{Z}_i^{(\ell)}\right) + \epsilon_i \tag{3.1}$$

where $\boldsymbol{X}_i$ are covariates for which we would like to adjust and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ follows a distribution with mean zero and general covariance $\boldsymbol{\Sigma}$. $\boldsymbol{Z}_i^{(\ell)}$ represents the vector of measurements corresponding to the ℓ^{th} data type, e.g. the SNPs, methylation marks, transcripts, etc. for a particular gene. Each $f_\ell(\cdot)$ is a generally specified function lying in a reproducing kernel Hilbert space $\mathcal{H}_\ell$ defined by a corresponding positive definite kernel function $k_\ell(\cdot,\cdot)$.

The kernel function $k_\ell(\cdot,\cdot)$ is a function that measures similarity between two subjects based on their inputs, e.g. their SNP profiles. Importantly, $k_\ell(\cdot,\cdot)$ can be tailored to capture important characteristics of the data. For example, if using methylation data, one may want to employ kernel functions that incorporate both the methylation values and the CpG location [32], while a linear kernel function may suffice for SNP count data.

Due to the relationship between kernel machine models and mixed models [27, 10], we can jointly model all subjects together such that instead of (3.1) we have the equivalent linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{\ell=1}^{L} \boldsymbol{f}_\ell + \boldsymbol{\epsilon} \tag{3.2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$, and $\boldsymbol{f}_\ell$ is a vector of subject specific random effects with mean zero and variance $\tau_\ell \boldsymbol{K}_\ell$ with $\boldsymbol{K}_\ell$ an $n \times n$ kernel matrix with $(i, i')^{\text{th}}$ entry equal to $k_\ell(\boldsymbol{Z}_i^{(\ell)}, \boldsymbol{Z}_{i'}^{(\ell)})$.

Alternatively, we could also rewrite (3.2) as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\epsilon} \tag{3.3}$$

where we have $\boldsymbol{f} \sim F(\boldsymbol{0}, \tau\boldsymbol{K})$, and $\boldsymbol{K} = \sum_{\ell=1}^{L} \omega_\ell \boldsymbol{K}_\ell$ for some weights $\omega_\ell \in [0,1]$ with $\sum_{\ell=1}^{L} \omega_\ell = 1$.

Then under this formulation, in order to evaluate the joint effect of all data types on a quantitative trait, we can test the equivalent null hypotheses that

$$H_0 : \boldsymbol{f} = \mathbf{0} \Leftrightarrow H_0 : \tau = 0$$

against the general alternative that features of one or more data type are related to the outcome.

If the ω_ℓ are determined or fixed without regard to the outcomes in the present study, then we can construct the variance component score test statistic

$$Q = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ are estimated from the null model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Under the null, Q asymptotically follows a mixture of chi-square distribution. This distribution can be approximated using exact methods like Davies' method [15] or via moment matching methods [28], allowing for analytic p-value calculation. The mixture weights are determined by the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1/2}$, where $\boldsymbol{P}_0 = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{X}\left(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$.

In practice, however, the ω_ℓ are unknown quantities that depend on the data at hand. Thus, we consider a few different strategies for p-value calculation in the subsequent subsections.

3.2.2 Standardization of Kernel Matrices

Differing data types may be on vastly different scales. For example, we typically codify genotype data as counts of a reference allele (counts take on values of 0, 1, or 2), whereas gene expression values may take on values ranging several orders of magnitude, depending on the normalization technique employed [35, 12]. If ω_ℓ are known, then these quantities should take into account the inherent difference of scale present in the individual data type kernel matrices. However, when ω_ℓ are unknown *a priori*, we first proceed by standardizing

kernel matrices, so that no one data type dominates information conveyed in the composite kernel matrix.

We standardize an individual data type's kernel matrix, $\boldsymbol{K}_\ell$, by pre-multiplying $\boldsymbol{K}_\ell$ by a constant η_ℓ. Let σ_ℓ be the standard error of $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}_\ell\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, and let $\gamma_\ell = \prod_{\substack{i\in\{1,\ldots,L\}\\ i\neq\ell}} \sigma_i$. Then we choose

$$\eta_\ell = \gamma_\ell \left(\sum_{j=1}^{L}\gamma_j\right)^{-1}.$$

Now, all L quadratics, $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\eta_\ell\boldsymbol{K}_\ell\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, have the same standard error.

3.2.3 Multi-Dimensional Grid Search with Cauchy Combination Test

Let $\omega_\ell = \rho_\ell$, for $\ell < L$, and set $\omega_L = 1 - \sum_{\ell=1}^{L-1}\rho_\ell$. We restrict all ω_ℓ to be between 0 and 1. We consider a grid of values for each of the ρ_ℓ. We then exhaustively generate composite kernel matrices based on the values of the ρ_ℓs such that we have G different composite kernel matrices. For example, we might consider a grid of values for ρ_ℓ of $\{0, 0.5, 1\}$ for $L = 3$ individual data type kernel matrices. Then we would generate $G = 6$ different composite kernel matrices:

	ω_1	ω_2	ω_3	$\boldsymbol{K}$
1	1	0	0	$\eta_1 \boldsymbol{K}_1$
2	0	1	0	$\eta_2 \boldsymbol{K}_2$
3	0	0	1	$\eta_3 \boldsymbol{K}_3$
4	0.5	0.5	0	$\eta_1 \boldsymbol{K}_1 + 0.5\eta_2 \boldsymbol{K}_2$
5	0.5	0	0.5	$\eta_1 \boldsymbol{K}_1 + 0.5\eta_3 \boldsymbol{K}_3$
6	0	0.5	0.5	$\eta_2 \boldsymbol{K}_2 + 0.5\eta_3 \boldsymbol{K}_3$

Table 3.1: Composite kernel matrices generated by $\rho_1 = \rho_2 = \{0, 0.5, 1\}$ when $L = 3$.

For each composite kernel matrix, we then compute a p-value for the association as in section 3.2.1. Finally, we aggregate the multiple effects of all G tests using the Cauchy combination test or the truncated Cauchy combination test to get a single p-value for the association.

Cauchy Combination Test

Our G p-values for association between the quantitative outcome and L data types are correlated quantities, as the set of G composite kernel matrices are correlated to one another. The Cauchy combination test (CCT) takes into account the correlation structure amongst these p-values, as we combine their effects, which controls the type I error of the test [29].

The observed Cauchy combination test statistic is

$$t_{\text{CCT}} = \frac{1}{G} \sum_{j=1}^{G} \tan\{(0.5 - p_j)\pi\},$$

and the final p-value can be approximated as

$$p_{\text{CCT}} = \frac{1}{2} - \frac{\arctan\left(t_{\text{CCT}}\right)}{\pi}$$

Truncated Cauchy Combination Test

We have observed that the CCT can suffer severe power loss in the case where many p-values are close to 1. Thus, an alternative is to use a truncated Cauchy combination test (TCCT) where we are effectively winsorizing large p-values. Specifically, for each of the G p-values, if the p-value is greater than some number $\alpha + \nu$, where α is the type I error rate (e.g. 0.05) and $0 < \nu < 1 - \alpha$, then we set the p-value to be equal to $\alpha + \nu$. Specifically, the observed TCCT statistic is computed as

$$t_{\text{TCCT}} = \frac{1}{G} \sum_{j=1}^{G} I(p_j < \alpha + \nu) \tan\{(0.5 - p_j)\pi\} + I(p_j \geq \alpha + \nu) \tan\{(0.5 - (\alpha + \nu))\pi\}$$

One choice of ν is to just choose α. Finally, the p-value can be approximated as

$$p_{\text{TCCT}} = \frac{1}{2} - \frac{\arctan\left(t_{\text{TCCT}}\right)}{\pi}.$$

3.2.4 Simulation Studies

Simulations when data types are uncorrelated

First, we assessed performance integrating simulated genotype, methylation, and expression data for sample sizes $n = \{1000, 2500, 5000\}$. These three data types were simulated uncorrelated. We evaluated type I error of three approaches: (1) the CCT, (2) the TCCT, and (3) the Omnibus-Fisher test [55]. The Omnibus-Fisher test is a combination test that jointly tests association between three data types and a quantitative outcome, but it does not accommodate correlation of outcomes. Thus, we compare the CCT and TCCT to the Omnibus-Fisher test to demonstrate the utility of modeling correlation amongst outcomes, when present in the study. We evaluated power of the CCT and the TCCT to demonstrate increased power to detect signals when using the TCCT.

Genotype data were simulated using cosi2 [42]. We simulated $2n$ haplotypes for a 1Mb region to mimic the linkage disequilibrium (LD) pattern, local recombination rate, and the

coalescent population history of European, African, and East Asian populations. Specifically, we designated three populations to be similar to three 1000 Genomes populations: Yoruba in Ibadan, Nigeria (YRI) of size $14,474$; Utah residents with Northern and Western European ancestry (CEU) of size $338,000$; and Han Chinese in Beijing, China (CHB) and Japanese in Tokyo, Japan (JPT) of size $454,000$ [4]. In the coalescent simulator, demographic events mimic those of these three populations. The simulated sample consists of 30% YRI haplotypes, 40% CEU haplotypes, and 30% CHB/JPT haplotypes. Of the $2n$ haplotypes, 30% were designated as a genetically unrelated subset of the population. We randomly created diploids with the remaining 70% of the haplotypes to give rise to generation 1. We randomly paired all diploids of generation 1 - denoting one in the pair female and the other male. We then randomly sampled haplotypes from each individual in the pair to give rise to 2 diploid offspring for each pair, again, denoting one offspring female and the other male. This is generation 2. We aggregated sibling pairs from generation 2 into groups of 5 sibling pairs. We paired the female from pair 1 with the male from pair 5, the female from pair 2 with the male from pair 1, the female from pair 3 with the male from pair 2, etc. Then we randomly sample haplotypes from each pair to create 2 diploid offspring, giving rise to generation 3. Ultimately, a sample of size n is comprised of

1. all $0.3n$ unrelated diploids,

2. a random subset of $0.1n$ diploids from generation 1,

3. a random subset of $0.2n$ diploids from generation 2, and

4. a random subset of $0.4n$ diploids from generation 3.

This mimics a study in which a modest proportion of subjects are closely related. These simulated samples feature 10%, 10% and 14% of the sample being close relatives (3rd degree or more closely related) for the samples of size 1000, 2500, and 5000, respectively.

Using variants from the 1Mb region with minor allele frequency of at least 0.1%, we estimated a genetic relatedness matrix, $\mathbf{\Phi}$, using the GCTA method from the SNPRelate R package [61]. We defined a gene as a 5kb region, giving rise to 200 genes. We only considered variants within the gene with minor allele frequency of at least 1%. Genes contained between 5 and 39 variants.

Methylation data were randomly sampled as multivariate normal with mean 0 and covariance matrix $\mathbf{\Sigma}_M$ for each simulation. Expression data from 30 probes were randomly sampled as multivariate normal with mean zero and covariance $\mathbf{\Sigma}_E$ for each simulation. The covariance matrices $\mathbf{\Sigma}_M$ and $\mathbf{\Sigma}_E$ were estimated from publicly available methylation data for 21 and 30 CpG sites, respectively [2].

We performed $10,000$ simulations to assess type I error at various α levels. Quantitative traits were generated as follows:

$$\boldsymbol{y} = \boldsymbol{X} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{I})$ and $\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \frac{3}{7}\mathbf{\Phi} + \boldsymbol{I}\right)$. Each simulated gene was used 50 times across the simulations, whereas unique methylation and expression variables were produced for each simulation. For the CCT and TCCT we embedded genotype, methylation, and expression data types into kernel matrices using the linear kernel function, e.g. the kernel matrix, $\boldsymbol{K}_G$, for genotype data matrix $\boldsymbol{G}$ is $\boldsymbol{K}_G = \boldsymbol{G}\boldsymbol{G}'$. For the CCT and TCCT we used ρ_ℓ grid values of $\{0, 0.25, 0.5, 0.75, 1\}$. We calculated type I error as the proportion of data sets whose hypothesis test results in p-value less than the significance threshold.

We performed $1,000$ simulations to evaluate power under various settings. For a gene, quantitative traits were generated according to the following model:

$$\boldsymbol{y} = \boldsymbol{X} + \beta_G \sum_{j \in J_1} \boldsymbol{G}_j + \beta_M \sum_{j \in J_2} \boldsymbol{M}_j + \beta_E \sum_{j \in J_3} \boldsymbol{E}_j + \boldsymbol{\epsilon}, \tag{3.4}$$

where $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are defined as above, $\boldsymbol{G}_j$ is a column vector of the genotypes for the j^{th} variant in the gene, $\boldsymbol{M}_j$ is a column vector of methylation values for the j^{th} CpG site, and

$\boldsymbol{E}_j$ is a column vector for the j^{th} expression probe.

Denote q_1 the number of columns of the genotype matrix. If the gene contains fewer than 10 variants, then J_1 is one randomly selected column index of the genotype matrix, and if the gene contains 10 or more variants, then J_1 is a set of $\lfloor \frac{q_1}{10} \rfloor$ randomly selected column indices of the genotype matrix. $J_2 = \{10, 20\}$, and $J_3 = \{10, 20, 30\}$. We calculated power as the proportion of data sets whose hypothesis test results in p-value less than 0.05.

Simulations when data types are correlated

Next, we assessed performance when genotype, methylation, and expression data are simulated to be correlated. Genes were defined the same as in subsection 3.2.4. We induced correlation between genotypes and methylation:

$$\boldsymbol{M}_j = \boldsymbol{A}_j + \boldsymbol{\upsilon}_M.$$

Here, $\boldsymbol{M}_j$ is a column vector of methylation values for the j^{th} CpG site, $\boldsymbol{A}_j$ is the j^{th} column vector of a $n \times 21$ matrix. We define $\boldsymbol{A}_j = 0.4 \sum_{j \in J} \boldsymbol{G}_j$, where J is a set of two randomly selected indices of the columns of the $\boldsymbol{G}$ matrix. Moreover, $\boldsymbol{\upsilon}_M \sim N_{21}(\boldsymbol{0}, \boldsymbol{\Sigma}_M)$.

Next, we induce correlation between methylation and the continuous variable. Define

$$\boldsymbol{B} = 0.1 \left(\begin{array}{cc|c} \boldsymbol{1}_{[8\times 8]} & \boldsymbol{0}_{[8\times 13]} & \boldsymbol{0}_{[4\times 9]} \\ \boldsymbol{0}_{[13\times 8]} & \boldsymbol{1}_{[13\times 13]} & \boldsymbol{1}_{[11\times 9]} \\ & & \boldsymbol{0}_{[6\times 9]} \end{array} \right).$$

Let $\boldsymbol{E}_j = \boldsymbol{B}_j + \boldsymbol{\upsilon}_E$, where $\boldsymbol{E}_j$ is a column vector for the j^{th} expression probe, and $\boldsymbol{B}_j$ is the j^{th} column of $\boldsymbol{B}$. Moreover, $\boldsymbol{\upsilon}_E \sim N_{30}(\boldsymbol{0}, \boldsymbol{\Sigma}_E)$.

We carried out the simulation procedure the same way as in section 3.2.4.

3.2.5 Real Data Application

We analyzed data from NHLBI's Trans-Omics for Precision Medicine (TOPMed) initiative to conduct genome-wide gene-based association analysis of platelet count (PLT). We use data from $1,240$ women from the Women's Health Initiative (WHI) [18]. For each of the $18,589$ gene transcripts tested, we integrated whole genome sequencing (WGS), gene expression as measured by RNAseq, and methylation data, and tested their association with PLT.

We fit one common null model to be used for all of the genes' association tests. We modeled the null model as a linear mixed model, so that we could account for the correlation due to relatedness amongst individuals in the study. We adjusted for age and 11 principal components (PCs) to account for population structure, since our sample is multi-ancestral. We estimated PCs using LD-pruned genetic variants on chromosomes 1-22 of minor allele frequency of at least 0.01. We first use these variants to preliminarily estimate PCs using PC-AiR [13]. To ensure the PCs take into account admixture of individuals in the study, we use the preliminary PCs to estimate kinship coefficients using PC-Relate [14]. In turn, we use these kinship coefficients to estimate PCs a final time, again using PC-AiR. Using variants of minor allele frequency of at least 0.001, we estimated the genetic related matrix (GRM) using the GCTA method from the SNPRelate R package [61]. Finally, we used GENESIS to fit the null model using Average Information REML [17].

For each gene, we used common variants (minor allele frequency of at least 0.05) that are within the boundaries of the gene transcript, methylation of CpG sites within the gene, and gene expression of the gene transcript. We restrict attention to gene transcripts that have at least one variant and at least one CpG site within the gene boundaries. We embedded genotype, methylation, and expression data into kernel matrices using a linear kernel function. We employed ρ_ℓ grid values of $\{0, 0.25, 0.5, 0.75, 1\}$.

3.3 Results

3.3.1 Simulation Studies

Type I error results are presented in Table 3.2 for uncorrelated and correlated data types. Note that the Omnibus Fisher test has poor type I error control when applied to correlated outcomes. The test results in inflation that is increasingly poorly controlled with increasing sample size. This demonstrates that we run the risk of obtaining several false positive hypothesis test results when we fail to account for correlated outcomes in multi-omics association analysis. However, we observe appropriate type I error of the CCT, TCCT, and single data type tests at all three significance thresholds, as these tests take into account the correlation amongst outcomes. When comparing the CCT directly to the TCCT at the 0.05 significance threshold, we often observe slightly larger type I error of the TCCT. This difference is less exaggerated at smaller significance thresholds. QQ plots of these type I error results for the CCT and the TCCT are presented in Figure 3.1.

Power results are presented in Table 3.3 for independent and correlated data types. We benefit in using the TCCT over the CCT, as it is the slightly more powerful of the two tests. When there is only an effect due to one data type, the most powerful approach is to test with solely that data type. The joint tests suffer a slight penalty in power due to integrating irrelevant information. On the other hand, when there is truly an effect due to two or more data types, we tend to benefit from an increased power to detect signals using the TCCT. As expected when data types are correlated, individual data type tests pick up signal due to the induced effect via the other data types.

Correlated Data types	n	Significance Threshold	Omnibus Fisher	CCT	TCCT	$\boldsymbol{G}$	$\boldsymbol{M}$	$\boldsymbol{E}$
No	1000	0.05	0.256	0.042	0.052	0.046	0.048	0.050
		0.005	0.0956	0.0038	0.0043	0.0045	0.0042	0.0060
		0.001	0.0518	0.0007	0.0007	0.0006	0.0008	0.0011
	2500	0.05	0.454	0.046	0.054	0.049	0.048	0.049
		0.005	0.2486	0.0037	0.0042	0.0048	0.0047	0.0046
		0.001	0.1735	0.0005	0.0006	0.0012	0.0008	0.0010
	5000	0.05	0.681	0.050	0.051	0.050	0.048	0.047
		0.005	0.4919	0.0053	0.0053	0.0059	0.0051	0.0047
		0.001	0.3964	0.0017	0.0017	0.0010	0.0017	0.0013
Yes	1000	0.05	0.325	0.052	0.055	0.047	0.047	0.047
		0.005	0.1533	0.0050	0.0051	0.0051	0.0043	0.0050
		0.001	0.0989	0.0008	0.0008	0.0012	0.0005	0.0005
	2500	0.05	0.543	0.045	0.050	0.047	0.044	0.047
		0.005	0.3309	0.0041	0.0045	0.0046	0.0035	0.0045
		0.001	0.2494	0.0012	0.0013	0.0009	0.0006	0.0013
	5000	0.05	0.736	0.050	0.054	0.046	0.047	0.049
		0.005	0.5637	0.0060	0.0060	0.0055	0.0040	0.0056
		0.001	0.4663	0.0010	0.0010	0.0014	0.0011	0.0013

Table 3.2: Empirical type 1 error results of the Omnibus Fisher test, the Cauchy Combination Test (CCT), and the Truncated Cauchy Test (TCCT), which integrate three data types, and individual data type tests. $\boldsymbol{G}$ refers to the test using only genotype data, $\boldsymbol{M}$ refers to the test using only methylation data, and $\boldsymbol{E}$ refers to the test using only the continuous data. n is the number of samples in each simulation. The simulations for independent data types simulated all three data types independently of each other, while the simulations for correlated data types simulated all three data types correlated to one another. Empirical type I error is reported as the proportion of $10,000$ hypothesis tests from $10,000$ simulations that attained p-value less than the specified significance threshold.

				Independent Data Types					Correlated Data Types				
n	β_G	β_M	β_E	CCT	TCCT	$\boldsymbol{G}$	$\boldsymbol{M}$	$\boldsymbol{E}$	CCT	TCCT	$\boldsymbol{G}$	$\boldsymbol{M}$	$\boldsymbol{E}$
	0.5	0	0	0.438	0.503	0.529	0.036	0.049	0.509	0.517	0.548	0.308	0.169
	0	0.05	0	0.131	0.158	0.050	0.191	0.048	0.125	0.137	0.054	0.184	0.070
	0	0	0.05	0.072	0.078	0.047	0.052	0.084	0.206	0.219	0.087	0.141	0.234
1000	0.5	0.05	0	0.484	0.543	0.521	0.191	0.044	0.571	0.582	0.556	0.500	0.233
	0.5	0	0.05	0.463	0.534	0.556	0.045	0.108	0.655	0.665	0.594	0.470	0.430
	0	0.05	0.05	0.144	0.166	0.053	0.199	0.085	0.494	0.519	0.122	0.498	0.403
	0.5	0.05	0.05	0.532	0.586	0.548	0.192	0.091	0.801	0.808	0.616	0.725	0.648
	0.5	0	0	0.455	0.525	0.557	0.054	0.044	0.523	0.550	0.594	0.303	0.109
	0	0.05	0	0.259	0.308	0.050	0.427	0.049	0.342	0.371	0.073	0.463	0.129
	0	0	0.05	0.110	0.126	0.045	0.063	0.181	0.399	0.424	0.092	0.252	0.462
2500	0.5	0.05	0	0.641	0.714	0.581	0.458	0.054	0.728	0.739	0.585	0.698	0.271
	0.5	0	0.05	0.535	0.578	0.573	0.044	0.171	0.753	0.760	0.589	0.529	0.629
	0	0.05	0.05	0.358	0.406	0.044	0.444	0.175	0.854	0.900	0.160	0.874	0.810
	0.5	0.05	0.05	0.714	0.741	0.580	0.436	0.175	0.964	0.969	0.633	0.942	0.869
	0.5	0	0	0.643	0.645	0.668	0.627	0.506	0.632	0.656	0.688	0.326	0.132
	0	0.05	0	0.952	0.954	0.881	0.958	0.920	0.632	0.693	0.080	0.789	0.234
	0	0	0.05	1	1	0.969	0.997	1	0.680	0.721	0.099	0.466	0.787
5000	0.5	0.05	0	0.990	0.990	0.975	0.990	0.963	0.908	0.912	0.715	0.891	0.416
	0.5	0	0.05	1	1	0.985	0.999	0.999	0.920	0.922	0.723	0.716	0.854
	0	0.05	0.05	1	1	0.992	1	1	0.953	0.999	0.182	0.995	0.990
	0.5	0.05	0.05	1	1	0.999	1	1	0.994	0.999	0.752	1	0.993

Table 3.3: Empirical power results of the Cauchy Combination Test (CCT) and the Truncated Cauchy Test (TCCT), which integrate three data types, and individual data type tests. $\boldsymbol{G}$ refers to the test using only genotype data, $\boldsymbol{M}$ refers to the test using only methylation data, and $\boldsymbol{E}$ refers to the test using only the continuous data. n is the number of samples in each simulation. The simulations for independent data types simulated all three data types independently of each other, while the simulations for correlated data types simulated all three data types correlated to one another. Empirical power is calculated as the proportion of $1,000$ hypothesis tests from $1,000$ simulations that attained p-value less than 0.05.

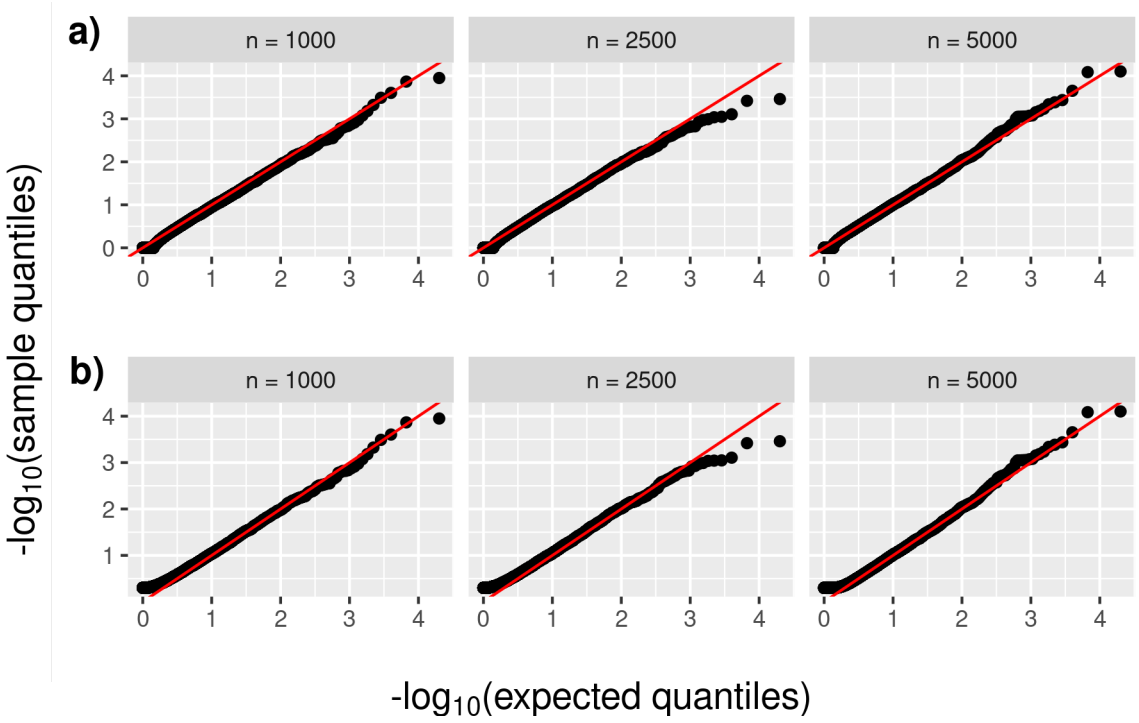


Figure 3.1: QQ plots under the null simulation setting. Panel a is the Cauchy Combination test results and panel b is the truncated Cauchy Combination test results.

3.3.2 Real Data Application

Joint association analysis via the TCCT using common genetic variation, methylation, and gene expression identified more than 100 gene transcripts associated with PLT on nearly all autosomal chromosomes, as we see in the Manhattan plot in Figure 3.2. We identified three signals using the joint test that the single data type tests did not identify (Figure 3.3). These transcripts come from long intergenic non-protein coding RNA *LINC00853* on chromosome 1, gene *ACCSL* on chromosome 11, and gene *TPST2* on chromosome 22 (Table 3.4). These three hits are primarily driven by the association with gene expression and modest association with methylation. None of these signals show evidence of association with common genetic variation within the gene. The signals due to expression were not strong enough to attain genome-wide significance on their own, but when we integrate additional information about the methylation within the gene, the aggregate signal of both data types is strong enough to

attain genome-wide significance for the joint test.

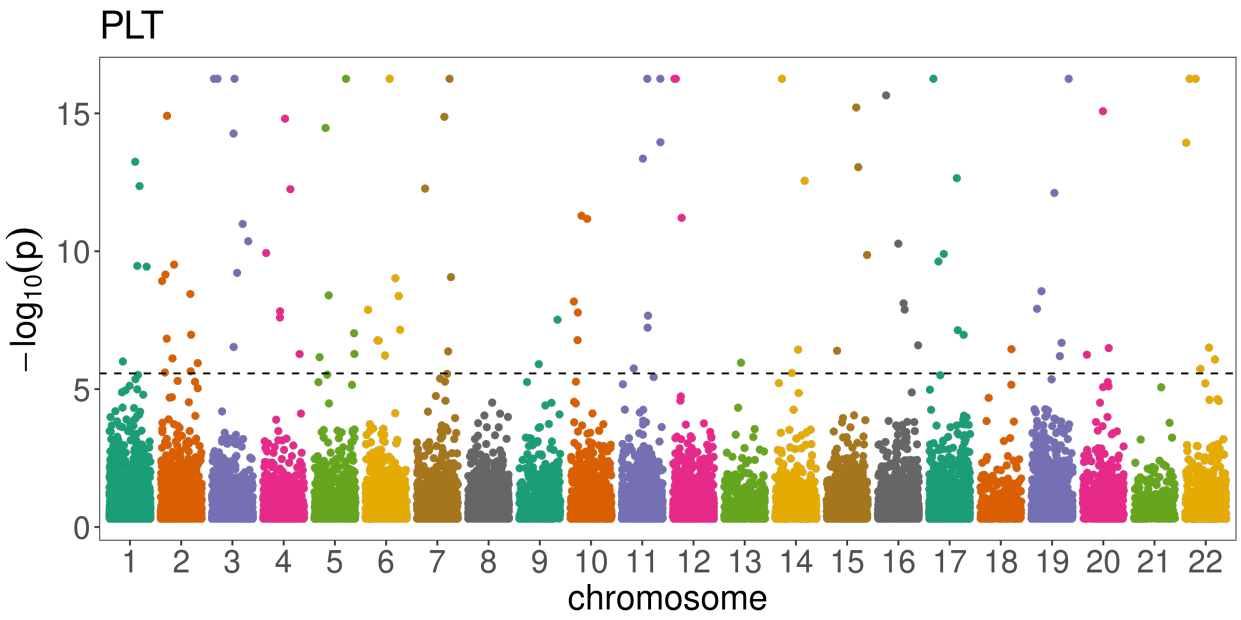


Figure 3.2: Manhattan plot of the joint test via TCCT. The height of the dashed black line is at the significance threshold. The Bonferroni corrected significance threshold is $0.05/18589 = 2.69 \times 10^{-6}$.

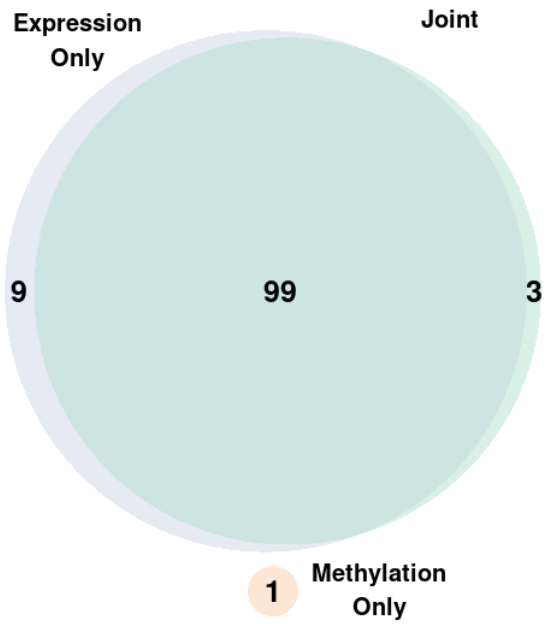


Figure 3.3: Venn diagram of number of significant hits of the TCCT (green), methylation only test (red), and gene expression only test (blue). The genotype only test is excluded because there were no significant hits using this test.

Gene	Transcript	$(\omega_1, \omega_2, \omega_3)$	TCCT	$\boldsymbol{G}$	$\boldsymbol{M}$	$\boldsymbol{E}$
LINC00853	ENSG00000224805.2	(0, 0.5, 0.5)	9.935E-07	3.307E-01	6.902E-04	4.874E-06
ACCSL	ENSG00000205126.2	(0, 0.5, 0.5)	1.779E-06	5.001E-01	4.252E-03	3.937E-06
TPST2	ENSG00000128294.16	(0, 0.5, 0.5)	1.860E-06	3.488E-01	6.531E-03	3.531E-06

Table 3.4: Signals identified by the TCCT that were not detected by individual data type tests. TCCT refers to the joint test that uses the truncated Cauchy combination test, $\boldsymbol{G}$ refers to the test using genotype data only, $\boldsymbol{M}$ refers to the test using methylation data only, and $\boldsymbol{E}$ refers to the test using gene expression data only. Weights, $(\omega_1, \omega_2, \omega_3)$, are the set of weights corresponding to the lowest p-value.

3.4 Discussion

Our proposed approach integrates two or more omics data types while taking into account the architecture of the specific data types, including high dimensional data and omics that may

interact with the trait in a non-linear fashion. We accommodate complex study designs, including relatedness or dependence amongst outcomes via linear mixed models (LMM). Making use of family-based studies and multi-ethnic studies is prudent, as in doing so, we may capture some "missing heritability" [31]. We demonstrated the utility of this approach on a multi-ethnic sample, including related individuals. Hopefully, identifying associations with health-implicating traits may aid in future disease prevention and/or treatment by learning about additional omics involved in the pathology.

We caution that this integrative approach may not be appropriate to use instead of association analysis with a single data type, such as GWAS. Here, we are assessing if there is joint association with at least one of the data types tested. Note, that if a signal is significant, this does not imply that all data types are indeed associated with the outcome. It is possible that a subset of data types are driving the association, while some data types are completely null. Moreover, if the joint test incorporates some null signals, we may dampen power to detect true associations. For example, we see in the TOPMed data application that there were 10 signals that were detected by gene expression or methylation alone, but were not captured by the joint test. Accordingly, we may want to extend this joint test to consider several different outcome-data type models and appropriately choose the model that has most evidence of association.

While this chapter mainly discussed how to apply this method in a gene-based association analysis setting, we are not limited to testing at the gene level. We may be interested in performing pathway-based analysis, for example, in which we regress a quantity related to a pathway on various facets of that pathway. Here, since we are not limited in the number of data types we want to use, we may also consider incorporating kernel matrices for the interaction of our data types, as in [60].

Chapter 4

PSEUDO-PERMUTATION FOR GENERAL KERNEL MACHINE ASSOCIATION TESTING OF SMALL SAMPLES

4.1 Introduction

Kernel-based association tests have been widely used in genetic and and other omic applications due to their modeling flexibility, computational efficiency, and ability to accommodate complex study designs while delivering statistical power. The sequence kernel association test (SKAT) has traditionally been used to assess association between common or rare genetic variants in a gene and a quantitative or binary outcome of interest, while accounting for covariates. Importantly, SKAT doesn't make assumptions about the directions of effect or the effect sizes of variants in the region. SKAT fits a null model, including covariates, which typically only needs to be fit once per outcome, resulting in a computationally efficient test procedure when applied genome-wide. This test employs a score-based variance-component hypothesis test to calculate a p-value for the regression of the outcome on a set of genetic variants [51, 52].

SKAT assumes independence amongst measured outcomes, however, genetic studies often feature related individuals. In this case, SKAT is inapplicable, as type I error tends to be when clustering amongst outcomes is not appropriately modeled. For quantitative traits, SKAT has been extended, as in the family-based SKAT (famSKAT), to accommodate genetic relatedness. Operationally, the famSKAT exploits the relationship between kernel machine regression and linear mixed models (LMM) to model the familial correlation of outcomes as a random effect with covariance proportional to the kinship matrix [10].

Due to the ability of SKAT to jointly test high-dimensional features with small to modest effects, the approach is an attractive choice for use with data of other omic modalities. For

example, the microbiome regression-based kernel association test (MiRKAT) extends SKAT to regress an outcome on a microbiome profile. MiRKAT exploits the flexibility of kernel-machine regression by non-parametrically relating the microbiome to the outcome of interest using a kernel that incorporates phylogenetic distance [58].

While these kernel-based association testing methods have been successful in their respective domains, they are not particularly statistically powerful when applied to small samples. SKAT, famSKAT, and MiRKAT all calculate p-values using the limiting distribution of their test statistics, and while these tests perform well for studies with large samples, we observe deflated type I error and a loss of power when these tests are used for studies with small samples. Motivated by the tendency of microbiome studies to be small and due to the tendency of genetic and epigenetic studies to feature longitudinal and/or family-based data collection, we seek a test that accounts for dependent outcomes and is powerful for small sample sizes. Moreover, we wish to benefit from the various attributes of kernel-based association tests: aggregation of small to modest effects of multiple features, like SKAT; accounting for correlation amongst outcomes, like famSKAT; and interrogation of association with various data modalities, as with MiRKAT.

We implement a pseudo-permutation approach to p-value calculation of kernel-based association tests, rather than relying on asymptotic results of existing approaches, so it the test is more powerful in small samples. Employing a pseudo-permutation test for small samples allows us to exploit the benefit of increased statistical power of a permutation test without the computational burden of a permutation test. Direct permutation may also be difficult if there are correlations in the features, as we would expect. Moreover, our approach accommodates dependence amongst outcomes, so it may be applied to genetic and epigenetic studies with relatedness and longitudinal study designs. In particular, we utilize a model similar to that of famSKAT, but we further allow for general covariance structure. Under this framework, the effects of the features on the outcomes are embedded within devices called kernels which are measures of similarity based on particular data types. We propose this test operate at the gene or pathway level, as this represents a common and natural unit of analysis

for many different data modalities. Then, one can assess the association between the features by using the pseudo-permutation distribution of the test statistic to calculate a p-value for the association. Intuitively, the approach compares similarity between subjects based on the specific data type's features (as measured through the kernel) to similarity between subjects based on the outcome, while adjusting for covariates and further accounting for effects due to dependence of the outcomes.

This approach addresses some challenges of association testing with high-dimensional omics data, including large numbers of features, small to modest effect sizes, complex (non-linear or interactive) effects, while also accommodating the characteristics of a data type including structure intrinsic to the data (LD for SNPs, phylogeny for microbiome data, etc.). It addresses stringent type I error control levels by employing a pseudo-permutation approach to p-value calculation. Finally, the flexibility of the model also may account for complexities of the study design.

The major contribution of this work is the development of a powerful test for small samples that accommodates general covariance structure. We further find that appropriately accounting for dependence of outcomes is critical to protecting type I error. In addition, we apply our method to a small multi-ethnic sample from the NHLBI's Trans-Omics for Precision Medicine initiative (TOPMed) to investigate genes associated with a common quantitative clinical measures: platelet count ($n = 151$). We also applied our method to a small vaginal microbiome study to investigate the association between the global microbiome and metabolites from a pathway ($n = 126$).

For the remainder of this chapter, we first introduce the kernel regression framework for dependent outcomes, and discuss the pseudo-permutation approach for p-value calculation for our association test. We then assess how the pseudo-permutation approach performs in simulation.

4.2 Methods

4.2.1 General Kernel Model Testing Framework

Assume a study in which there are n dependent quantitative outcomes. For example, our quantitative outcome may be a measured for n potentially related individuals, thereby inducing clustering amongst related subsets of the study. Or we may have repeated measures on some m individuals, amounting to n total measurements aggregated across all individuals. For each outcome, we have a set of features for an "-omic" data type collected, e.g. genetic data, methylation, gene expression, etc. Suppose we are interested in assessing association between. For the i^{th} outcome, we can then relate the "-omic" data type to a quantitative outcome through the additive linear kernel machine model:

$$y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + f(\boldsymbol{Z}_i) + \epsilon_i \tag{4.1}$$

where $\boldsymbol{X}_i$ are covariates for which we would like to adjust and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)'$ follows a distribution with mean zero and general covariance $\boldsymbol{\Sigma}$. $\boldsymbol{Z}_i$ represents the vector of measurements corresponding to the omics data type. The function $f(\cdot)$ is a generally specified function lying in a reproducing kernel Hilbert space $\mathcal{H}$ defined by a corresponding positive definite kernel function $k(\cdot, \cdot)$.

The kernel function $k_\ell(\cdot, \cdot)$ is a function that measures similarity between two subjects based on their inputs, e.g. their SNP profiles. Importantly, $k_\ell(\cdot, \cdot)$ can be tailored to capture important characteristics of the data. For example, if using methylation data, one may want to employ kernel functions that incorporate both the methylation values and the CpG location [32], while a linear kernel function may suffice for SNP count data.

Due to the relationship between kernel machine models and mixed models [27, 10], we can jointly model all subjects together such that instead of (4.3) we have the equivalent linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\epsilon} \tag{4.2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$, and $\boldsymbol{f}$ is a vector of subject specific random effects with mean zero and variance $\tau \boldsymbol{K}$ with $\boldsymbol{K}$ an $n \times n$ kernel matrix with $(i, i')^{\text{th}}$ entry equal to $k(\boldsymbol{Z}_i, \boldsymbol{Z}_{i'})$.

Leveraging the linear mixed model representation of the model, we can derive a variance component score test statistic to test the equivalent null hypotheses that

$$H_0 : \boldsymbol{f} = \boldsymbol{0} \Leftrightarrow H_0 : \tau = 0$$

against the general alternative that features of the omics data are related to the outcome. Following [10], the variance component score test statistic is

$$Q = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ are estimated from the null model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Under the null, Q asymptotically follows a mixture of chi-square distribution, allowing for analytic p-value calculation. The mixture weights are determined by the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1/2}$, where $\boldsymbol{P}_0 = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{X}\left(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$.

However, for small sample sizes, it may be inappropriate to rely on asymptotic results. Direct permutation of the data may be difficult, due to the correlation structure of the outcomes, especially since we are allowing for general covariance structure of outcomes. Moreover, we expect additional correlation structure to be present in the omics data, so any permutation procedure must account for correlation in the omics data as well. Even for small samples, direct permutation of the data may be computationally intensive. Thus, to overcome these challenges, we propose a pseudo-permutation computation of p-values for the variance component score test.

4.2.2 Pseudo-Permutation P-value Calculation

If we perform a permutation test when $\boldsymbol{y}$ doesn't have an exchangeable covariance structure (as is the case when outcomes are dependent), the p-values of the test are smaller than those of the actual null distribution, and we observe inflated type I error of the test [1]. The strategy is to overcome this issue of non-exchangeability of the correlation structure by permuting transformed residuals that do have an exchangeable covariance structure.

Under the null hypothesis, our residuals $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ have covariance equal to $\boldsymbol{P}_0$. Components $\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n$ don't have the same distribution, since $\boldsymbol{\Sigma}$ is allowed to be completely unstructured, thus, we do not have exchangability of the covariance structure. We aim to transform residuals so that transformed residuals have exchangeable covariance structure.

Decompose $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\Sigma}}^{1/2}$, and transform the null model by $\widehat{\boldsymbol{\Sigma}}^{-1/2}$:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{y} &= \widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{X}\boldsymbol{\beta} + \widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\epsilon} \implies \\ \boldsymbol{z} &= \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{e} \end{aligned}$$

The transformed residuals $\widehat{\boldsymbol{e}} = \boldsymbol{z} - \boldsymbol{W}\widehat{\boldsymbol{\beta}}$ have covariance

$$\boldsymbol{V} \equiv \boldsymbol{I} - \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'.$$

$\boldsymbol{V}$ is a projection matrix, so it can be decomposed as

$$\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}',$$

where the columns of $\boldsymbol{U}$ are orthonormal eigenvectors of $\boldsymbol{V}$. Denote p as the rank of matrix $\boldsymbol{X}$. Then $\boldsymbol{\Lambda}$ is a diagonal matrix with $n - p$ 1's and p 0's – eigenvalues of $\boldsymbol{V}$. Let $\boldsymbol{U}_1$ be a matrix of eigenvectors that correspond to the eigenvalues of 1, and let $\boldsymbol{U}_0$ be a matrix of eigenvectors that correspond to the eigenvalues of 0. Then $\boldsymbol{U} = (\boldsymbol{U}_1 \ \boldsymbol{U}_0)$, $\boldsymbol{U}_1\boldsymbol{U}_1' = \boldsymbol{V}$, and

$\boldsymbol{U}_1'\boldsymbol{U}_1 = \boldsymbol{I}_{n-p}$. Now, $\boldsymbol{U}_1'\widehat{\boldsymbol{e}}$ has covariance

$$\boldsymbol{U}_1'\boldsymbol{V}\boldsymbol{U}_1 = \boldsymbol{I}_{n-p}.$$

$\boldsymbol{U}_1'\widehat{\boldsymbol{e}}$ has exchangeable covariance structure, so these transformed residuals are permutable.

We re-express the test statistic in terms of the transformed residuals:

$$\begin{aligned} Q &= \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\epsilon}} \\ &= \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\epsilon}} \\ &= \widehat{\boldsymbol{e}}'\boldsymbol{U}_1\boldsymbol{U}_1'\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{U}_1\boldsymbol{U}_1'\widehat{\boldsymbol{e}} \\ &= \operatorname{tr}\left(\boldsymbol{U}_1'\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{U}_1\boldsymbol{U}_1'\widehat{\boldsymbol{e}}\left(\boldsymbol{U}_1'\widehat{\boldsymbol{e}}\right)'\right) \end{aligned}$$

We center the columns of $\boldsymbol{U}_1'\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{U}_1$ and $\boldsymbol{U}_1'\widehat{\boldsymbol{e}}$ using centering matrix $\boldsymbol{H}_{n-p} = \boldsymbol{I}_{n-p} - \frac{1}{n}\boldsymbol{J}_{n-p}$, where $\boldsymbol{J}_m$ is a $m \times m$ matrix of ones. Now,

$$\begin{aligned} Q &= \operatorname{tr}\left(\boldsymbol{H}_{n-p}\boldsymbol{U}_1'\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{U}_1\boldsymbol{H}_{n-p}\boldsymbol{U}_1'\widehat{\boldsymbol{e}}\left(\boldsymbol{U}_1'\widehat{\boldsymbol{e}}\right)'\boldsymbol{H}_{n-p}\right) \\ &\equiv \operatorname{tr}\left(\boldsymbol{R}\boldsymbol{S}\right) \end{aligned}$$

with $\boldsymbol{R} = \boldsymbol{H}_{n-p}\boldsymbol{U}_1'\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{U}_1$ and $\boldsymbol{S} = \boldsymbol{H}_{n-p}\boldsymbol{U}_1'\widehat{\boldsymbol{e}}\left(\boldsymbol{U}_1'\widehat{\boldsymbol{e}}\right)'\boldsymbol{H}_{n-p}$. Note that the statistic $\boldsymbol{s} = \boldsymbol{H}_{n-p}\boldsymbol{U}_1'\widehat{\boldsymbol{e}}$ has covariance equal to $\boldsymbol{H}_{n-p}$, so the distribution of $\boldsymbol{s}$ is permutable over the statistical units.

Now, we turn our attention to the permutation distribution of the test statistic – the distribution of the $n!$ values of Q obtained by permutation of the entries of vector $\boldsymbol{s}$. Rather than doing direct permutation to ascertain the permutation distribution of the test statistic Q, we approximate the permutation distribution using a Pearson type III distribution. The Pearson type III distribution has been shown to be an efficiently implemented and accurate approximation to the permutation distribution of RV test statistics, which take on a similar form to our test statistic, Q [21]. The Pearson type III distribution can be fully specified using the first three moments of Q under its permutation distribution. We can calculate

these moments using results found in the literature [22]. Finally, we use the Pearson type II distribution to calculate tail probabilities to get a p-value for the association.

4.2.3 Simulation Studies

We assess performance of the pseudo-permutation approach to p-value calculation via simulation under various settings that mimic real data applications. First, we simulated genotypes to create a sample with relatedness amongst subjects for sample sizes of $n = \{50, 100\}$. Using cosi2 [42], we simulated $2n$ haplotypes for a 20Mb region to mimic the linkage disequilibrium (LD) pattern, local recombination rate, and the coalescent population history of European, African, and East Asian populations. Specifically, we designated three populations to be similar to three 1000 Genomes populations: Yoruba in Ibadan, Nigeria (YRI) of size $14,474$; Utah residents with Northern and Western European ancestry (CEU) of size $338,000$; and Han Chinese in Bejing, China (CHB) and Japanese in Tokyo, Japan (JPT) of size $454,000$ [4]. In the coalescent simulator, demographic events mimic those of these three populations. The simulated sample consists of 30% YRI haplotypes, 40% CEU haplotypes, and 30% CHB/JPT haplotypes. We randomly chose 28% of the haplotypes and paired them off to create $0.28n$ unrelated haploids. We randomly paired the remaining 72% of haplotypes to give rise to $0.72n$ haploids, thereby creating generation 1. Then, we randomly paired all diploids of generation 1, denoting one in the pair female and the other male, for mating. Each pair produces two offspring, one male and one female. We create each offspring by drawing a random haplotype from the female parent and a random haplotype from the male parent. This set of offspring is considered generation 2. We aggregated sibling pairs from generation 2 into groups of 6. We mated the female from sibling pair 1 with the male from sibling pair 6, the female from sibling pair 2 with the male from sibling pair 1, the female from sibling pair 3 with the male from sibling pair 2, etc. Then, we randomly sample haplotypes from each mating pair to create 2 diploid offspring, giving rise to generation 3. Finally, a sample of size n is comprised of

1. all $0.28n$ unrelated diploids,

2. a random subset of $0.12n$ diploids from generation 1,

3. a random subset of $0.2n$ diploids from generation 2, and

4. a random subset of $0.4n$ diploids from generation 3.

This mimics a study in which a modest proportion of subjects are closely related. These simulated samples feature 6% and 5% of the pairs in the sample being close relatives (3rd degree or more closely related) for the samples of size 50 and 100, respectively.

We estimated a genetic relatedness matrix, $\mathbf{\Phi}$, using the GCTA method from the SNPRelate R package [61], only using variants from the 20Mb region with minor allele frequency of at least 0.1%.

Association testing of genotype data

In the first setting we perform association testing between a set of common variants and a continuous outcome, featuring correlation amongst outcomes. We simulate dependent continuous outcomes from common variants within simulated genes. For each subject in the sample, we restricted attention to the first half of the simulated 20Mb region and partitioned it into "genes" of length 50kb, giving rise to 200 genes. Next we created 200 gene sets for testing, only considering variants within the gene with minor allele frequency of at least 10%. Simulated genes contain between 12 and 186 common variants.

We performed $100,000$ simulations for each of three testing scenarios using three different kernel functions to assess type I error at various α levels. Each simulated gene was used in 100 simulations, while we simulated a unique outcome for each simulation. Dependent continuous outcomes were generated according to the following model:

$$\boldsymbol{y} = \boldsymbol{X} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{\epsilon} \sim N\left(\boldsymbol{0}, \frac{5}{7}\boldsymbol{\Phi} + \boldsymbol{I}\right)$. For each test, we embedded the common variant set into a genotype kernel matrix, $\boldsymbol{K}_G$. Let $\boldsymbol{G}$ have m_G columns. Then, we embedded genotype data into kernel matrices using three different kernel functions:

1. Gaussian: $\boldsymbol{K}_G(i,j) = \exp\left(-\frac{1}{m_G}\left(\boldsymbol{G}'_i - \boldsymbol{G}'_j\right)'\left(\boldsymbol{G}'_i - \boldsymbol{G}'_j\right)\right)$, where $\boldsymbol{G}'_i$ is the i^{th} row of $\boldsymbol{G}$
2. linear: $\boldsymbol{K}_G = \boldsymbol{G}\boldsymbol{G}'$, and
3. quadratic: $\boldsymbol{K}_G = (\boldsymbol{G}\boldsymbol{G}') \circ (\boldsymbol{G}\boldsymbol{G}')$.

We calculated type I error as the proportion of simulations that result in hypothesis tests with p-value less than α.

To assess power, we performed $10,000$ simulations for each of three kernel types, under two different data generating mechanisms. Each simulated gene from section 4.2.3 was used in 50 simulations, while we generated a unique outcome for each simulation. For a given variant set, $\boldsymbol{G}$, dependent continuous outcomes were generated under a linear (4.3) and quadratic (4.4) generating mechanism:

$$\boldsymbol{y} = \boldsymbol{X} + 0.5\sum_{j\in J}\boldsymbol{G}_j + \boldsymbol{\epsilon}, \tag{4.3}$$

$$\boldsymbol{y} = \boldsymbol{X} + 0.5\sum_{j\in J}\boldsymbol{G}_j^2 + \boldsymbol{\epsilon}, \tag{4.4}$$

where $\boldsymbol{G}_j$ is a column vector of the genotypes for the j^{th} variant in the gene. $J = \{10, 20, \dots\}$. We calculated power as the proportion of simulations that result in hypothesis tests with p-value less than 5×10^{-3}.

Here, we compare empirical type I error of 3 approaches: (1) the pseudo-permutation approach assuming independence amongst outcomes, (2) the pseudo-permutation approach, and (3) the asymptotic distribution approach. We compare empirical power of approaches (2) and (3).

Association testing of multivariate continuous data

Now, we consider association testing of a dependent continuous outcome with a multivariate continuous predictor.

We embedded the multivariate continuous predictors into a kernel matrix using three different kernel functions, and we performed $100,000$ simulations for each scenario to assess type I error. Dependent continuous outcomes were generated according to the following model:

$$\boldsymbol{y} = \boldsymbol{X} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are defined the same as they were in section 4.2.3. For each simulation, we generated a set of 21 correlated continuous predictors for each subject in our study, giving rise to $n \times 21$ matrix, $\boldsymbol{M}$. These predictors follow a multivariate normal distribution with mean 0 and covariance $\boldsymbol{\Sigma}_M$, which was estimated using publicly available methylation data [2]. Then, we pre-multiplied $\boldsymbol{M}$ by $\left(\frac{5}{7}\boldsymbol{\Phi} + \boldsymbol{I}\right)^{1/2}$ to induce genetic relatedness amongst these continuous predictors. We embedded the continuous data into a kernel matrix using a Gaussian, linear, or quadratic kernel function, as in section 4.2.3. We report empirical type I error as the proportion of $100,000$ simulations that result in hypothesis tests with p-value less than α.

To assess power, we performed $10,000$ simulations for each of the three kernel types, under two different data generating mechanisms. For a simulated set of continuous predictors, $\boldsymbol{M}$, dependent continuous outcomes were generated under a linear (4.5) and quadratic (4.6) generating mechanism:

$$\boldsymbol{y} = \boldsymbol{X} + \sum_{j \in J} \boldsymbol{M}_j + \boldsymbol{\epsilon}, \tag{4.5}$$

$$\boldsymbol{y} = \boldsymbol{X} + \sum_{j \in J} \boldsymbol{M}_j^2 + \boldsymbol{\epsilon}, \tag{4.6}$$

where $\boldsymbol{M}_j$ is a column vector of the j^{th} continuous predictor. $J = \{10, 20\}$. We report empirical power as the proportion of $10,000$ simulations that result in hypothesis tests with

p-value less than 5×10^{-3}.

Here, we compare empirical type I error of 3 approaches: (1) the pseudo-permutation approach assuming independence amongst outcomes, (2) the pseudo-permutation approach, and (3) the asymptotic distribution approach. We compare empirical power of approaches (2) and (3).

Association testing of microbiome data

Finally, we consider association testing of a dependent continuous outcome with a set of microbiotic taxa.

We simulate counts of 856 operational taxonomic units (OTUs) via a Dirichlet distribution, generating $1,000$ total counts per sample. We specify the Dirichlet distribution by using parameters estimated from throat microbiome data, publicly available via the GUniFrac R package [9]. Then, we pre-multiplied the counts matrix by $\left(\frac{5}{7}\boldsymbol{\Phi} + \boldsymbol{I}\right)^{1/2}$ to induce genetic relatedness. We embedded the microbiome data matrix, $\boldsymbol{B}$, into a kernel matrix using three different kernel functions: (1) Bray-Curtis dissimilarity function [6] adapted to a similarity function, as in the MiRKAT R package, (2) linear, and (3) Gaussian. We performed $100,000$ simulations for each scenario to assess type I error. Dependent continuous outcomes were generated according to the following model:

$$\boldsymbol{y} = \boldsymbol{X} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are defined the same as they were in section 4.2.3. We report empirical type I error as the proportion of $100,000$ simulations that result in hypothesis tests with p-value less than α.

To assess power, we performed $10,000$ simulations for each of the three kernel types, under two different data generating mechanisms. For a simulated set of continuous predictors, $\boldsymbol{B}$, dependent continuous outcomes were generated under a linear (4.5) and quadratic (4.6)

generating mechanism:

$$\boldsymbol{y} = \boldsymbol{X} + 0.5 \sum_{j \in J} \boldsymbol{B}_j + \boldsymbol{\epsilon}, \tag{4.7}$$

$$\boldsymbol{y} = \boldsymbol{X} + 0.5 \sum_{j \in J} \boldsymbol{B}_j^2 + \boldsymbol{\epsilon}, \tag{4.8}$$

where $\boldsymbol{B}_j$ is a column vector of the j^{th} OTU predictor. $J = \{10, 20, \dots\}$. We report empirical power as the proportion of $10,000$ simulations that result in hypothesis tests with p-value less than 5×10^{-3}.

Here, we compare empirical type I error of 3 approaches: (1) the pseudo-permutation approach assuming independence amongst outcomes, (2) the pseudo-permutation approach, and (3) the asymptotic distribution approach. We compare empirical power of approaches (2) and (3).

Various error distributions

Next, we consider correlated continuous outcomes that do not necessarily follow a normal distribution. Since our procedure for p-value calculation is based on permutation of errors, the test does not require any distributional assumptions of the errors. We demonstrate this in simulation by simulating correlated outcomes that follow a Normal, Cauchy, and Student-t distributions.

We performed $100,000$ simulations for each of three testing scenarios to assess type I error at various α levels. Each simulated gene from section 4.2.3 was used in 100 simulations, while we simulated a unique outcome for each simulation. Dependent continuous outcomes were generated according to the following model:

$$\boldsymbol{y} = \boldsymbol{X} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I})$. The error term, $\boldsymbol{\epsilon} = \left(\frac{3}{7}\boldsymbol{\Phi} + \boldsymbol{I}\right)^{1/2} \boldsymbol{e}$, where $\boldsymbol{e}$ follows one of the 3 following distributions:

1. $N(0, 1)$

2. Cauchy$(x_0 = 0, \gamma = 0.03)$

3. t_2.

For each test, we embedded the common variant set, $\boldsymbol{G}$, into a genotype kernel matrix, $\boldsymbol{K}_G$, using a linear kernel function. We report empirical type I error as the proportion of $100,000$ simulations that result in hypothesis tests with p-value less than α.

To assess power, we performed $10,000$ simulations for each of three kernel types, under two different data generating mechanisms. Each simulated gene (from the previous section) was used in 50 simulations, while we generated a unique outcome for each simulation. For a given variant set, $\boldsymbol{G}$, dependent continuous outcomes were generated as follows:

$$\boldsymbol{y} = \boldsymbol{X} + 0.5 \sum_{j \in J} \boldsymbol{G}_j + \boldsymbol{\epsilon}.$$

We report empirical power as the proportion of $10,000$ simulations that result in hypothesis tests with p-value less than 5×10^{-3}.

Here, we compare empirical type I error and empirical power of (1) the pseudo-permutation approach and (2) the asymptotic distribution approach.

4.2.4 Data Analysis

TOPMed

We applied this approach to conduct genome-wide gene-based association analysis of platelet count (PLT) using data from a multi-ethnic population, featuring some related individuals. We here utilized gene transcript expression predicted for whole blood [33] to a small sample of 151 subjects from NHLBI's Trans-Omics for Precision Medicine initiative (TOPMed).

We fit a null model of PLT adjusting for age, sex, study, and the first 11 PCs accounting for population structure, while taking into account relatedness. We estimated a genetic

relatedness matrix as in [25].

Gene-based tests were performed using the predicted gene expression for $12,424$ transcripts. We applied a linear kernel function to predicted gene expression data and compared p-values using the pseudo-permutation approach and the asymptotic approach. We used a Bonferroni corrected genome-wide significance threshold of $0.05/12424 = 4.02 \times 10^{-6}$.

MsFLASH Vaginal Health Trial

We applied our proposed framework to perform association analysis of 171 metabolites explained by 381 operational taxonomic units (OTUs) using data from the Menopause Strategies: Finding Lasting Answers for Symptoms and Health (MsFLASH) Vaginal Health Trial. The trial aimed to identify microbial, immune, or metabolic markers associated with response to topical treatment for postmenopausal symptoms of vaginal discomfort. Over the course of a 12-week randomized trial, postmenopausal women were randomly assigned to a vaginal discomfort treatment of vaginal estradiol (arm 1), moisturizer (arm 2), or placebo (arm 3). At 0, 4, and 12 weeks, they measured vaginal fluid metabolites via broad-based metabolomic profiling and vaginal microbiota via 16S ribosomal RNA gene sequencing [34].

We performed association analysis pooling all patients and stratifying by treatment arm. For each of the 171 vaginal fluid metabolites, we assessed association between the metabolite and the vaginal microbiota, adjusting for time of measurements and age. We also adjusted for arm in the pooled analysis. We restricted analysis to patients who had complete data measured at all three visits. We excluded OTUs that had zero counts for all patient measurements, leaving us with 373 OTUs for analysis.

For the pooled analysis, we used the following null model to model repeated measures of vaginal fluid metabolites, adjusting for age, time of measurement, and trial arm:

$$y_{ij}^{(k)} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{time}_{ij} + \beta_3 \text{I}(\text{arm}_i = 2) + \beta_4 \text{I}(\text{arm}_i = 3) + \epsilon_{ij}^{(k)}, \tag{4.9}$$

where $y_{ij}^{(k)}$ is the k^{th} metabolite value for the i^{th} individual, measured at the j^{th} time point.

Moreover, $\boldsymbol{\epsilon}^{(k)} = \left(\epsilon_{11}^{(k)}, \epsilon_{12}^{(k)}, \epsilon_{13}^{(k)}, \epsilon_{21}^{(k)}, \epsilon_{22}^{(k)}, \epsilon_{23}^{(k)}, \ldots, \epsilon_{n1}^{(k)}, \epsilon_{n2}^{(k)}, \epsilon_{n3}^{(k)}\right)' \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}^{(k)}\right)$, where

$$\boldsymbol{\Sigma}^{(k)} = \begin{pmatrix} \boldsymbol{R}^{(k)} & & & \\ & \boldsymbol{R}^{(k)} & & \\ & & \ddots & \\ & & & \boldsymbol{R}^{(k)} \end{pmatrix}, \text{ and } \boldsymbol{R}^{(k)} = \begin{pmatrix} \rho_{11}^{(k)} & \rho_{12}^{(k)} & \rho_{13}^{(k)} \\ \rho_{21}^{(k)} & \rho_{22}^{(k)} & \rho_{23}^{(k)} \\ \rho_{31}^{(k)} & \rho_{32}^{(k)} & \rho_{33}^{(k)} \end{pmatrix}.$$

For the analyses stratified by trial arm, we used the following null model to model repeated measures of vaginal fluid metabolites, adjusting for age, and time of measurement:

$$y_{ij}^{(k)} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{time}_{ij} + \epsilon_{ij}^{(k)}, \tag{4.10}$$

where $y_{ij}^{(k)}$ is the k^{th} metabolite value for the i^{th} individual, measured at the j^{th} time point. The error term, $\epsilon_{ij}^{(k)}$, is modeled similarly as before.

We embedded OTU count data into a kernel matrix using the Bray-Curtis distance function, adapted to a similarity metric. We perform the association test using the pseudo-permutation test and the asymptotic test. We consider an association significant if the resulting p-value is less than the Bonferroni corrected significance threshold of $0.05/171 = 2.92 \times 10^{-4}$. We applied the proposed pseudo-permutation approach to calculate p-values for the association between each vaginal fluid metabolite and the full set of OTUs.

4.3 Results

4.3.1 Simulation Studies

Type I Error

Type I error results for genotype, multivariate continuous, and microbiome data simulations are presented in Table 4.1, Table 4.2, and Table 4.3, respectively. Note that using the test that assumes independence of outcomes has poor type I error control. This indicates that if we don't properly accommodate the dependence of outcomes, we run the risk of encountering

many false positive results in association test settings like these. For example, in Table 4.3, using the Bray-Curtis kernel function when the sample size is 100, the empirical type I error is about 5 times, 18 times, and 50 times the nominal level of 0.05, 0.005, and 0.001, respectively. Generally, the fold change in empirical type I error compared to the nominal level increases as the nominal level decreases. Therefore, if we were to apply the independent test in such a setting, the vast majority of significant signals would be false positives.

On the other hand, the test that assumes the asymptotic distribution of the test statistic is overly-conservative. As anticipated, the conservativeness of the test is less pronounced at the higher sample size of 100. This demonstrates the inappropriateness of the asymptotic assumption for small samples. However, we observe type I error closer to the nominal level at all three significance thresholds when we use the pseudo-permutation hypothesis test.

Type I error results for simulations of error terms of various distributions are presented in Table 4.4. We see a similar phenomenon in type I error of the pseudo-permutation test vs. the asymptotic test – the asymptotic test is overly-conservative while the pseudo-permutation test is closer to the nominal significance threshold. The asymptotic test is more conservative when the error structure is non-normal. The asymptotic test assumes normality of the residuals in constructing the sampling distribution.

n	α	Kernel Type	Independent	Pseudo-Permutation	Asymptotic
50	0.050	Gaussian	0.1060	0.0452	0.0368
		Linear	0.1022	0.0475	0.0417
		Quadratic	0.1002	0.0459	0.0394
	0.005	Gaussian	0.0177	0.0052	0.0026
		Linear	0.0170	0.0047	0.0027
		Quadratic	0.0171	0.0044	0.0024
	0.001	Gaussian	0.0052	0.0010	0.0003
		Linear	0.0049	0.0009	0.0004
		Quadratic	0.0050	0.0008	0.0003
100	0.050	Gaussian	0.1084	0.0466	0.0423
		Linear	0.1052	0.0478	0.0453
		Quadratic	0.1002	0.0472	0.0450
	0.005	Gaussian	0.0195	0.0050	0.0033
		Linear	0.0191	0.0051	0.0038
		Quadratic	0.0183	0.0053	0.0037
	0.001	Gaussian	0.0062	0.0009	0.0003
		Linear	0.0062	0.0010	0.0006
		Quadratic	0.0059	0.0012	0.0006

Table 4.1: Empirical type I error results of the genotype simulation setting. Simulations either used a sample size, n, of 50 or 100. Genotype data were embedded in Gaussian, linear, or quadratic kernel functions. We compare hypothesis test results for a pseudo-permutation test that assumes independence of outcomes (Independent), a pseudo-permutation test that accounts for dependence of outcomes (Pseudo-Permutation), and a test that uses asymptotic results (Asymptotic). Empirical type I error is reported as the proportion of $100,000$ simulated data sets whose hypothesis test results in a p-value less than the specified significance threshold, α.

n	α	Kernel Type	Independent	Pseudo-Permutation	Asymptotic
50	0.05	Gaussian	0.0733	0.0464	0.0305
		Linear	0.0724	0.0480	0.0393
		Quadratic	0.0778	0.0457	0.0307
	0.005	Gaussian	0.0092	0.0050	0.0016
		Linear	0.0094	0.0053	0.0025
		Quadratic	0.0118	0.0046	0.0018
	0.001	Gaussian	0.0023	0.0011	0.0002
		Linear	0.0023	0.0011	0.0003
		Quadratic	0.0030	0.0011	0.0001
100	0.05	Gaussian	0.0700	0.0473	0.0400
		Linear	0.0683	0.0487	0.0457
		Quadratic	0.0721	0.0466	0.0383
	0.005	Gaussian	0.0089	0.0052	0.0031
		Linear	0.0089	0.0054	0.0035
		Quadratic	0.0098	0.0048	0.0024
	0.001	Gaussian	0.0024	0.0012	0.0005
		Linear	0.0022	0.0009	0.0004
		Quadratic	0.0027	0.0010	0.0004

Table 4.2: Empirical type I error results of the multivariate continuous simulation setting. Simulations either used a sample size, n, of 50 or 100. Multivariate continuous data were embedded in Gaussian, linear, or quadratic kernel functions. We compare hypothesis test results for a pseudo-permutation test that assumes independence of outcomes (Independent), a pseudo-permutation test that accounts for dependence of outcomes (Pseudo-Permutation), and a test that uses asymptotic results (Asymptotic). Empirical type I error is reported as the proportion of $100,000$ simulated data sets whose hypothesis test results in a p-value less than the specified significance threshold, α.

n	α	Kernel Type	Independent	Pseudo-Permutation	Asymptotic
50	0.05	Bray-Curtis	0.2153	0.0376	0.0026
		Linear	0.1208	0.0439	0.0121
		Quadratic	0.1177	0.0418	0.0093
	0.005	Bray-Curtis	0.0650	0.0035	0
		Linear	0.0194	0.0047	0.0003
		Quadratic	0.0182	0.0044	0.0003
	0.001	Bray-Curtis	0.0301	0.0008	0
		Linear	0.0054	0.0011	0
		Quadratic	0.0054	0.0010	0
100	0.05	Bray-Curtis	0.2412	0.0452	0.0109
		Linear	0.1077	0.0466	0.0241
		Quadratic	0.1071	0.0458	0.0198
	0.005	Bray-Curtis	0.0901	0.0043	0.0001
		Linear	0.0170	0.0054	0.0014
		Quadratic	0.0160	0.0057	0.0012
	0.001	Bray-Curtis	0.0500	0.0008	0
		Linear	0.0047	0.0011	0.0001
		Quadratic	0.0046	0.0015	0.0001

Table 4.3: Empirical type I error results of the microbiome simulation setting. Simulations either used a sample size, n, of 50 or 100. Microbiome data were embedded in a Bray-Curtis, linear, or quadratic kernel function. We compare hypothesis test results for a pseudo-permutation test that assumes independence of outcomes (Independent), a pseudo-permutation test that accounts for dependence of outcomes (Pseudo-Permutation), and a test that uses asymptotic results (Asymptotic). Empirical type I error is reported as the proportion of $100,000$ simulated data sets whose hypothesis test results in a p-value less than the specified significance threshold, α.

n	α	Error Distribution	Pseudo-Permutation	Asymptotic
50	0.05	Cauchy	0.0459	0.0319
		Normal	0.0467	0.0412
		Student-t	0.0469	0.0383
	0.005	Cauchy	0.0045	0.0018
		Normal	0.0043	0.0025
		Student-t	0.0042	0.0022
	0.001	Cauchy	0.0009	0.0001
		Normal	0.0008	0.0003
		Student-t	0.0007	0.0003
100	0.05	Cauchy	0.0540	0.0398
		Normal	0.0483	0.0459
		Student-t	0.0503	0.0450
	0.005	Cauchy	0.0049	0.0027
		Normal	0.0048	0.0036
		Student-t	0.0050	0.0036
	0.001	Cauchy	0.0009	0.0004
		Normal	0.0010	0.0007
		Student-t	0.0010	0.0006

Table 4.4: Empirical type I error results using dependent outcomes that are simulated from genotype data, using three different error distributions: Cauchy, Normal, and Student-t. Simulations either used a sample size, n, of 50 or 100. We compare hypothesis test results for a pseudo-permutation test that accounts for dependence of outcomes (Pseudo-Permutation) and a test that uses asymptotic results (Asymptotic). Empirical type I error is reported as the proportion of $100,000$ simulated datasets whose hypothesis test results in a p-value less than the specified significance threshold, α.

Power

Power results for genotype, multivariate continuous, and microbiome data simulations are presented in Table 4.5, Table 4.6, and Table 4.7, respectively. As expected, since the pseudo-permutation test is less conservative than the asymptotic test, this translates to increased power of the test across all scenarios. Across the board, power increases as sample size increases.

We also observe the implication of tailoring the kernel function to best accommodate the data type and the data generating mechanism. If we restrict our attention to the multivariate continuous simulations in Table 4.6, we see that when the multivariate continuous features are related to the outcome in a linear fashion, the Gaussian and linear kernel functions result in higher powered tests than the test using a quadratic kernel function. We observe the opposite when the features are related to the outcome quadratically – the test with the quadratic kernel function is higher-powered than the tests with the Gaussian and linear kernel functions.

Power results for simulations of error terms of various distributions are presented in Table 4.8

		Linear		Quadratic	
n	Kernel Type	Pseudo-Permutation	Asymptotic	Pseudo-Permutation	Asymptotic
50	Gaussian	0.4954	0.4515	0.6744	0.6313
	Linear	0.4946	0.4643	0.6395	0.6081
	Quadratic	0.4555	0.4223	0.6876	0.6519
100	Gaussian	0.7323	0.7087	0.8935	0.8819
	Linear	0.7239	0.7116	0.8504	0.8411
	Quadratic	0.6954	0.6799	0.8808	0.8701

Table 4.5: Empirical power results of the genotype simulation setting. Simulations either used a sample size, n, of 50 or 100. Dependent outcomes were related to the data type in a linear or quadratic fashion. Genotype data were embedded in Gaussian, linear, or quadratic kernel functions. We compare hypothesis test results for a pseudo-permutation test that accounts for dependence of outcomes (Pseudo-Permutation) and a test that uses asymptotic results (Asymptotic). Empirical power is reported as the proportion of $10,000$ simulated data sets whose hypothesis test results in a p-value less than 5×10^{-3}.

		Linear		Quadratic	
n	Kernel Type	Pseudo-Permutation	Asymptotic	Pseudo-Permutation	Asymptotic
50	Gaussian	0.4341	0.3501	0.0255	0.0100
	Linear	0.4546	0.3996	0.0126	0.0068
	Quadratic	0.0136	0.0050	0.2674	0.1999
100	Gaussian	0.7246	0.6831	0.0656	0.0461
	Linear	0.7123	0.6864	0.0120	0.0084
	Quadratic	0.0192	0.0113	0.3798	0.3422

Table 4.6: Empirical power results of the multivariate continuous simulation setting. Simulations either used a sample size, n, of 50 or 100. Dependent outcomes were related to the data type in a linear or quadratic fashion. Multivariate continuous data were embedded in Gaussian, linear, or quadratic kernel functions. We compare hypothesis test results for a pseudo-permutation test that accounts for dependence of outcomes (Pseudo-Permutation) and a test that uses asymptotic results (Asymptotic). Empirical power is reported as the proportion of $10,000$ simulated datasets whose hypothesis test results in a p-value less than 5×10^{-3}.

		Linear		Quadratic	
n	Kernel Type	Pseudo-Permutation	Asymptotic	Pseudo-Permutation	Asymptotic
50	Bray-Curtis	0.3393	0.0740	0.2438	0.0891
	Linear	0.3533	0.1894	0.3486	0.2156
	Quadratic	0.2338	0.1185	0.2980	0.1825
100	Bray-Curtis	0.8039	0.5946	0.5846	0.3871
	Linear	0.7220	0.6050	0.6918	0.5525
	Quadratic	0.5217	0.3647	0.5978	0.4554

Table 4.7: Empirical power results of the microbiome simulation settings. Simulations either used a sample size, n, of 50 or 100. Dependent outcomes were related to the data type in a linear or quadratic fashion. Microbiome data were embedded in a Bray-Curtis, linear, or quadratic kernel function. We compare hypothesis test results for a pseudo-permutation test that accounts for dependence of outcomes (Pseudo-Permutation) and a test that uses asymptotic results (Asymptotic). Empirical power is reported as the proportion of $10,000$ simulated data sets whose hypothesis test results in a p-value less than 5×10^{-3}.

n	Error Structure	Pseudo-Permutation	Asymptotic
50	Cauchy	0.7903	0.7720
	Normal	0.4987	0.4648
	Student-t	0.1866	0.1595
100	Cauchy	0.8550	0.8396
	Normal	0.7259	0.7102
	Student-t	0.3299	0.3117

Table 4.8: Empirical power results using dependent outcomes that are simulated from genotype data, using three different error distributions: Cauchy, Normal, and Student-t. Simulations either used a sample size, n, of 50 or 100. Dependent outcomes were related to the data type in a linear or quadratic fashion. We compare hypothesis test results for a pseudo-permutation test that accounts for dependence of outcomes (Pseudo-Permutation) and a test that uses asymptotic results (Asymptotic). Empirical power is reported as the proportion of $10,000$ simulated datasets whose hypothesis test results in a p-value less than 5×10^{-3}.

4.3.2 Data Analysis

TOPMed

Results of genome-wide association analysis of predicted gene expression using the pseudo-permutation test and the asymptotic test are presented in Figure 4.1. While no signals attained genome-wide significance, we notice a few meaningful results from the direct comparison of these two hypothesis tests. The results indicate that both tests are valid – the distribution of the p-values don't significantly deviate from the expected distribution for both tests (Figure 4.2). However, we observe a slight deflation of p-values in the asymptotic test. Again, this is consistent with the asymptotic test being overly conservative for small sample sizes.

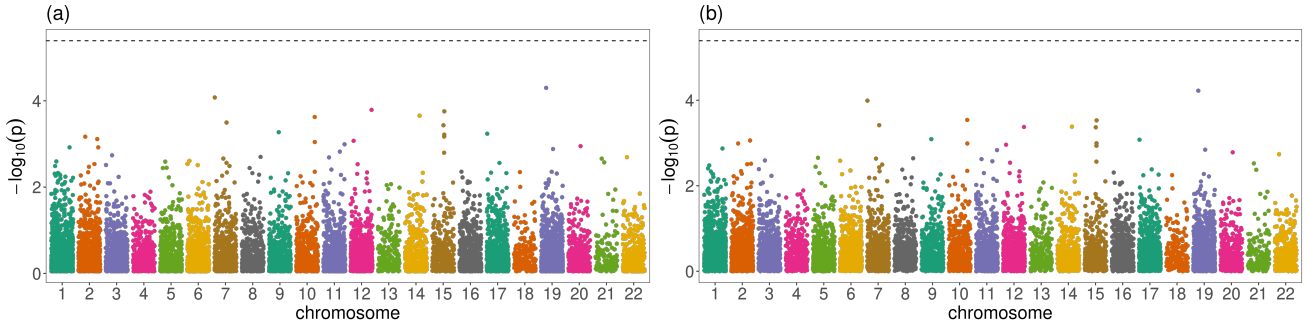


Figure 4.1: Manhattan plot of the joint test. The height of the dashed black line is at the significance threshold. The Bonferroni corrected significance threshold is $0.05/12424 = 4.02 \times 10^{-6}$

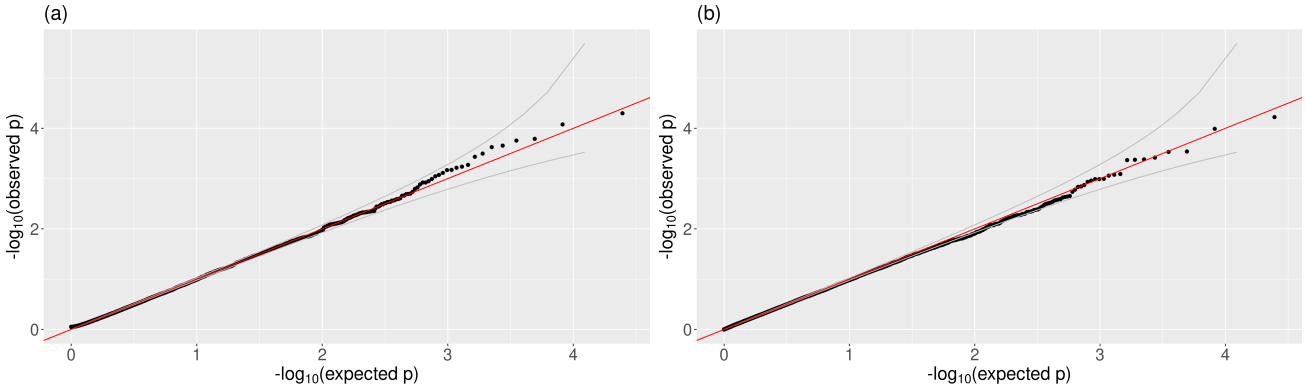


Figure 4.2: QQ plot. Gray lines specify the 95% confidence intervals.

MsFLASH Vaginal Health Trial

Association analysis of vaginal metabolites explained by vaginal microbiota demonstrate a similar phenomenon – we observe fewer signals identified by the asymptotic test compared to the pseudo-permutation test. In this case, we most metabolites to be truly associated with the microbiome, as the microbiome give rise to the metabolites. In the pooled analysis, we identify 5 more signals using the pseudo-permutation test compared to using the asymptotic test. In the analyses restricted one arm, the discrepancy in number of signals identified is more pronounced. We identify 12, 18, and 9 more signals using the pseudo-permutation test

compared to the asymptotic test compared to using the asymptotic test for arm 1, 2, and 3, respectively. This demonstrates the increased utility of the pseudo-permutation test over the asymptotic test as sample size decreases.

Arm	n	Pseudo-permutation	Asymptotic
1	38	35 (0.205)	23 (0.135)
2	44	56 (0.327)	38 (0.222)
3	44	45 (0.263)	34 (0.199)
pooled	126	142 (0.830)	137 (0.801)

Table 4.9: Association analysis of 171 metabolites and vaginal microbiota from the MsFLASH Vaginal Health Trial; Number and proportion of 171 tests that resulted in p-value less than 2.92×10^{-4}. The pooled analysis uses data of all patients from all three arms of the trial, whereas the first three rows are analysis results from one of three arms of the trial. Each patient has data on three repeated measures, taken at weeks 0, 4, and 12.

4.4 Discussion

Our proposed approach approximates a permutation distribution for a kernel based association test while accounting for dependent outcomes. The pseudo-permutation approach benefits from the advantages of using a permutation test without dealing with the computational intensity of a permutation test. This approach also takes into account the architecture of the specific data types, including high dimensional data allowing them to relate to the trait in a non-linear fashion. This approach overcomes the challenges associated with using a the more conservative approach of assuming an asymptotic distribution of the test statistic.

We demonstrated the utility of this approach on a small, multi-ethnic sample, including related individuals from TOPMed. This demonstrates the utility of this approach in performing a gene-based genome-wide association analysis. This test is also useful for other modalities of grouping omic information as we interrogate associations. As demonstrated

with the data application to a small clinical trial, we may also use this test to interrogate the association between quantitative outcomes, as functionally grouped by a pathway.

This approach is not limited to analysis of one data type. This approach integrates cleanly with projects 1 and 2. We may integrate multiple data types to interrogate association with a quantitative dependent outcome using the same approach as presented in the previous projects, replacing the asymptotic approach to p-value calculation with the pseudo-permutation approach to p-value calculation.

Moreover, we are not limited to using this approach for small samples. An advantage of employing a permutation-type test is that we do not have to make an assumption about the distribution of the error term in order to ascertain the distribution of the test statistic. This could be useful if the distribution of the residuals is non-normal, even if the sample size is large. However, we note that with large sample sizes ($n > 1000$), the computation of the pseudo-permutation test tends to take longer than the computation of an asymptotic test.

This framework can be extended to accommodate various other data and study considerations. For example, this method can be extended to perform association analysis of different types of dependent outcomes (binary, counts, etc.) by employing a generalized linear mixed model in place of the LMM used here. We may also extend this framework to accommodate association analysis of multiple outcomes. This is similar to the DKAT method [57], however DKAT does not account for clustering amongst the units of measurement in the study. Thus, we may extend the pseudo-permutation approach in studies that have multiple outcomes measured repeatedly on the same individuals or measured on related individuals.

Chapter 5
DISCUSSION

5.1 Summary

The integration of multiple biological resources with GWAS results can aid in uncovering causal relationships between traits and molecules, thereby identifying etiology and greater understanding of complex traits. GWAS mainly identify loci in non-coding regions [30], and we do not yet fully understand their biological impacts. This leaves requires extensive follow-up for GWAS hits in order to interpret their results. Therefore, incorporating information of intermediate cellular processes that influence genomics is a desirable direction for enhancing interpretations of genomic studies. To this end, in this dissertation, we contributed methods for association testing of quantitative traits and multi-omics.

In Chapter 2, we extended an existing method that integrates two multi-omics data types for association testing of a quantitative outcome to accommodate relatedness amongst study subjects. We proposed perturbation and kernel PCA hypothesis testing methods, which allow us to consider multiple effect models, thereby increasing statistical power to detect joint signal, if it exists. Simulations demonstrated that it is inappropriate to use a test that assumes independence of outcomes when there is indeed relatedness amongst study subjects. We demonstrated that the presented methods for integrative analysis attain higher power for detection of signals when both data types are implicated than a test with one data type, while appropriately controlling type I error. Finally, in application to a TOPMed sample, a method for integrative association testing identified three genes associated with MPV and three genes associated with PLT that were not identified using single data type tests.

In Chapter 3, we extended the work in Chapter 2 to integrate two or more multi-omics data types for association testing of a quantitative outcome, while accounting for relatedness

amongst study outcomes. The approaches presented in Chapter 2 are inappropriate for use with more than two data types, as they quickly become computationally infeasible as the number of data types grows. Therefore, we propose a Cauchy combination test and a truncated version of the Cauchy combination test for hypothesis testing, which remain computationally efficient, even as the number of tested data types grows. In simulations we integrated three data types for association testing. We observed drastically inflated type I error when using a test that assumes independence of outcomes when there is indeed relatedness amongst study subjects. We demonstrated that the presented combination tests for integrative analysis attain higher power for detection of signals when more than one data type is implicated than a test with one data type, while appropriately controlling type I error. Finally, in application to a TOPMed sample, we integrated genotype, methylation, and expression data for association testing of PLT. In doing so, the joint test identified three signals associated with PLT that were not identified using single data type tests.

In Chapter 4, we proposed a pseudo-permutation approach for association testing of a quantitative outcome for small samples, while accommodating dependence amongst the outcomes. The tests in the Chapters 2 and 3 assume an asymptotic distribution of the test statistic, which we show to be overly-conservative when applied to small samples. The pseudo-permutation approach address this conservativeness while accounting for the dependence present in the residuals of the null model. We applied this approach to a small TOPMed sample to perform association testing of predicted gene expression and PLT, which demonstrated deflation of p-values using the asymptotic distribution compared to using the pseudo-permutation distribution. We also observed a loss in power to detect signal when using the asymptotic distribution in data application to the MsFLASH Vaginal Health Trial.

In this dissertation, we demonstrated the benefit of integrating multi-omics via a KMR framework. In doing so, we address various challenges in such analyses. We are able to integrate high-dimensional multi-omics data, allowing us to aggregate several features with small effects. The KMR framework provides tailored embedding of each data type, such that we can accentuate relevant scientific information about each data type, thereby increasing

statistical power to detect association. Moreover, with the proposed methods, we are readily able to account for various complexities that typically arise in study designs, such as relatedness amongst study subjects, repeated measures on study subjects, and population structure within our sample. We contribute association testing methods that are valid for integrating an arbitrary number of multi-omics data types and for sample sizes of any size.

5.2 Future Work

Methods presented within this dissertation substantially generalize existing methods, such that we may perform association analysis in varying study designs. Yet, we may further generalize to cover even more data applications. One obvious extension is to relax the model such that it can accommodate outcomes that are not quantitative. Rather than the LMMs used in Chapters 2, 3, and 4, we may instead use a generalized LMM (GLMM). There is already precedent for variance component score testing derived from GLMMs under the KMR framework [26, 51, 59].

One may also apply these methods to other data domains. Here we mostly focused on integrating multi-omics of molecules from single cells. However, the KMR framework is flexible enough to accommodate other data types, such as high dimensional environmental exposures. Because the methods in Chapter 3 can accommodate an arbitrary number of data types, we could also consider creating interaction kernel matrices, in case there is suspected effect modification of any subset of data types. There is precedent for incorporating environmental and interactions (specifically gene-environment interactions) under the KMR framework [20, 60], but these methods can be further generalized.

Finally, the pseudo-permutation method that accounts for relatedness in Chapter 4 may be extended in order to perform association testing of multiple outcomes and multi-omics. An existing method does this for one data type (specifically genetic data) without accounting for dependence of outcomes [57], thus there is opportunity for extension. While the method presented in Chapter 4 was presented for small samples, it is still valid for larger samples, albeit, more computationally burdensome as sample size increases.

BIBLIOGRAPHY

[1] Mark Abney. Permutation testing in the presence of polygenic variation. *Genetic Epidemiology*, 39(4):249–258, may 2015.

[2] Reid S. Alisch, Benjamin G. Barwick, Pankaj Chopra, Leila K. Myrick, Glen A. Satten, Karen N. Conneely, and Stephen T. Warren. Age-associated DNA methylation in pediatric populations. *Genome Research*, 22(4):623–632, 2012.

[3] William Astle and David J. Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471, 2009.

[4] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie Laure Yaspo, Lucinda Fulton, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping

Zhan, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Aniko Sabo, Zhuoyi Huang, Lachlan J.M. Coin, Lin Fang, Qibin Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Erik P. Garrison, Deniz Kural, Wan Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, Eric Banks, Gaurav Bhatia, Guillermo Del Angel, Giulio Genovese, Heng Li, Seva Kashin, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Kathryn Beal, Avik Datta, Javier Herrero, Graham R.S. Ritchie, Daniel Zerbino, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Herwig, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, Anand Swaroop, Emily Chew, Tuuli Lappalainen, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Jared T. Simpson, Mark D. Shriver, Jeffrey A. Rosenfeld, Carlos D. Bustamante, Stephen B. Montgomery, Francisco M. De La Vega, Jake K. Byrnes, Andrew W. Carroll, Marianne K. DeGorter, Phil Lacroute, Brian K. Maples, Alicia R. Martin, Andres Moreno-Estrada, Suyash S. Shringarpure, Fouad Zakharia, Eran Halperin, Yael Baran, Eliza Cerveira, Jaeho Hwang, Ankit Malhotra, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, Fiona C.L. Hyland, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Kevin Squire, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Madhusudan Gujral, Amina Noor, Kenny Ye, Esteban G. Burchard, Ryan D. Hernandez, Christopher R. Gignoux, David Haussler, Sol J. Katzman, W. James Kent, Bryan Howie, Andres Ruiz-Linares, Emmanouil T. Dermitzakis, Scott E. Devine, Hyun Min Kang, Jeffrey M. Kidd, Tom Blackwell, Sean Caron, Wei Chen, Sarah Emery, Lars Fritsche, Christian Fuchsberger, Goo Jun, Bingshan Li, Robert Lyons, Chris Scheller, Carlo Sidore, Shiya Song, Elzbieta Sliwerska, Daniel Taliun, Adrian Tan, Ryan Welch, Mary Kate Wing, Xiaowei Zhan, Philip Awadalla, Alan Hodgkinson, Yun Li, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Jonathan L. Marchini, Simon Myers, Claire Churchhouse, Olivier Delaneau, Anjali Gupta-Hinch, Warren Kretzschmar, Zamin Iqbal, Iain Mathieson, Androniki Menelaou, Andy Rimmer, Dionysia K. Xifara, Taras K. Oleksyk, Yunxin Fu, Xiaoming Liu, Momiao Xiong, Lynn Jorde, David Witherspoon, Jinchuan Xing, Brian L. Browning, Sharon R. Browning, Fereydoun Hormozdiari, Peter H. Sudmant, Ekta Khurana, Chris Tyler-Smith, Cornelis A. Albers, Qasim Ayub, Yuan Chen, Vincenza Colonna, Luke Jostins, Klaudia Walter, Yali Xue, Mark B. Gerstein, Alexej Abyzov, Suganthi Balasubramanian, Jieming Chen, Declan Clarke, Yao Fu, Arif O. Harmanci, Mike Jin,

Donghoon Lee, Jeremy Liu, Xinmeng Jasmine Mu, Jing Zhang, Yan Zhang, Chris Hartl, Khalid Shakir, Jeremiah Degenhardt, Sascha Meiers, Benjamin Raeder, Francesco Paolo Casale, Oliver Stegle, Eric Wubbo Lameijer, Ira Hall, Vineet Bafna, Jacob Michaelson, Eugene J. Gardner, Ryan E. Mills, Gargi Dayama, Ken Chen, Xian Fan, Zechen Chong, Tenghui Chen, Mark J. Chaisson, John Huddleston, Maika Malig, Bradley J. Nelson, Nicholas F. Parrish, Ben Blackburne, Sarah J. Lindsay, Zemin Ning, Yujun Zhang, Hugo Lam, Cristina Sisu, Danny Challis, Uday S. Evani, James Lu, Uma Nagaswamy, Jin Yu, Wangshen Li, Lukas Habegger, Haiyuan Yu, Fiona Cunningham, Ian Dunham, Kasper Lage, Jakob Berg Jespersen, Heiko Horn, Donghoon Kim, Rob Desalle, Apurva Narechania, Melissa A.Wilson Sayres, Fernando L. Mendez, G. David Poznik, Peter A. Underhill, David Mittelman, Ruby Banerjee, Maria Cerezo, Thomas W. Fitzgerald, Sandra Louzada, Andrea Massaia, Fengtang Yang, Divya Kalra, Walker Hale, Xu Dan, Kathleen C. Barnes, Christine Beiswanger, Hongyu Cai, Hongzhi Cao, Brenna Henn, Danielle Jones, Jane S. Kaye, Alastair Kent, Angeliki Kerasidou, Rasika Mathias, Pilar N. Ossorio, Michael Parker, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Zhongming Tian, Sarah Tishkoff, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Zhiming Cai, Yang Gao, Jiayou Chu, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Rasika A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Xiaoyan Deng, Danny Asogun, Onikepe Folarin, Christian Happi, Omonwunmi Omoniwa, Matt Stremlau, Ridhi Tariyal, Muminatou Jallow, Fatoumatta Sisay Joof, Tumani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Tran Tinh Hien, Sarah J. Dunstan, Nguyen ThuyHang, Richard Fonnie, Robert Garry, Lansana Kanneh, Lina Moses, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yekaterina Vaydylevich, Audrey Duncanson, Michael Dunn, and Jeffery A. Schloss. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[5] Bradley E. Bernstein, Alexander Meissner, and Eric S. Lander. The Mammalian Epigenome. *Cell*, 128(4):669–681, 2007.

[6] J. Roger Bray and J. T. Curtis. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, 27(4):325–349, 1957.

[7] T. Burdett, E. Hastings, and D. Welter. SPOT. *EMBL-EBI, and NHGRI (GWAS Catalog)*, 2018.

[8] Cara L. Carty, Nicholas A. Johnson, Carolyn M. Hutter, Alexander P. Reiner, Ulrike Peters, Hua Tang, and Charles Kooperberg. Genome-wide association study of body height in African Americans: The Women's Health Initiative SNP Health Association Resource (SHARe). *Human Molecular Genetics*, 21(3):711–720, 2012.

[9] Emily S. Charlson, Jun Chen, Rebecca Custers-Allen, Kyle Bittinger, Hongzhe Li, Rohini Sinha, Jennifer Hwang, Frederic D. Bushman, and Ronald G. Collman. Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS ONE*, 5(12):1–10, 2010.

[10] Han Chen, James B. Meigs, and Josée Dupuis. Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genetic Epidemiology*, 37(2):196–204, 2013.

[11] Melina Claussnitzer, Simon N. Dankel, Kyoung-Han Kim, Gerald Quon, Wouter Meuleman, Christine Haugen, Viktoria Glunk, Isabel S. Sousa, Jacqueline L. Beaudry, Vijitha Puviindran, Nezar A. Abdennur, Jannel Liu, Per-Arne Svensson, Yi-Hsiang Hsu, Daniel J. Drucker, Gunnar Mellgren, Chi-Chung Hui, Hans Hauner, and Manolis Kellis. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine*, 373(10):895–907, sep 2015.

[12] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michal Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):1–19, 2016.

[13] Matthew P. Conomos, Alex P. Reiner, Mary Sara McPeek, and Timothy A. Thornton. Genome-Wide Control of Population Structure and Relatedness in Genetic Association Studies via Linear Mixed Models with Orthogonally Partitioned Structure. *bioRxiv*, page 409953, sep 2018.

[14] Matthew P Conomos, Alexander P Reiner, Bruce S Weir, and Timothy A Thornton. Model-free Estimation of Recent Genetic Relatedness. 2016.

[15] Robert B. Davies. Algorithm AS 155: The Distribution of a Linear Combination of χ 2 Random Variables. *Applied Statistics*, 29(3):323, 1980.

[16] Eric R. Gamazon, Heather E. Wheeler, Kaanan P. Shah, Sahar V. Mozaffari, Keston Aquino-Michaels, Robert J. Carroll, Anne E. Eyler, Joshua C. Denny, Dan L. Nicolae, Nancy J. Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 2015.

[17] Stephanie M. Gogarten, Tamar Sofer, Han Chen, Chaoyu Yu, Jennifer A. Brody, Timothy A. Thornton, Kenneth M. Rice, and Matthew P. Conomos. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*, 35(24):5346–5348, dec 2019.

[18] Jennifer Hays, Julie R Hunt, F.Allan Hubbell, Garnet L Anderson, Marian Limacher, Catherine Allen, and Jacques E Rossouw. The women's health initiative recruitment methods and results. *Annals of Epidemiology*, 13(9):S18–S77, 2003.

[19] Yao Hu, Adrienne M. Stilp, Caitlin P. McHugh, Shuquan Rao, Deepti Jain, Xiuwen Zheng, John Lane, Sébastian Méric de Bellefon, Laura M. Raffield, Ming Huei Chen, Lisa R. Yanek, Marsha Wheeler, Yao Yao, Chunyan Ren, Jai Broome, Jee Young Moon, Paul S. de Vries, Brian D. Hobbs, Quan Sun, Praveen Surendran, Jennifer A. Brody, Thomas W. Blackwell, Hélène Choquet, Kathleen Ryan, Ravindranath Duggirala, Nancy Heard-Costa, Zhe Wang, Nathalie Chami, Michael H. Preuss, Nancy Min, Lynette Ekunwe, Leslie A. Lange, Mary Cushman, Nauder Faraday, Joanne E. Curran, Laura Almasy, Kousik Kundu, Albert V. Smith, Stacey Gabriel, Jerome I. Rotter, Myriam Fornage, Donald M. Lloyd-Jones, Ramachandran S. Vasan, Nicholas L. Smith, Kari E. North, Eric Boerwinkle, Lewis C. Becker, Joshua P. Lewis, Goncalo R. Abecasis, Lifang Hou, Jeffrey R. O'Connell, Alanna C. Morrison, Terri H. Beaty, Robert Kaplan, Adolfo Correa, John Blangero, Eric Jorgenson, Bruce M. Psaty, Charles Kooperberg, Russell T. Walton, Benjamin P. Kleinstiver, Hua Tang, Ruth J.F. Loos, Nicole Soranzo, Adam S. Butterworth, Debbie Nickerson, Stephen S. Rich, Braxton D. Mitchell, Andrew D. Johnson, Paul L. Auer, Yun Li, Rasika A. Mathias, Guillaume Lettre, Nathan Pankratz, Cathy C. Laurie, Cecelia A. Laurie, Daniel E. Bauer, Matthew P. Conomos, and Alexander P. Reiner. Whole-genome sequencing association analysis of quantitative red blood cell phenotypes: The NHLBI TOPMed program. *American Journal of Human Genetics*, 108(5):874–893, 2021.

[20] Yen-Tsung Huang, Tyler J. VanderWeele, and Xihong Lin. Joint Analysis of Snp and Gene Expression Data in. *Annals of Applied Statistics*, 8(1):1–24, 2014.

[21] J Josse, J Pagès, and F Husson. Testing the significance of the RV coefficient. *Computational Statistics and Data Analysis*, 53(1):82–91, 2008.

[22] Frédérique Kazi-Aoual, Simon Hitier, Robert Sabatier, and Jean Dominique Lebreton. Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis*, 20(6):643–656, 1995.

[23] Lydia Coulter Kwee, Dawei Liu, Xihong Lin, Debashis Ghosh, and Michael P. Epstein. A Powerful and Flexible Multilocus Association Test for Quantitative Traits. *American Journal of Human Genetics*, 82(2):386–397, feb 2008.

[24] Jin Li, Leslie A. Lange, Jeremy Sabourin, Qing Duan, William Valdar, Monte S. Willis, Yun Li, James G. Wilson, and Ethan M. Lange. Genome- and exome-wide association study of serum lipoprotein (a) in the Jackson Heart Study. *Journal of Human Genetics*, 60(12):755–761, 2015.

[25] Amarise Little, Yao Hu, Quan Sun, Deepti Jain, Jai Broome, Ming Huei Chen, Florian Thibord, Caitlin Mchugh, Praveen Surendran, Thomas W. Blackwell, Jennifer A. Brody, Arunoday Bhan, Nathalie Chami, Paul S. De Vries, Lynette Ekunwe, Nancy Heard-Costa, Brian D. Hobbs, Ani Manichaikul, Jee Young Moon, Michael H. Preuss, Kathleen Ryan, Zhe Wang, Marsha Wheeler, Lisa R. Yanek, Goncalo R. Abecasis, Laura Almasy, Terri H. Beaty, Lewis C. Becker, John Blangero, Eric Boerwinkle, Adam S. Butterworth, Hélène Choquet, Adolfo Correa, Joanne E. Curran, Nauder Faraday, Myriam Fornage, David C. Glahn, Lifang Hou, Eric Jorgenson, Charles Kooperberg, Joshua P. Lewis, Donald M. Lloyd-Jones, Ruth J.F. Loos, Yuan I. Min, Braxton D. Mitchell, Alanna C. Morrison, Deborah A. Nickerson, Kari E. North, Jeffrey R. O'connell, Nathan Pankratz, Bruce M. Psaty, Ramachandran S. Vasan, Stephen S. Rich, Jerome I. Rotter, Albert V. Smith, Nicholas L. Smith, Hua Tang, Russell P. Tracy, Matthew P. Conomos, Cecelia A. Laurie, Rasika A. Mathias, Yun Li, Paul L. Auer, Timothy Thornton, Alexander P. Reiner, Andrew D. Johnson, and Laura M. Raffield. Whole genome sequence analysis of platelet traits in the NHLBI Trans-Omics for Precision Medicine (TOPMed) initiative. *Human Molecular Genetics*, 31(3):347–361, 2022.

[26] Dawei Liu, Debashis Ghosh, and Xihong Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9(1):292, jun 2008.

[27] Dawei Liu, Xihong Lin, and Debashis Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, 2007.

[28] Huan Liu, Yongqiang Tang, and Hao Helen Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics and Data Analysis*, 53(4):853–856, 2009.

[29] Yaowu Liu and Jun Xie. Cauchy Combination Test: A Powerful Test With Analytic p-Value Calculation Under Arbitrary Dependency Structures. *Journal of the American Statistical Association*, 115(529):393–402, 2020.

[30] Teri A. Manolio. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine*, 363(2):166–176, 2010.

[31] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L.

Haines, Trudy F.C. MacKay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases, 2009.

[32] Tom R. Mayo, Gabriele Schweikert, and Guido Sanguinetti. M 3 D: A kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*, 31(6):809–816, 2015.

[33] Anna V Mikhaylova. *Statistical Methods for Transcriptome-Wide Association Studies in Ancestrally Diverse Populations*. PhD thesis, University of Washington, 2022.

[34] Caroline M. Mitchell, Nanxun Ma, Alissa J. Mitchell, Michael C. Wu, D. J. Valint, Sean Proll, Susan D. Reed, Katherine A. Guthrie, Andrea Z. Lacroix, Joseph C. Larson, Robert Pepin, Daniel Raftery, David N. Fredricks, and Sujatha Srinivasan. Association between postmenopausal vulvovaginal discomfort, vaginal microbiota, and mucosal inflammation. *American Journal of Obstetrics and Gynecology*, 225(2):159.e1–159.e15, 2021.

[35] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, jul 2008.

[36] Indranil Mukhopadhyay, Eleanor Feingold, Daniel E. Weeks, and Anbupalam Thalamuthu. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epidemiology*, 34(3):213–221, 2010.

[37] Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, and Toshihiro Tanaka. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4):650–654, 2002.

[38] Yakir A. Reshef, Hilary K. Finucane, David R. Kelley, Alexander Gusev, Dylan Kotliar, Jacob C. Ulirsch, Farhad Hormozdiari, Joseph Nasser, Luke O'Connor, Bryce van de Geijn, Po Ru Loh, Sharon R. Grossman, Gaurav Bhatia, Steven Gazal, Pier Francesco Palamara, Luca Pinello, Nick Patterson, Ryan P. Adams, and Alkes L. Price. Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nature Genetics*, 50(10):1483–1493, 2018.

[39] G. K. Robinson. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, 6(1):15–32, 1991.

[40] Aswin Sekar, Allison R Bialas, Heather De Rivera, Avery Davis, Timothy R Hammond, Nolan Kamitaki, Katherine Tooley, Jessy Presumey, Matthew Baum, Vanessa Van

Doren, Giulio Genovese, Samuel A Rose, Robert E Handsaker, Mark J Daly, Michael C Carroll, Beth Stevens, and Steven A Mccarroll. Schizophrenia risk from complex variation of complement component 4 Schizophrenia Working Group of the Psychiatric Genomics Consortium HHS Public Access. *Nature*, 11(5307589):177–183, 2016.

[41] So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, Klaudia Walter, Cristina Menni, Lu Chen, Louella Vasquez, Ana M Valdes, Craig L Hyde, Vicky Wang, Daniel Ziemek, Phoebe Roberts, Li Xi, Elin Grundberg, Melanie Waldenberger, J Brent Richards, Robert P Mohney, Michael V Milburn, Sally L John, Jeff Trimmer, Fabian J Theis, John P Overington, Karsten Suhre, M Julia Brosnan, Christian Gieger, Gabi Kastenmüller, Tim D Spector, and Nicole Soranzo. An atlas of genetic influences on human blood metabolites. *Nature Genetics*, 46(6):543–550, jun 2014.

[42] Ilya Shlyakhter, Pardis C. Sabeti, and Stephen F. Schaffner. Cosi2: An efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30(23):3427–3429, 2014.

[43] Karsten Suhre, Matthias Arnold, Aditya Mukund Bhagwat, Richard J. Cotton, Rudolf Engelke, Johannes Raffler, Hina Sarwath, Gaurav Thareja, Annika Wahl, Robert Kirk Delisle, Larry Gold, Marija Pezer, Gordan Lauc, Mohammed A.El Din Selim, Dennis O. Mook-Kanamori, Eman K. Al-Dous, Yasmin A. Mohamoud, Joel Malek, Konstantin Strauch, Harald Grallert, Annette Peters, Gabi Kastenmüller, Christian Gieger, and Johannes Graumann. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications*, 8, 2017.

[44] Benjamin B. Sun, Joseph C. Maranville, James E. Peters, David Stacey, James R. Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, Clare Oliver-Williams, Mihir A. Kamat, Bram P. Prins, Sheri K. Wilcox, Erik S. Zimmerman, An Chi, Narinder Bansal, Sarah L. Spain, Angela M. Wood, Nicholas W. Morrell, John R. Bradley, Nebojsa Janjic, David J. Roberts, Willem H. Ouwehand, John A. Todd, Nicole Soranzo, Karsten Suhre, Dirk S. Paul, Caroline S. Fox, Robert M. Plenge, John Danesh, Heiko Runz, and Adam S. Butterworth. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018.

[45] Atsushi Takata, Naomichi Matsumoto, and Tadafumi Kato. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature Communications*, 8, 2017.

[46] Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A.Gagliano Taliun, André Corvelo, Stephanie M. Gogarten, Hyun Min Kang, Achilleas N. Pitsillides, Jonathon LeFaive, Seung been Lee, Xiaowen

Tian, Brian L. Browning, Sayantan Das, Anne Katrin Emde, Wayne E. Clarke, Douglas P. Loesch, Amol C. Shetty, Thomas W. Blackwell, Albert V. Smith, Quenna Wong, Xiaoming Liu, Matthew P. Conomos, Dean M. Bobo, François Aguet, Christine Albert, Alvaro Alonso, Kristin G. Ardlie, Dan E. Arking, Stella Aslibekyan, Paul L. Auer, John Barnard, R. Graham Barr, Lucas Barwick, Lewis C. Becker, Rebecca L. Beer, Emelia J. Benjamin, Lawrence F. Bielak, John Blangero, Michael Boehnke, Donald W. Bowden, Jennifer A. Brody, Esteban G. Burchard, Brian E. Cade, James F. Casella, Brandon Chalazan, Daniel I. Chasman, Yii Der Ida Chen, Michael H. Cho, Seung Hoan Choi, Mina K. Chung, Clary B. Clish, Adolfo Correa, Joanne E. Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn L. DeMeo, Susan K. Dutcher, Patrick T. Ellinor, Leslie S. Emery, Celeste Eng, Diane Fatkin, Tasha Fingerlin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie M. Fullerton, Soren Germer, Mark T. Gladwin, Daniel J. Gottlieb, Xiuqing Guo, Michael E. Hall, Jiang He, Nancy L. Heard-Costa, Susan R. Heckbert, Marguerite R. Irvin, Jill M. Johnsen, Andrew D. Johnson, Robert Kaplan, Sharon L.R. Kardia, Tanika Kelly, Shannon Kelly, Eimear E. Kenny, Douglas P. Kiel, Robert Klemmer, Barbara A. Konkle, Charles Kooperberg, Anna Köttgen, Leslie A. Lange, Jessica Lasky-Su, Daniel Levy, Xihong Lin, Keng Han Lin, Chunyu Liu, Ruth J.F. Loos, Lori Garman, Robert Gerszten, Steven A. Lubitz, Kathryn L. Lunetta, Angel C.Y. Mak, Ani Manichaikul, Alisa K. Manning, Rasika A. Mathias, David D. McManus, Stephen T. McGarvey, James B. Meigs, Deborah A. Meyers, Julie L. Mikulla, Mollie A. Minear, Braxton D. Mitchell, Sanghamitra Mohanty, May E. Montasser, Courtney Montgomery, Alanna C. Morrison, Joanne M. Murabito, Andrea Natale, Pradeep Natarajan, Sarah C. Nelson, Kari E. North, Jeffrey R. O'Connell, Nicholette D. Palmer, Nathan Pankratz, Gina M. Peloso, Patricia A. Peyser, Jacob Pleiness, Wendy S. Post, Bruce M. Psaty, D. C. Rao, Susan Redline, Alexander P. Reiner, Dan Roden, Jerome I. Rotter, Ingo Ruczinski, Chloé Sarnowski, Sebastian Schoenherr, David A. Schwartz, Jeong Sun Seo, Sudha Seshadri, Vivien A. Sheehan, Wayne H. Sheu, M. Benjamin Shoemaker, Nicholas L. Smith, Jennifer A. Smith, Nona Sotoodehnia, Adrienne M. Stilp, Weihong Tang, Kent D. Taylor, Marilyn Telen, Timothy A. Thornton, Russell P. Tracy, David J. Van Den Berg, Ramachandran S. Vasan, Karine A. Viaud-Martinez, Scott Vrieze, Daniel E. Weeks, Bruce S. Weir, Scott T. Weiss, Lu Chen Weng, Cristen J. Willer, Yingze Zhang, Xutong Zhao, Donna K. Arnett, Allison E. Ashley-Koch, Kathleen C. Barnes, Eric Boerwinkle, Stacey Gabriel, Richard Gibbs, Kenneth M. Rice, Stephen S. Rich, Edwin K. Silverman, Pankaj Qasba, Weiniu Gan, Namiko Abe, Laura Almasy, Seth Ament, Peter Anderson, Pramod Anugu, Deborah Applebaum-Bowden, Tim Assimes, Dimitrios Avramopoulos, Emily Barron-Casella, Terri Beaty, Gerald Beck, Diane Becker, Amber Beitelshees, Takis Benos, Marcos Bezerra, Joshua Bis, Russell Bowler, Ulrich Broeckel, Jai Broome, Karen Bunting, Carlos Bustamante, Erin Buth, Jonathan Cardwell, Vincent Carey, Cara Carty, Richard Casaburi, Peter Castaldi, Mark Chaffin, Christy Chang, Yi Cheng Chang, Sameer Chavan, Bo Juen Chen, Wei Min Chen, Lee Ming Chuang,

Ren Hua Chung, Suzy Comhair, Elaine Cornell, Carolyn Crandall, James Crapo, Jeffrey Curtis, Coleen Damcott, Sean David, Colleen Davis, Lisa de las Fuentes, Michael DeBaun, Ranjan Deka, Scott Devine, Qing Duan, Ravi Duggirala, Jon Peter Durda, Charles Eaton, Lynette Ekunwe, Adel El Boueiz, Serpil Erzurum, Charles Farber, Matthew Flickinger, Myriam Fornage, Chris Frazar, Mao Fu, Lucinda Fulton, Shanshan Gao, Yan Gao, Margery Gass, Bruce Gelb, Xiaoqi Priscilla Geng, Mark Geraci, Auyon Ghosh, Chris Gignoux, David Glahn, Da Wei Gong, Harald Goring, Sharon Graw, Daniel Grine, C. Charles Gu, Yue Guan, Namrata Gupta, Jeff Haessler, Nicola L. Hawley, Ben Heavner, David Herrington, Craig Hersh, Bertha Hidalgo, James Hixson, Brian Hobbs, John Hokanson, Elliott Hong, Karin Hoth, Chao Agnes Hsiung, Yi Jen Hung, Haley Huston, Chii Min Hwu, Rebecca Jackson, Deepti Jain, Min A. Jhun, Craig Johnson, Rich Johnston, Kimberly Jones, Sekar Kathiresan, Alyna Khan, Wonji Kim, Greg Kinney, Holly Kramer, Christoph Lange, Ethan Lange, Leslie Lange, Cecelia Laurie, Meryl LeBoff, Jiwon Lee, Seunggeun Shawn Lee, Wen Jane Lee, David Levine, Joshua Lewis, Xiaohui Li, Yun Li, Henry Lin, Honghuang Lin, Keng Han Lin, Simin Liu, Yongmei Liu, Yu Liu, James Luo, Michael Mahaney, Barry Make, Jo Ann Manson, Lauren Margolin, Lisa Martin, Susan Mathai, Susanne May, Patrick McArdle, Merry Lynn McDonald, Sean McFarland, Daniel McGoldrick, Caitlin McHugh, Hao Mei, Luisa Mestroni, Nancy Min, Ryan L. Minster, Matt Moll, Arden Moscati, Solomon Musani, Stanford Mwasongwe, Josyf C. Mychaleckyj, Girish Nadkarni, Rakhi Naik, Take Naseri, Sergei Nekhai, Bonnie Neltner, Heather Ochs-Balcom, David Paik, James Pankow, Afshin Parsa, Juan Manuel Peralta, Marco Perez, James Perry, Ulrike Peters, Lawrence S. Phillips, Toni Pollin, Julia Powers Becker, Meher Preethi Boorgula, Michael Preuss, Dandi Qiao, Zhaohui Qin, Nicholas Rafaels, Laura Raffield, Laura Rasmussen-Torvik, Aakrosh Ratan, Robert Reed, Elizabeth Regan, Muagututi'a Sefuiva Reupena, Carolina Roselli, Pamela Russell, Sarah Ruuska, Kathleen Ryan, Ester Cerdeira Sabino, Danish Saleheen, Shabnam Salimi, Steven Salzberg, Kevin Sandow, Vijay G. Sankaran, Christopher Scheller, Ellen Schmidt, Karen Schwander, Frank Sciurba, Christine Seidman, Jonathan Seidman, Stephanie L. Sherman, Aniket Shetty, Wayne Hui Heng Sheu, Brian Silver, Josh Smith, Tanja Smith, Sylvia Smoller, Beverly Snively, Michael Snyder, Tamar Sofer, Garrett Storm, Elizabeth Streeten, Yun Ju Sung, Jody Sylvia, Adam Szpiro, Carole Sztalryd, Hua Tang, Margaret Taub, Matthew Taylor, Simeon Taylor, Machiko Threlkeld, Lesley Tinker, David Tirschwell, Sarah Tishkoff, Hemant Tiwari, Catherine Tong, Michael Tsai, Dhananjay Vaidya, Peter VandeHaar, Tarik Walker, Robert Wallace, Avram Walts, Fei Fei Wang, Heming Wang, Karol Watson, Jennifer Wessel, Kayleen Williams, L. Keoki Williams, Carla Wilson, Joseph Wu, Huichun Xu, Lisa Yanek, Ivana Yang, Rongze Yang, Norann Zaghloul, Maryam Zekavat, Snow Xueyan Zhao, Wei Zhao, Degui Zhi, Xiang Zhou, Xiaofeng Zhu, George J. Papanicolaou, Deborah A. Nickerson, Sharon R. Browning, Michael C. Zody, Sebastian Zöllner, James G. Wilson, L. Adrienne Cupples, Cathy C. Laurie, Cashell E. Jaquish, Ryan D. Hernandez, Timothy D. O'Connor, and Gonçalo R. Abecasis. Sequencing of

53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299, 2021.

[47] Ashley K. Tehranchi, Marsha Myrthil, Trevor Martin, Brian L. Hie, David Golan, and Hunter B. Fraser. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell*, 165(3):730–741, 2016.

[48] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, Johan L.M. Björkegren, Hae Kyung Im, Bogdan Pasaniuc, Manuel A. Rivas, and Anshul Kundaje. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*, 51(4):592–599, 2019.

[49] Kai Wang, Haitao Zhang, Subra Kugathasan, Vito Annese, Jonathan P. Bradfield, Richard K. Russell, Patrick M.A. Sleiman, Marcin Imielinski, Joseph Glessner, Cuiping Hou, David C. Wilson, Thomas Walters, Cecilia Kim, Edward C. Frackelton, Paolo Lionetti, Arrigo Barabino, Johan Van Limbergen, Stephen Guthery, Lee Denson, David Piccoli, Mingyao Li, Marla Dubinsky, Mark Silverberg, Anne Griffiths, Struan F.A. Grant, Jack Satsangi, Robert Baldassano, and Hakon Hakonarson. Diverse Genome-wide Association Studies Associate the IL12/IL23 Pathway with Crohn Disease. *American Journal of Human Genetics*, 84(3):399–405, 2009.

[50] Yan Wang, Jing Liu, Bo Huang, Yan-Mei Xu, Jing Li, Lin-Feng Huang, Jin Lin, Jing Zhang, Qing-Hua Min, Wei-Ming Yang, and Xiao-Zhong Wang. Mechanism of alternative splicing and its regulation. *Biomedical Reports*, 3(2):152–158, mar 2015.

[51] Michael C Wu, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin. Powerful SNP-set analysis for case-control genome-wide association studies. *American journal of human genetics*, 86(6):929–42, jun 2010.

[52] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93, 2011.

[53] Michael C. Wu, Arnab Maity, Seunggeun Lee, Elizabeth M. Simmons, Quaker E. Harmon, Xinyi Lin, Stephanie M. Engel, Jeffrey J. Molldrem, and Paul M. Armistead. Kernel machine SNP-set testing under multiple candidate kernels. *Genetic Epidemiology*, 37(3):267–275, 2013.

[54] Xiaoji Wu and Yi Zhang. TET-mediated active DNA demethylation: Mechanism, function and beyond. *Nature Reviews Genetics*, 18(9):517–534, 2017.

[55] Qi Yan, Nianjun Liu, Erick Forno, Glorisa Canino, Juan C. Celedón, and Wei Chen. An integrative association method for omics data based on a modified Fisher's method with application to childhood asthma. *PLoS Genetics*, 15(5):1–18, 2019.

[56] Jian Yang, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82, jan 2011.

[57] Xiang Zhan, Ni Zhao, Anna Plantinga, Timothy A Thornton, Karen N Conneely, Michael P Epstein, and Michael C Wu. Powerful Genetic Association Analysis for Common or Rare Variants with High-Dimensional Structured Traits. 206(August):1779–1790, 2017.

[58] Ni Zhao, Jun Chen, Ian M. Carroll, Tamar Ringel-Kulka, Michael P. Epstein, Hua Zhou, Jin J. Zhou, Yehuda Ringel, Hongzhe Li, and Michael C. Wu. Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. *American Journal of Human Genetics*, 96(5):797–807, 2015.

[59] Ni Zhao, Xiang Zhan, Yen Tsung Huang, Lynn M. Almli, Alicia Smith, Michael P. Epstein, Karen Conneely, and Michael C. Wu. Kernel machine methods for integrative analysis of genome-wide methylation and genotyping studies. *Genetic Epidemiology*, 42(2):156–167, 2018.

[60] Ni Zhao, Haoyu Zhang, Jennifer J. Clark, Arnab Maity, and Michael C. Wu. Composite kernel machine regression based on likelihood ratio test for joint testing of genetic and gene–environment interaction effect. *Biometrics*, 75(2):625–637, 2019.

[61] Xiuwen Zheng, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie, and Bruce S. Weir. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24):3326–3328, 2012.

Appendix A

CONNECTION BETWEEN KERNEL MACHINE REGRESSION MODEL AND LINEAR MIXED MODEL

According to Mercer's Theorem, a continuous positive semidefinite kernel k has an associated orthonormal basis $\{e_i\}_i$ of L^2 consisting of eigenfunctions of the Hilbert-Schmidt integral operator. By the Representer theorem, f can be expressed as a linear combination of the orthonormal basis functions. We can use the dual representation of $f(\boldsymbol{Z}) = \sum_{i=1}^{B} \alpha_i k(\boldsymbol{Z}_i^*, \boldsymbol{Z})$, for some integer B. Now we use penalized generalized least squares to estimate kernel machine regression (KMR) parameters.

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}} &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\operatorname{argmax}} \left\{ -\tfrac{1}{2} \left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{\alpha})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}' \boldsymbol{K} \boldsymbol{\alpha} \right] \right\} \\
&= (\lambda\boldsymbol{\Sigma} + \boldsymbol{K})^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \right)
\end{aligned}$$

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\gamma}\in\mathbb{R}^q}{\operatorname{argmax}} \left\{ -\tfrac{1}{2} \left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{\alpha})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}' \boldsymbol{K} \boldsymbol{\alpha} \right] \right\} \\
&= \left\{ \boldsymbol{X}' (\lambda\boldsymbol{\Sigma} + \boldsymbol{K})^{-1} \boldsymbol{X} \right\}^{-1} \boldsymbol{X}' (\lambda\boldsymbol{\Sigma} + \boldsymbol{K})^{-1} \boldsymbol{y}
\end{aligned}$$

$$\begin{aligned}
\widehat{\boldsymbol{f}} &= \boldsymbol{K}\widehat{\boldsymbol{\alpha}} \\
&= \boldsymbol{K} (\lambda\boldsymbol{\Sigma} + \boldsymbol{K})^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})
\end{aligned}$$

Now consider the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{A.1}$$

where $\boldsymbol{f}$ is an $n \times 1$ vector of random effects with distribution $N(\mathbf{0}, \lambda^{-1}\boldsymbol{K})$, for some non-negative constant λ^{-1}. Then,

$$\text{var}\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \frac{1}{\lambda}\boldsymbol{K} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \end{bmatrix}.$$

Assume $\boldsymbol{K}$ and $\boldsymbol{\Sigma}$ are known positive definite matrices. The BLUP estimates [39] $\widehat{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\beta}}$ are given by

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{\Sigma}^{-1} + \lambda\boldsymbol{K}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{f}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{y} \end{bmatrix}.$$

Rearranging,

$$\widehat{\boldsymbol{\beta}} = \left\{\boldsymbol{X}'\left(\lambda\boldsymbol{\Sigma} + \boldsymbol{K}\right)^{-1}\boldsymbol{X}\right\}^{-1}\boldsymbol{X}'\left(\lambda\boldsymbol{\Sigma} + \boldsymbol{K}\right)^{-1}\boldsymbol{y}$$

and

$$\widehat{\boldsymbol{f}} = \boldsymbol{K}\left(\lambda\boldsymbol{\Sigma} + \boldsymbol{K}\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}).$$

The kernel regularized least squares problem results in the same BLUPs of the corresponding linear mixed model.

Let $\tau = \lambda^{-1}$. Leveraging the linear mixed model representation of the problem, we can derive a variance component score test statistic for the null hypothesis that $\boldsymbol{y}$ does not depend on $\boldsymbol{f}$. This is equivalent to testing

$$H_0 : \tau = 0 \text{ vs. } H_1 : \tau > 0.$$

The variance of $\boldsymbol{y}$ is given by

$$\text{cov}(\boldsymbol{y}) = \tau\boldsymbol{K} + \boldsymbol{\Sigma} \equiv \boldsymbol{V}.$$

Then, the log likelihood is

$$l(\tau; \boldsymbol{\beta}, \sigma_a^2, \sigma_e^2) = C - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Then, the score for τ is

$$U_\tau(\tau, \boldsymbol{\beta}, \sigma_a^2, \sigma_e^2) = -\frac{1}{2}\text{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{K}\right) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}\boldsymbol{K}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Thus, under the null hypothesis that $\tau = 0$,

$$U_\tau(0, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}_e^2, \widehat{\sigma}_a^2) \;\; = -\tfrac{1}{2}\text{tr}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\right) + (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

where Σ and $\widehat{\boldsymbol{\beta}}$ are estimated under the null hypothesis. Since the first term of the score for τ doesn't depend on the data, $\boldsymbol{y}$, it can be dropped.

Under the null hypothesis,

$$\begin{aligned} Q &= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'\widehat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad \text{(Spectral Decomposition)} \end{aligned}$$

where $\boldsymbol{P}_0 = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'$. Let $\boldsymbol{z} = \boldsymbol{Q}'\widehat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Then $\boldsymbol{z} \sim N_n(\boldsymbol{0}, \boldsymbol{Q}'\boldsymbol{Q}) \overset{d}{=} N_n(\boldsymbol{0}, \boldsymbol{I})$ since $\boldsymbol{Q}$ is an orthogonal matrix. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of

$$\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1/2}.$$

$$\begin{aligned} Q &= \boldsymbol{z}'\boldsymbol{\Lambda}\boldsymbol{z} \\ &= \begin{pmatrix} z_1 & z_2 & \cdots & z_n \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \\ &= \sum_{i=1}^{n} \lambda_i z_i^2 \\ &\sim \sum_{i=1}^{n} \lambda_i \chi_{1,i}^2 \end{aligned}$$

Asymptotically, $Q \xrightarrow{d} \sum_{i=1}^{n} \lambda_i \chi^2_{1,i}$.

Appendix B

P-VALUE CALCULATION USING THE PERTURBATION APPROACH

The perturbation p-value calculation hinges on the observation that the test statistic consists of two normal distributed variables sandwiching a real, symmetric matrix. For a fixed value of $\omega_j \in \{\omega_1, \ldots, \omega_L\}$:

$$\begin{aligned} Q(\omega_j) &= \widehat{\boldsymbol{\epsilon}}' \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\epsilon}} \\ &= \boldsymbol{\epsilon}' \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\epsilon} \\ &\equiv \boldsymbol{w}' \widehat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{w} \end{aligned} \tag{B.1}$$

We decompose the matrix $\widehat{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{K}(\omega_j) \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{P}_0 \widehat{\boldsymbol{\Sigma}}^{-1/2}$ as $\boldsymbol{V}_j \boldsymbol{\Lambda}_j \boldsymbol{V}_j'$. Now,

$$Q(\omega_j) = \boldsymbol{w}' \boldsymbol{V}_j \boldsymbol{\Lambda}_j \boldsymbol{V}_j' \boldsymbol{w}. \tag{B.2}$$

Denote

$$\boldsymbol{\Omega} \equiv \operatorname{cov} \begin{pmatrix} \boldsymbol{V}_1' \boldsymbol{w} \\ \boldsymbol{V}_2' \boldsymbol{w} \\ \vdots \\ \boldsymbol{V}_N' \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{V}_1' \boldsymbol{V}_2 & \ldots & \boldsymbol{V}_1' \boldsymbol{V}_N \\ \boldsymbol{V}_2' \boldsymbol{V}_1 & \boldsymbol{I} & \ldots & \boldsymbol{V}_2' \boldsymbol{V}_N \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{V}_N' \boldsymbol{V}_1 & \boldsymbol{V}_N' \boldsymbol{V}_2 & \ldots & \boldsymbol{I} \end{pmatrix}. \tag{B.3}$$

Let m_j be the number of columns of $\boldsymbol{V}_j$, and let $m \equiv \sum_{j=1}^{L} m_j$. Decompose $\boldsymbol{\Omega} = \boldsymbol{R}\boldsymbol{R}'$. The perturbation sampling procedure proceeds as follows:

1. Sample a m length normal distributed vector, $\boldsymbol{r}$.

2. Rotate the vector $\boldsymbol{r}$ by $\boldsymbol{R}$. In other words, construct $\boldsymbol{r}^* = \boldsymbol{R}\boldsymbol{r}$. Now, $\boldsymbol{r}^*$ has covariance

$\boldsymbol{\Omega}$.

3. Construct $Q_j^* = \boldsymbol{r}_j^{*\prime} \boldsymbol{\Lambda}_j \boldsymbol{r}_j^*$. Where $\boldsymbol{r}_j^*$ is a m_j length vector defined as follows:

$$\boldsymbol{r}_j^* = \begin{pmatrix} r_k^* & r_{k+1}^* & \dots & r_{k+m_j-1}^* \end{pmatrix}',$$

where $k = \sum_{i=1}^{j-1} m_i + 1$. Obtain the p-value, p_j^* corresponding to Q_j^*. Note that Q_j^* shares the same asymptotic distribution as Q_j, so the mixture probabilities for the mixture of chi-square distribution are based on the eigenvalues in $\boldsymbol{\Lambda}_j$. Set $p_0^* = \min_{1 \le j \le L} p_j^*$.

4. Repeat steps 1-3 a large number of times, B, to obtain $\left\{ p_{0(1)}^*, p_{0(2)}^*, \dots, p_{0(B)}^* \right\}$.

5. The final p-value for significance is

$$p = \frac{1}{B} \sum_{b=1}^{B} I\left(p_0 \ge p_{0(b)}^* \right).$$

Appendix C

P-VALUE CALCULATION USING KERNEL PCA

The kernel PCA p-value calculation construction a test statistic that is the sum of two independent terms. For a fixed value of $\omega_j \in \{\omega_1, \ldots, \omega_L\}$

$$Q^*(\omega_j) = \omega_j Q_1 + (1 - \omega_j) Q_2^*.$$

We select the minimum p-value amongst the L candidate kernel matrices:

$$p_0 = \min_{1 \le j \le L} p_j.$$

Now, to get a final p-value for the hypothesis test, we compare p_0 to the distribution of minimum p-values amongst L candidate kernel matrices arising from weights $\{\omega_1, \ldots, \omega_L\}$ when the null hypothesis is true. Let random variable $p_{H_0,j}$ be the p-value from a hypothesis test using composite kernel matrix $K^*(\omega_j)$, when the omics data are truly not associated with the trait. Then, the final p-value for our hypothesis test is

$$\begin{aligned} p &= P\left[\min_{1 \le j \le L} p_{H_0,j} \le p_0\right] \\ &= 1 - P\left[\min_{1 \le j \le L} p_{H_0,j} > p_0\right] \\ &= 1 - P\left[p_{H_0,1} > p_0, \ldots, p_{H_0,L} > p_0\right] \end{aligned}$$

Remember for $\boldsymbol{K}^*(\omega_j)$, if $q^*_{H_0,j}$ is the corresponding test statistic, we calculate $p_{H_0,j}$ as

$$p_{H_0,j} = P\left[Q_j^* > q^*_{H_0,j}\right],$$

where, when the null hypothesis is true, Q_j^* follows a mixture of chi-square distribution with mixture probabilities depending on the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}^*(\omega_j)\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1/2}$. Therefore, $P[p_{H_0,j} > p_0] = P[q^*_{H_0,j} < q_j]$, where q_j is the $(1-p_0)^{\text{th}}$ quantile of the distribution of Q_j^*. Thus,

$$\begin{aligned}
p &= 1 - P\left[Q_1^* < q_1, \ldots, Q_L^* < q_L\right] \\
&= 1 - P\left[\omega_1 k_1 + (1-\omega_1)k_2 < q_1, \ldots, Q_L^* < q_L\right] \\
&= 1 - P\left[k_1 < \tfrac{q_1-(1-\omega_1)k_2}{\omega_1}, \ldots, k_1 < \tfrac{q_L-(1-\omega_L)k_2}{\omega_L}\right] \\
&= 1 - P\left[k_1 < \min_{1\le j\le L} \tfrac{q_j-(1-\omega_j)k_2}{\omega_j}\right] \\
&= 1 - \mathrm{E}\left[P\left[k_1 < \min_{1\le j\le L} \tfrac{q_j-(1-\omega_j)k_2}{\omega_j}\,\middle|\, k_2\right]\right]
\end{aligned}$$

Here, k_1 follows a mixture of chi-square distribution when the null hypothesis is true, with mixture probabilities based on the eigenvalues of $\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{K}_1\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{P}_0\widehat{\boldsymbol{\Sigma}}^{-1/2}$.