THE UNIVERSITY OF CHICAGO

ASSESSMENTS OF SIGNIFICANCE FOR GENETIC ASSOCIATION ANALYSIS IN STRUCTURED SAMPLES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
JOELLE MBATCHOU

CHICAGO, ILLINOIS
JUNE 2019

To the many empowering women in my life

There is one quality which one must possess to win, and that is definiteness of purpose, the knowledge of what one wants, and a burning desire to possess it.

– Napoleon Hill

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This has been quite a journey of personal and professional growth, in which I had the opportunity to cross paths with many individuals who helped push me further along. As it is often said that it takes a village to raise a child, I believe that my growth has been the combined effort of many individuals whom I want to acknowledge and give thanks to here.

I want to thank my advisor Professor Mary Sara McPeek, who has been a great mentor to me over the years and continuously encouraged me in my growth as a scholar. I am deeply grateful for the many hours she made available to me either for meetings or to read over my research papers. Her incredible work ethics, keen insights and attention to details have inspired me in my growth as a research scholar.

I want to thank Professors Dan Nicolae and Mark Abney for being members of my dissertation committee and for their valuable comments. My growth as a scholar would not have been possible without the teachings from the faculty members in the department. Indeed, through the many classes I have taken, there has always been an emphasis on rigorous learning, which has been invaluable in this doctoral journey. I want to thank Professor Mei Wang for her continuous words of encouragement and support, as well as Dr. Linda Collins for the great guidance and support she gave me over the many times I worked with her both as an instructor and as a teaching assistant. I also want to thank Dr. Ed Friedman and John Zekos for all their quick and invaluable technical support. I am also grateful to the members of the departmental administrative staff for the time, patience and support they have given me over the years.

It has also been my privilege over the years to meet and learn with many fellow students. I would like to thank Cheng Gao, Swati, Marc Goessling, Andrew Poppick, Rishideep Roy, Vivak Patel, Sayar Karmakar, Kathleen Hui, Miaoyan Wang, Duo Jiang, Sheng Zhong, among many others, for the insightful discussions and supportive camaraderie.

I want to thank the many friends who have supported me on this journey, including Isra Omar, Sahima Ameer, Barakat Olorunoje, Marc Ndueyap, and Minyu Lv. I am so grateful

to Saritha Teralandur, Anna Delia McDonald, and Uyiosa Ruth Obaseki, who have all been pillars of support from the start of this PhD journey and have always encouraged me and believed in my potential.

My family has been my anchor through this chapter of my life. I am so grateful for the constant pouring of love and support I've received, often in moments where I was doubting myself and my strength. I want to thank my two aunts Christine and Catherine, who have been mothers to me as I stayed in Chicago. I also want to thank my cousin Samuela for her support and reminding me to stay balanced. I am forever grateful to my sister Sonya and my mom, for always empowering me to strive for better and to never give up.

My God is good.

ABSTRACT

In this dissertation, we develop methods to address several problems that arise in the assessment of significance for genetic association analysis of complex traits in samples with population structure and/or relatedness. In Chapter 2, we focus on phenotype resampling methods for binary trait analysis in structured samples. We develop BRASS, a permutation-based approach to testing association between a binary trait and an arbitrary predictor in samples with population structure and/or related individuals. This method is applicable in various contexts, including (1) correction for multiple comparisons when testing for region-wide or genome-wide significance and (2) assessment of significance for tests that combine test statistics that perform well in different scenarios. Previous methods are applicable only to analysis of a quantitative trait and do not perform well for a binary trait. BRASS allows for covariates, ascertainment and simultaneous testing of multiple markers, and it does not place strong restrictions on the test statistic used. We use an estimating equation approach that can be viewed as a hybrid of logistic regression and linear mixed-effects model methods, and we use a combination of principal components and a genetic relatedness matrix to account for sample structure. In simulation studies, we demonstrate that BRASS maintains correct control of type 1 error. We illustrate the proposed approach in two genome-wide analyses of binary traits in domestic dog.

In Chapter 3, we focus on assessment of significance in genetic association analysis of single or multi-dimensional phenotypes, including high-dimensional phenotypes, where we consider test statistics of a certain form, allow association to be tested with either a single genetic marker or with multiple genetic markers simultaneously, and where there is population structure and/or relatedness in the sample. Existing approaches that can be used to assess significance in this context are either computationally burdensome (permutation-based approaches), or do not perform well in settings such as small samples, high-dimensional traits, or misspecified phenotype model (asymptotic approximations based on prospective models), or require unrealistically accurate estimation of the correlation structure among

a set of rare variants in order to be applied with rare variants (retrospective methods), or require an assumption of second-order exchangeability of individuals' genotypes, possibly after correction for a few ancestry-informative covariates (existing moment-matching methods for detecting association of two matrices). We develop JASPER, which can be viewed as an extension of existing moment-matching methods for detecting association of two matrices, to allow very general population structure and relatedness in the sample. JASPER can be used for a reasonably broad class of test statistics currently used in genetic association analysis, including most linear mixed model-based score tests and kernel-based test statistics. Notable features of JASPER are that it (1) is insensitive to misspecification of the phenotype model, (2) does not require knowledge of the distribution of the test statistic under the null hypothesis, (3) allows population structure, related individuals, covariates, ascertainment, rare variants, and multiple traits, and (4) with rare variant mapping, it does not require knowledge of the correlation structure among the rare variants. Through simulation studies for a range of settings and test statistics, including for high-dimensional traits, we demonstrate that JASPER properly controls type 1 error in the presence of sample structure and can in some cases provide substantial power gains compared to large-sample-based assessments of significance. JASPER is applied in a study of the genetic regulation of gene expression levels within biological pathways in data from the Framingham Heart Study.

CHAPTER 1

INTRODUCTION

1.1 Genetic Association Studies

We first introduce some genetic terminology relevant for the rest of this dissertation.

The human genome is organized into 23 paired **chromosomes**, with 22 pairs of autosomal chromosomes and one pair of sex chromosomes. For the autosomal chromosome pairs, one is inherited from the father and the other is inherited from the mother, and for the pair of sex chromosomes, females inherit a copy of the X chromosome from each parent while males inherit only one copy of the X chromosome from the mother and one copy of the Y chromosome from the father. Each chromosome consists of two **DNA** strands that are sequences of **nucleotides**. A nucleotide contains one of four distinct bases, represented by A, T, C, and G, and the human genome overall contains over 3 billion bases. A **gene** is a sequence of nucleotides and constitutes the functional unit of heredity. In humans, genes vary in size from a few hundred bases to over 2 million bases, and there are about 20,000 genes in the human genome.

The major part of the human DNA sequence is identical across the population (pairs of individuals share about 99.5% of their DNA sequence). For the remaining segments of the DNA which vary among individuals and are called **polymorphic**, a change in the sequence across individuals in a population is termed a **genetic variant**, while the alternate forms of the sequence are known as **alleles**. For each individual, the collection of alleles at a position in the genome is known as the individual's **genotype**. A common type of genetic variant is referred to as **SNP** (single nucleotide polymorphism), which is a single base position mutation that differentiates people in the population. This is the most common type of genetic variation, and the majority of SNPs are bi-allelic, meaning there are only two different alleles in the population. For such SNPs, the allele that occurs at a lower frequency in the population is called the **minor allele**, and the population frequency of the minor allele is

called the minor allele frequency (**MAF)**. Based on their MAF, genetic variants are classified as either common or rare, where 1% is a common threshold used to define the two classes (i.e. variants with MAF less than 1% are classified as rare). A collection of loci (i.e. positions in the genome) is characterized as being in linkage disequilibrium (**LD**) if some combinations of their alleles tend to occur together more or less often than expected by chance.

A **phenotype** is a measured characteristic of an individual and can be either quantitative (e.g. height, weight) or categorical (e.g. presence or absence of some disease). Many phenotypes have been found to be influenced by both genetic and non-genetic factors. The proportion of the phenotypic variance in a population that can be attributed to genetic factors is called the **heritability**. This can be inferred from data using family-based study designs or using mixed effects regression models [Abney et al., 2000; Vattikuti et al., 2012; Yang et al., 2010].

The main objective of genetic association studies is to understand the genetic basis of complex human traits. Many of the studies aim to identify SNPs that are statistically associated with a phenotype of interest, where association is sometimes tested for each SNP one at a time, or for a set of SNPs simultaneously. More recently, interest has also been given to analyses involving multiple traits in order to explore the pleiotropic effects of genetic variants. The very large number of SNPs, the LD structure between them, the structure between individuals in the sample, which could be hidden, and, in some cases, the large number of traits, all constitute important challenges when testing for association. The goal of this dissertation is to address a few statistical challenges that arise in association studies.

1.2 Population and Pedigree Structure

A challenge commonly encountered in genetic association studies is the presence of confounding factors, which can lead to spurious association results. A main source of confounding is sample structure, which refers to relatedness present in the sample that is due to either known and/or unknown structure. An example of known sample structure is the inclusion of

related individuals in the sample, where Mendelian segregation induces correlation between the genotypes of individuals in the sample, and polygenic effects can create correlation in the phenotype of interest. Another important example of sample structure (either known or unknown) is population stratification, which arises from having individuals in the sample from multiple population subgroups. Alternatively, sample structure could also arise from the inclusion of individuals who genetically are varying mixtures of multiple ancestral populations, referred to as population admixture. Finally, an example of unknown sample structure is cryptic relatedness, where related individuals are present but the family relationships are not explicitly observed in the sample.

Sample structure produces dependency among observations in the sample, which can be among the genotypes, among the phenotypes, and/or between genotypes and phenotypes of different individuals. If ignored, it can compromise the performance of methods for association testing that assume that observations are independent. Incorporating the sample structure in an association testing method can improve its performance, both in the control of the type 1 error and in the statistical power [Jakobsdottir and McPeek, 2013; Jiang et al., 2016; Thornton and McPeek, 2007; Zhong et al., 2016]. In Chapter 2, we develop a novel method which can be used for binary trait mapping and in Chapter 3, we develop a method which can be used for either marginal or joint analysis of binary or quantitative traits, where both methods proposed adjust for sample structure which includes population structure, admixture and/or cryptic relatedness.

1.3 Linear Mixed Model to Adjust for Sample Structure

We give a brief review of a very common approach to adjust for the genetic similarities between individuals in the sample, which is the use of linear mixed models (LMM). Indeed, they have proven to be quite adept at handling different types of confounding, including relatedness and population structure [Kang et al., 2008; Lippert et al., 2011; Price et al., 2010b; Zhou and Stephens, 2012]. In these models, the structure is incorporated through the

use of random effects whose covariance matrix reflects the dependency present in the sample.

For a sample of n individuals, which can include known or unknown structure, in the absence of major genes, the value of the trait for individual i, denoted by Y_i, is modeled as

$$Y_i = \mathbf{X}_i^T \boldsymbol{\gamma} + u_i + e_i, \tag{1.1}$$

where $\mathbf{X}_i$ is a length k vector of covariates, $\boldsymbol{\gamma}$ is a length k vector of unknown fixed covariate effects, $\mathbf{u} = (u_1, \ldots, u_n)^T$ is a vector of random effects that account for correlation in the phenotype values due to relatedness, and the e_i's are i.i.d. $N(0, \sigma^2)$ random effects that represent environmental effects. It is common for LMMs to assume that $\text{Var}(\mathbf{u}) = \sigma^2 \boldsymbol{\Phi}$, where the $n \times n$ matrix $\boldsymbol{\Phi}$ is referred to as a genetic relatedness matrix (GRM) or kinship matrix.

Alternatively, this model can also be derived from Fisher's polygenic model, which considers the effects of "many small, equal and additive loci" on the phenotype distribution [Fisher, 1919]. More precisely, for a set of M independent markers, the phenotype value of individual i is modeled as

$$Y_i = \mathbf{X}_i^T \boldsymbol{\gamma} + \sum_{j=1}^{M} \beta_j \widetilde{G}_{ij} + e_i, \tag{1.2}$$

where $\widetilde{G}_{ij}$ is a standardized version of G_{ij}, which is the genotype of individual i at the j^{th} marker (i.e. $\widetilde{G}_{ij} = \frac{G_{ij} - \mu_j}{\sigma_j}$ where μ_j and σ_j are the mean and standard deviation of G_{ij}, respectively), β_j's are independent random effects with $\text{E}(\beta_j)$=0 and $\text{Var}(\beta_j) = \sigma_a^2/M$, and the e_i's are independent random effects with $e_i \sim N(0, \sigma^2)$. When the number of loci M is large, the distribution of the combined genetic effect $\sum_{j=1}^{M} \beta_j \widetilde{G}_{ij}$ can be approximated by a $N(0, \sigma_a^2 \boldsymbol{\Phi})$ random variable, with

$$\boldsymbol{\Phi} = \lim_{M \to \infty} \frac{1}{M} \sum_{j=1}^{M} \frac{(\mathbf{G}_j - \mu_j \mathbf{1}_n)(\mathbf{G}_j - \mu_j \mathbf{1}_n)^T}{\sigma_j^2}, \tag{1.3}$$

where $\mathbf{G}_j = (G_{1j}, \ldots, G_{nj})^T$. The matrix $\boldsymbol{\Phi}$ can be thought of as a covariance matrix based

on genotypes in the whole genome that measures genetic similarities between individuals in the sample. In practice, the number of independent SNPs is not infinite, and so $\mathbf{\Phi}$ is unknown. However, when the samples consists of pedigrees, $\mathbf{\Phi}$ can be obtained from the expected amount of genetic sharing given the pedigree structure, which requires knowing kinship and inbreeding coefficients for the pedigree. The kinship coefficient between individuals i and j, denoted by ϕ_{ij}, is a measure of the relatedness between the two individuals and is defined as the probability that at an arbitrary location in the genome, an allele randomly selected from i and one from j are inherited copies of the same founder allele. The inbreeding coefficient of an individual i, denoted by h_i, is a measure of relatedness within a single individual and is defined as the probability that an individual's two alleles at an arbitrary locus are copies of the same founder allele. Given the pedigree structure, recursive algorithms are available to obtain these coefficients [Boyce, 1983].

When the sample structure is unknown, genotype data at a large number of genetic variants can be used to obtain an empirical estimate of $\mathbf{\Phi}$. For example, a commonly used estimate for $\mathbf{\Phi}$ based on genome-scan data is

$$\widehat{\mathbf{\Phi}} = \frac{1}{M}\sum_{j=1}^{M}\frac{(\mathbf{G}_j - 2\hat{p}_j\mathbf{1}_n)(\mathbf{G}_j - 2\hat{p}_j\mathbf{1}_n)^T}{2\hat{p}_j(1-\hat{p}_j)}, \tag{1.4}$$

where $\hat{p}_j$ is an unbiased estimator of p_j, which is the population allele frequency at marker j (e.g. $\hat{p}_j = \bar{\mathbf{G}}_j/2$, which is the sample mean estimator), and we assume that the two alleles of an individual are independent draws from the same distribution (i.e. Hardy-Weinberg equilibrium) so that $\mathrm{Var}(G_{ij}) = 2p_j(1-p_j)$ for $i = 1, \ldots, n$. Dependence in the sample can also be modeled through fixed effects (e.g. with ancestry-informative covariates) or a mixture of both fixed and random effects. We explore the latter option in Chapter 2 of the thesis.

CHAPTER 2

PERMUTATION METHODS FOR ASSESSING SIGNIFICANCE IN BINARY TRAIT ASSOCIATION MAPPING WITH STRUCTURED SAMPLES

2.1 Introduction

To elucidate the genetic architecture of complex traits in either human populations or model organisms such as mouse, dog or cattle, many studies have used genome-wide association (GWA) analyses. In these GWA studies, the primary objective has been to identify associations between a phenotype of interest and genetic markers, usually SNPs. This involves assessing the statistical significance of a given test statistic by deriving its null distribution or an asymptotic approximation to it. However, this is not always feasible as the distribution may be intractable. Such a scenario can arise in region-based tests where association signals over multiple sites are combined (e.g. rare variant tests [Lee et al., 2014]), or when the test statistic involves data-adaptive weights [Price et al., 2010a], or for tests that combine test statistics that perform well in different alternative models [Jiang and Mcpeek, 2014; Lee et al., 2012a]. A further limitation can arise when, even if the distribution (or the asymptotic distribution) of the test statistic is known for single tests, significance needs to be assessed for the maximum of many correlated tests. This occurs in genome scans to establish a genome-wide significance threshold, where the linkage disequilibrium present between the markers induces correlation between the association tests [Dudbridge and Gusnanto, 2008].

To overcome these limitations, a common approach is to perform permutation testing so as to obtain replicates of the data under the null hypothesis from which an empirical null distribution can be derived. A fundamental assumption underlying permutation testing is exchangeability of the subjects in the sample, which might be approximately satisfied in some cases. However, this assumption can be violated in the presence of population structure

and/or related individuals because they introduce correlation in the sample, including in the phenotype values through polygenic effects [Abney et al., 2002; Astle and Balding, 2009]. In such cases, naive application of permutation testing will usually not preserve the correlation structure and can result in inflated type 1 error rates [Churchill and Doerge, 2008] (though this can be avoided if, for example, all subjects in the sample are equally related [Abney, 2015]). We consider here the problem of permutation testing for a binary trait in the presence of polygenic effects in a sample with population structure, cryptic and/or family relatedness. While an exact permutation test may not be feasible, a permutation-based test that adjusts for the correlation structure is feasible and has been proposed for quantitative traits [Abney, 2015; Abney et al., 2002]. The approach is based on a linear mixed model (LMM) and is asymptotically valid for multivariate-normal data. It incorporates genetic relatedness through the inclusion of random effects in the LMM and can also adjust for covariates. However, as it is primarily designed for quantitative traits that are multivariate-normal, there will be model misspecification when applied to binary traits as the binary nature of the data is not incorporated into the LMM.

We propose BRASS (for "Binary trait Resampling method Adjusting for Sample Structure"), a permutation procedure for a binary trait which incorporates both covariates and the correlation structure present in the sample. In contrast to the LMM-based approach, it accommodates the binary nature of the trait through a quasi-likelihood framework that considers the effect of covariates on a logit scale in the mean structure as well as the relationship between the trait mean and its variance, both of which are important features of binary data. Here, we show that the permutation-based replicates obtained with BRASS are able to maintain correct type 1 error control and compare it with several alternative approaches. As the use of our method requires accounting for the phenotypic correlation structure, we go over modeling choices for the relatedness and how it can be incorporated in the quasi-likelihood framework used. We demonstrate the use of our method in a context of assessing genome-wide significance in domestic dog GWA studies. In animal studies,

permutation testing is commonly used when there is not a consensus threshold for genome-wide significance [Bianchi et al., 2015; Melin et al., 2016; Safra et al., 2013; Tengvall et al., 2013]. We apply BRASS to perform association mapping in two GWA studies of domestic dogs, one for elbow dysplasia (ED) and and one for idiopathic epilepsy (IE).

2.2 Methods

2.2.1 Quasi-likelihood model

We consider the problem of resampling for correlated binary data, where the correlation can arise from various sources such as population structure and/or relatedness. Our aim is to derive replicates of the trait under the null hypothesis of no association while accounting for the correlation that is present in the sample but whose structure is unknown. To model the response, we use a recently described quasi-likelihood framework for correlated binary traits [Jiang et al., 2016; Zhong et al., 2016],

$$\boldsymbol{\mu} := \mathbb{E}(\mathbf{Y}|\mathbf{X},\mathbf{G}) = \text{logit}^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\gamma), \text{ and} \tag{2.1}$$

$$\boldsymbol{\Omega} := \text{Var}(\mathbf{Y}|\mathbf{X},\mathbf{G}) = \boldsymbol{\Gamma}^{1/2}\,\boldsymbol{\Sigma}\,\boldsymbol{\Gamma}^{1/2}. \tag{2.2}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ denotes the phenotype vector for n subjects, $\mathbf{X}$ is the $n \times k$ matrix of k covariates (including an intercept term), $\mathbf{G} = (G_1, \ldots, G_n)^T$ is the vector of genotypes at the marker of interest, where G_i denotes the minor allele count for the i-th individual, $\boldsymbol{\beta}$ is a k-length vector representing the unknown effects of covariates, γ represents the unknown effect of the tested marker, $\boldsymbol{\Gamma}$ is a diagonal matrix with i^{th} diagonal element $\mu_i(1-\mu_i)$, and $\boldsymbol{\Sigma} = \xi\,\boldsymbol{\Phi} + (1-\xi)\,\mathbf{I}_n$, where the unknown scalar parameter $\xi \in [0,1]$ allows for the incorporation of sample correlation as determined by the elements of the matrix $\boldsymbol{\Phi}$ which is treated as known in the model, but which in practice is taken to be a GRM estimated from genome-wide data.

The framework characterized by (2.1) and (2.2) allows adjustment for important covariates in the mean structure, such as biological or ancestry informative covariates, and residual correlation between subjects is captured directly within the variance structure. Furthermore, by including the matrix $\mathbf{\Gamma}$ in (2.2), this framework specifies the dependence of the variance on the phenotypic mean, a key feature of binary data.

2.2.2 Derivation of trait replicates

A permutation test based on the phenotype involves generating replicates under the null hypothesis of no association ($H_0 : \gamma = 0$). Under this null, the form of the quasi-likelihood framework proposed in (2.1) and (2.2) contains unknown parameter $(\boldsymbol{\beta}, \xi)$, which needs to be estimated from the data. The null estimate $(\widehat{\boldsymbol{\beta}}, \hat{\xi})$ is obtained by iteratively solving the following system of estimating equations [Jiang et al., 2016] under the constraint $\gamma = 0$,

$$\boldsymbol{D}_{\boldsymbol{\beta}}^T \mathbf{\Omega}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0}, \tag{2.3}$$

$$(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{\Gamma}^{-1/2} \mathbf{\Sigma}^{-1} (\mathbf{\Phi} - \mathbf{I}) \mathbf{\Sigma}^{-1} \mathbf{\Gamma}^{-1/2} (\mathbf{Y} - \boldsymbol{\mu}) = \text{trace}(\mathbf{\Sigma}^{-1}(\mathbf{\Phi} - \mathbf{I})), \tag{2.4}$$

where $\boldsymbol{D}_{\boldsymbol{\beta}} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$, is a Jacobian of the conditional mean of the trait with respect to $\boldsymbol{\beta}$. In our model with a logit link function, $\boldsymbol{D}_{\boldsymbol{\beta}} = \mathbf{\Gamma X}$. Unlike with logistic mixed models (LogMM) which are a natural choice for correlated binary data but are computationally challenging to fit as they involve high dimensional integrals, obtaining parameter estimates here is computationally efficient because it only involves solving a system of estimating equations.

Under the constraint $\gamma = 0$, let $U(\boldsymbol{\beta}) = \boldsymbol{D}_{\boldsymbol{\beta}}^T \mathbf{\Omega}^{-1}(\mathbf{Y} - \boldsymbol{\mu})$, which corresponds to the quasi-score function for $\boldsymbol{\beta}$ with fixed ξ. As this quasi-score function involves both the unknown vector $\boldsymbol{\beta}$ as well as the unknown scalar ξ through $\boldsymbol{\mu}$ and $\mathbf{\Omega}$, respectively, we first consider the case where ξ is known. Assuming that H_0 is true, let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$, and let $\boldsymbol{\mu}_0, \mathbf{\Gamma}_0$, $\mathbf{D}_0$ and $\mathbf{\Omega}_0$ correspond to $\boldsymbol{\mu}, \mathbf{\Gamma}, \mathbf{D}_\beta$ and $\mathbf{\Omega}$, respectively, evaluated at $\boldsymbol{\beta}_0$ and $\gamma = 0$. Similarly, let $\widehat{\boldsymbol{\mu}}, \widehat{\mathbf{\Gamma}}$, $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{\Omega}}$ be the same quantities evaluated at $\widehat{\boldsymbol{\beta}}$ and $\gamma = 0$.

We take an approach that is an extension of previous work [Abney et al., 2002] on deriving permutation-based replicates for a quantitative response in the presence of correlation. More precisely, we aim to obtain a vector with entries that are uncorrelated and have the same mean and variance under the null hypothesis, and then permutation will be performed under second-order exchangeability. In order to remove the mean effect, we first consider the residual vector $(\mathbf{Y} - \widehat{\boldsymbol{\mu}})$, which asymptotically has zero mean under H_0. We then aim to derive a closed-form expression of its second moment, so that we can determine a linear transformation that would result in a vector with uncorrelated entries. The mean $\widehat{\boldsymbol{\mu}}$ depends on the fixed effects estimate $\widehat{\boldsymbol{\beta}}$, which we can express in terms of the data by applying a Taylor series expansion to $U(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$, and evaluated at $\widehat{\boldsymbol{\beta}}$,

$$U(\widehat{\boldsymbol{\beta}}) \approx U(\boldsymbol{\beta}_0) + \left.\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta_0}} \cdot (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{2.5}$$

Using the fact that $U(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, and replacing the Jacobian in (2.5) by its expectation (similar to Fisher scoring), we get,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &\approx - \left\{ \mathbb{E}\left[\frac{\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta_0}} \right\}^{-1} U_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0), \\ &= (\mathbf{D}_0^T\, \boldsymbol{\Omega}_0^{-1}\, \mathbf{D}_0)^{-1} \mathbf{D}_0^T\, \boldsymbol{\Omega}_0^{-1} (\mathbf{Y} - \boldsymbol{\mu}_0), \end{aligned} \tag{2.6}$$

As $\widehat{\boldsymbol{\mu}}$ is a non-linear function of $\widehat{\boldsymbol{\beta}}$, a Taylor series expansion is performed on $\boldsymbol{\mu}$ around $\boldsymbol{\beta}_0$, and evaluated at $\widehat{\boldsymbol{\beta}}$,

$$\widehat{\boldsymbol{\mu}} \approx \boldsymbol{\mu}_0 + \mathbf{D}_0\, (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0), \tag{2.7}$$

Combining (2.6) and (2.7), we get for the residual vector $(\mathbf{Y} - \widehat{\boldsymbol{\mu}})$,

$$\mathbf{Y} - \widehat{\boldsymbol{\mu}} = \mathbf{Y} - \boldsymbol{\mu}_0 - (\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)$$

$$\approx \left[\mathbf{I} - \mathbf{D}_0 \left(\mathbf{D}_0^T \, \boldsymbol{\Omega}_0^{-1} \, \mathbf{D}_0\right)^{-1} \mathbf{D}_0^T \, \boldsymbol{\Omega}_0^{-1}\right] (\mathbf{Y} - \boldsymbol{\mu}_0), \tag{2.8}$$

As a result, we can approximate the covariance matrix of the residuals as,

$$\mathrm{Var}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \approx \boldsymbol{\Omega}_0 - \boldsymbol{\Gamma}_0 \mathbf{X} \left(\mathbf{X}^T \boldsymbol{\Gamma}_0 \, \boldsymbol{\Omega}_0^{-1} \, \boldsymbol{\Gamma}_0 \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Gamma}_0. \tag{2.9}$$

The correlation present in $(\mathbf{Y} - \widehat{\boldsymbol{\mu}})$ arises from two sources: that introduced by $\boldsymbol{\Phi}$ through $\boldsymbol{\Omega}_0$ and that introduced from using the estimated mean $\widehat{\boldsymbol{\mu}}$ instead of the true unknown mean $\boldsymbol{\mu}$. We use a factorization $\mathbf{C}$ of $\boldsymbol{\Omega}_0$, with $\boldsymbol{\Omega}_0 = \mathbf{C}^T\mathbf{C}$, to remove the correlation due to $\boldsymbol{\Phi}$ and obtain,

$$\mathrm{Var}[\mathbf{C}^{-T}(\mathbf{Y} - \widehat{\boldsymbol{\mu}})] \approx \mathbf{I} - \mathbf{W}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T = \boldsymbol{\Psi}_0, \tag{2.10}$$

where $\mathbf{W} = \mathbf{C}^{-T}\boldsymbol{\Gamma}_0\mathbf{X}$. The matrix $\boldsymbol{\Psi}_0$ in (2.10) is symmetric and idempotent and thus can be expressed as $\boldsymbol{\Psi}_0 = \mathbf{V}\mathbf{V}^T$, where the columns of $\mathbf{V}$ contain the eigenvectors of $\boldsymbol{\Psi}_0$ corresponding to the eigenvalue 1 with $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{n-k}$ (assuming $\mathbf{X}$ is full rank). The linear transformation $\hat{\boldsymbol{\zeta}} = \mathbf{V}^T\mathbf{C}^{-T}(\mathbf{Y} - \widehat{\boldsymbol{\mu}})$ has for covariance matrix,

$$\mathrm{Var}(\hat{\boldsymbol{\zeta}}) \approx \mathbf{V}^T\boldsymbol{\Psi}_0\mathbf{V} = \mathbf{I}_{n-k}. \tag{2.11}$$

Hence, we obtain a transformation of the residuals with approximately uncorrelated entries. With $\boldsymbol{\Pi}$ denoting a random permutation matrix, we generate a new trait replicate $\mathbf{Y}_\pi$ as,

$$\mathbf{Y}_\pi = \widehat{\boldsymbol{\mu}} + \mathbf{C}^T\mathbf{V}\boldsymbol{\Pi}\mathbf{V}^T\mathbf{C}^{-T}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \tag{2.12}$$

In (2.12), $\boldsymbol{\Gamma}_0$ (the matrix used to get $\mathbf{V}$), and $\mathbf{C}$, both depend on the unknown vector $\boldsymbol{\beta}_0$ and so we replace these by $\widehat{\boldsymbol{\Gamma}}$ (to get $\widehat{\mathbf{V}}$) and $\widehat{\mathbf{C}}$ respectively, where $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Gamma}}^{1/2}\,\boldsymbol{\Sigma}\,\widehat{\boldsymbol{\Gamma}}^{1/2} = \widehat{\mathbf{C}}^T\widehat{\mathbf{C}}$.

We note that $\widehat{\mathbf{C}}$ can be obtained based on a factorization of $\boldsymbol{\Sigma}$ (e.g. using Cholesky or eigen-decomposition). This derivation relies on the assumption that the parameter ξ in $\boldsymbol{\Sigma}$ is known, yet in practice, ξ is also estimated from the data under H_0. In this case, we compute $\widehat{\mathbf{C}}$ based on the factorization of $\widehat{\boldsymbol{\Sigma}} = \hat{\xi}\,\boldsymbol{\Phi} + (1-\hat{\xi})\,\mathbf{I}$, where $\hat{\xi}$ represents the null estimate of ξ. Hence, a trait replicate is obtained as,

$$\mathbf{Y}_\pi = \widehat{\boldsymbol{\mu}} + \widehat{\mathbf{C}}^T\widehat{\mathbf{V}}\boldsymbol{\Pi}\widehat{\mathbf{V}}^T\widehat{\mathbf{C}}^{-T}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \tag{2.13}$$

Under the identity permutation, meaning that $\boldsymbol{\Pi} = \mathbf{I}$, we are able to recover the original response vector. Examining the form of (2.13), there are three main steps used to obtain a transformation of the residuals with second-order exchangeable entries. The first is to center the response by using the estimated phenotypic mean. The second is to remove the correlation present due to polygenic effects, which is done by the pre-multiplication by $\widehat{\mathbf{C}}^{-T}$. The last step is to remove the correlation that is generated from using parameter estimates instead of the true values when centering, and is represented by the pre-multiplication by $\widehat{\mathbf{V}}^T$. We expect this last step to have a minor impact when the sample size is large and the number of parameters is low. The replicates generated from this approach are quantitative, and so the binary nature of the response is not preserved during resampling. However, the first two moments assumed in the quasi-likelihood model and estimated from the data are approximately preserved.

2.2.3 Other resampling methods considered

In simulations, we compare BRASS to six other resampling methods. MVNpermute [Abney, 2015; Abney et al., 2002] is an existing method developed for quantitative traits in which residuals from a LMM are transformed to be uncorrelated, then are permuted and then reconstituted into trait replicates. MVNpermute is asymptotically exact for multivariate normal data, as the entries of the transformed residuals being uncorrelated

implies independence under the normality assumption, and hence exchangeability. In contrast, our method BRASS is approximate as it relies on second-order exchangeability rather than full exchangeability; the framework we use only considers the first two moments of the response. However, the covariate effects being modeled on a logit scale, rather than a linear scale as in MVNpermute, is more appropriate for binary data as the mean is constrained within $(0, 1)$. More importantly, the covariance matrix assumed in our method allows for dependence on the mean function, which is an inherent feature of binary data; this is absent in MVNpermute. Thus, we expect these two methods to differ the most when a major portion of the trait variability is attributable to covariate rather than polygenic effects. In addition to MVNpermute, we consider five approaches that have not been previously described.

In the method we call LogMM-PQL, we fit a logistic mixed model using penalized-quasi-likelihood (PQL), where the model form under the null is $\text{logit}(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{u}$, with $\mathbf{Z}$ being the covariate matrix and $\mathbf{u} \sim \text{MVN}(0, \sigma^2\boldsymbol{\Phi})$, and then we sample trait replicates from the model using the estimates for $(\boldsymbol{\alpha}, \sigma^2)$. In the method we call Naive, we fit a LMM whose form under the null is $\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e}$, with $\mathbf{Z}$ being the covariate matrix and $\mathbf{e} \sim \text{MVN}(0, \sigma_1^2\boldsymbol{\Phi} + \sigma_2^2\mathbf{I})$, and then permute the residuals $(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\alpha}})$, and add back $\mathbf{Z}\widehat{\boldsymbol{\alpha}}$. We also consider modified versions of BRASS, MVNpermute and Naive, in which the resulting trait values in each replicate are converted to binary values using the procedure described below. We call these methods BRASS_{mod}, MVNpermute_{mod}, and Naive_{mod}, respectively. Properties of the seven methods are summarized in Table 2.1.

To convert a simulated quantitative trait replicate $\mathbf{Y}_q$ to binary version $\mathbf{Y}_b$, a threshold is set and the values of $\mathbf{Y}_q$ above and below that threshold are converted to 1 and 0, respectively, where the threshold is chosen so that the number of cases (1 values) in $\mathbf{Y}_b$ is the same as that in the original trait vector $\mathbf{Y}$.

We use GMMAT [Chen et al., 2016a] to fit the null model used in LogMM-PQL, and use GEMMA [Zhou and Stephens, 2012] to fit the null model used in Naive, Naive_{mod}, MVNpermute and MVNpermute_{mod}.

Table 2.1: Properties of the seven resampling methods used to obtain trait replicates.

	Feature[a]			
Method	Binary trait replicate?	Trait variance is function of mean?	Resampling accounts for correlation in $\mathbf{Y}$?	Resampling accounts for correlation due to parameter estimation?
BRASS	–	√	√	√
BRASS$_{mod}$	√	√	√	√
LogMM-PQL	√	√	√	–
MVNpermute	–	–	√	√
MVNpermute$_{mod}$	√	–	√	√
Naive	–	–	–	–
Naive$_{mod}$	√	–	–	–

[a] Check marks (√) indicate that the method incorporates the specified feature.

2.2.4 Adaptive resampling procedure

Given the computationally intensive nature of permutation-based approaches, we use an adaptive procedure when simulating replicates to evaluate the type 1 error rate of our method for multiple testing correction at level α. Our chosen stopping criteria are checked initially with 1000 simulated replicates and then in increments of 5000 starting at 5000 replicates. More specifically, we continue generating replicates until either (1) N_{max}, the maximum number of replicates allowed, has been reached (where $N_{max} \geq 1000$); (2) A hypothesis test for $H_0 : p = \alpha$ against $H_a : p > \alpha$ is rejected at significance level 0.01, where the estimate for the p-value p is the proportion of replicates with test statistic at least as extreme as the one observed. An exact test for H_0 is used when the number of replicates is 1000; otherwise, a z-test is performed. If N_{max} has been reached, the p-value estimated from the N_{max} replicates is compared to α.

2.2.5 Incorporating the sample structure in the null model

Our approach to addressing sample structure is to model genetic relatedness as both fixed and random effects in the null model, where the fixed component is represented by the inclusion

of covariates that represent major axes of genetic variation, and the random component is represented by the inclusion of a GRM that reflects the leftover structure. We use these modeling components in all seven resampling methods considered.

We use a previously proposed method, PC-Relate [Conomos et al., 2016], where we first estimate the GRM $\mathbf{\Phi}$ using population estimates for the minor allele frequencies (MAF), resulting in a matrix $\widehat{\mathbf{\Phi}}$ with entries

$$\widehat{\mathbf{\Phi}}_{ij} = \frac{1}{L}\sum_{l=1}^{L}\frac{(G_{il}-2\hat{p}_l)(G_{jl}-2\hat{p}_l)}{2\hat{p}_l(1-\hat{p}_l)}, \tag{2.14}$$

where L is the number of markers, G_{il} is the minor allele count for the i-th individual at marker l, and $\hat{p}_l = 0.5 \cdot \overline{G}_l$, where $\overline{G}_l$ is the sample average estimator of the population MAF for marker l. We obtain the top D PCs from $\widehat{\mathbf{\Phi}}$, which are included as ancestry informative covariates in the null model. We then build a GRM estimate $\tilde{\mathbf{\Phi}}$ to capture the remaining genetic similarities among subjects that are not reflected by the top PCs. The entries of $\tilde{\mathbf{\Phi}}$ are

$$\tilde{\mathbf{\Phi}}_{ij} = \frac{\sum_{l=1}^{L}(G_{il}-2\widetilde{p}_{il})(G_{jl}-2\widetilde{p}_{jl})}{2\sum_{l=1}^{L}\left[\widetilde{p}_{il}(1-\widetilde{p}_{il})\widetilde{p}_{jl}(1-\widetilde{p}_{jl})\right]^{1/2}}, \tag{2.15}$$

where G_{il} and $\widetilde{p}_{il}$ are the minor allele count and a predicted subject-specific MAF, respectively, for the i-th individual at marker l. For each marker l, the subject-specific MAF estimate is obtained as half of the fitted values from a linear regression of $\mathbf{G}_l$ on the top D PCs. So in the numerator of (2.15), the centered genotype values represent the residuals from this linear regression, effectively removing the effects captured through the top D PCs. The resulting matrix $\tilde{\mathbf{\Phi}}$ is used as the GRM estimate in all seven resampling methods.

2.3 Simulation studies and data analysis

2.3.1 Simulation study design

We perform simulation studies to evaluate the effectiveness of permutation-based trait replicates in the context of correcting for multiple testing, with both population structure and pedigree structure present in the sample. We simulate non-causal markers that are tested for association with a binary trait, and estimate the type 1 error rate for the top signal amongst them using the empirical distribution from trait replicates generated under the null hypothesis of no association. Genotype, trait and covariates are generated under multiple simulation settings. Since in real applications it is usually not known a priori which variables should be included as covariates in the null model for the phenotype, we also consider the case of misspecification in the fitted null model due to the exclusion of a relevant covariate. In addition, we investigate the impact of ascertainment in the sample as case-control studies often involve phenotype-based ascertainment which introduces misspecification in the model if not properly accounted for. We simulate $M_c = 10,000$ causal SNPs which are used to generate polygenic effects in the trait generative model as well as to correct for correlation in the sample when simulating trait replicates from the seven methods.

We consider a setting in which a binary trait is tested for association with each of m markers and the aim is to assess the significance of the smallest p-value out of the m association tests. This context would arise in testing significance of a genomic region or in assessing genome-wide significance. Owing to the computational intensive nature of the simulations, the trait vector is simulated 200 times, and 100 independent marker panels, each consisting of $m = 100$ non-causal markers, are simulated for each trait replicate. Retaining only the top association signal in each panel, this results in 100 association test statistics which overall amounts to 20,000 replicates used for type 1 error estimation. To assess significance in each data simulation, up to 10,000 trait replicates are generated under the null using each of the seven resampling methods. For each marker panel, the trait replicates are individually tested

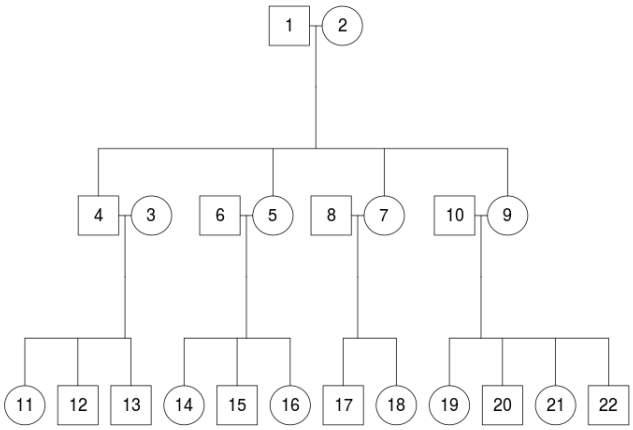
for association with the non-causal markers, and the smallest p-value for each replicate is compared with the one from the original data simulation. We use CARAT [Jiang et al., 2016] for single-SNP association testing; it is a retrospective association test for binary traits in structured samples.

The Balding-Nichols model [Balding and Nichols, 1995] is used to draw allele frequencies for 2 sub-populations with the fixation index F set to 0.01 which is representative of the population structure seen in humans within continents [Marchini et al., 2004]. For each marker s, given the ancestral allele frequency p_s drawn from a uniform distribution on $[0.2, 0.8]$ (and independently across markers), the allele frequency in sub-population $k = 1, 2$ is drawn independently from a Beta distribution with parameters $p_s(1-F)/F$ and $(1-p_s)(1-F)/F$. An equal number of pedigrees, each with structure given in Figure 2.1, is assigned to the 2 sub-populations with all founders within a given pedigree being assigned to the same sub-population. For each marker, genotypes for a pedigree's founders are simulated as independent Bernoulli draws using the corresponding sub-population allele frequency, and gene dropping is used to determine genotypes for other pedigree members. The M_c causal markers are simulated only once, and they are used to compute the GRM estimate $\tilde{\mathbf{\Phi}}$ in (2.15), which is then re-used in all simulations. In addition, independent non-causal markers are simulated for each simulation replicate using the same Balding-Nichols model, where the ancestral allele frequencies are drawn from a uniform distribution on $[0.2, 0.8]$, and these are the markers tested for association.

We simulate 46 pedigrees of the same configuration (Figure 2.1), each consisting of 22 individuals. For each individual, three covariates are simulated: sex, age and an i.i.d. standard normal covariate. Sex is determined by the pedigree configuration and age is drawn uniformly and independently within 1.5 years of $73, 75, 46, 43, 40, 46, 40, 43, 47, 51, 18, 21, 15,\ 15, 12, 9, 13, 17, 24, 21, 18$, and 14 years respectively for individuals labeled $1-22$ in the pedigree. All the covariates (besides sex) are re-generated for each simulation replicate.

We consider two generative models for the trait. First, a logistic model is used in which

Figure 2.1: Three-generation pedigree used in the simulation studies.



Y_i is given by

$$Y_i|\mathbf{X}_i, \mathbf{W}_i, \boldsymbol{\alpha} \sim \text{Bernoulli}(p_i), \text{ independently, with } \text{logit}(p_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\alpha}, \tag{2.16}$$

where $\mathbf{X}_i$ is the covariate row vector for the i-th individual; $\boldsymbol{\beta}$ are the fixed effects for the covariates; $\mathbf{W}_i$ is a row vector representing the SNP information for the i-th individual corresponding to the M_c causal markers standardized to have mean 0 and variance 1; $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{M_c})^T \sim MVN(0, \sigma^2\mathbf{I})$ represent the random effects of the M_c causal markers with polygenic variance $\sigma_a^2 = \sigma^2 M_c$. The values for $\boldsymbol{\beta}$ and σ_a^2 are chosen such that: (1a) If all covariates are included in the fitted model, they each explain an equal amount of the variation on the logit scale due to covariates; (1b) If the standard normal covariate is excluded, its effect size is set to correspond to an average p-value of 0.05 in a LMM Wald test (as determined by simulations), and age and sex equally explain the remaining portion of the variability due to covariates; (2) Considering the total variance on the logit scale due to covariate effects and polygenic random effects, the fraction of this variance due to covariates is fixed at 20%, 40%, 60% or 80%; (3) Bernoulli error explains on average about 20% of the total phenotypic variability; (4) The prevalence is approximately 30%. When ascertainment is included in the

sample, we simulate data under the model in (2.16) and select 500 cases and 500 controls at random to be retained in the sample.

We also consider generating the trait using a liability threshold model,

$$Y_i = 1 \iff L_i > 0, \text{ with } L_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\alpha} + \epsilon_i, \tag{2.17}$$

where L_i is the latent liability for individual i. The values of $\boldsymbol{\beta}$ and σ_a^2 are chosen using the same conditions as for the logistic model, where conditions (1a), (1b), (2) and (4) are enforced on the liability scale instead of the logit scale, and condition (3) is used for the random error ϵ_i in (2.17), rather than for the Bernoulli error.

Larger values of σ_a^2 will correspond to more severe structure in the sample (from both population structure and family relatedness). We chose $D = 1$ for the number of top PCs of the GRM $\widehat{\boldsymbol{\Phi}}$ in (2.14) to include as covariates when fitting the null models in all seven resampling methods and when performing association testing with CARAT. We note that the pedigree information is only needed to simulate the data; it is not used when fitting the null models as the realized rather than the expected identity by descent (IBD) sharing across the genome is used to model correlation in the trait values.

2.3.2 Real Data Application

We illustrate the use of BRASS in the context of multiple testing correction in a GWAS of domestic dogs [Hayward et al., 2016a]. Unlike with humans, the genome-wide threshold for significance has not been well determined. This is one of the largest dog genotyping studies with 4,224 dogs genotyped at 185,805 SNPs and twelve clinical and morphological phenotypes recorded. We analyze two binary traits from this study (1) elbow dysplasia (ED), which is an ensemble of abnormalities that affect the articular surfaces of the elbow; (2) idiopathic epilepsy (IE), which is diagnosed when there is no identifiable cause for seizures in affected dogs. Data was obtained from the Dryad online repository [Hayward et al., 2016b].

The dogs in the sample represent over 150 breeds, with an additional 170 from mixed breeds and 350 village dogs (i.e. dogs from free-breeding human-commensal populations). For the ED trait, the data contain 113 cases and 633 controls among 82 breeds, and for IE, there are 34 cases and 168 controls from the Irish Wolfhound breed. After quality control for the markers [Hayward et al., 2016a], there are 150,418 and 98,350 autosomal markers (MAF $> 1\%$) for ED and IE respectively. For ED, we exclude dogs that come from breeds not well represented in the data (fewer than 10 members). We also exclude dogs that appear more closely related to dogs of a different breed than to dogs of the same breed using hierarchical clustering with UPGMA [Murtagh, 1984; Sokal, 1958] (Table 2.2). This results in 93 cases and 445 controls amongst 10 breeds. We compute the GRM estimate $\widehat{\mathbf{\Phi}}$ in (2.14) from 138,192 markers with MAF $> 5\%$ (Figure 2.2), and determine the number of top PCs that are sufficient to capture the breed structure. Since there are 10 breeds in the data, it would seem sensible for the top 9 PCs to capture genetic similarities due to breed membership, which is corroborated by plotting the top 10 PCs and the eigen spectrum (Figures 2.3 and 2.4). As in some instances it may not be possible to have prior population membership information (e.g. breed membership), we use either the top 9 or the top 13 PCs to estimate subject-specific allele frequencies, from which we compute the GRM estimate $\tilde{\mathbf{\Phi}}$ in (2.15) which will mostly represent within-breed genetic similarities. These top PCs are also included as covariates in the analysis of ED. For both traits, sex is included as a covariate.

2.3.3 Modification of CARAT

In order to assess the variance under the null hypothesis of no association, CARAT uses a retrospective model for the tested marker $\mathbf{G}$ (see equation (14) in [Jiang et al., 2016]). When we include the top PCs as covariates in our analysis, the GRM $\tilde{\mathbf{\Phi}}$ only reflects the leftover genetic correlation that is not captured by the top PCs. To incorporate all of the genetic

Table 2.2: Classification of dogs based on breed and clade membership for the top 10 breeds in the domestic dog data with elbow dysplasia.

	Breed[a]									
Clade[b]	Border Collie	English Setter	German Shepherd	Golden Retriever	Gordon Setter	Labrador Retriever	Mixed Breed	Newfound-land	Rottweiler	Vizsla
1	.[c]	79	.	.	.	.	.	.	.	.
2	.	.	.	.	.	**1**[d]	.	.	.	.
3	.	.	.	.	.	17	20	.	.	.
4	.	.	.	.	.	188	.	.	.	.
5	.	.	.	57	.	**1**	.	.	.	.
6	.	.	55	.	.	**1**	.	.	.	.
7	.	.	.	.	.	**2**	.	.	41	.
8	.	.	.	.	.	.	.	30	.	.
9	18	.	.	.	.	.	.	.	.	.
10	.	.	.	.	13	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.	20

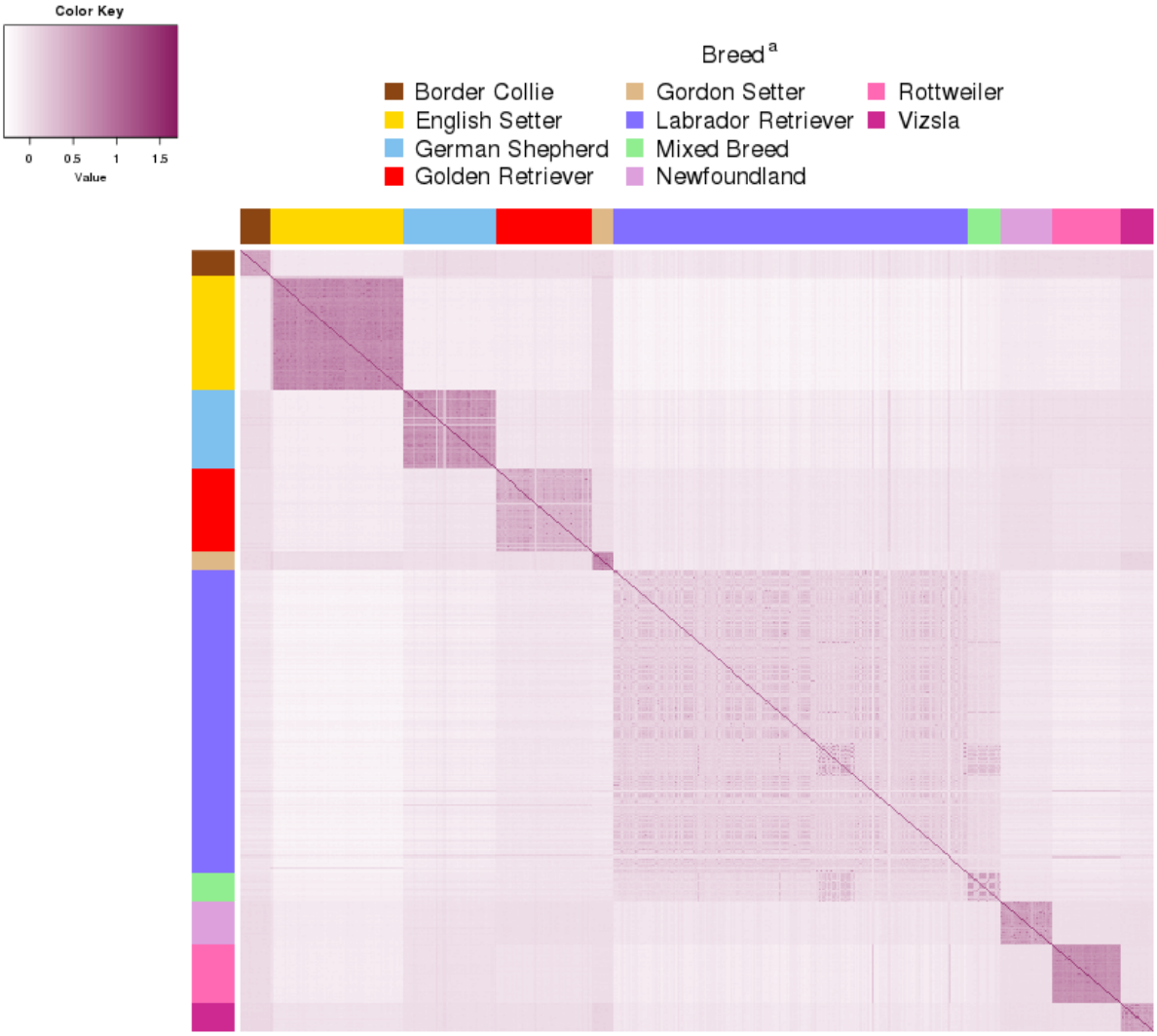
[a] Breed labeling information was provided in the data set.

[b] Clade membership was predicted from hierarchical clustering using UPGMA, where the dissimilarity measure used is $(1 - \hat{\rho}_{ij})$ with $\hat{\rho}_{ij} = \widehat{\mathbf{\Phi}}_{ij}/\sqrt{\widehat{\mathbf{\Phi}}_{ii}\widehat{\mathbf{\Phi}}_{jj}}$ for all $i, j = 1, \ldots, 543$, and $\widehat{\mathbf{\Phi}}$ is the genetic relatedness matrix in Equation (2.14).

[c] This notations means no dogs are in the cell.

[d] Bolded entries correspond to dogs that are not in the same clade as the majority of dogs of the same breed (excluding mixed breed dogs).

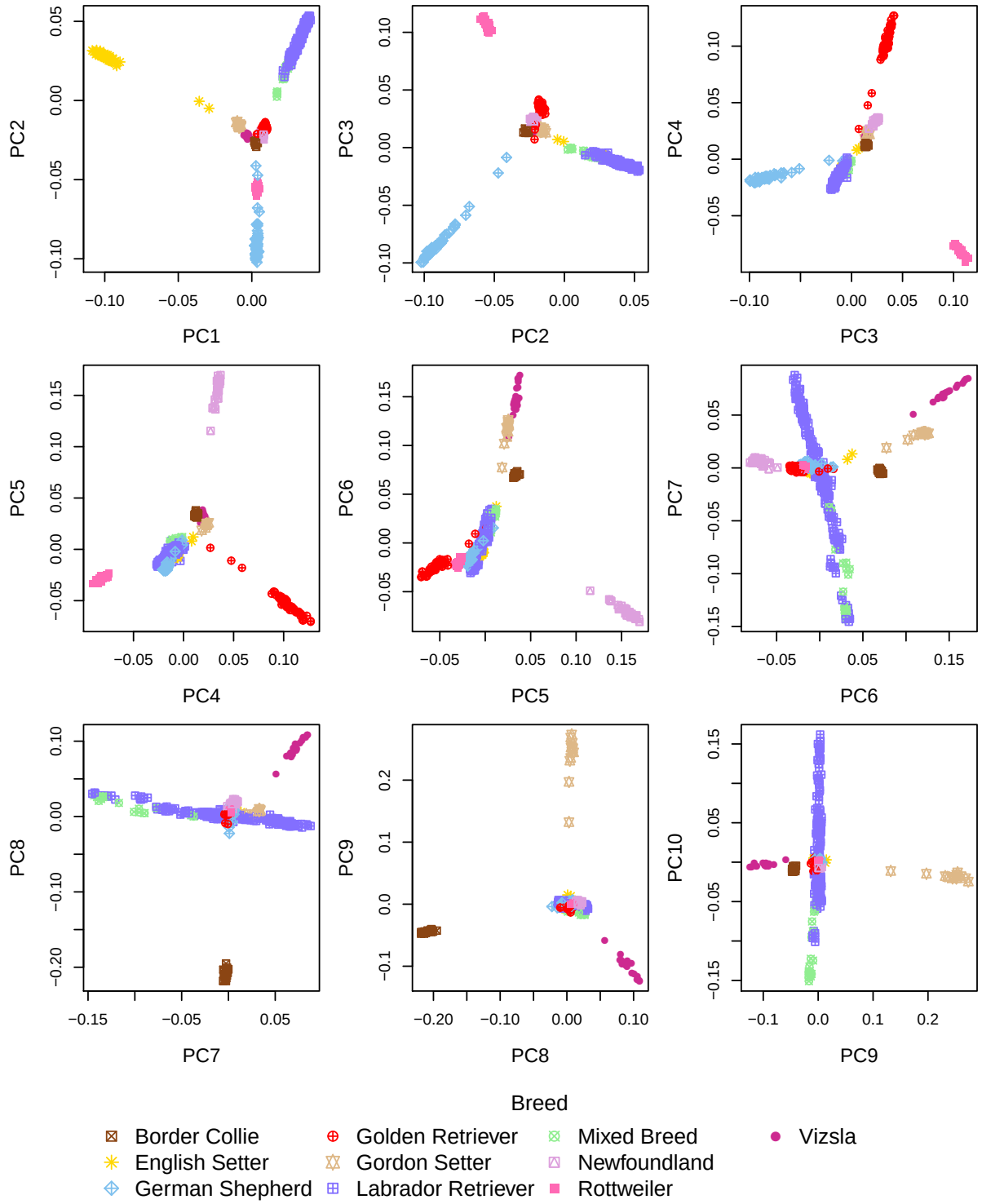
Figure 2.2: Level plot of the GRM $\widehat{\mathbf{\Phi}}$ in the domestic dog data with elbow dysplasia.



The plot is based on 543 dogs and the adjacent color bars represent the breed information for the pair of dogs in the corresponding GRM entry. We observe stronger patterns of genetic relatedness within breeds, which are represented by darker shades in the plot, compared to those across breeds.

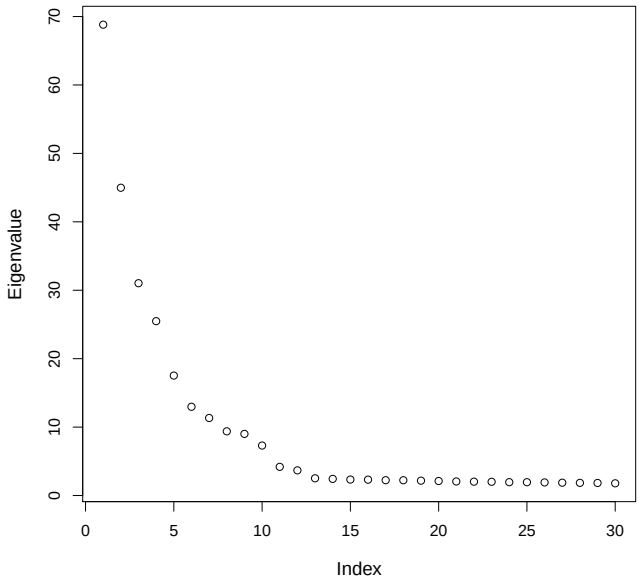
[a] Breed labeling information was provided in the data set.

Figure 2.3: Plot of the top 10 principal components extracted from the GRM $\widehat{\mathbf{\Phi}}$ in the domestic dog data with elbow dysplasia after quality control.



538 dogs are present in total among 10 breeds (breed information is provided in the data set). We see that breed membership is well captured within the top 9 principal components.

Figure 2.4: Eigen spectrum of the GRM $\widehat{\mathbf{\Phi}}$ in the domestic dog data with the elbow dysplasia.



The largest 30 eigenvalues of the GRM are displayed in decreasing order. We observe that past the 13^{th} largest eigenvalue there is very little change in the magnitude of the values.

correlation, we consider the following retrospective model,

$$\mathbb{E}(\mathbf{G}|\mathbf{Y},\mathbf{Z}) = \mathbf{Z}\boldsymbol{\alpha} \text{ and } \text{Var}(\mathbf{G}|\mathbf{Y},\mathbf{Z}) = \sigma_G^2\tilde{\boldsymbol{\Phi}}, \tag{2.18}$$

where $\mathbf{Z}$ is a matrix containing in its columns the top D PCs along with an intercept, and $\boldsymbol{\alpha}$ is an unknown vector for their effects. The model in (2.18) allows for different MAFs conditional on each subject's genetic ancestry captured in the top PCs. Hence, we modify CARAT and use $\widetilde{\sigma}_G^2 = \mathbf{G}^T\mathbf{P}\mathbf{G}/(n-D-1)$, with $\mathbf{P} = \tilde{\boldsymbol{\Phi}}^{-1} - \tilde{\boldsymbol{\Phi}}^{-1}\mathbf{Z}(\mathbf{Z}^T\tilde{\boldsymbol{\Phi}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\tilde{\boldsymbol{\Phi}}^{-1}$, as an estimate of the variance parameter.

2.4 Results

2.4.1 Assessing the type 1 error rate

We begin by examining the performance of the proposed resampling methods in the context of correcting for multiple testing, with both population structure and pedigree structure present in the sample in addition to important covariates. We expect that by estimating the structure present in the sample in a way that incorporates the binary nature of the trait, our BRASS resampling method will lead to correct calibration of the type 1 error. As the PQL fitting algorithm used in LogMM-PQL is known to lead to underestimation of the variance component parameter when it is high [Jang and Lim, 2009; Rodríguez and Goldman, 2001], we expect this approach to result in inflated type 1 error when the sample contains a high amount of correlation. Naive and MVNpermute are expected to perform worse than BRASS and LogMM-PQL in the presence of covariates as they do not incorporate the dependence of the variance on the mean and hence, on the covariates, and they model covariate effects on a linear scale. Furthermore, we expect the Naive method to perform worse with increasing sample correlation as the dependence is unaccounted for when generating replicates.

We consider the setting in which a binary trait is measured in 46 equal-sized families

in 2 sub-populations, and association is tested with each of $m = 100$ null markers (i.e. not associated with the binary trait) using the test statistic CARAT [Jiang et al., 2016]. The aim is to sample trait replicates in order to determine the significance of the smallest p-value among the m tests. Correlation is introduced by the presence of population structure and family relatedness, which are incorporated in the trait generative model through polygenic effects. We compare seven resampling methods: (1) BRASS, our proposed approach based on a quasi-likelihood framework; (2) LogMM-PQL, which generates binary replicates by simulating from a fitted LogMM; (3) MVNpermute, a resampling method for correlated data with quantitative phenotypes; (4) Naive, which is based on permuting residuals from a LMM; and (5) BRASS$_{mod}$; (6) MVNpermute$_{mod}$; and (7) Naive$_{mod}$, which are modified versions of BRASS, MVNpermute and Naive, respectively, in which the resulting trait replicates are converted to binary. Each of these seven methods will lead to a p-value estimate corresponding to the significance of the smallest p-value among the m tests conducted with the observed trait, and the type 1 error for each method is estimated as the proportion of simulations where the estimated p-value is less than the nominal level.

All the resampling strategies involve fitting a trait model that incorporates the correlation present in the sample. A common approach is to estimate the genetic correlation present by the use of an empirical genetic relatedness matrix (GRM) so as to model the structure using random effects [Thornton, 2015]. Alternatively, the top principal components of a given GRM are used as covariates to model the structure present using fixed effects [Price et al., 2006]. Here, we combine the two approaches and use both fixed and random effects to capture the sample correlation in the null model. We use the top PCs from an empirical GRM as fixed effects in the null model and then build a new GRM matrix based on the genotype information adjusting for the effects of these top PCs in order to capture the leftover structure as random effects (see Methods for details).

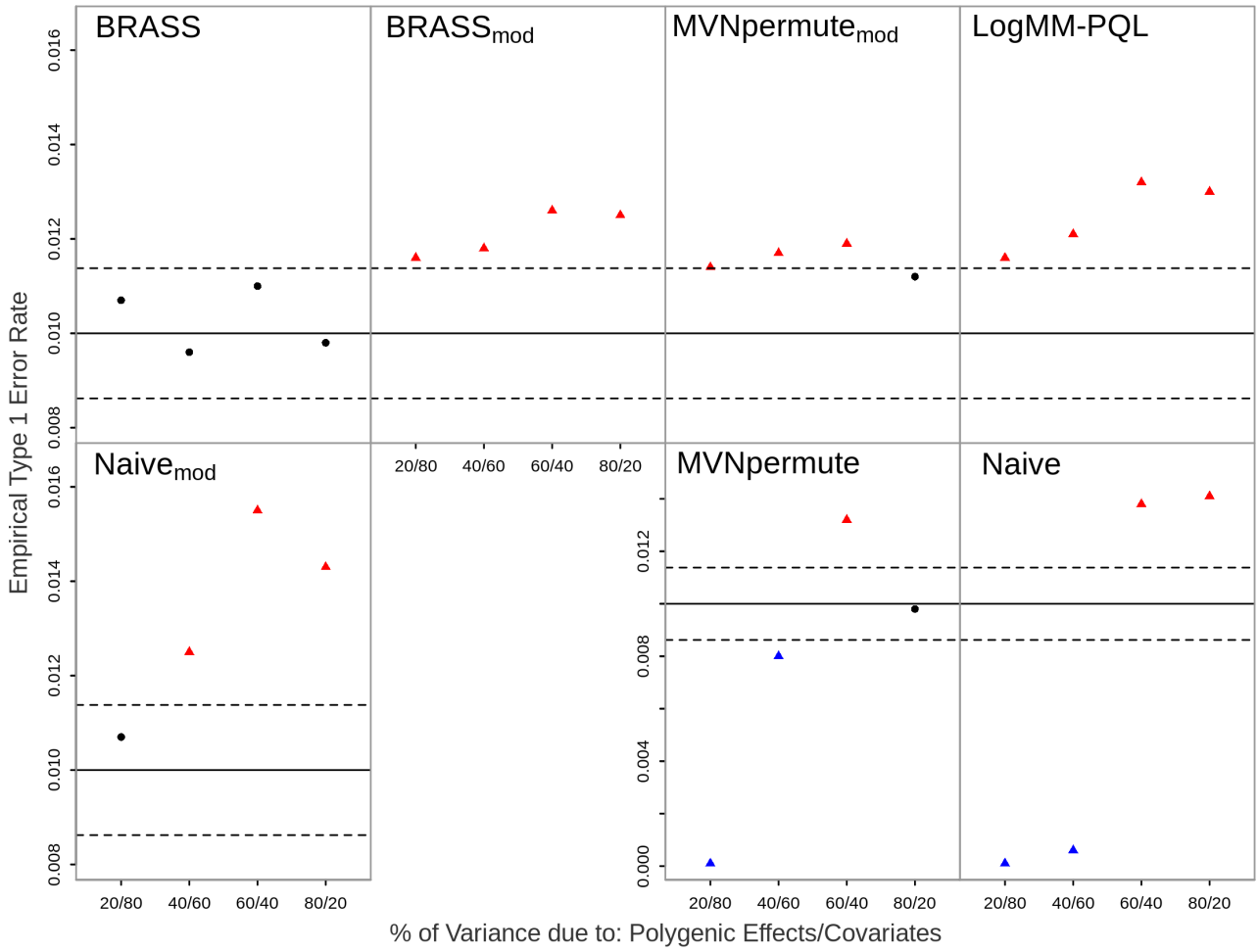
As the true model for the trait is generally unknown in an association study, we simulate the data using both a logistic model as well as a liability threshold model (see Methods),

where in the latter case, the models fit by all seven resampling methods are misspecified. The results based on a logistic generating model are shown in Figures 2.5 and 2.6 and those based on a liability threshold generating model are shown in Figures 2.7. and 2.8.

We obtain similar results for both generating models, where as expected we observe that permuting the residuals from a linear mixed model (Naive) leads in all settings to lack of control of the type 1 error. More precisely, when covariates explain most of the variation on the logit scale, the method is overly conservative and when instead polygenic effects explain most of the variability, the type 1 error rate is significantly inflated. Moreover, we also find that adjusting for sample correlation prior to permutation based on LMM residuals, as done in MVNpermute, still leads to a very conservative test when covariates highly influence the variability on the logit scale. This is most likely because a linear mixed model does not allow any dependence between the trait variance and the mean, which is influenced by the covariates. As covariates become less important, we do see better control of the type 1 error. In contrast to the former two approaches, BRASS maintains control of the type 1 error in all settings considered. For LogMM-PQL, we observe an inflation in the type 1 error rate as the correlation present in the sample becomes more important. This approach was expected to lead to an inflation in the empirical error rate since the PQL model fitting approach tends to underestimate the variance components and hence, would result in less correlation present in the simulated data compared to that in the original data. For Naive$_{mod}$ and MVNpermute$_{mod}$, control of the type 1 error is improved compared to the corresponding original methods when covariates strongly influence the phenotypic mean, and for BRASS$_{mod}$, type 1 error control is worse compared to BRASS.

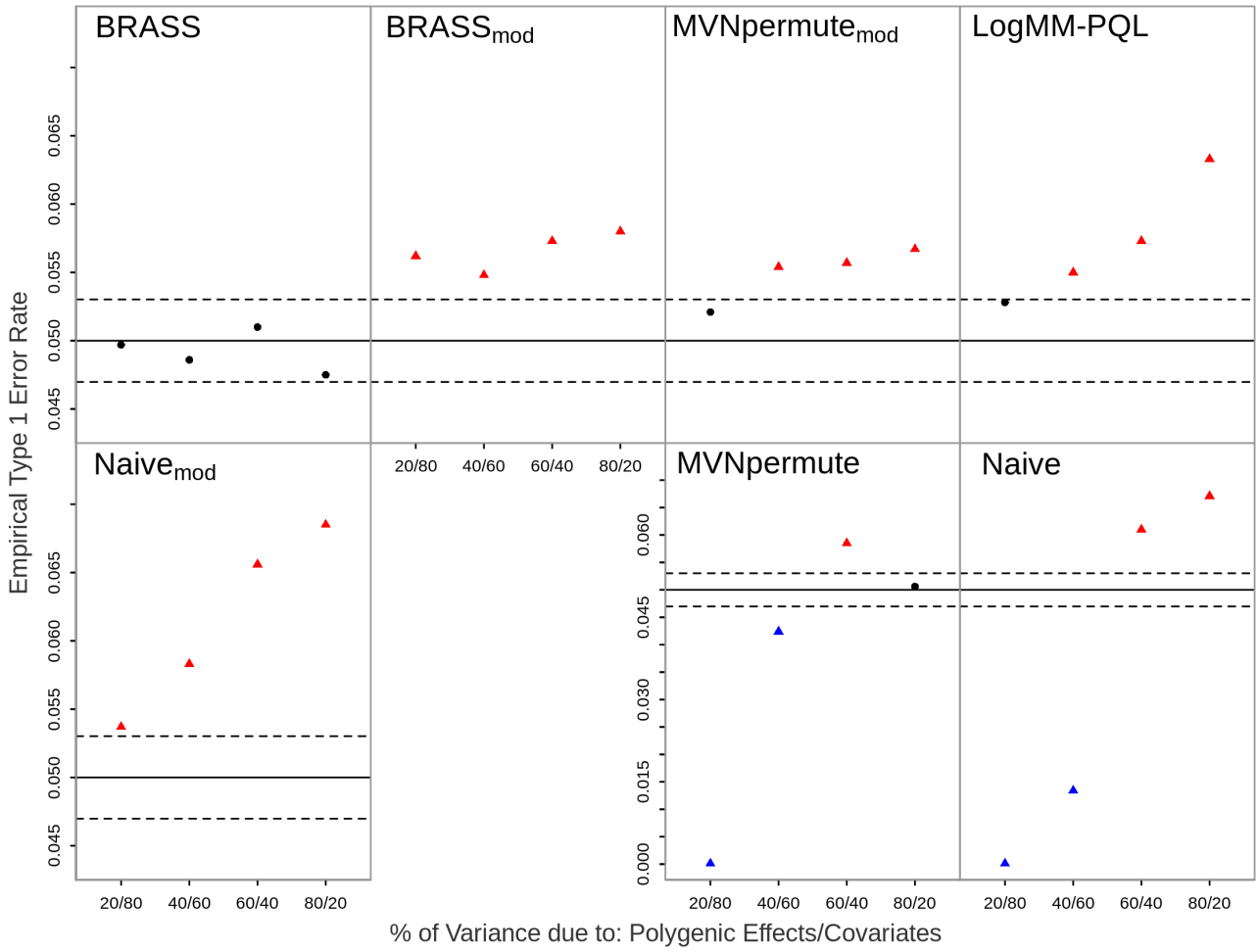
We find that none of the resampling methods that simulate binary replicates consistently control the type 1 error rate, and MVNpermute and Naive lead to significantly deflated type 1 error in the presence of strong covariate effects relative to the polygenic effects. However, BRASS consistently results in well-calibrated type 1 error, regardless of the amount of population and family structure present in the sample relative to the covariate effects. This

Figure 2.5: Empirical Type 1 Error Rates with Logistic Model Including all Covariates at Nominal Level 0.01.



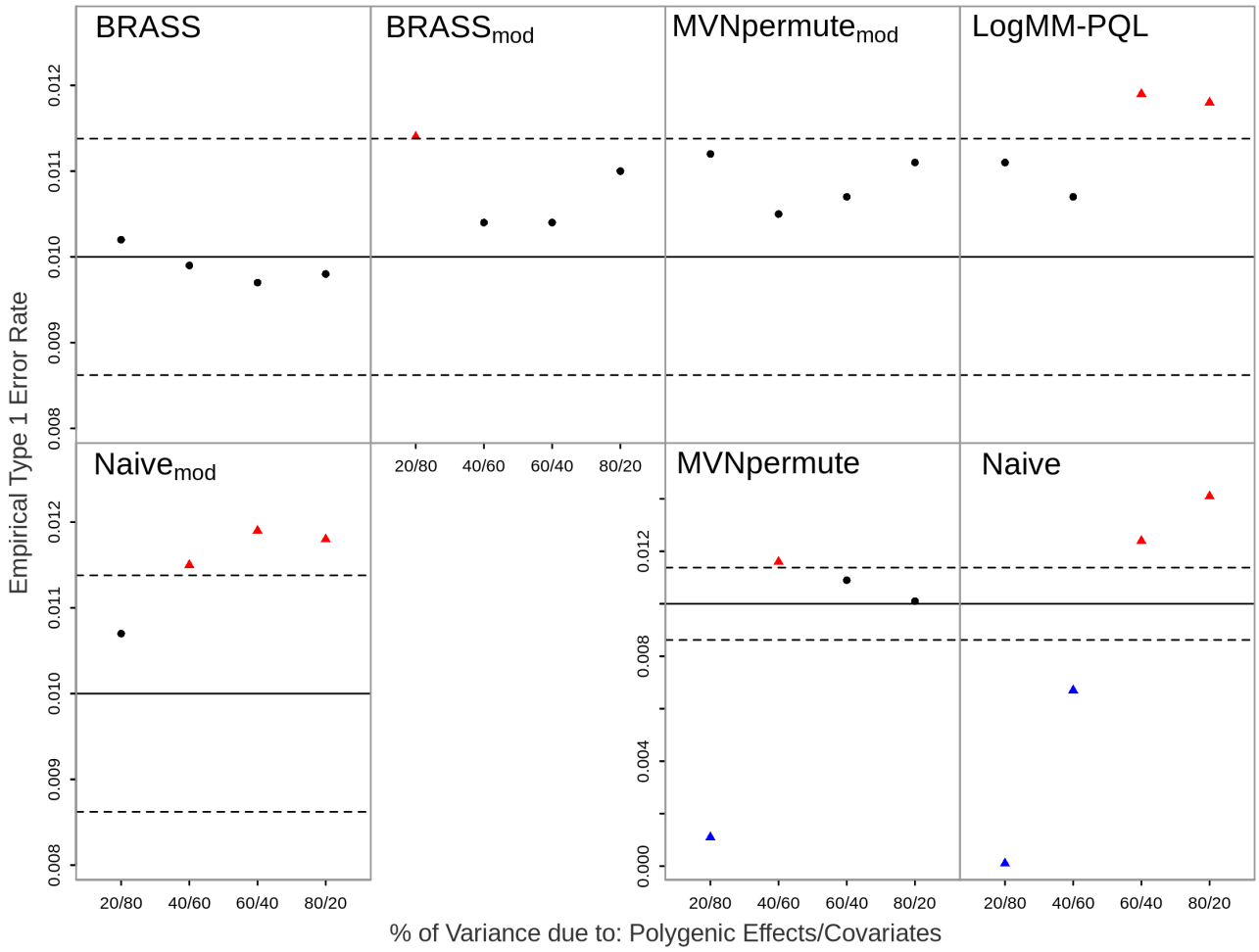
The error rate is based on 20,000 simulated replicates. The solid horizontal line represents the nominal level and the dashed lines represent rejection bounds outside of which the z-test comparing the estimated type 1 error to the nominal level is rejected at level .05. Estimates inside the rejections bounds are represented by circles and those outside the bounds are represented by triangles. Red, black and blue symbols represent liberal, well-controlled and conservative type 1 error rates, respectively. The proportion of variability on the logit scale attributable to polygenic effects vs. covariates is varied from 20 to 80% in increments of 20%.

Figure 2.6: Empirical Type 1 Error Rates with Logistic Model Including all Covariates at Nominal Level 0.05.



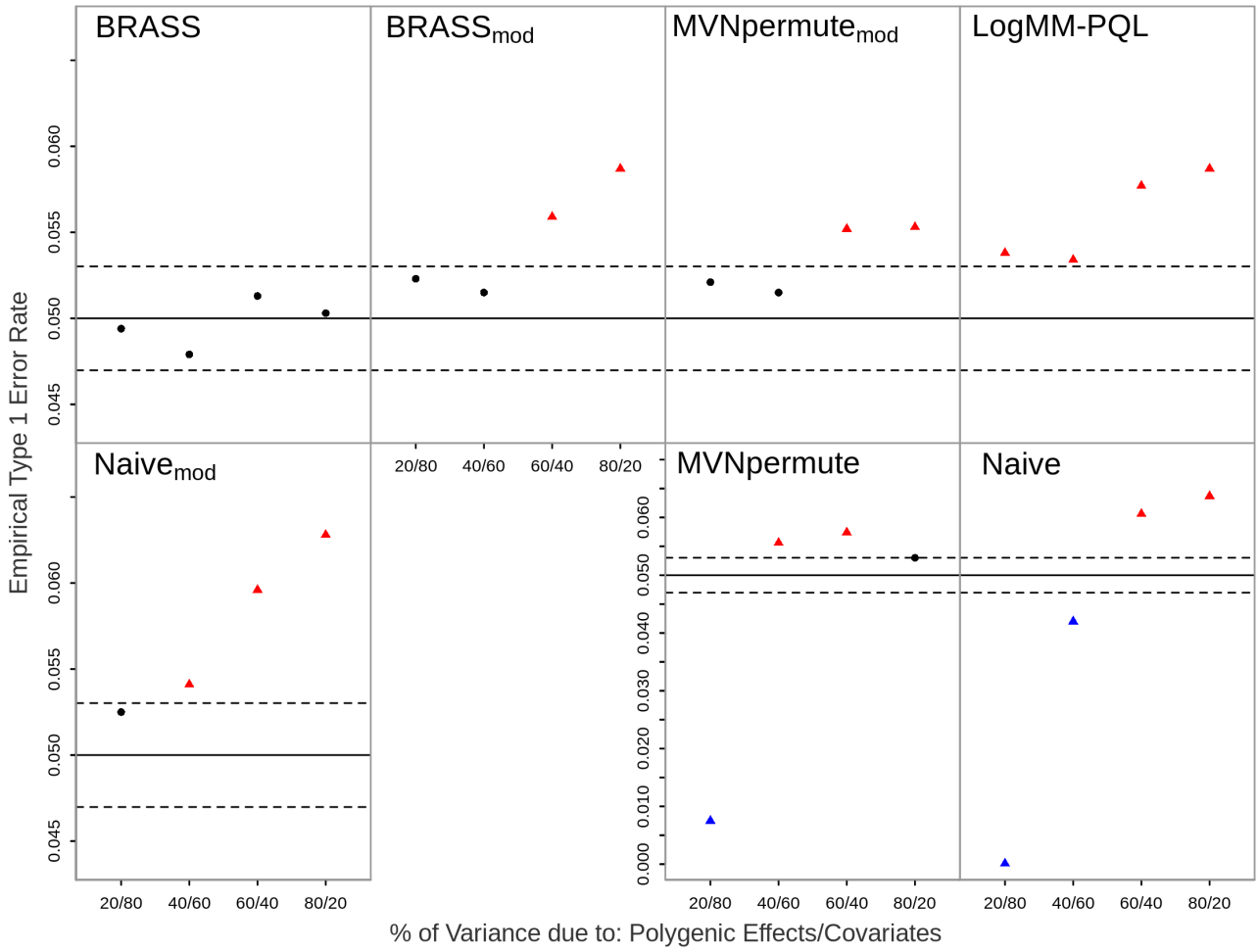
The error rate is based on 20,000 simulated replicates. The solid horizontal line represents the nominal level and the dashed lines represent rejection bounds outside of which the z-test comparing the estimated type 1 error to the nominal level is rejected at level .05. Estimates inside the rejections bounds are represented by circles and those outside the bounds are represented by triangles. Red, black and blue symbols represent liberal, well-controlled and conservative type 1 error rates, respectively. The proportion of variability on the logit scale attributable to polygenic effects vs. covariates is varied from 20 to 80% in increments of 20%.

Figure 2.7: Empirical Type 1 Error Rates with Liability Threshold Model at Nominal Level 0.01.



The error rate is based on 20,000 simulated replicates. The solid horizontal line represents the nominal level and the dashed lines represent rejection bounds outside of which the z-test comparing the estimated type 1 error to the nominal level is rejected at level .05. Estimates inside the rejections bounds are represented by circles and those outside the bounds are represented by triangles. Red, black and blue symbols represent liberal, well-controlled and conservative type 1 error rates, respectively. The proportion of variability on the liability scale attributable to polygenic effects vs. covariates is varied from 20 to 80% in increments of 20%.

Figure 2.8: Empirical Type 1 Error Rates with Liability Threshold Model at Nominal Level 0.05.



The error rate is based on 20,000 simulated replicates. The solid horizontal line represents the nominal level and the dashed lines represent rejection bounds outside of which the z-test comparing the estimated type 1 error to the nominal level is rejected at level .05. Estimates inside the rejections bounds are represented by circles and those outside the bounds are represented by triangles. Red, black and blue symbols represent liberal, well-controlled and conservative type 1 error rates, respectively. The proportion of variability on the liability scale attributable to polygenic effects vs. covariates is varied from 20 to 80% in increments of 20%.

suggests that (1) incorporating the dependence between the trait mean and variance, as well the effects of covariates on a logit scale, lead to robustness of the method in the presence of important covariates; (2) converting the replicates from BRASS to binary, as done with BRASS_{mod}, does not well preserve the structure in the observed data that was captured in the quantitative replicates and performs only about as well as generating replicates based on a LMM with MVNpermute_{mod}.

Next, we assess the robustness of the proposed methods when important covariates are omitted from the fitted model. As it is usually not known a priori which variables should be kept in the analysis, one would usually try including different combinations of covariates to finally determine the ones to include in the final model. It may occur that one of the covariates has a moderate effect on the trait and leads to a p-value close to the significance threshold (e.g. 0.05). Hence, a judgment call would be required for whether to keep the covariate in the model, and one might decide to exclude it. It is thus of interest to see how the proposed methods would fare in such a scenario as the replicates generated would come from a more misspecified model.

To illustrate this situation, we simulated a binary trait using a logistic model with both covariates and polygenic effects present. However, among the covariates, we selected one of them to have an effect on the trait which corresponded to a p-value of about 0.05 when fitting a LMM with all the covariates included. In all resampling methods used, we excluded this covariate when fitting the null model. Since these methods all have misspecified first moment, the estimates of the covariate effects will have some amount of bias. We assessed the resulting type 1 error rate when correcting for multiple tests, where we used the same simulation design as previously mentioned to generate the genetic markers.

The simulation results are displayed in Figures 2.9 and 2.10 We see that BRASS retains good calibration of the type 1 error in all settings. When covariates explain most of the variability in the trait mean, MVNpermute and Naive lead to a highly conservative type 1 error rate, with barely any false rejections being made. The other methods which generate

binary replicates lead to significantly inflated type 1 error rates as the amount of correlation present in the sample increases.

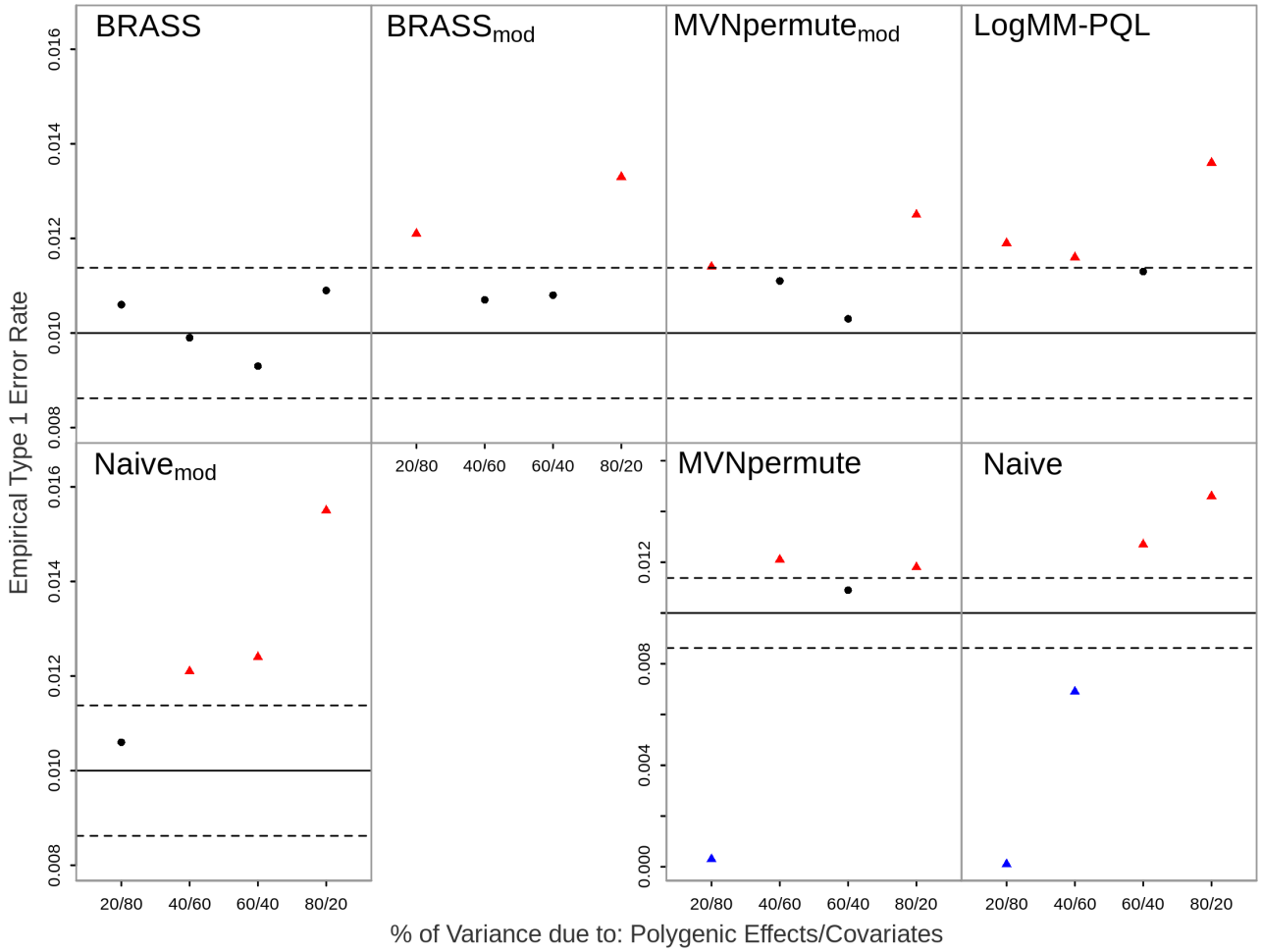
We assess the robustness of the proposed methods when trait-based ascertainment has been applied to the sample. This is commonly used in case-control studies in which individuals are included in the sample based on their disease affection status, particularly when the prevalence of the trait in the population is too low to obtain sufficient power by sampling uniformly at random. We simulated a binary trait under the same settings as for the logistic model with all covariates included. To introduce ascertainment, we selected at random 500 cases to be retained in the sample, and an additional 500 controls so that the the sample had a 1:1 case-control ratio. The results are shown in Figures 2.11 and 2.12. We again see that BRASS retains good control of the type 1 error rate in all settings considered, and is slightly conservative when covariates explain most of the variability in the trait mean. MVNpermute and Naive fail to control the type 1 error, as they give overly conservative rates when covariates have a major impact on the trait. The remaining methods all exhibit too liberal error rates as polygenic effects become more important in the trait model.

These results demonstrate the robustness of BRASS under various sources of model misspecification, including the use of a quasi-likelihood model that does not match to any true likelihood, the exclusion of covariates with moderate effects from the fitted null model and the presence of trait-based ascertainment.

2.4.2 Application to Domestic Dog Data

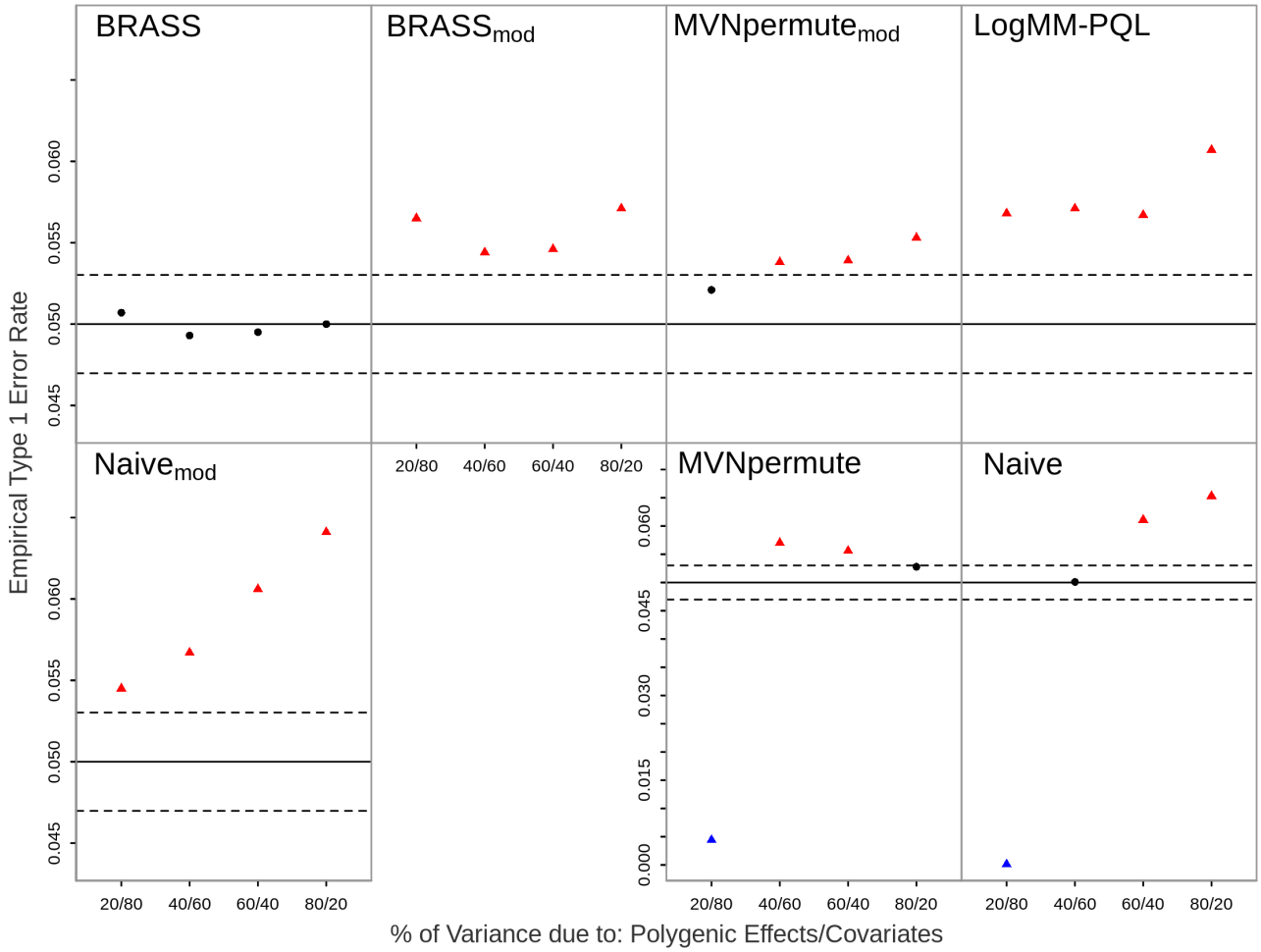
We illustrate the use of our resampling method for multiple testing correction in a GWAS of domestic dogs [Hayward et al., 2016a]. The data for ED contained dogs of various breeds while for IE we had dogs from a single breed. Similarly to our simulations, we incorporated the structure present in the ED data as both fixed and random effects in our model. We used either the top 9 or 13 PCs to estimate a GRM based on genotype residuals removing the effects of these top PCs. Hence for the analysis of ED, we included the top PCs as

Figure 2.9: Empirical Type 1 Error Rates Omitting a Relevant Covariate at Nominal Level 0.01.



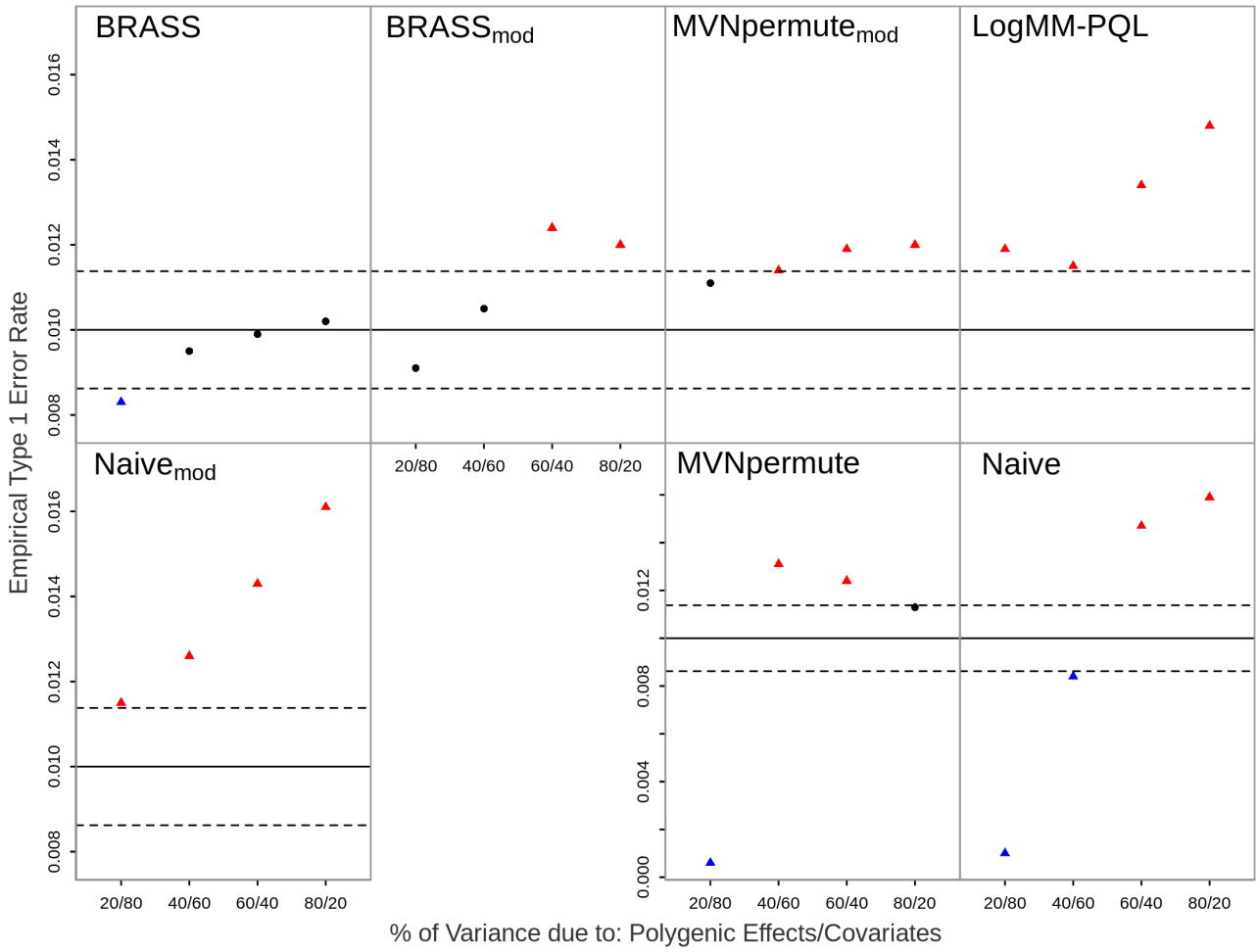
The error rate is based on 20,000 simulated replicates. The solid horizontal line represents the nominal level and the dashed lines represent rejection bounds outside of which the z-test comparing the estimated type 1 error to the nominal level is rejected at level .05. Estimates inside the rejections bounds are represented by circles and those outside the bounds are represented by triangles. Red, black and blue symbols represent liberal, well-controlled and conservative type 1 error rates, respectively. The proportion of variability on the logit scale attributable to polygenic effects vs. covariates is varied from 20 to 80% in increments of 20%. The effect of the omitted covariate on the trait correspond to a Wald test p-value of .05 using a linear mixed model.

Figure 2.10: Empirical Type 1 Error Rates Omitting a Relevant Covariate at Nominal Level 0.05.



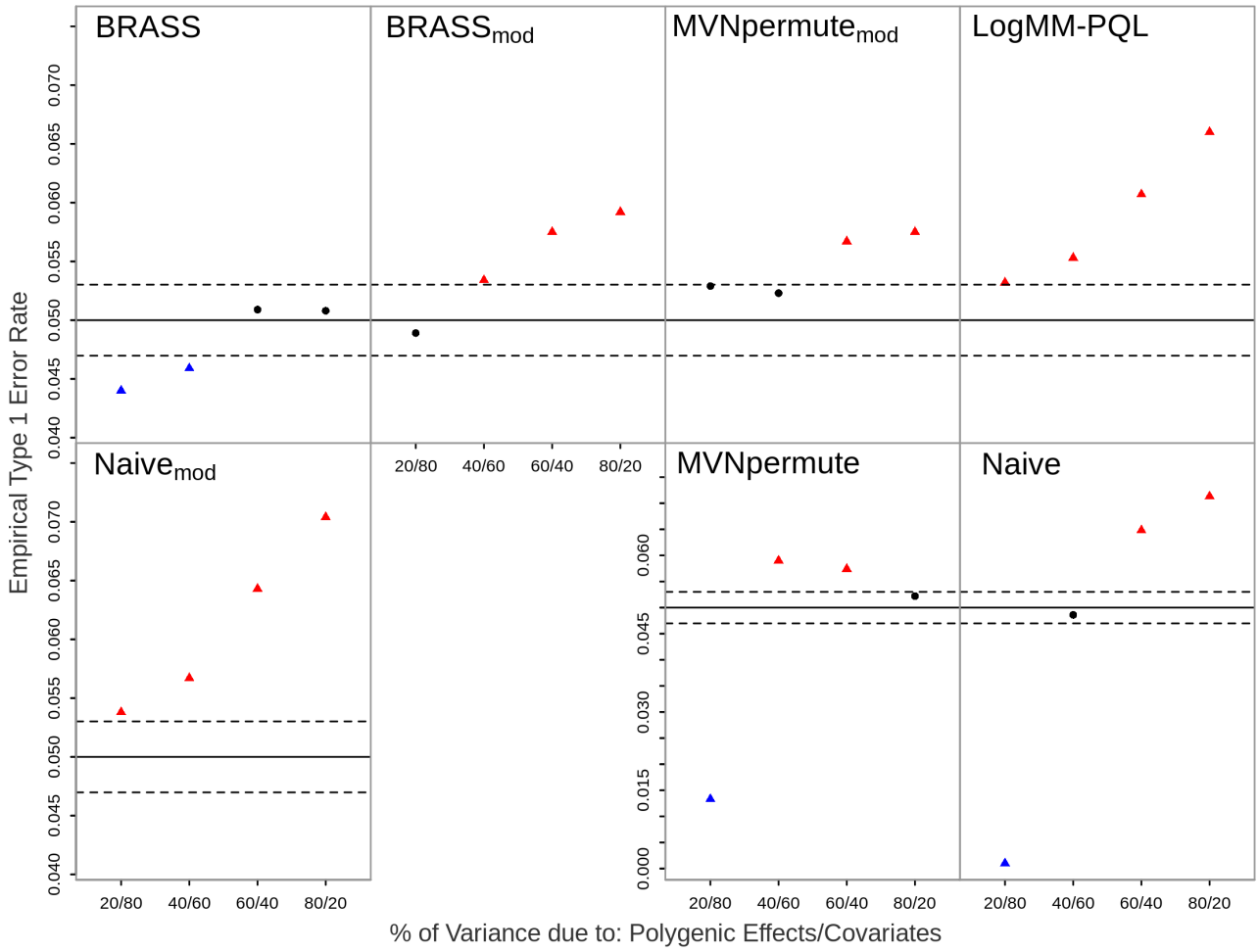
The error rate is based on 20,000 simulated replicates. The solid horizontal line represents the nominal level and the dashed lines represent rejection bounds outside of which the z-test comparing the estimated type 1 error to the nominal level is rejected at level .05. Estimates inside the rejections bounds are represented by circles and those outside the bounds are represented by triangles. Red, black and blue symbols represent liberal, well-controlled and conservative type 1 error rates, respectively. The proportion of variability on the logit scale attributable to polygenic effects vs. covariates is varied from 20 to 80% in increments of 20%. The effect of the omitted covariate on the trait correspond to a Wald test p-value of .05 using a linear mixed model.

Figure 2.11: Empirical Type 1 Error Rates with Ascertainment Present at Nominal Level 0.01.



The error rate is based on 20,000 simulated replicates. The solid horizontal line represents the nominal level and the dashed lines represent rejection bounds outside of which the z-test comparing the estimated type 1 error to the nominal level is rejected at level .05. Estimates inside the rejections bounds are represented by circles and those outside the bounds are represented by triangles. Red, black and blue symbols represent liberal, well-controlled and conservative type 1 error rates, respectively. The proportion of variability on the logit scale attributable to polygenic effects vs. covariates is varied from 20 to 80% in increments of 20%.

Figure 2.12: Empirical Type 1 Error Rates with Ascertainment Present at Nominal Level 0.05.



The error rate is based on 20,000 simulated replicates. The solid horizontal line represents the nominal level and the dashed lines represent rejection bounds outside of which the z-test comparing the estimated type 1 error to the nominal level is rejected at level .05. Estimates inside the rejections bounds are represented by circles and those outside the bounds are represented by triangles. Red, black and blue symbols represent liberal, well-controlled and conservative type 1 error rates, respectively. The proportion of variability on the logit scale attributable to polygenic effects vs. covariates is varied from 20 to 80% in increments of 20%.

covariates in the null model and used the estimated GRM in the variance structure. For IE, we used an empirical GRM based on the genotype information to capture the genetic relatedness. We used CARAT to perform single-SNP association tests. For genome-wide significance assessment, 100,000 trait replicates were generated under the null hypothesis of no association using both BRASS and MVNpermute$_{mod}$, the latter of which is arguably the second-best method based on type 1 error performance in simulations, although it is often too liberal. Genome-wide p-values were estimated using the empirical distribution of the test statistics based on the trait replicates. The parameter estimates from the fitted models of both BRASS and MVNpermute$_{mod}$ (excluding effects for the PCs) are reported in Table 2.3 for both traits analyzed.

The genomic control inflation factors for ED with 9 and 13 PCs as well as for IE are $\lambda_{\text{GC}} = 0.98, 1.01$ and 0.94, respectively. The genome-wide significance threshold is estimated at nominal level 0.05 from the empirical distribution of the top association signal using the trait replicates. Manhattan plots of the p-values of the single-SNP tests for the observed data are presented in Figure 2.13 for both ED and IE phenotypes, along with the estimated genome-wide significance thresholds. In the analysis of ED using the top 9 PCs, the strongest association signal is found on chromosome 26 for SNP rs9000666, and it fails to reach the genome-wide significance threshold of 3.4×10^{-8} estimated using BRASS and 2.9×10^{-7} using MVNpermute$_{mod}$ (Table 2.4). Similar results are obtained using the top 13 PCs (not included). The genome-wide thresholds of BRASS and MVNpermute$_{mod}$ are quite different, with MVNpermute$_{mod}$ giving a more liberal threshold. This could partly be driven by the inflated type 1 error of the test when high amounts of structure are present in the sample, which could lead to a lower significance threshold. For IE, we find a 12Mb region on chromosome 4 (position 7.5-19.3 Mb) that reaches the genome-wide significance thresholds estimated with both BRASS and MVNpermute$_{mod}$, which are 6.4×10^{-7} and 4.7×10^{-6}, respectively (Table 2.5). The candidate genes in this region are: *DISC1* [NCBI Gene: 488970], which has been associated with various neuronal abnormalities in humans [Hodgkinson et al.,

Table 2.3: Parameter null estimates in the domestic dog data for elbow dysplasia (ED) and idiopathic epilepsy (IE).

Parameter	Estimate	SE
VC ξ	0.76	–
Intercept	−1.2	0.38
Sex	−0.43	0.24

(a) ED: Quasi-likelihood model

Parameter	Estimate	SE
Additive variance σ_a^2	0.036	–
Total variance σ_T^2	0.130	–
Heritability σ_a^2/σ_T^2	0.28	–
Intercept	0.27	0.05
Sex	−0.06	0.03

(b) ED: Linear mixed model

Parameter	Estimate	SE
VC ξ	0.09	–
Intercept	0.08	0.59
Sex	−1.11	0.39

(c) IE: Quasi-likelihood model

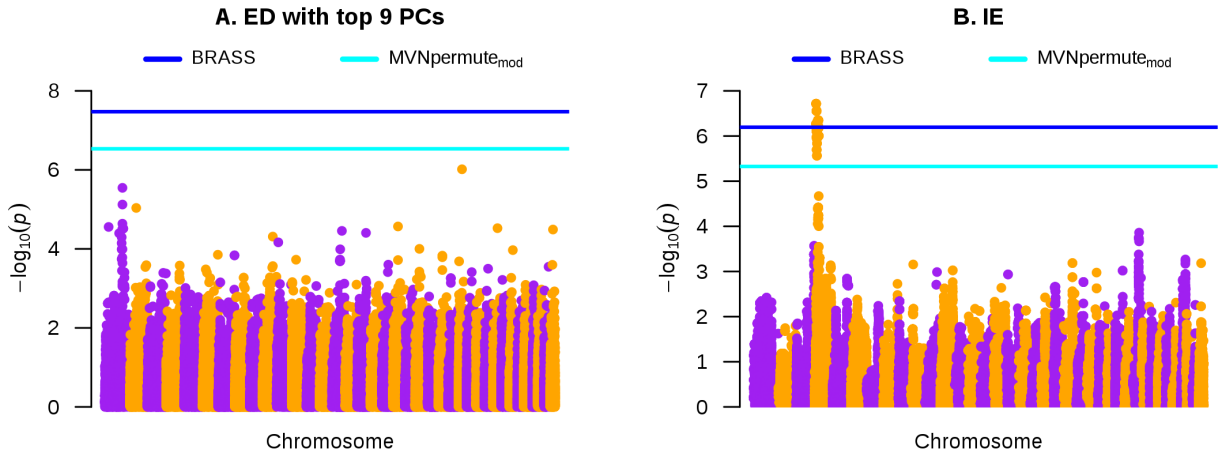
Parameter	Estimate	SE
Additive variance σ_a^2	0.011	–
Total variance σ_T^2	0.135	–
Heritability σ_a^2/σ_T^2	0.081	–
Intercept	0.42	0.09
Sex	−0.16	0.05

(d) IE: Linear mixed model

The estimates are obtained using: (a) a quasi-likelihood model for ED fitted with CARAT; (b) a linear mixed model for ED fitted with GEMMA; (c) a quasi-likelihood model for IE fitted with CARAT; (d) a linear mixed model for IE fitted with GEMMA. For ED, the results are shown for the analysis including top 9 PCs as covariates (corresponding null estimates are omitted); the parameter null estimates using the top 13 PCs are similar and omitted. Sex is coded as 1=male, 2 = female.

2004; Iqbal et al., 2013]; *MYPN* [NCBI Gene: 489013], which has been associated with dilated cardiomyopathy [MIM: 615248] in humans [Duboscq-Bidot et al., 2008; Meyer et al., 2013]; *ARV1* [NCBI Gene: 488975], which has been associated with epileptic encephalopathy [MIM: 617020] in humans [Palmer et al., 2016] , and *ACTA1* [NCBI Gene: 488984], which has been associated with scapulohumeroperoneal myopathy [MIM: 616852], a neurological disease, in humans [Zukosky et al., 2015] .

Figure 2.13: Genome screen results in the domestic dog data.



Manhattan plots of the single-SNP association p-values using CARAT (on $-\log_{10}$ scale) for elbow dysplasia (ED) and idiopathic epilepsy (IE), with genomic position on the x-axis. The horizontal lines represent estimated genome-wide thresholds at nominal level 0.05 using 100,000 trait replicates: (A) 3.4×10^{-8} (BRASS) and 2.9×10^{-7} (MVNpermute$_{mod}$); (B) 6.4×10^{-7} (BRASS) and 4.7×10^{-6} (MVNpermute$_{mod}$). The Manhattan plot for the analysis of ED using the top 13 PCs is similar to that in (A) and omitted.

Table 2.4: Top association signal for elbow dysplasia.

					Genome-wide p-value	
SNP	Chr	Position	MAF	Single SNP p-value	BRASS	MVNpermute$_{mod}$
rs9000666	26	16554631	0.017	9.6×10^{-7}	0.322	0.146

The results are shown for the analysis including the top 9 PCs; the results with top 13 PCs are similar (not included).

Table 2.5: Top 3 association signals for idiopathic epilepsy in domestic dog data.

						Genome-wide p-value	
SNP[a]	Chr	Position (Mb)	MAF	Nearest Genes[b]	P-value	BRASS	MVNpermute$_{mod}$
{rs9032722, rs24053634, rs24080710, rs24054032, rs24068259}	4	7.59-8.39	0.016	*ARV1, DISC1, FAM89A*	1.9×10^{-7}	0.024	0.002
{rs24158651, rs24077813, rs24089517, and 12 others}	4	9.5-10.1	0.16	*ACTA1, CCSAP, NUP133*	2.5×10^{-7}	0.030	0.002
{rs24072714, rs24111345, rs24058877, and 55 others}	4	16.5-19.4	0.18	*CTNNA3, DNAJC12, LRRTM3, MYPN*	4.1×10^{-7}	0.040	0.004

A 12Mb region on chromosome 4 (position 7.5-19.3 Mb) attained genome-wide significance with BRASS (the 7 remaining SNPs in that region are omitted from the table).

[a]SNPs within the brackets resulted in the same single-SNP test p-value.

[b]Top 3 nearest genes to SNPs in the brackets (using CanFam3.1). For the least significant SNP set, at least one of the SNPs was located within the listed genes. NCBI Gene ID for genes not mentioned in the text: $FAM89A$ (NCBI Gene: 488974), $CCSAP$ (NCBI Gene: 479209), $NUP133$ (NCBI Gene: 479208), $CTNNA3$ (NCBI Gene: 489008), $DNAJC12$ (NCBI Gene: 479224), $LRRTM3$ (NCBI Gene: 489009).

2.4.3 Computation time

In order to generate replicates from BRASS, the null prospective model of CARAT needs to be fitted, which includes specifying covariates and a GRM $\mathbf{\Phi}$. If not provided, obtaining the eigendecomposition of $\mathbf{\Phi}$ is the main computationally intensive step involved in fitting the null model [Jiang et al., 2016]. However, this step (or one of equivalent computational complexity) is also performed when using methods based on a linear or logistic mixed model. Once parameter null estimates are obtained, the transformed residuals $\hat{\boldsymbol{\zeta}}$ as well as the linear map $\widehat{\mathbf{C}}^T\widehat{\mathbf{V}}$ in (2.13) need to be computed; however, once obtained, they can be re-used when generating trait replicates. As the eigendecomposition of $\mathbf{\Phi}$, and hence $\widehat{\mathbf{\Sigma}}$, is obtained prior to fitting the null model, we use it to compute $\widehat{\mathbf{C}}$. Therefore, the most computationally intensive step is to obtain the eigenvectors of $\widehat{\mathbf{\Psi}}$ (the estimate used for $\mathbf{\Psi}_0$ in (2.10)), which involves a time complexity of $O(n^3)$. By using the singular value decomposition of the matrix $\mathbf{W}$ in (2.10), which involves a time complexity of $O(nk^2)$ (assuming the number of covariates $k < n$), we are able to obtain the eigenvectors of $\widehat{\mathbf{\Psi}}$ without having to construct it and compute its eigendecomposition. Past this step, the additional cost for generating L replicates is $O(n^2L)$; this computation can easily be parallelized to generate sets of traits replicates independently.

We report run times for BRASS in real and simulated data. Using a single processor on a machine with 6 core Intel Xeon 3.50 GHz CPUs and 32 GB RAM, it takes 1s to fit the null model and generate 1,000 trait replicates for the dog data with ED and 538 dogs and 0.16s for IE in 202 dogs. On simulated data with 1,000, 5,000 and 10,000 individuals, it takes 1.6s, 1 min and 4.6 min (275s), respectively, to generate 1,000 trait replicates. As the use of BRASS will vary based on the context (e.g. multiple testing correction, region-based association testing), the computational burden involved in comparing the test statistic between the observed trait and its replicates will vary based on what statistic is being considered. We note that since the trait replicates are generated under the null hypothesis of no association, they can be re-used when performing separate analyses on a large number of markers.

2.5 Discussion and Future Work

In genetic association analysis of a complex trait, permutation testing can be a useful tool in many contexts in which the approximate null distribution of the test statistic is not available. In the presence of population structure, cryptic and/or family relatedness, which are all common sources of confounding in genetic association studies, exact permutation tests are usually not feasible as they disregard the patterns of genetic relatedness among the subjects, and thus fail to retain the correlation structure present in the data. The previously proposed MVNpermute method addresses this problem only for quantitative traits. For binary traits, we have introduced BRASS, a novel permutation-based resampling procedure for generating trait replicates in samples with population structure, cryptic and/or family relatedness. BRASS allows for covariates, ascertainment, and simultaneous testing of multiple markers, and it accommodates a wide range of test statistics. Using a quasi-likelihood model that incorporates the correlation present in the data, we obtain transformed phenotypic residuals that are approximately second-order exchangeable. From the permuted values, new trait replicates are obtained, and the assumed correlation in the phenotype values as well as the genotype structure present in the original data are maintained. Compared to MVNpermute, the model assumed by BRASS incorporates the effects of covariates on a logit scale and accounts for the dependence of the trait variance on the mean, both of which reflect key aspects of binary data. Consequently, BRASS benefits from less model misspecification when the phenotype is a binary trait and the covariates effects are substantial.

We demonstrate the validity of our approach in simulation studies. Specifically, we find that BRASS maintains correct control of the type 1 error under varying amounts of population structure, familial relatedness, ascertainment and sources of trait model misspecification. We compare BRASS to two resampling approaches based on a linear mixed model, where in the first the raw residuals are permuted and in the second, MVNpermute, permutation is applied after removing the correlation in the residuals. We find that both of the LMM-based approaches fail to maintain correct control the type 1 error when covariate effects

are substantial, suggesting that only accounting for the genetic relatedness patterns present without incorporating the key features associated with the binary nature of the trait (i.e. the dependence of the phenotypic variance on the mean) is not sufficient to obtain replicates that correctly estimate the null distribution of the statistic. We also find that when polygenic effects are important, accounting for the genetic relatedness in the sample prior to permutation as done in BRASS and MVNpermute leads to better control of the type 1 error compared to ignoring it, as done when permuting the raw residuals from a LMM. As the replicates simulated from all three approaches are quantitative, yet the original trait is binary, we also considered converting the simulated quantitative replicates to binary. We find that this additional step results in an improvement of the type 1 error rate calibration for the LMM-based methods when the polygenic component has a low impact on the trait distribution relative to covariates, but does not control the type 1 error when more structure is present in the sample.

Finally, we evaluated an approach based on sampling trait replicates from a fitted LogMM using PQL. We find that this approach leads to worse control of the type 1 error rate the more the polygenic component of the trait is important, a result likely explained by the fitting algorithm used in LogMM-PQL. Indeed, while a logistic mixed model is a natural choice for binary data that contains correlation, fitting such a model is computationally intensive and often requires some approximation of the high-dimensional integral involved. LogMM-PQL uses PQL which is known to give biased estimates in highly correlated data. We could have considered fitting algorithms with higher accuracy (e.g. using Laplace approximation [Raudenbush et al., 2000] or Gauss-Hermite quadrature [Rabe-hesketh et al., 2002]) but these also lead to increased computational burden, which may not scale well for large GWA studies.

We applied BRASS in the context of multiple testing correction in association mapping studies of ED and IE in domestic dogs. In the analysis of IE which involved dogs from a single breed, we found a 12 Mb region on chromosome 4 that reached the genome-wide significance threshold estimated using the replicates from BRASS, and has been previously associated

with IE [Hayward et al., 2016a]. On the other hand, in the ED data set which contained multiple breeds, we did not find any region that was close to the estimated genome-wide significance threshold. In the latter analysis, we chose to include PCs as covariates and these mostly captured genetic differences across breeds. Alternatively, as we had breed labeling information provided in the data set, we could have included that as an ancestry informative covariate instead of the PCs. However, in the presence of mislabeling in the breed information, using breed labels instead of PCs would result in inaccurate modeling of the breed structure present. In addition, as we had dogs of mixed breed in the sample (akin to admixture), the use of breed labels instead of PCs would not differentiate between these dogs (i.e. they could come from different breed combinations but would be treated as originating from the same breed combination). Notably, the genome-wide p-value threshold estimated using replicates from BRASS was an order of magnitude different for the ED trait in a sample of 82 breeds than for the IE trait in a single breed. This is expected given the extremely different population structure between the two samples.

BRASS assumes that an estimate of the phenotypic covariance matrix is available in order to capture polygenic effects, which can arise through various sources including population structure, cryptic and family relatedness. Here, we chose to use both fixed effects and random effects to capture the structure that is present in the data. This modeling choice was taken as a means to combine two common approaches to correct for population structure, which are either the use of PCs as covariates or the use of random effects in the trait model. Our approach does not attempt to distinguish whether distant (recent) sources of genetic relatedness are captured in the fixed (random) components of the model; we simply partition the relatedness into random and fixed effects and include both in our model. To the extent that population membership is discrete, using PCs that reflect that structure as covariates would provide more confidence that population structure has been well corrected for in the analysis. A recently proposed method, PC-Air [Conomos et al., 2015], obtains PCs from a derived subset of mutually unrelated subjects that are representative of the ancestral diversity

present in the sample, so as to ensure that the top PCs will only capture distant genetic relatedness (i.e. population structure). However, we did not investigate the use of such an approach in our analyses.

An alternative resampling strategy not considered here is to obtain replicates for the genotypes by using gene dropping. In this approach, which is applicable to samples that contain families, the genotypes for the pedigree founders are dropped down the pedigree assuming Mendelian segregation. Hence, the correlation present between the trait values of different individuals is preserved. The resulting simulated replicate can be used to get an empirical distribution of the test statistic considered under the null hypothesis of no association. A major caveat of gene dropping is that along with the presence of families in the sample, both the complete pedigree structure and the appropriate distribution from which to draw founders' genotypes have to be known in order to simulate transmissions down the pedigree; this is infeasible when cryptic relatedness is present in the sample or when the genealogy in earlier generations is unknown. Another resampling strategy we did not explore is based on permuting the genotypes. Since an exact permutation of the genotypes would not preserve the sample structure, we could propose a model based on the first two moments of the genotypes and obtain a linear map that projects them onto an orthonormal space, similarly to the approach in BRASS. This could be useful when looking at a smaller region in the genome as it involves the joint distribution of multiple markers, though with rare variants it would be challenging to obtain accurate estimate of the correlation among them.

We chose here to consider the application of BRASS in the context of performing a scan over a genetic region with multiple tests being performed and significance being assessed after correcting for the simultaneous tests. More generally, BRASS can be applied to situations where either the null distribution or the asymptotic null distribution of the statistic is intractable. Example of such situations arise in the analysis of rare variants, where the statistic is obtained by combining signals over multiple sites, when the statistic includes weights that depend on the phenotype, or when test statistics that perform well in different

scenarios are combined. Depending on the application, one can consider various approaches to capturing the structure present in the sample, as the model used by BRASS can accommodate both covariates and random effects, where the latter are represented by the inclusion of an additional matrix besides the identity in the trait covariance structure. We have shown here one example of how the structure can be incorporated in the model.

CHAPTER 3

A FAST METHOD FOR ASSESSING SIGNIFICANCE FOR A CERTAIN CLASS OF ASSOCIATION TESTS IN THE PRESENCE OF POPULATION STRUCTURE

3.1 Introduction

Advances in sequencing technologies have made it possible to obtain genetic information in large samples of individuals and identify many genetic variants, common and rare, across a wide range of complex human traits and disorders [Stranger et al., 2011; Visscher et al., 2012]. Many studies initially looked for associations between a single genetic variant and a single trait. Although such studies have been successful at building associations between various complex traits and genetic variants, many found only a moderate proportion of the narrow-sense heritability being explained by the GWA loci [Manolio et al., 2009]. Among the many explanations suggested, one was the low power involved when association is tested with rare variants. To overcome this limitation, gene or region-based tests have been proposed for association testing with a single trait, and are able to combine signals across multiple variants to improve power over single variant analyses [Lee et al., 2014]. Recent GWA studies have found genetic regions associated with multiple traits [Sivakumaran et al., 2011; Wang et al., 2015]. In such cases of genetic pleiotropy, joint analysis of the multiple traits may be more advantageous compared to combining the univariate analysis results of each single trait. For example, the joint analysis can provide better power gains by leveraging the dependence between the traits and thus boosting modest marginal association signals [Allison et al., 1998]. Many association tests have been developed in this context with some applicable for association testing with multiple variants [Broadaway et al., 2016; Galesloot et al., 2014; Wang et al., 2018; Wu and Pankow, 2016; Zhou and Stephens, 2014] .

In order to assess statistical significance, most association tests are either permutation-

based [Hua and Ghosh, 2015; Joo et al., 2016; Kim et al., 2016] or large sample-based [Broadaway et al., 2016; Dutta et al., 2018; Wu and Pankow, 2016; Wu et al., 2011]. The first approach relies on listing either all possible permutations or some fixed number of permutations, and using the empirical permutation distribution to estimate p-values. While it benefits from being robust to misspecifications in the underlying model, listing all permutations is computationally infeasible for large sample sizes while using Monte Carlo simulations can become computationally burdensome for very small p-value calculations. The second approach for significance assessment is based on large sample theory, where the asymptotic distribution of the test statistic is used to calculate p-values. While this approach is more computationally efficient than the permutation-based approach as the p-values can be obtained analytically, a major drawback is that it can lead to poor control of the type 1 error rate in small samples or when the number of traits is large [Chen et al., 2016b; Wu et al., 2011; Zhan et al., 2017b]. To obtain a method that has the strengths of both approaches, several association tests have been proposed based on a fast approximation to the permutation approach to assess significance in unrelated samples [Minas et al., 2013; Zhan et al., 2017a,b]. More precisely, a moment-matching procedure is used to approximate the empirical distribution of all possible permutations of the test statistic under the null hypothesis of no association. As this fast approximation approach is based on the null permutation distribution, it is not sensitive to violations of the assumptions underlying large sample-based methods. In addition, analytical expressions of the first three moments needed for moment-matching have been previously derived [Josse et al., 2008], which makes the fast approximation approach computationally efficient compared to standard permutation-based methods.

We address the problem of assessing significance of multivariate association tests in the presence of population structure, as it is well-known that failing to adjust for it, when present, can lead to inflated type 1 error rates as well as power loss. Most approaches used to account for populations structure either incorporate the structure through fixed effects or through random effects. The fixed effects approach usually consists of representing the

ancestry differences in the sample through a fixed set of covariates, which can be inferred from genome-wide data or ancestry informative markers. For example, one could extract the top l PCs of a GRM matrix estimate, where l is chosen to be low, and include those as covariates when testing for associations. Using residuals that remove covariate effects in the test statistic, either the large sample-based or the fast moment-matching approximation approach could be used to assess significance. However, a drawback of the fixed effects approach is that it assumes that the inferred covariates capture most of the sample structure and thus, are sufficient to correctly control the type 1 error rate. This could fail if there are strong patterns of relatedness, such as when related individuals are included in the sample, which would lead to inflated type 1 errors [Hoffman, 2013; Jiang et al., 2016].

Alternatively, the random effects approach estimates a dependence structure that reflects the genetic similarities in the sample and models it as correlation between individuals. For example, a GRM estimate based on genome scan data can be computed and incorporated in the model through a variance component, as is popularly done in linear mixed models. The test statistic could then be based on a transformation of the residuals from the null model that removes the estimated dependence structure. A multivariate test, Multi-SKAT [Dutta et al., 2018], has recently been proposed that is based on this approach and assesses the significance of the resulting test statistic using large sample theory. When the sample size is small or the number of traits is large, such tests may not work well for correct control of the type 1 error as the large-sample assumptions involved may not hold [Chen et al., 2016b; Lee et al., 2012a; Zhan et al., 2017b]. Furthermore, with high dimensional traits, these tests can become computationally burdensome as they often require the eigen-decomposition of matrices whose dimension depends on the number of traits.

We propose a very general approach to assessment of significance in structured samples, JASPER, (Joint Association analysis in Structured samples based on approximating a PERmutation distribution), which can be used to assess significance of multivariate tests by extending the fast moment-matching approximation method to account for population

structure. Similarly to the random effects approach, JASPER uses a variance component to incorporate sample structure in a model which considers the genotypes as random. An important feature of JASPER is that, unlike other approaches that consider the genotypes as random [Mbatchou et al., 2018; Schaid et al., 2013; Thornton and McPeek, 2007], JASPER can be used with multiple genetic markers simultaneously without the need to specify or estimate the LD structure among the markers. We derive the correction for population structure and detail the moment-matching approximation used for analytical p-value calculations. We demonstrate through simulation studies with various test statistics that JASPER effectively corrects for population structure and provides similar or better power to large sample-based assessments of significance across a wide range population structure settings and trait models. We also show in our simulations that, by using a model based on the genotypes to assess significance, JASPER is robust against various misspecifications in the model for the traits in terms of type 1 error control, and it provides better power performance than large sample-based assessments of significance, even more so with high dimensional traits. Finally, we illustrate the application of JASPER by evaluating the genetic regulation of gene expression levels within biological pathways in data from the Framingham Heart Study.

3.2 Methods

3.2.1 Notation

We consider a relatively general context in which multiple traits (either binary or quantitative) have been measured on n individuals (e.g. gene expression values in a biological pathway) and association is being tested with a group of genetic variants (e.g. variants within a single gene), where the genetic variants can be common or rare. Population substructure, admixture and relatedness, either known or cryptic, can be present in the sample. The method proposed to assess significance relies on two components. The first is a transformation of the test statistic based on a null model proposed for the genetic markers, where both the mean and

variance structure are specified and incorporate the correlation present due to individuals with similar genetic backgrounds. The second is an approximation of the null distribution of the transformed test statistic based on its first three moments.

Let $\mathbf{Y}$ denote the $n \times k$ matrix of k traits among n subjects, $\mathbf{G}$ denote the $n \times m$ genotype matrix for the m genetic variants, and $\mathbf{G}_{.j} = (G_{1j}, \ldots, G_{nj})^T$ denote the j^{th} column of $\mathbf{G}$, which is the genotype vector of length n for marker j, with G_{ij} being the minor allele count (0,1, or 2) for the i-th individual. We focus on association test statistics of the form

$$Q_T = \text{tr}\left(\widetilde{\mathbf{G}}\mathbf{K}_G\widetilde{\mathbf{G}}^T \cdot \widetilde{\mathbf{Y}}\mathbf{K}_Y\widetilde{\mathbf{Y}}^T\right) \tag{3.1}$$

where tr(.) denotes the trace operator for a matrix, $\widetilde{\mathbf{Y}}$ is a matrix obtained from $\mathbf{Y}$ and possibly covariates (e.g. a linear transformation of $\mathbf{Y}$ with $\widetilde{\mathbf{Y}} = \mathbf{AYB}$, for some $n \times n$ and $k \times k$ matrices $\mathbf{A}$ and $\mathbf{B}$, respectively, which could depend on covariates), $\widetilde{\mathbf{G}}$ is either $\mathbf{G}$ or some type of column-centered version of $\mathbf{G}$, possibly involving ancestry-informative vectors (more on this below), $\mathbf{K}_G$ and $\mathbf{K}_Y$ are positive semi-definite (p.s.d.) symmetric matrices (could be the identity matrix), and we assume $\mathbf{K}_G$ is independent of $\mathbf{Y}$ under the null hypothesis. We note that the matrix $\widetilde{\mathbf{Y}}$ could correspond to some transformation of residuals from a linear mixed model for the multiple traits $\mathbf{Y}$ with covariates such as sex or age (and possibly ancestry-informative covariates, where we could estimate these using existing software packages [Alexander et al., 2009; Conomos et al., 2015; Liu et al., 2013; Tang et al., 2005]), and a variance component that captures the genetic similarities present between individuals in the sample. Many test statistics are of the form in (3.1), such as SKAT [Wu et al., 2011], DKAT [Zhan et al., 2017b], MSKAT [Wu and Pankow, 2016], GAMuT [Broadaway et al., 2016], Multi-SKAT [Dutta et al., 2018] and CARAT [Jiang et al., 2016].

We further assume that at least one of $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{Y}}$ is mean-centered, i.e. either $\mathbf{1}_n^T\widetilde{\mathbf{G}} = \mathbf{0}$ or $\mathbf{1}_n^T\widetilde{\mathbf{Y}} = \mathbf{0}$, or more generally, when ancestry-informative covariates are involved, we assume that either $\mathbf{X}_G^T\widetilde{\mathbf{G}} = \mathbf{0}$ or $\mathbf{X}_G^T\widetilde{\mathbf{Y}} = \mathbf{0}$, where $\mathbf{X}_G$ is a $n \times q$ matrix of ancestry-

informative covariates that has $\mathbf{1}_n$ in its column space with $\text{rank}(\mathbf{X}_G) = q < n$. The special case of no ancestry-informative covariates corresponds to $\mathbf{X}_G = \mathbf{1}_n$. Note that in the case where $\mathbf{X}_G^T\widetilde{\mathbf{Y}} = \mathbf{0}$ and $\mathbf{X}_G^T\widetilde{\mathbf{G}} \neq \mathbf{0}$, if we let $\mathbf{J}_{n\times n} = \mathbf{I}_n - \mathbf{X}_G(\mathbf{X}_G^T\mathbf{X}_G)^{-1}\mathbf{X}_G^T$, then $Q_T = \text{tr}[(\mathbf{J}\widetilde{\mathbf{G}})\mathbf{K}_G(\mathbf{J}\widetilde{\mathbf{G}})^T\widetilde{\mathbf{Y}}\mathbf{K}_Y\widetilde{\mathbf{Y}}^T]$ since $\mathbf{J}\widetilde{\mathbf{Y}} = \widetilde{\mathbf{Y}}$. Hence, without loss of generality, we can take each column of $\widetilde{\mathbf{G}}$ to be centered by ancestry-informative covariates, if any, and to be mean-centered, i.e. we can assume $\widetilde{\mathbf{G}} = \mathbf{J}\mathbf{G}$. In the following derivations, we concentrate on the multivariate case where multiple traits and multiple markers are considered, but we note that our method also applies to the univariate case (i.e. single trait and/or single marker).

3.2.2 Null model for first and second moments of genotypic random variables

The approximation used to estimate the p-value for the test statistic Q_T would rely on the rows of either the genotype matrix $\widetilde{\mathbf{G}}$ or the phenotype matrix $\widetilde{\mathbf{Y}}$ being exchangeable, which in general is not true. In order to achieve robustness to trait model misspecification while properly controlling the type 1 error rate, we consider a null model that views the genotypes as random and use it to transform the genotype matrix $\widetilde{\mathbf{G}}$ so as to obtain exchangeable rows.

We use the following quasi-likelihood model for the genotypes under the null hypothesis of no association

$$\begin{aligned} \text{E}[\text{vec}(\mathbf{G})|\mathbf{Y}] &= \text{vec}(\mathbf{X}_G\mathbf{A}), \text{ and} \\ \text{Var}[\text{vec}(\mathbf{G})|\mathbf{Y}] &= \mathbf{F} \otimes \mathbf{\Phi}, \end{aligned} \tag{3.2}$$

where $\text{vec}(\mathbf{M}_{r\times c})$ represents the vectorization of the matrix $\mathbf{M}$ by its columns into a rc-length vector, $\mathbf{A}$ is a $q \times m$ matrix of unknown coefficients, and in the special case where $\mathbf{X}_G = \mathbf{1}_n$, $\mathbf{A}^T/2$ is a length m vector of unknown allele frequencies for the m tested markers, $\mathbf{F} = (F_{ij})$ is an unknown $m \times m$ p.s.d. matrix representing the LD structure among the markers, and $\mathbf{\Phi}$ is a $n \times n$ p.s.d. matrix representing the kinship or genetic relatedness between pairs of

individuals in the sample. The covariance structure presented in (3.2) assumes the correlation structure between individuals is the same for all m tested markers, and similarly, that the correlation structure between different markers is the same for all individuals. For a single marker j, (3.2) is equivalent to

$$\begin{aligned} \mathrm{E}(\mathbf{G}_{\cdot j}|\mathbf{Y}) &= \mathbf{X}_G\mathbf{A}_j, \text{ and} \\ \mathrm{Var}(\mathbf{G}_{\cdot j}|\mathbf{Y}) &= F_{jj}\cdot\mathbf{\Phi}, \end{aligned} \tag{3.3}$$

where $\mathbf{A}_j$ is the j^{th} column of $\mathbf{A}$. The moment conditions in (3.3) have previously been used in retrospective association tests with single markers such as MASTOR [Jakobsdottir and McPeek, 2013] and CERAMIC [Zhong et al., 2016].

From the moment conditions for $\mathbf{G}$, we obtain the following moment conditions for $\widetilde{\mathbf{G}} = \mathbf{J}\mathbf{G}$,

$$\begin{aligned} \mathrm{E}[\mathrm{vec}(\widetilde{\mathbf{G}})|\mathbf{Y}] &= \mathbf{0}, \text{ and} \\ \mathrm{Var}[\mathrm{vec}(\widetilde{\mathbf{G}})|\mathbf{Y}] &= \mathbf{F}\otimes\widetilde{\mathbf{\Phi}}, \end{aligned} \tag{3.4}$$

where $\widetilde{\mathbf{\Phi}} = \mathbf{J}\mathbf{\Phi}\mathbf{J}$. Note that $\widetilde{\mathbf{\Phi}}\mathbf{X}_G = \mathbf{0}$. In what follows, we usually take $\widetilde{\mathbf{\Phi}}$ to be a GRM estimated from genome-wide data, where $\widetilde{\mathbf{\Phi}}$ is symmetric p.s.d. and $\widetilde{\mathbf{\Phi}}\mathbf{1}_n = \mathbf{0}$ (a property that commonly arises in GRMs estimated with no missing marker data, or else we enforce this condition by pre- and post-multiplying the GRM by $\mathbf{J}$). More generally, with ancestry-informative covariates $\mathbf{X}_G$, we require $\widetilde{\mathbf{\Phi}}\mathbf{X}_G = \mathbf{0}$, which would arise under certain approaches to estimating the GRM such as those used in Chapter 2 of this thesis, or else we could enforce this condition as above. Alternatively, if there was a known kinship matrix $\mathbf{\Phi}$, we could use $\widetilde{\mathbf{\Phi}} = \mathbf{J}\mathbf{\Phi}\mathbf{J}$.

We take the eigen-decomposition of $\widetilde{\mathbf{\Phi}}$ to be $\widetilde{\mathbf{\Phi}} = \mathbf{V}\mathbf{D}\mathbf{V}^T$, where $\mathbf{D} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ is a diagonal matrix containing the sorted (in decreasing order) eigenvalues of $\widetilde{\mathbf{\Phi}}$, and $\mathbf{V}$ is an orthogonal matrix containing the corresponding eigenvectors in its columns. Since

$\widetilde{\mathbf{\Phi}}$ is orthogonal to $\mathbf{1}_n$ (or, more generally, to $\mathbf{X}_G$), we have that 0 is an eigenvalue of $\widetilde{\mathbf{\Phi}}$, and so, for the desired transformation $\widetilde{\mathbf{\Phi}}^{-1/2}\widetilde{\mathbf{G}}$, we take $\widetilde{\mathbf{\Phi}}^{-1/2} = (\mathbf{D}^-)^{1/2}\mathbf{V}^T$, with $\mathbf{D}^- = \text{diag}\{1/\lambda_1, \ldots, 1/\lambda_{n_+}, 0, \ldots, 0\}$ where $\lambda_1, \ldots, \lambda_{n_+}$ are the n_+ positive eigenvalues of $\widetilde{\mathbf{\Phi}}$, with $n_+ \leq n-1$ (e.g. when the $n \times q$ matrix $\mathbf{X}_G$ of ancestry-informative covariates is used to center $\mathbf{\Phi}$, $n_+ = n - q$). We have

$$\text{E}[\text{vec}(\widetilde{\mathbf{\Phi}}^{-1/2}\widetilde{\mathbf{G}})|\mathbf{Y}] = \mathbf{0}, \tag{3.5}$$

$$\text{Var}[\text{vec}(\widetilde{\mathbf{\Phi}}^{-1/2}\widetilde{\mathbf{G}})|\mathbf{Y}] = \mathbf{F} \otimes \begin{pmatrix} \mathbf{I}_{n_+} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{3.6}$$

with the last $(n - n_+)$ rows of $\widetilde{\mathbf{\Phi}}^{-1/2}\widetilde{\mathbf{G}}$ being zero. As the first n_+ rows of $\widetilde{\mathbf{\Phi}}^{-1/2}\widetilde{\mathbf{G}}$ all have mean $\mathbf{0}$ and covariance matrix $\mathbf{F}$, and hence, are second-order exchangeable, we re-write the test statistic in (3.1) as

$$Q_T = \text{tr}\left(\check{\mathbf{G}}\check{\mathbf{G}}^T \cdot \check{\mathbf{Y}}\check{\mathbf{Y}}^T\right), \tag{3.7}$$

where $\check{\mathbf{G}} = \widetilde{\mathbf{\Phi}}^{-1/2}\widetilde{\mathbf{G}}\mathbf{K}_G^{1/2}$, $\check{\mathbf{Y}} = \widetilde{\mathbf{\Phi}}^{1/2}\widetilde{\mathbf{Y}}\mathbf{K}_Y^{1/2}$, with $\widetilde{\mathbf{\Phi}}^{1/2} = \mathbf{D}^{1/2}\mathbf{V}^T$, and the equality in (3.7) holds because $(\widetilde{\mathbf{\Phi}}^{1/2})^T\widetilde{\mathbf{\Phi}}^{-1/2} = \mathbf{J}$ and $\mathbf{J}\widetilde{\mathbf{G}} = \widetilde{\mathbf{G}}$ was assumed. Similarly to $\widetilde{\mathbf{\Phi}}^{-1/2}\widetilde{\mathbf{G}}$, the first n_+ rows of $\check{\mathbf{G}}$ remain second-order exchangeable because

$$\text{E}[\text{vec}(\check{\mathbf{G}})|\mathbf{Y}] = \mathbf{0}, \text{ and, } \text{Var}[\text{vec}(\check{\mathbf{G}})|\mathbf{Y}] = \mathbf{K}_G^{1/2}\mathbf{F}\mathbf{K}_G^{1/2} \otimes \begin{pmatrix} \mathbf{I}_{n_+} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

i.e. multiplication by $\mathbf{K}_G^{1/2}$ affects the covariance structure only between the columns of $\widetilde{\mathbf{\Phi}}^{-1/2}\widetilde{\mathbf{G}}$ for any given row, where this structure is the same for every row.

3.2.3 Novel moment-matching approximation for the null permutation distribution of Q_T

In principle, since the first n_+ rows of $\check{\mathbf{G}}$ are second-order exchangeable, we could obtain the null distribution of Q_T by assessing its value on each of the $n_+!$ possible permutations of these rows, or on a random sample of the permutations. As this approach is computationally intensive, we propose to use a moment-matching procedure based on approximating this null distribution with a Pearson type III distribution using exact analytical calculations of the first three moments of Q_T. Previous work [Josse et al., 2008] has demonstrated the effectiveness of the Pearson type III distribution to adequately capture the skewness component of permutation distributions. A key advantage of this approach is that it eliminates the need to explicitly carry out the permutations themselves and thus drastically decreases the computational cost involved in assessing significance.

Closed-form expressions for the first three moments of Q_T have been previously derived [Kazi-Aoual et al., 1995] under more restrictive assumptions: (1) $\mathbf{1}_n^T\check{\mathbf{G}}_+ = \mathbf{0}$, and (2) $\mathbf{1}_n^T\check{\mathbf{Y}}_+ = \mathbf{0}$, where $\check{\mathbf{G}}_+$ and $\check{\mathbf{Y}}_+$ refer to the n_+ non-zero rows of $\check{\mathbf{G}}$ and $\check{\mathbf{Y}}$ in (3.7), respectively. However, both of these two assumptions generally fail to hold unless the correlation between sampled individuals is fully captured through a set of ancestry-informative covariates, i.e. it fails to hold unless $\mathbf{D}$ is proportional to

$$\begin{pmatrix} \mathbf{I}_{n_+} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

which is not very realistic in practice. In the case of unrelated samples (or when the structure is assumed to be captured by a set of ancestry-informative covariates in the trait model), the Pearson type III approximation has been used to assess significance of test statistics of the form of Q_T using the closed-forms expressions of the first three moments [Minas et al., 2013; Zhan et al., 2017b]. However, with arbitrary structure present in the sample, more general

analytical expressions of the three moments of Q_T are needed. We relax the exchangeability assumption of the n_+ non-zero rows of $\check{\mathbf{G}}$ to second-order exchangeability and thus assume that all $n_+!$ permutations of the rows are equally likely. The analytical expression for the first moment of Q_T is

$$\mathrm{E}(Q_T) = \frac{1}{n_+} T_1^Y T_1^G + \frac{1}{n_+(n_+-1)}(S_1^Y - T_1^Y)(S_1^G - T_1^G), \tag{3.8}$$

where $T_1^Y = \mathrm{tr}(\mathbf{W}^Y)$ and $S_1^Y = \sum_{ij} W_{ij}^Y$ for $\mathbf{W}^Y = \check{\mathbf{Y}}_+ \check{\mathbf{Y}}_+^T$, and the corresponding quantities T_1^G, S_1^G and $\mathbf{W}^G$ are computed based on $\check{\mathbf{G}}_+$. Similarly, closed-form expressions for the second and third moment of Q_T are derived in terms of a small number of quantities that are easily computed from $\mathbf{W}^G$ and $\mathbf{W}^Y$ (see Appendix). The mean, variance and skewness of Q_T are computed as

$$\mu = \mathrm{E}(Q_T), \tag{3.9}$$

$$\sigma^2 = \mathrm{E}(Q_T^2) - \mu^2, \tag{3.10}$$

$$\gamma = \frac{\mathrm{E}(Q_T^3) - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \tag{3.11}$$

We approximate the null permutation distribution of Q_T using a Pearson type III distribution with scale parameter $a = \gamma\sigma/2$, , shape $b = 4/\gamma^2$, and location parameter $c = \mu - \frac{2\sigma}{\gamma}$, where the probability density function is

$$f(x) = \frac{1}{|a|^b \Gamma(b)} |x-c|^{b-1} e^{-\frac{x-c}{a}},$$

for $a \neq 0, b > 0$ and $\frac{x-c}{a} \geq 0$. The p-value is computed analytically using the observed value of Q_T and the approximated Pearson type III distribution.

3.3 Simulation studies

The JASPER method for assessing significance is potentially applicable to a wide range of contexts for association testing. To evaluate the method in simulations, we consider two different types of association testing: (1) multivariate association testing between a set of quantitative traits and a set of SNPs, and (2) association testing between a single binary trait and a set of SNPS. We perform simulation studies to (1) evaluate the impact of our assessment of significance on the type I error rate of different association test statistics, (2) determine how robust our method is to various sources of misspecification in the trait model, (3) compare the performance of the selected association tests when our method and a large sample approximation are used to assess significance for high dimensional traits. Genotype, trait and covariates are simulated under multiple population structure settings and trait models, where we simulate $k = 1$, 2, 5, 50 or 100 traits.

3.3.1 Simulation study design

The models of both settings (multiple quantitative traits and a single binary trait) have a number of common features such as three covariates, two unobserved major genes, effects of multiple tested SNPs, and effects of sample structure including an ancestry effect and additive polygenic variance.

In each simulation setting, we simulate 100,000 markers that are used to correct for population structure. The genotypes are simulated using the Balding-Nichols model [Balding and Nichols, 1995], where for each marker, the ancestral allele frequency p is drawn independently from a uniform distribution on $[0.2, .0.8]$, and given p, the allele frequency in sub-population t $(t = 1, 2)$ is drawn from a Beta distribution with parameters $p\frac{1-F}{F}$ and $(1-p)\frac{1-F}{F}$ (independently across subpopulations), with the fixation index F set to 0.01 to reflect closely related populations. In simulation settings with population admixture, the allele frequency for an individual is given by $(a_1 p_1 + a_2 p_2)$, where $a = (a_1, a_2)$ is the

admixture vector of the individual and $a_1 + a_2 = 1$. Two causal genes are also simulated using the same Balding-Nichols model, with the ancestral allele frequencies simulated independently from a uniform distribution on $[0.2, 0.8]$.

Three covariates are simulated: sex, age and an i.i.d. standard normal trait. The first two are commonly used as covariates in association studies and the last one is chosen to represent an environmental covariate. When no familial relatedness is present in the sample, sex is generated as a Bernoulli(0.5) variable and age is a variable uniformly distributed between 20 and 60. When the sample consists of related individuals, sex is determined by the pedigree structure shown in Figure 2.1, with a circle representing a female, and a square representing a male. In addition, age is a variable uniformly distributed within 1.5 years of 73, 75, 46, 43, 40, 46, 40, 43, 47, 51, 18, 21, 15, 15, 12, 9, 13, 17, 24, 21, 18, and 14 years respectively for individuals labeled $1 - 22$ in a given pedigree whose structure is shown in Figure 2.1. For all simulation scenarios, the covariate values are assumed to be independent across individuals and are regenerated for each replicate.

We consider two models to generate the traits given the genotype and covariate data. In the first model, referred to as Model I, we generate k quantitative traits as

$$Y_{ij} = \mathbf{X}_i^T \boldsymbol{\beta} + \alpha_1 M_{1,i} + \alpha_2 M_{2,i} + \mathbf{G}_i^T \boldsymbol{\gamma}_j + \rho A_i + u_{ij} + e_{ij} \tag{3.12}$$

where Y_{ij} is the value of the j-th trait for the i-th individual; $\mathbf{X}_i$ is the length p covariate vector for individual i; $\boldsymbol{\beta}$ is a length p vector of fixed covariate effects assumed to be the same across all traits; $M_{1,i}$ and $M_{2,i}$ are the genotypes of individual i at the two unobserved major genes coded as 0,1 or 2; α_1 and α_2 are their effects on the trait and are assumed to be the same across all traits; $\mathbf{G}_i$ is the genotype vector at the m tested genetic markers; $\boldsymbol{\gamma}_j$ is a length m vector of fixed effects for trait j; A_i is the proportion of ancestry from population 1 for individual i; ρ is the fixed effect of ancestry on a trait and is assumed to be the same

across all traits; u_{ij} and e_{ij} are random effects distributed as

$$\begin{aligned}\text{vec}(\mathbf{u}) &\sim \text{MVN}_{nk\times nk}(\mathbf{0}, \boldsymbol{\Sigma}_u \otimes \mathbf{K}), \text{ for } \mathbf{u} := (u_{ij}),\\ \text{vec}(\mathbf{e}) &\sim \text{MVN}_{nk\times nk}(\mathbf{0}, \boldsymbol{\Sigma}_e \otimes \mathbf{I}_n), \text{ for } \mathbf{e} := (e_{ij}),\end{aligned}$$

where $\mathbf{u}$ represents additive polygenic effects due to pedigree relatedness; $\mathbf{K}$ is the kinship matrix resulting from the pedigree relatedness, and $\boldsymbol{\Sigma}_u$ introduces correlation between the k traits due to the additive polygenic effects; $\mathbf{e}$ represents random errors that are independent across individuals; and $\boldsymbol{\Sigma}_e$ represents the residual correlation between traits for a given individual. The term u_{ij} is (3.12) is only included for some simulation settings as a way to introduce correlation between traits due to pedigree relatedness, when present.

We also consider the simulation of a single binary trait using a liability threshold model, referred to as Model II, where the phenotype value Y_i for individual i is given by

$$\begin{aligned}&Y_i = 1 \iff L_i > 0,\\ &L_i = \mathbf{X}_i^T\boldsymbol{\beta} + \alpha_1 M_{1,i} + \alpha_2 M_{2,i} + \mathbf{G}_i^T\boldsymbol{\gamma} + \rho A_i + u_i + e_i, \qquad (3.13)\end{aligned}$$

where $Y_i = 1$ indicates that individual i is affected (i.e. case subject) and $Y_i = 0$ that i is unaffected (i.e. control subject); L_i is the underlying liability value for individual i; $\boldsymbol{\gamma}$ is a length m vector of fixed effects for the tested markers; $(u_1, \ldots, u_n)^T \sim MVN(0, \sigma_u^2\,\mathbf{K})$ is the random vector of additive polygenic effects, with σ_u^2 being a scalar parameter that reflects the additive polygenic variance; and the random errors $e_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2)$.

We consider two different covariance structures for $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ depending on how many traits are being simulated. In the case where $k \leq 10$, we use the commonly used compound

symmetric structure [Dutta et al., 2018; Zhan et al., 2017b], which is

$$\boldsymbol{\Sigma}_u = \sigma_u^2 \begin{pmatrix} 1 & \rho_u & \dots & \rho_u \\ & \ddots & \ddots & \vdots \\ & & 1 & \rho_u \\ & & & 1 \end{pmatrix}, \text{ and } \boldsymbol{\Sigma}_e = \sigma_e^2 \begin{pmatrix} 1 & \rho_e & \dots & \rho_e \\ & \ddots & \ddots & \vdots \\ & & 1 & \rho_e \\ & & & 1 \end{pmatrix}, \tag{3.14}$$

where ρ_u and ρ_e reflect the correlation across the traits due to additive polygenic effects and random errors, respectively, and we assume the variance due to additive polygenic effects and random errors (i.e. σ_u^2 and σ_e^2) are assumed to be the same for all traits. We fix $\rho_u = 0.3$ and consider two levels for the correlation parameter $\rho_e \in \{0.2, 0.5\}$ to reflect low and moderate amounts of correlation between the k traits due to the random errors. In higher dimensional settings (k=50 or 100), we use a different structure that is based on having clusters within the k traits, with the correlation within clusters being higher than that between different clusters. We use non-overlapping clusters of 10 traits each, where for a given cluster, the correlation structure is the same as in (3.14), where we also fix $\rho_u = 0.3$ and have $\rho_e \in \{0.2, 0.5\}$. However, we set the correlation between traits in different clusters to be 0.05 for the polygenic effects and for the random errors, we set it to 0.01 if ρ_e is 0.2 and to 0.1 if ρ_e is 0.5 within clusters.

For our population structure scenarios, we consider two settings where in the first, which we refer as "2 Subpopulations", individuals are sampled from a stratified population with two ancestral subpopulations, and in the second, which we refer as "Admixture", individuals are sampled from an admixed population with two ancestral subpopulations. In the "2 Subpopulations" setting without relatedness, 500 individuals are sampled from each subpopulation to obtain a sample of 1,000 individuals. When relatedness is present, 23 pedigrees of size 22 (as in Figure 2.1) are sampled from each subpopulation to obtain a sample of 1012 individuals. In the "Admixture" setting without pedigree relatedness, 1000 individuals are simulated with i.i.d. admixture proportions sampled from a uniform distribution on

$[0, 1]$. In the presence of relatedness, 46 pedigrees of size 22 are simulated with an admixture proportion sampled independently for each pedigree from a uniform distribution on $[0, 1]$.

Population structure is introduced in the generating models of the traits in (3.12) and (3.13) by the ancestry effect ρA_i and the polygenic additive effect u_{ij} (or u_i for the model in (3.13)), where the parameter $\rho \neq 0$ implies different phenotypic means in each ancestral subpopulation. Increases in ρ and σ_u^2 correspond to more severe population structure in the sample. In the linear mixed model of Equation (3.1), the parameters $\boldsymbol{\beta}, \alpha_1, \alpha_2, \rho, \sigma_u^2$ and σ_e^2 are chosen to satisfy the following conditions for each trait: (1) The total trait variance is set to 1 and the mean to 0; (2) The three covariates explain an equal amount of the total trait variance; (3) The unobserved major genes each explain 1.5% of the total trait variance; (4) The total variance explained by the ancestry effect (ρA_i) and additive polygenic effect (u_{ij}) is set to 20% of the total trait variance; (5) For the variance in (4), either it is completely explained by the ancestry fixed effect (in the case of no relatedness), or half of it is explained by the ancestry effect with the remaining half explained by the polygenic additive effect (in the case of relatedness) (6) The random error e_{ij} explains 20% of the total trait variance; (7) Given the variance explained by $\mathbf{G}_i^T \boldsymbol{\gamma}_j$, the details of which is explained in the next two paragraphs, covariates explain the remaining amount of trait variance. For the liability threshold model in (3.13), we modify condition (6) so that the random error e_i explains 60% of the total trait variance (on the liability scale), and keep the other conditions the same but applied on the liability scale.

For each given setting, the 100,000 markers used to correct for population structure are re-simulated three times and in each simulation, Equation (2.14) is used to obtain a GRM estimate $\hat{\boldsymbol{\Phi}}$. The markers are split into 2,000 non-overlapping panels containing $m = 50$ markers each. In type 1 error simulations, phenotypes are re-simulated 100 times and 100 marker panels are tested for association with the simulated phenotype. Overall, 5 tests are performed for each of the 6,000 marker panels, which results in 30,000 replicates used for type 1 error estimation. Model I in (3.12) and II in (3.13) are used to generate the traits

with $\boldsymbol{\gamma}_j = \mathbf{0}$ and $\boldsymbol{\gamma} = \mathbf{0}$, respectively.

In power simulations, for each given setting, we only simulate the 100,000 markers once and select a panel of $m = 50$ markers to be causal and tested for association with the traits. Model I is solely used to generate k traits, where only a subset of the traits are associated with the selected marker panel. This is because it is unlikely in real applications for all the traits to be associated with the same set of markers. For $k = 2, 5, 50, 100$, we randomly choose 1, 2, 5 and 10 of the traits to be associated with the chosen marker panel, respectively. For each associated trait j, $\boldsymbol{\gamma}_j$ in Model 1 is chosen using the following conditions (independently across associated traits): (1) The overall effect of the m markers explain 2.5% (if $k \leq 5$) or 3% (if $k > 5$) of the total trait variance; (2) Half of the m markers are chosen at random to be causal, the rest are null; (3) Among the causal markers, half of the markers have positive effects and the other half have negative effects. For the remaining traits, which are not associated with the marker panel, $\boldsymbol{\gamma}_j$ is set to $\mathbf{0}$. Phenotypes are re-simulated 1,000 times and the marker panel is tested for association with each phenotype replicate. This results in 1,000 replicates used for power estimation. For both type 1 error and power simulations, the two major genes included in the Models I and II are assumed to be unobserved and so are not included as covariates when testing for association.

3.3.2 Association test statistics considered

In the multivariate quantitative trait simulations, we consider two different test statistics, T_1 and T_2, which represent variance component score tests for two different linear mixed models. For the binary trait simulations, we consider test statistic T_3 which is a score test in a logistic mixed model. All three test statistics have the required form of (3.1). Test statistics T_1 and T_2 are based on the following multivariate quantitative trait model,

$$\mathbf{Y} = \mathbf{XB} + \tau_a \mathbf{GA} + \tau_u \mathbf{U} + \mathbf{E}, \tag{3.15}$$

where $\mathbf{Y}$ is a $n \times k$ phenotype matrix; $\mathbf{X}$ is a $n \times p$ covariate matrix; $\mathbf{B}$ a $p \times k$ matrix of fixed effects; $\mathbf{G}$ is a $n \times m$ matrix of genotypes at tested markers; τ_a is an unknown scalar; $\mathbf{A}$ is a $m \times k$ matrix of random effects for the tested markers, where we do not specify the full distribution for $\mathbf{A}$ but only require that $\mathrm{E}[\mathrm{vec}(\mathbf{A})] = \mathbf{0}$ and $\mathrm{Var}[\mathrm{vec}(\mathbf{A})] = \mathbf{C}_e \otimes \mathbf{I}_m$; τ_a is an known scalar set to 0 for the model used for T_1 and set to 1 for the model used for T_2; $\mathrm{vec}(\mathbf{U}) \sim MVN(0, \mathbf{C}_u \otimes \mathbf{K})$ where $\mathbf{U}$ is a $n \times k$ matrix of random effects for the additive polygenic effects; and $\mathrm{vec}(\mathbf{E}) \sim MVN(0, \mathbf{C}_e \otimes \mathbf{I}_n)$ where $\mathbf{E}$ is a $n \times k$ matrix of random errors. For both T_1 and T_2, the matrix $\widetilde{\mathbf{Y}}$ in (3.1) will have the form,

$$\mathrm{vec}(\widetilde{\mathbf{Y}}) = \mathbf{V}_e^{-1}\mathbf{H}_X\mathrm{vec}(\mathbf{Y}), \tag{3.16}$$

where $\mathbf{V}_e$ and $\mathbf{H}_X$ are $nk \times nk$ matrices that are functions of $\mathbf{X}$ and $\mathbf{Y}$ (but not $\mathbf{G}$). T_1 is derived as a score test statistic for the null hypothesis $H_0 : \tau_a = 0$ in the reduced model in which τ_u is set to 0, which results in $\mathbf{H}_X = \mathbf{I}_k \otimes \mathbf{P}_X$ and $\mathbf{V}_e = \mathbf{C}_e \otimes \mathbf{I}_n$, with $\mathbf{P}_X = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. As the $k \times k$ covariance matrix $\mathbf{C}_e$ is unknown, it is replaced by the estimate $\widehat{\mathbf{C}}_e = \mathbf{Y}^T\mathbf{P}_X\mathbf{Y}/(n-p)$. The second statistic T_2 is obtained as the score test statistic for the null hypothesis $H_0 : \tau_a = 0$ in the full model of (3.15) with τ_u set to 1, and results in $\mathbf{H}_X = \mathbf{I}_{nk} - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}^T\mathbf{V}_e^{-1}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{V}_e^{-1}$ and $\mathbf{V}_e = (\mathbf{C}_u \otimes \mathbf{K} + \mathbf{C}_e \otimes \mathbf{I}_n)$, with $\widetilde{\mathbf{X}} = \mathbf{I}_k \otimes \mathbf{X}$. The $k \times k$ covariance matrices $\mathbf{C}_u$ and $\mathbf{C}_e$ are unknown, and they estimated under the null hypothesis ($\tau_a = 0$) giving $\widehat{\mathbf{C}}_u$ and $\widehat{\mathbf{C}}_e$, respectively, which depend on $\mathbf{X}$ and $\mathbf{Y}$ (but not $\mathbf{G}$). To obtain $\widehat{\mathbf{C}}_u$ and $\widehat{\mathbf{C}}_e$, we use GEMMA [Zhou and Stephens, 2012, 2014] in the case where $k \leq 5$, and PHENIX [Dahl et al., 2016] in the case where $k > 5$. For both T_1 and T_2, when they are written in the form of (3.1), the resulting matrices $\mathbf{K}_G$ and $\mathbf{K}_Y$ will be $\mathbf{I}_m$ and $\widehat{\mathbf{C}}_e$, respectively.

For a single binary trait, we use the test statistic T_3 which is based on a logistic mixed model [Lee et al., 2012a; Lin, 1997] in which,

$$Y_i|\mathbf{X}, \mathbf{G}, \boldsymbol{\gamma} \sim \text{Bernouilli}(\pi_i), \text{ independently for } 1 \leq i \leq n, \text{ and,}$$

$$\text{logit}(\pi_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{j=1}^{m} \mathbf{G}_{ij}\gamma_j, \tag{3.17}$$

where $\mathbf{X}_i$ is the covariate vector for individual i with fixed effects vector $\boldsymbol{\beta}$, G_{ij} is the genotype at the j-th tested marker, and $\gamma_1 \ldots, \gamma_m$ are independent from some distribution F with mean 0 and variance τ. T_3 corresponds to a score test for $H_0 : \tau = 0$ and has the form,

$$T_3 = \text{tr}\left[\mathbf{G}\mathbf{G}^t \cdot (\mathbf{Y} - \widehat{\boldsymbol{\pi}})(\mathbf{Y} - \widehat{\boldsymbol{\pi}})^T\right], \tag{3.18}$$

where $\widehat{\boldsymbol{\pi}}$ is the estimate of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^T$ obtained by fitting the logistic model in (3.17) under H_0. In the form of (3.1), the resulting matrices $\mathbf{K}_G$ and $\mathbf{K}_Y$ will be $\mathbf{I}_m$ and $\mathbf{I}_k$, respectively.

The models used to derive the test statistics T_1 and T_3 assume conditional independence of the trait values for different individuals under the null hypothesis given the covariates, which is not the case in our simulations (i.e. the phenotype models are misspecified as they ignore population structure). Hence, we consider including the top l PCs of the GRM estimate as covariates in the null models for T_1 and T_3, with l being either 1 or 10. In simulations, T_1 and T_2 are used with quantitative traits simulated from Model I, and T_3 is used with binary traits simulated with Model II.

We compare different methods to assess significance for T_1, T_2 and T_3. For T_1, we use for comparison MSKAT [Wu and Pankow, 2016], which uses Davies' method to analytically compute asymptotic p-values of T_1. The resulting procedure is referred to as T_1-asymp in what follows. For T_1, we also use DKAT [Zhan et al., 2017b], which, like JASPER, uses a Pearson type III approximation to obtain p-values, but, unlike JASPER, assumes exchangeability of the rows of either $\check{\mathbf{G}}$ or $\check{\mathbf{Y}}$ under the null, which would hold if, for example, there was no population structure or if all the structure was captured by the PCs included as covariates (or if there was a very simple type of structure that preserved exchangeability). The resulting procedure is referred to as T_1-perm in what follows. For T_2, we compare our

method to Multi-SKAT [Dutta et al., 2018], which also uses Davies' method to compute asymptotic p-values. For high dimensional traits ($k \geq 50$), we were forced to reduce the number of replicates used to assess the type 1 error rate of Multi-SKAT to 1,000 owing to the computational burden of the Multi-SKAT method. The resulting procedure is referred to as T_2-asymp. Lastly, for T_3, we compare our method to SKAT adapted to binary traits with a moment matching adjustment to compute asymptotic p-values [Lee et al., 2016]. The resulting procedure is referred to as T_3-asymp. When we apply JASPER to assess significance of test statistic T_i $(i = 1, 2, 3)$, we refer to the resulting procedure as T_i-newperm.

3.3.3 Real Data Application

We illustrate the use of our method by analyzing data from the Framingham Heart Study (FHS) [Feinleib et al., 1975], which is a large longitudinal observational study that includes both unrelated individuals and individuals from multi-generation pedigrees. The goal of our data analysis is to investigate the local (i.e. *cis*) genetic architecture of gene expression in select biological pathways that have previously been associated with a wide range of disorders including inflammatory bowel disease, coronary heart disease, and type 2 diabetes, all of which are highly investigated complex diseases. We focus on Cohort 1 (i.e. offspring cohort) and 11 gene pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [Kanehisa and Goto, 2000] (see Table 3.1).

Gene expression profiling was carried out on a sample of 2442 offspring subjects during examination 8 (2005-2008) using the Affymetrix Human Exon 1.0 ST Gene Chip platform, resulting in the profiling of 17,873 measured transcripts clusters. After removing transcripts not mapping to RefSeq genes, 17,601 probes corresponding to 17,379 distinct genes remained. The expression values were normalized and residuals were obtained from a linear mixed model which included ten technical covariates (chip batch, RNA quality and 8 quality control metrics provided by Affymetrix APT program) [Joehanes et al., 2013]. We exclude HLA genes from the analysis, as these are the most polymorphic genes in the human genome and

Table 3.1: List of KEGG pathways analyzed with the FHS data.

Pathway name	KEGG ID	Number of Genes[a]
Antigen processing and presentation	hsa04612	42
Cytokine-cytokine receptor interaction	hsa04060	247
JAK-STAT signaling pathway	hsa04630	141
Inflammatory bowel disease	hsa05321	49
Autophagy	hsa04140	106
cGMP-PKG signaling pathway	hsa04022	151
T cell receptor signaling pathway	hsa04660	94
Neurotrophin signaling pathway	hsa04722	111
Wnt signaling pathway	hsa04310	144
PPAR signaling pathway	hsa03320	66
Notch signaling pathway	hsa04330	43

[a] Excludes HLA genes, genes not on chromosomes 1-22 and genes that don't have transcripts in the FHS data

known hotspots for disease associations [Kennedy et al., 2017], and also exclude genes not located on autosomal chromosomes 1 to 22. Both age at the time of the exam and sex are included as covariates in the analysis.

Genotype data is obtained using the Affymetrix 500K mapping array set and we exclude from the analysis individuals who (1) have completeness (proportion of markers with successful genotype calls) $< 96\%$; (2) have more than 10% missing genotypes in any of the genetic regions (i.e. collection of *cis*-SNPs for a single gene) analyzed; (3) have empirical self-kinship coefficient greater than $> .525$ (i.e. empirical inbreeding coefficient $> .05$). We also exclude a few individuals whose off-diagonal empirical kinship coefficient values appear inconsistent with the given pedigree information. This results in 1894 individuals with covariate, genotype and expression data that are included in the analysis.

For each gene in a given KEGG pathway, we extract all the polymorphic sites on the Affymetrix 500K chip that are within 500kb of the gene. We exclude sites that have (1) call rate $\leq 96\%$, and (2) Mendelian error rate $> 2\%$. For the remaining sites, we impute any missing genotypes using IMPUTE2 [Howie et al., 2009]. This yields a total of 182,447 SNPs mapped to 1,194 genes across the 11 KEGG pathways, with a median number of *cis*-SNPs

for any given gene of 152. We conduct 3 sets of association analyses in which we evaluate the association between *cis*-SNPs of a given gene with (1) the expression values of the gene; (2) the expression values of all genes in the pathway excluding the gene; (3) the expression values of all genes in the pathway including the gene. The three sets of analyses are referred to as 'SINGLE' (for single gene), 'LOGO' (for leave one gene out), and 'ALL' (for all genes in the pathway), respectively.

3.4 Results

3.4.1 Type 1 error and power studies

In simulations, we evaluate the JASPER method for assessing significance in two different types of association testing: (1) multivariate association testing between a set of quantitative traits and a set of SNPs (for which we consider two different test statistics), and (2) association testing between a single binary trait and a set of SNPs. We compare the type 1 error of the JASPER method to that of two other general approaches to assess significance: (1) use of an asymptotic approximation to the null distribution of the test statistic, and (2) use of a Pearson III approximation to the permutation distribution, as in JASPER, but performed under the assumption that the rows of either $\check{\mathbf{G}}$ or $\check{\mathbf{Y}}$ in (3.7) are exchangeable under the null hypothesis. Empirical type 1 error rates are evaluated at nominal levels 0.05 and 0.005 based on 30,000 replicates.

For a single trait (Tables 3.2 and 3.3), all three statistics assessed using JASPER are well-calibrated. The other methods of assessing significance lead to inflated type 1 error rates for T_1 and T_3, and slightly conservative tests for T_2. Table 3.4 shows the empirical type 1 error rates for multi-trait mapping with various numbers of traits. There is no significant difference from the nominal level for both T_1 and T_2 assessed using JASPER. The test T_2-asymp, which is T_2 assessed using large-sample theory, is conservative in some simulations with small k, with the type 1 error rate being highly conservative in high dimensional settings. These

Table 3.2: Empirical Type 1 Error in Model I (Linear Mixed Model) with 2 Subpopulations, Relatedness, and a Single Quantitative Trait ($k = 1$) based on 30,000 Replicates.

		Empirical Type 1 Error of [b]				
# PCs[a]	Nominal Level (SD)	T_1-newperm	T_1-perm	T_1-asymp	T_2-newperm	T_2-asymp
0	0.05 (.001)	0.0500	**1**	**1**	0.0480	**0.0428**
0	0.005 (.0004)	0.0043	**1**	**1**	0.0045	**0.0035**
1	0.05 (.001)	0.0503	**0.6761**	**0.6893**	-	-
1	0.005 (.0004)	0.0051	**0.3615**	**0.3652**	-	-
10	0.05 (.001)	0.0514	**0.5370**	**0.6360**	-	-
10	0.005 (.0004)	0.0054	**0.2344**	**0.3125**	-	-

The quantitative trait is generated using the linear mixed model with parameters chosen as explained in Section 3.3.1.
[a] The number of top principal components of the estimated GRM included as covariates in the trait model for the test statistic T_1.
[b] A number in bold indicates a type I error rate that is significantly different from the nominal level, using the z-test at level .01.

results verify that JASPER is correctly calibrated, demonstrate the appropriateness of our Pearson III approximation to the null permutation distribution, and suggest that JASPER is robust to various sources of misspecification in the trait null model including unobserved causal genes or using a limited number of PCs to correct for sample structure in the trait model.

To compare the power for association, we simulate k traits under Model I (a linear mixed model), with 23 families sampled from 2 sub-populations each. For each simulation scenario, we use T_1 and T_2 assessed using JASPER and a large-sample based approximation, where for T_1, we do not include the large-sample approximation as we have already seen that it results in highly inflated type 1 error rates. Also for T_1, we include the top 10 PCs of the estimated GRM as covariates in the trait model. We test the set of k traits for association with a set of 50 unlinked markers, where only a subset of the traits are truly associated with the tested region. Empirical power is computed at the significance level of $0.05/20,000 = 2.5 \times 10^{-6}$, which corresponds to a Bonferroni correction for 20,000 genes.

Table 3.3: Empirical Type 1 Error in Model II (Liability Threshold Model) with a Single Binary Trait ($k = 1$) based on 30,000 Replicates

Population structure setting	Relatedness	Nominal Level (SD)	Empirical Type 1 Error of [a]			
			T_3-newperm (1PC)	T_3-newperm (10PC)	T_3-asymp (1PC)	T_3-asymp (10PC)
2 Subpopulations	no	0.05 (.001)	0.0500	0.0503	0.0509	**0.0577**
	no	0.005 (.0004)	0.0051	0.0051	0.0053	0.0060
	yes	0.05 (.001)	0.0482	0.0490	**0.1833**	**0.1825**
	yes	0.005 (.0004)	0.0046	0.0044	**0.0367**	**0.0351**
Admixture	no	0.05 (.001)	0.0510	0.0519	**0.0534**	**0.0601**
	no	0.005 (.0004)	0.0052	0.0051	0.0053	0.0059
	yes	0.05 (.001)	0.0514	0.0513	**0.2123**	**0.2066**
	yes	0.005 (.0004)	0.0051	0.0048	**0.0453**	**0.0440**

The binary trait is generated using the liability threshold model with parameters chosen as explained in Section 3.3.1. We include either the top 1 or 10 principal components (PC) of the estimated GRM as covariates in the trait model.
[a] A number in bold indicates a type I error rate that is significantly different from the nominal level, using the z-test at level .01.

Empirical power results based on 1,000 replicates are shown in Table 3.5. In the low dimensional setting ($k < 50$), comparing the T_1 test to T_2 tests, we see that correcting for structure in the trait model using the whole GRM (T_2) compared to only using the top 10 PCs of the GRM (T_1), leads to substantial increase in power, with the difference being less important as the number of traits increases. Also, T_2 evaluated with JASPER consistently has higher power than T_2-asymp, which is the assessment based on large-sample approximations. The difference in power in the low dimensional settings ($k < 50$) could be due to T_2-asymp being slightly conservative as shown in Table 3.4.

In the case of high dimensional traits, T_2 assessed with JASPER leads to substantial increase in power compared to T_2 assessed with large-sample approximations, which barely has any power. The lack of power of T_2-asymp could be due to the estimation error incurred by estimating the two $k \times k$ covariance matrices between traits in the null model of (3.15), which corresponds to 2,550 and 10,100 parameters for $k = 50$ and 100, respectively, and is much higher than the sample size of 1,012. As the assessment of significance for T_2-asymp relies on how accurate the estimates of the two covariance matrices are, this could explain the highly conservative type 1 error results obtained, which could lead to substantial loss in

Table 3.4: Empirical Type 1 Error in Model I (Linear Mixed Model) with 2 Subpopulations and Relatedness based on 30,000 Replicates.

			Empirical Type 1 Error of [d]		
k	ρ_e [a]	Nominal Level (SD)	T_1-newperm[b]	T_2-newperm	T_2-asymp[c]
2	0.20	0.05 (.001)	0.0491	0.0485	**0.0424**
		0.005 (.0004)	0.0054	0.0048	**0.0036**
	0.50	0.05 (.001)	0.0511	0.0509	**0.0460**
		0.005 (.0004)	0.0048	0.0052	0.0043
5	0.20	0.05 (.001)	0.0516	0.0506	**0.0438**
		0.005 (.0004)	0.0056	0.0058	0.0041
	0.50	0.05 (.001)	0.0509	0.0494	**0.0440**
		0.005 (.0004)	0.0049	0.0050	0.0040
50	0.20	0.05 (.001)	0.0492	0.0510	**0.0003**
		0.005 (.0004)	0.0049	0.0045	0.0003
	0.50	0.05 (.001)	0.0511	0.0527	**0.0003**
		0.005 (.0004)	0.0051	0.0058	0.0003
100	0.20	0.05 (.001)	0.0525	0.0501	**0.0008**
		0.005 (.0004)	0.0055	0.0050	0.0008
	0.50	0.05 (.001)	0.0526	0.0496	**0.0008**
		0.005 (.0004)	0.0060	0.0048	0.0008

The k traits are generated using the linear mixed model with parameters chosen as explained in Section 3.3.1.
[a] The residual correlation between all traits (for $k < 50$) or between traits within clusters of size 10 (for $k \geq 50$).
[b] The top 10 PCs are included as covariates in the trait model for T_1-newperm.
[c] For T_2-asymp with $k \geq 50$, the empirical type 1 error rate is based on 1,000 replicates.
[d] A number in bold indicates a type I error rate that is significantly different from the nominal level, using the z-test at level .01.

power. Comparing T_1 to T_2 using JASPER when $k = 50$ or 100, the comparison reverses with T_1 consistently having higher power and with the power difference increasing as the number of traits gets larger. This likely highlights the cost in estimating an additional $\binom{k+1}{2}$ parameters in T_2 (due to $\mathbf{C}_u$ in (3.15)) compared to T_1.

Overall, we find that, in addition to providing robustness in type 1 error control, assessing significance using JASPER can greatly increase power compared to an asymptotic assessment of significance, particularly when the number of traits analyzed is large relative to the sample size. These results demonstrate the superiority of JASPER over both the large-sample based approximations and the Pearson type III approximation assuming unrelated samples, in terms of type 1 error control and statistical power for association testing in structured samples.

3.4.2 Analysis on the Framingham Heart Study data

In the SINGLE, LOGO and ALL analyses, T_1-newperm, T_2-newperm, T_1-asymp, and T_2-asymp are used for association testing between expression values and *cis*-SNPs. The statistic T_2-asymp is only used in the SINGLE analysis, which corresponds to a single gene's expression values (i.e. $k = 1$) being tested with its *cis*-SNPs, as it is computationally challenging to use it for the other two analyses which involve a much larger number of traits ($k \geq 41$). For T_1-asymp and T_1-newperm, we include the top 10 PCs of the estimated GRM as covariates in the trait model. Figure 3.1 displays the eigenvalue spectrum of the estimated GRM, where there are a couple of large leading values.

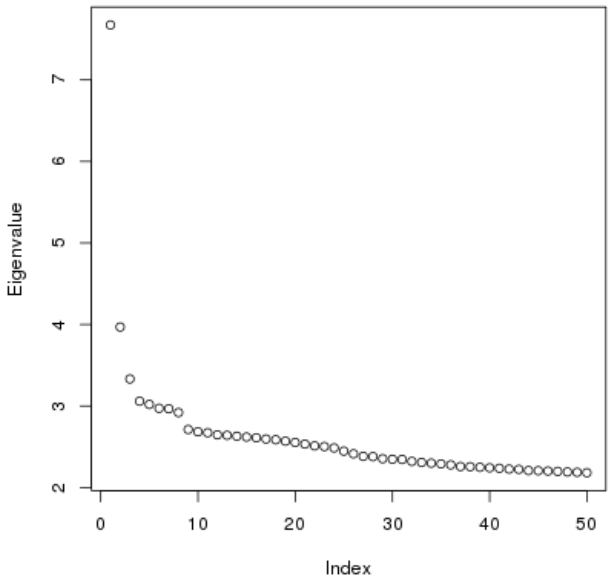
Tables 3.6-3.9 report the gene regions for which the p-value for T_2-newperm in the ALL analysis is below 4.2×10^{-5}, which corresponds to a Bonferroni correction based on the 1,194 genes being tested across the 11 pathways. We find many genes with highly significant p-values in the both the SINGLE and ALL analyses with T_1-newperm and T_2-newperm, such as *KLRC3* in pathway hsa04612, *SCP2* in pathway hsa03320, *IL18RAP* in pathway hsa04060, and *DAPK1* in pathway hsa04140. The results from the LOGO analysis help us

Table 3.5: Empirical Power in Model I (Linear Mixed Model) with 2 Subpopulations and Relatedness based on 1,000 Replicates.

		Empirical Power of [c]		
k	ρ_e [a]	T_1-newperm[b]	T_2-newperm	T_2-asymp
1	1	0.248	**0.475**	0.398
2	0.20	0.141	**0.327**	0.221
	0.50	0.252	**0.514**	0.411
5	0.20	0.432	**0.632**	0.54
	0.50	0.722	**0.91**	0.849
50	0.20	**0.49**	0.382	0.002
	0.50	**0.694**	0.585	0.056
100	0.20	**0.561**	0.407	0
	0.50	**0.934**	0.692	0

The k traits are generated using the linear mixed model with parameters chosen as explained in Section 3.3.1.
[a] The residual correlation between all traits (for $k < 50$) or between traits within clusters of size 10 (for $k \geq 50$).
[b] The top 10 PCs are included as covariates in the trait model for T_1-newperm.
[c] For each simulation setting, the highest empirical power among all the methods is in bold, as well as any empirical power not significantly different from the highest one.

Figure 3.1: Eigen-spectrum of the FHS estimated GRM.



elucidate whether the signal detected in the ALL analysis is driven mostly by the effect of a set of *cis*-SNPs on the gene it is nearest to or if it is also driven by the association of the *cis*-SNPs with the expression of other genes in the same pathway. Among the genes where both the SINGLE and ALL p-values were significant, only a few genes also had significant p-values in the LOGO analysis: *KLRC1, KLRD1* and *KLRC3* in pathway hsa04612, *IL1R2, IL1RL1, IL18R1, IL18RAP, CCR2, CCR1*, and *CCR3* in pathway hsa04060, *IL18RAP* and *IL18R1* in pathway hsa05321, and *CALML4* in pathway hsa04722. Eight of these genes have previously been associated with a disease that has been linked to the corresponding KEGG pathway [Stelzer et al., 2016]. For the remaining five genes, the analysis suggests that the corresponding KEGG pathway is a candidate pathway for association with the diseases the genes have been previously linked to. For example, the gene *IL1RL1* in the 'Cytokine-cytokine receptor interaction' pathway has been previously associated with asthma [Ferreira et al., 2011; Grotenboer et al., 2013], and *CCR1* in the same pathway has been previously associated with Behcet's disease [Hou et al., 2012; Meguro et al., 2013], a form of vasculitis, but neither of the two diseases has previously been associated with the KEGG pathway.

One could also look at genes that have a significant p-value in the LOGO analysis but not in the SINGLE analysis, as this would suggest that the *cis*-SNPs of the gene have *trans* effects in the pathway (i.e. affecting genes other than the nearest one). Examples of such genes are *HSPA1A* in pathway hsa04612, *CPT2* in pathway hsa03320, *NUMB* in pathway hsa04330, *IL1RL2* and *IL1R1* in pathway hsa04060, and *IL23R* in pathway hsa05321. Six out of twelve of such genes have been previously associated with a disease linked to the KEGG pathway. The remaining genes could be candidate genes to study in a GWA analysis with the diseases linked to the corresponding KEGG pathway. Overall, seven out of the eleven KEGG pathways analyzed have, among the top 3 signals, genes that were previously associated with diseases linked to the pathway.

We observe that in the low dimensional analysis (i.e. SINGLE), all the four test statistics give mostly similar p-values, with the large-sample based tests resulting in the largest p-values for the top association signals. In the high dimensional analyses (LOGO and ALL), we observe that T_2-newperm almost always gives larger p-values than T_1-newperm for the top association signals. This agrees with our simulation results, where we also observe that T_1-newperm had higher power than T_2-newperm for large k. More precisely, in the ALL analysis, 24 genes result in a p-value for T_1-newperm below the significance threshold and a p-value for T_2-newperm above the threshold. Genes included are *IFNG* in pathway hsa04060, which has been previously linked to aplastic anemia [Dufour et al., 2004] and *FZD6* in pathway hsa04310, which has been previously associated with nail disease [Fröjmark et al., 2011]. We also note that among the non-significant signals (i.e. corresponding to p-values above the significance threshold), T_1-asymp usually results in smaller p-values than T_1-newperm, which also agrees with the simulation results.

Table 3.6: Genes with the Strongest Associations with Expression Levels across the 11 KEGG Pathways in Framingham Heart Study.

			SINGLE				LOGO			ALL		
Pathway	Gene[a]	Chr	T_2-newperm	T_2-asymp	T_1-newperm	T_1-asymp	T_2-newperm	T_1-newperm	T_1-asymp	T_2-newperm	T_1-newperm	T_1-asymp
hsa04612	**KLRC1**	12	3.57e-43	1.57e-38	4.93e-43	1.00e-30	1.84e-31	6.84e-66	4.78e-60	4.74e-38	3.90e-73	5.49e-66
	KLRD1	12	3.21e-11	2.74e-10	1.73e-10	3.31e-09	1.07e-22	3.03e-43	1.50e-41	5.69e-37	8.28e-71	1.33e-64
	KLRC3	12	1.49e-78	9.12e-70	7.99e-79	8.68e-56	8.80e-16	2.25e-28	2.17e-28	1.14e-36	7.43e-71	1.88e-64
	HSP90AA1	14	3.13e-22	2.71e-20	7.85e-22	8.87e-18	5.06e-01	6.36e-01	3.91e-01	7.27e-14	4.87e-16	8.41e-17
	IFNG	12	1.13e-22	3.74e-19	3.24e-23	7.88e-17	4.01e-01	4.19e-01	2.12e-01	3.77e-07	1.06e-08	2.29e-09
	TNF	6	3.13e-06	6.41e-05	6.14e-06	7.81e-05	1.13e-04	1.22e-04	3.09e-03	1.36e-06	3.17e-06	2.33e-04
	HSPA1A	6	1.86e-01	2.57e-01	2.52e-01	2.97e-01	3.23e-06	5.45e-06	2.80e-04	4.53e-06	7.90e-06	3.86e-04
	HSPA1L	6	8.50e-03	2.15e-02	1.21e-02	2.40e-02	7.65e-05	9.96e-05	2.31e-03	4.67e-06	8.07e-06	3.93e-04
	HSPA1B	6	8.36e-03	2.28e-02	1.26e-02	2.45e-02	5.63e-05	7.51e-05	1.89e-03	8.30e-06	1.61e-05	6.40e-04
	IFI30	19	8.42e-12	4.29e-11	8.29e-12	1.68e-10	1.03e-01	3.46e-01	6.72e-02	2.94e-05	2.47e-07	1.10e-09
hsa03320	**SCP2**	1	9.78e-133	2.14e-111	1.80e-132	3.00e-95	9.99e-01	9.97e-01	9.78e-01	6.68e-26	6.42e-93	1.35e-74
	CPT2	1	5.40e-01	5.44e-01	5.19e-01	4.93e-01	1.89e-21	7.16e-82	1.87e-69	1.76e-21	1.17e-81	2.45e-69
	ACSL5	10	1.20e-42	7.70e-38	1.14e-41	3.89e-31	2.67e-01	1.97e-01	1.40e-01	1.91e-20	1.98e-27	1.39e-25
	CYP27A1	2	2.18e-141	3.80e-128	3.79e-153	2.36e-131	6.93e-01	6.92e-01	8.38e-01	3.39e-16	5.66e-155	6.52e-116
	FADS2	11	1.59e-112	1.80e-97	1.71e-103	8.40e-73	3.95e-01	4.36e-01	2.40e-01	1.79e-15	6.42e-49	7.81e-44
	ACSL6	5	3.98e-26	5.73e-21	1.71e-29	2.70e-21	1.68e-02	1.88e-02	2.22e-01	4.13e-09	2.96e-18	3.42e-11
hsa04330	**PSEN1**	14	2.56e-21	3.26e-19	9.19e-22	2.79e-17	1.32e-01	1.15e-01	6.79e-02	6.03e-14	5.93e-16	1.01e-15
	NUMB	14	4.56e-03	6.28e-03	1.77e-03	2.32e-03	8.24e-10	6.20e-11	2.28e-10	4.65e-11	7.24e-13	3.96e-12
	ADAM17	2	1.06e-27	6.74e-24	6.53e-28	2.56e-20	1.22e-02	2.50e-02	1.04e-02	4.14e-10	2.55e-17	1.16e-16
	NCSTN	1	2.74e-19	5.12e-18	2.52e-18	1.70e-15	7.21e-01	7.61e-01	4.90e-01	6.28e-09	1.39e-16	3.38e-18
	MAML3	4	2.61e-23	1.74e-21	2.20e-23	1.70e-18	8.21e-01	7.74e-01	4.49e-01	2.43e-05	2.86e-05	1.03e-06

[a] Genes in bold are those that have been previously associated with a disease that is among (or associated to) the ones listed for the corresponding pathway in the KEGG database. MIM numbers of genes: KLRC1 [MIM 161555], KLRD1 [MIM 602894], KLRC3 [MIM 602892], HSP90AA1 [MIM 140571], IFNG [MIM 147570], TNF [MIM 191160], HSPA1A [MIM 603012], HSPA1L [MIM 140559], HSPA1B [MIM 603012], IFI30 [MIM 604664], SCP2 [MIM 184755], CPT2 [MIM 600650], ACSL5 [MIM 605677], CYP27A1 [MIM 606530], FADS2 [MIM 606149], ACSL6 [MIM 604443], PSEN1 [MIM 104311], NUMB [MIM 603728], ADAM17 [MIM 603639], NCSTN [MIM 605254], MAML3 [MIM 608991] .

Table 3.7: Genes with the Strongest Associations with Expression Levels across the 11 KEGG Pathways in Framingham Heart Study.

			SINGLE				LOGO			ALL		
Pathway	Gene[a]	Chr	T_2-newperm	T_2-asymp	T_1-newperm	T_1-asymp	T_2-newperm	T_1-newperm	T_1-asymp	T_2-newperm	T_1-newperm	T_1-asymp
hsa04060	**IL1R2**	2	2.14e-09	7.84e-09	7.43e-10	6.26e-08	4.41e-20	3.31e-194	7.21e-136	2.99e-25	2.01e-212	3.62e-148
	IL1RL2	2	8.62e-01	8.77e-01	7.95e-01	8.18e-01	6.74e-24	1.43e-196	3.61e-135	4.71e-24	5.04e-196	1.09e-134
	IL1R1	2	3.03e-04	7.90e-04	2.09e-04	4.47e-04	3.62e-23	3.51e-195	1.69e-134	6.95e-24	1.18e-195	8.48e-135
	IL1RL1	2	1.15e-49	1.51e-42	8.79e-50	1.31e-37	1.17e-17	8.69e-162	3.68e-111	5.18e-23	8.50e-193	3.75e-132
	IL18R1	2	6.41e-21	1.94e-17	1.10e-19	1.63e-14	5.64e-19	1.62e-184	1.10e-126	6.05e-23	1.38e-192	4.68e-132
	IL18RAP	2	1.05e-234	4.24e-206	4.45e-240	5.24e-177	2.50e-11	1.27e-32	1.37e-21	7.34e-23	5.42e-192	1.29e-131
	CCR2	3	5.44e-12	1.23e-10	1.48e-12	2.67e-10	7.08e-14	3.28e-66	6.45e-51	6.63e-16	1.89e-71	7.52e-55
	CCR1	3	1.45e-33	3.40e-29	1.07e-32	1.20e-23	3.18e-10	1.31e-46	8.48e-36	2.18e-14	3.06e-66	1.21e-50
	XCR1	3	8.17e-01	8.29e-01	8.59e-01	8.52e-01	3.01e-13	8.08e-56	5.98e-42	4.76e-13	1.64e-55	1.04e-41
	CCR3	3	9.72e-105	3.13e-93	1.53e-101	1.16e-77	1.78e-05	1.87e-12	2.60e-09	1.32e-12	1.06e-57	7.06e-44
	CXCR6	3	2.85e-02	3.73e-02	2.98e-02	3.28e-02	3.39e-11	4.77e-49	5.66e-37	2.51e-11	2.39e-49	3.49e-37
	CCR9	3	1.90e-01	2.15e-01	3.62e-01	3.84e-01	6.20e-10	9.49e-44	5.61e-33	4.55e-10	5.81e-44	4.14e-33
	CXCL2	4	2.92e-01	2.92e-01	3.07e-01	2.89e-01	5.26e-10	5.62e-26	9.67e-25	5.04e-10	4.38e-26	7.77e-25
	LTA	6	3.67e-01	4.61e-01	2.23e-01	2.86e-01	2.08e-08	1.24e-08	2.70e-03	2.57e-08	1.10e-08	2.63e-03
	CXCL3	4	8.13e-01	8.24e-01	7.44e-01	7.56e-01	1.85e-08	5.18e-24	9.77e-22	2.71e-08	9.55e-24	1.66e-21
	LTB	6	5.50e-01	6.11e-01	4.06e-01	4.31e-01	3.50e-08	2.26e-08	3.07e-03	4.54e-08	2.96e-08	3.52e-03
	TNF	6	3.13e-06	6.41e-05	6.14e-06	7.81e-05	5.12e-07	2.36e-07	9.11e-03	4.56e-08	2.93e-08	3.48e-03
	IL7R	5	8.62e-60	3.89e-54	9.63e-59	9.56e-46	4.26e-01	7.15e-01	4.19e-01	7.75e-08	4.57e-21	8.81e-21
	CXCL5	4	3.66e-22	1.14e-19	7.23e-22	6.18e-17	2.01e-03	2.03e-13	1.60e-11	3.57e-06	1.24e-18	8.50e-16
	PPBP	4	3.18e-01	3.28e-01	2.38e-01	2.08e-01	9.13e-06	1.05e-17	6.28e-15	5.49e-06	4.14e-18	2.86e-15
	PF4	4	3.07e-01	3.18e-01	2.93e-01	2.87e-01	1.82e-05	9.04e-17	5.26e-14	1.39e-05	6.22e-17	4.01e-14
	IL22	12	1.82e-04	5.11e-04	2.12e-04	4.93e-04	3.86e-05	1.83e-08	2.45e-09	1.96e-05	1.20e-08	1.67e-09
	PF4V1	4	2.88e-57	9.34e-51	5.43e-54	6.62e-42	1.68e-01	5.41e-02	7.45e-02	2.53e-05	3.10e-16	1.85e-13
	CXCL1	4	5.40e-06	2.98e-05	1.08e-05	5.38e-05	3.71e-04	3.90e-14	8.72e-12	2.74e-05	3.28e-16	1.96e-13
	CXCL6	4	3.95e-01	4.21e-01	3.47e-01	3.64e-01	2.90e-05	5.73e-16	3.24e-13	3.30e-05	6.60e-16	3.75e-13
	IL26	12	1.58e-27	4.00e-23	3.57e-27	2.82e-19	9.67e-04	8.09e-06	7.76e-07	3.98e-05	3.15e-08	3.50e-09
	TNFSF8	9	2.13e-10	1.71e-09	1.34e-10	1.88e-09	6.38e-04	6.30e-03	1.48e-03	4.03e-05	2.58e-04	5.52e-05

[a] Genes in bold are those that have been previously associated with a disease that is among (or associated to) the ones listed for the corresponding pathway in the KEGG database. MIM numbers of genes: IL1R2 [MIM 147811], IL1RL2 [MIM 604512], IL1R1 [MIM 147810], IL1RL1 [MIM 601203], IL18R1 [MIM 604494], IL18RAP [MIM 604509], CCR2 [MIM 601267], CCR1 [MIM 601159], XCR1 [MIM 600552], CCR3 [MIM 601268], CXCR6 [MIM 605163], CCR9 [MIM 604738], CXCL2 [MIM 139110], LTA [MIM 153440], CXCL3 [MIM 139111], LTB [MIM 600978], TNF [MIM 191160], IL7R [MIM 146661], CXCL5 [MIM 600324], PPBP [MIM 121010], PF4 [MIM 173460], IL22 [MIM 605330], PF4V1 [MIM 173461], CXCL1 [MIM 155730], CXCL6 [MIM 138965], IL26 [MIM 605679], TNFSF8 [MIM 603875] .

Table 3.8: Genes with the Strongest Associations with Expression Levels across the 11 KEGG Pathways in Framingham Heart Study.

			SINGLE				LOGO			ALL		
Pathway	Gene[a]	Chr	T_2-newperm	T_2-asymp	T_1-newperm	T_1-asymp	T_2-newperm	T_1-newperm	T_1-asymp	T_2-newperm	T_1-newperm	T_1-asymp
hsa04630	PIK3R3	1	9.87e-64	2.14e-52	4.89e-65	8.74e-48	6.01e-03	8.97e-03	1.33e-01	6.21e-19	1.20e-32	1.38e-21
	IL7R	5	8.62e-60	3.89e-54	9.63e-59	9.56e-46	1.50e-01	9.95e-02	3.95e-02	3.20e-17	1.47e-45	6.60e-43
	IFNGR2	21	4.61e-06	1.40e-05	3.18e-05	5.03e-05	1.04e-03	8.00e-04	3.38e-06	2.96e-05	3.07e-06	3.57e-09
hsa05321	**IL18RAP**	2	1.05e-234	4.24e-206	4.45e-240	5.24e-177	1.70e-05	9.17e-07	1.72e-05	8.84e-35	9.78e-249	1.51e-182
	IL18R1	2	6.41e-21	1.94e-17	1.10e-19	1.63e-14	2.38e-22	3.60e-226	9.75e-167	1.52e-34	2.45e-248	2.54e-182
	NOD2	16	4.14e-163	3.72e-148	1.84e-152	1.10e-115	4.48e-01	5.28e-01	1.72e-01	1.16e-33	8.14e-99	1.16e-93
	TLR5	1	1.89e-24	1.46e-21	2.77e-24	9.01e-18	3.94e-01	6.93e-01	4.02e-01	9.47e-11	6.13e-07	3.60e-08
	TLR4	9	2.93e-34	3.56e-30	1.49e-33	1.26e-24	1.95e-01	2.31e-01	1.30e-01	7.35e-10	8.19e-18	1.05e-17
	IL23R	1	1.37e-03	2.36e-03	1.31e-03	1.40e-03	6.26e-06	1.10e-05	9.19e-06	4.28e-07	3.34e-07	3.27e-07
	IL12RB2	1	1.59e-11	2.27e-10	4.56e-11	1.13e-09	7.13e-03	1.52e-02	7.41e-03	5.20e-06	6.75e-06	3.07e-06
	IL22	12	1.82e-04	5.11e-04	2.12e-04	4.93e-04	3.82e-05	7.57e-07	1.11e-07	5.38e-06	1.37e-07	2.08e-08
	TNF	6	3.13e-06	6.41e-05	6.14e-06	7.81e-05	1.67e-03	1.84e-03	2.51e-02	1.66e-05	3.53e-05	1.67e-03
hsa04140	PIK3R3	1	9.87e-64	2.14e-52	4.89e-65	8.74e-48	1.60e-01	4.46e-02	2.29e-01	1.26e-19	5.83e-36	3.72e-25
	DAPK1	9	6.31e-149	1.79e-126	1.82e-149	1.85e-98	2.63e-01	1.81e-01	8.03e-02	1.88e-15	1.52e-81	2.07e-71
	VAMP8	2	3.07e-32	4.93e-27	2.31e-30	3.43e-22	6.87e-03	2.83e-03	3.28e-04	4.69e-12	4.49e-19	3.54e-19
	MAPK3	16	2.21e-17	2.59e-16	2.74e-17	2.97e-15	9.06e-01	8.02e-01	4.62e-01	2.41e-10	5.71e-19	1.47e-20
	CTSB	8	1.30e-04	1.37e-03	3.78e-07	8.82e-05	1.67e-04	7.03e-05	6.61e-02	1.19e-05	1.93e-08	1.95e-03
hsa04022	MAPK3	16	2.21e-17	2.59e-16	2.74e-17	2.97e-15	4.82e-01	7.13e-01	3.21e-01	9.96e-10	2.91e-15	4.80e-17
	TRPC6	11	1.57e-37	2.73e-31	9.81e-38	5.27e-26	7.57e-02	1.49e-01	2.98e-01	1.94e-07	5.76e-10	1.23e-07

[a] Genes in bold are those that have been previously associated with a disease that is among (or associated to) the ones listed for the corresponding pathway in the KEGG database. MIM numbers of genes: PIK3R3 [MIM 606076], IL7R [MIM 146661], IFNGR2 [MIM 147569], IL18RAP [MIM 604509], IL18R1 [MIM 604494], NOD2 [MIM 605956], TLR5 [MIM 603031], TLR4 [MIM 603030], IL23R [MIM 607562], IL12RB2 [MIM 601642], IL22 [MIM 605330], TNF [MIM 191160], DAPK1 [MIM 600831], VAMP8 [MIM 603177], MAPK3 [MIM 601795], CTSB [MIM 116810], TRPC6 [MIM 603652] .

Table 3.9: Genes with the Strongest Associations with Expression Levels across the 11 KEGG Pathways in Framingham Heart Study.

			SINGLE				LOGO			ALL		
Pathway	Gene[a]	Chr	T_2-newperm	T_2-asymp	T_1-newperm	T_1-asymp	T_2-newperm	T_1-newperm	T_1-asymp	T_2-newperm	T_1-newperm	T_1-asymp
hsa04660	PIK3R3	1	9.87e-64	2.14e-52	4.89e-65	8.74e-48	5.01e-03	2.10e-03	3.38e-02	5.52e-26	6.35e-41	7.90e-29
	MAPK3	16	2.21e-17	2.59e-16	2.74e-17	2.97e-15	5.00e-01	4.72e-01	1.88e-01	3.35e-14	9.82e-19	5.00e-20
	NCK2	2	1.01e-37	1.48e-32	1.22e-38	1.59e-26	5.49e-01	5.54e-01	2.73e-01	1.08e-06	2.43e-09	2.37e-10
	MAP3K14	17	8.19e-01	8.75e-01	7.35e-01	8.22e-01	1.41e-05	1.14e-09	7.56e-03	1.35e-05	8.17e-10	7.06e-03
hsa04722	MAP2K5	15	2.32e-70	5.61e-59	8.68e-72	9.05e-51	1.08e-03	3.85e-04	2.20e-03	6.50e-25	1.19e-40	2.13e-32
	PIK3R3	1	9.87e-64	2.14e-52	4.89e-65	8.74e-48	9.33e-02	3.01e-02	2.07e-01	6.27e-21	1.76e-36	4.75e-25
	CALML4	15	2.38e-07	1.93e-06	5.31e-07	2.68e-06	9.78e-12	3.56e-19	4.88e-18	1.69e-14	1.62e-22	5.37e-21
	MAPK3	16	2.21e-17	2.59e-16	2.74e-17	2.97e-15	1.52e-01	2.62e-01	7.15e-02	5.48e-13	1.73e-18	7.79e-20
	RAPGEF1	9	3.51e-46	5.27e-41	4.60e-50	2.52e-37	9.16e-01	7.61e-01	3.81e-01	3.13e-12	3.72e-24	1.51e-25
	MAP3K5	6	1.54e-32	3.11e-28	3.08e-33	9.31e-25	1.90e-01	1.17e-01	1.29e-01	3.74e-12	3.61e-22	2.72e-19
	PSEN1	14	2.56e-21	3.26e-19	9.19e-22	2.79e-17	9.64e-01	9.84e-01	9.39e-01	8.35e-06	2.39e-07	7.17e-08
hsa04310	**PRKCB**	16	4.87e-52	1.47e-41	6.67e-52	6.70e-33	4.44e-01	4.76e-01	1.76e-01	8.14e-09	4.14e-27	7.81e-26
	PSEN1	14	2.56e-21	3.26e-19	9.19e-22	2.79e-17	1.89e-01	4.55e-01	2.90e-01	2.24e-08	5.35e-09	3.19e-09
	CSNK2B	6	4.57e-21	2.60e-16	2.28e-20	7.07e-14	5.21e-02	1.38e-02	3.22e-01	2.33e-07	8.61e-09	1.71e-04
	FRAT1	10	5.43e-17	1.08e-15	3.47e-17	9.77e-15	5.56e-02	2.26e-02	5.36e-03	6.74e-06	1.34e-06	1.97e-07
	FRAT2	10	2.00e-11	1.15e-10	8.69e-12	1.38e-10	8.35e-05	1.52e-05	2.26e-06	6.76e-06	1.35e-06	1.96e-07

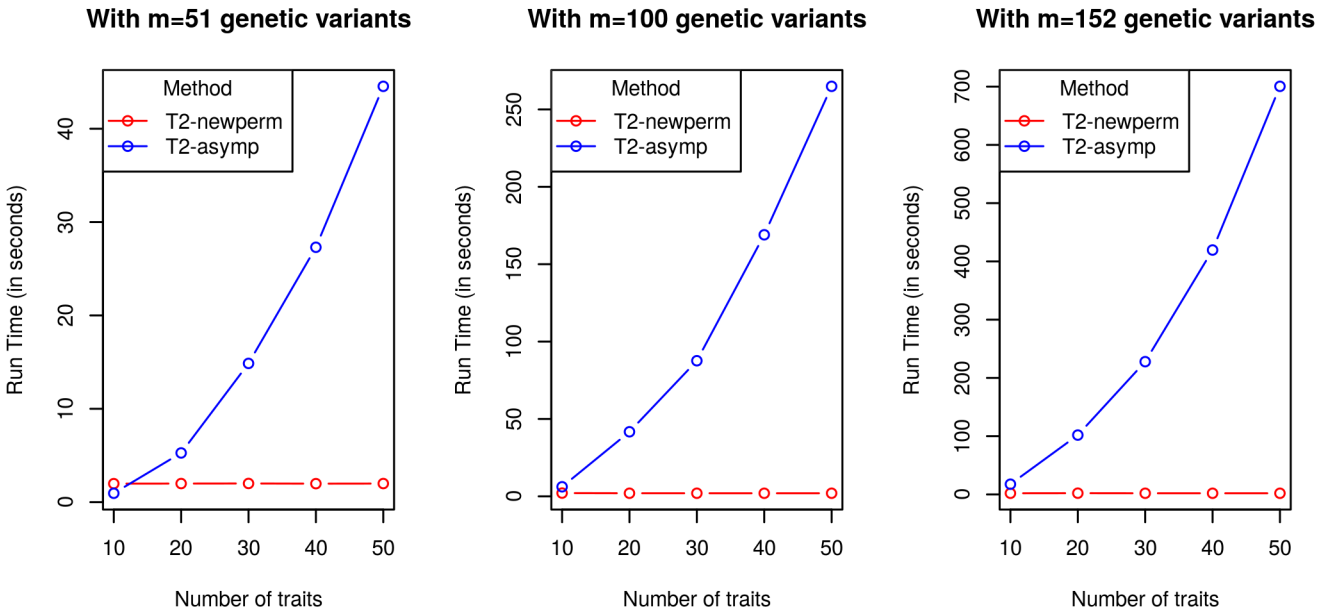
[a] Genes in bold are those that have been previously associated with a disease that is among (or associated to) the ones listed for the corresponding pathway in the KEGG database. MIM/HGNC numbers of genes: PIK3R3 [MIM 606076], MAPK3 [MIM 601795], NCK2 [MIM 604930], MAP3K14 [MIM 604655], MAP2K5 [MIM 602520], CALML4 [HGNC: 18445], RAPGEF1 [MIM 600303], MAP3K5 [MIM 602448], PSEN1 [MIM 104311], PRKCB [MIM 176970], CSNK2B [MIM 115441], FRAT1 [MIM 602503], FRAT2 [MIM 605006] .

3.4.3 Computation time of JASPER

The main computational burden with JASPER is from the eigenvalue decomposition of the $n \times n$ estimated GRM $\widetilde{\mathbf{\Phi}}$ in (3.4) which is needed to correct for population structure. This decomposition usually needs to be performed once for each tested region. However, if the different regions tested are all based on the same set of individuals, then the decomposition only needs to be performed once with the decomposition results being re-used for each region. We note that if the test statistic used with JASPER requires the eigen-decomposition of an estimated GRM $\widehat{\mathbf{\Phi}}$, which satisfies the condition $\widehat{\mathbf{\Phi}} = \mathbf{J}\widehat{\mathbf{\Phi}}\mathbf{J}$, then there is no extra cost to JASPER as the matrix $\widehat{\mathbf{\Phi}}$ can be used for $\widetilde{\mathbf{\Phi}}$, and the eigen-decomposition results obtained to compute the statistic can be used in JASPER.

Using a single processor (at 3.5GHz) of Intel Xeon CPU E5-2637 v3, we compare the runtime for assessing significance using JASPER and a large sample-based approximation for a single gene region in the FHS data. We pick the largest KEGG pathway 'Cytokine-cytokine receptor interaction' (hsa04060) containing 247 genes, and test the association between a gene region and gene expression values at k genes in the pathway. We pick 3 gene regions corresponding to 51 *cis*-SNPs, 100 *cis*-SNPs, and 152 *cis*-SNPs. For each region, we sample at random k genes in the pathway for which we obtain the expression values, and the statistic T_2 is used for association testing. We exclude the time taken to obtain the GRM eigen-decomposition (which is required in both methods). Results are shown in Figure 3.2, where we see that JASPER is computationally efficient as the number of tested variants m and the number of traits k increase. This is due to the assessment of significance of JASPER being based on two $n \times n$ matrices. The run time for the large sample-based method increases as either m or k increase. This is because the assessment of significance relies on the eigen-decomposition of $mk \times mk$ matrices. This demonstrates the computational efficiency of JASPER in high dimensional settings.

Figure 3.2: Run Time Comparison for JASPER and a Large Sample-based Method Using the Test Statistic T_2.



3.5 Discussion and Future Work

We have developed a novel method, JASPER, for assessing significance of a general class of association tests, in the presence of population substructure, admixture and/or relatedness. Compared to existing methods, JASPER is applicable to association tests involving single/multiple traits with single/multiple markers, is robust to various sources of misspecifications in the trait model, including when a large number of traits are analyzed, offers computational feasibility in the presence of high dimensional traits, and is applicable to samples that have cryptic relatedness in addition to population stratification and admixture. Through extensive simulation studies, we have demonstrated the validity of JASPER and the power improvement it can provide over other methods. We have shown for various statistics, including SKAT, MSKAT and Multi-SKAT (the first of which is highly popular in GWA studies with one trait), that significance assessment using JASPER provides better performance compared to assessments of significance based on large-sample theory for a wide range of population structure configurations and trait models.

To reduce the loss in robustness incurred by possible misspecification in the model for the traits, statistical significance in JASPER is assessed based on a model that views genotypes as random. We demonstrate through simulation studies that this approach of assessing p-values can provide robust control of the type 1 error in addition to substantial power improvement in situations where it is difficult to estimate the covariance matrix between traits due to the high dimensionality of the traits. An additional feature of JASPER is that it does not require for the LD structure across markers to be known for its validity. JASPER accounts for unobserved sample structure by modeling the dependency between individuals in the sample through the inclusion of a random effect in the model for the genotypes whose covariance structure, if unknown, involves a genetic relatedness matrix that can be estimated from genome scan data. This methodology has been used in several retrospective association tests (either with known or unknown sample structure) that are applicable only to single traits (e.g. MASTOR, CERAMIC, CARAT). We demonstrate in our simulations that by adjusting for the sample structure in the model for the genotypes, JASPER retains correct control of the type 1 error when the structure is partly incorporated through covariates in the model for the traits. Importantly, JASPER bypasses the high computational cost involved when using permutation-testing with stringent significance thresholds and instead uses a moment-matching approach to approximate the null permutation distribution. This renders JASPER not only computationally efficient but also valid for small sample sizes as it does not rely on asymptotic results. In addition, it makes JASPER computationally scalable as it involves $n \times n$ matrices regardless of the number of traits or markers being tested, and applicable to analyses with high-dimensional traits.

An important advantage of JASPER is that it is very flexible as it is applicable to a large number of test statistics that can involve single or multiple traits, single or multiple variants, and quantitative or binary traits. Statistical power to detect association signals can be highly improved by incorporating multiple traits and markers in the test, all the more when strong correlation exists between the traits or markers. Although we did not consider a large range

of structures between the traits or markers in our simulation studies, JASPER is applicable to statistics that incorporate more complex structures for the similarity matrices between traits or markers (i.e. $\mathbf{K}_Y$ and $\mathbf{K}_G$, respectively), such as the Wu weights [Wu et al., 2011].

In the context of high dimensional traits, it could happen that a large proportion of the traits are not associated with the set of variants being tested. Power could then be reduced as not much information is gained by including all of the traits in the test statistic. In such cases, adaptive strategies have been implemented to give higher weights to traits that are more likely to be associated [Kim et al., 2016; Pan et al., 2015]. Similarly, association tests have been previously proposed that implement adaptive strategies for increased power when combining the effects across genetic variants in a region [Jiang and Mcpeek, 2014; Lee et al., 2012b; Wang et al., 2018]. Hence, it is of interest to extend JASPER to include adaptive test statistics. The method proposed, JASPER, has been applied in the context of gene expression association analysis with local genetic variation within a biological network. However, we note applications of JASPER can go beyond this context, such as in GWA studies evaluating associations between single (or multiple) traits and genetic variants.

3.6 Appendix

We derive the first three moments of the null permutation distribution of $Q_T = \text{tr}(\mathbf{W}^G\mathbf{W}^Y)$, where $\mathbf{W}^G$ and $\mathbf{W}^Y$ are two $n \times n$ symmetric p.s.d. matrices with $\mathbf{W}^G = \check{\mathbf{G}}\check{\mathbf{G}}^T$, and we assume $\mathbf{W}^G$ is independent of $\mathbf{W}^Y$ under the null hypothesis. Let σ denote a permutation of the indices $\{1, \ldots, n\}$, S_n be the set of all permutations of the indices $\{1, \ldots, n\}$, and $f(\mathbf{W}^{G,\sigma})$ be the probability density function of $\mathbf{W}^{G,\sigma}$, where $\mathbf{W}^{G,\sigma} = \check{\mathbf{G}}_\sigma\check{\mathbf{G}}_\sigma^T$, with $\check{\mathbf{G}}_\sigma$ obtained by permuting the rows of $\check{\mathbf{G}}$ according to σ. Ignoring the higher-order moments beyond the first two moments, we obtain under second-order exchangeability of the rows of $\check{\mathbf{G}}$ that $f(\mathbf{W}^{G,\sigma}) = \dfrac{1}{n!}$.

Proposition 1. *The first moment of Q_T is given by,*

$$\mathrm{E}(Q_T) = \frac{1}{n} T_1^Y T_1^G + \frac{1}{n(n-1)} (S_1^Y - T_1^Y)(S_1^G - T_1^G), \tag{3.19}$$

where $T_1^Y = \mathrm{tr}(\mathbf{W}^Y)$, $S_1^Y = \sum_{ij} W_{ij}^Y$, $T_1^G = \mathrm{tr}(\mathbf{W}^G)$, and $S_1^G = \sum_{ij} W_{ij}^G$.

Proof. We have,

$$\begin{aligned}
\mathrm{E}(Q_T) &= \sum_{\sigma \in S_n} \mathrm{tr}(\mathbf{W}^{G,\sigma} \mathbf{W}^Y) f(\mathbf{W}^{G,\sigma}) \\
&= \frac{1}{n!} \sum_{ij} \sum_{\sigma \in S_n} W_{ij}^{G,\sigma} W_{ij}^Y \\
&= \frac{1}{n!} \sum_{i} W_{ii}^Y \sum_{\sigma \in S_n} W_{ii}^{G,\sigma} + \frac{1}{n!} \sum_{i \neq j} W_{ij}^Y \sum_{\sigma \in S_n} W_{ij}^{G,\sigma}.
\end{aligned}$$

Using Lemma 2, and noting that $\mathrm{tr}(\mathbf{A}) = \sum_i A_{ii}$, and $\sum_{i \neq j} A_{ij} = \sum_{ij} A_{ij} - \sum_i A_{ii}$, for an arbitrary matrix $\mathbf{A}$, the desired result follows. □

Lemma 2. *For an arbitrary symmetric p.s.d. matrix $\mathbf{A}$, and up to two distinct indices i and j,*

$$\begin{aligned}
\sum_{\sigma \in S_n} A_{ii}^\sigma &= (n-1)! \sum_i A_{ii}, \\
\sum_{\sigma \in S_n} A_{ij}^\sigma &= (n-2)! \sum_{i \neq j} A_{ij}.
\end{aligned}$$

For ease of notation, we define the symbol $\sum'$ to be the sum taken over the distinct indices (e.g. $\sum'_{ijk}$ is the sum taken over i, j and k all different).

Proposition 3. *The second moment of Q_T is given by,*

$$\begin{aligned}
n!\,\mathrm{E}(Q_T^2) &= (n-1)! \sum_i (W_{ii}^Y)^2 \cdot \sum_i (W_{ii}^G)^2 \\
&\quad + (n-2)! \Bigg\{ 4 \sum_{ij}{}' W_{ii}^Y W_{ij}^Y \cdot \sum_{ij}{}' W_{ii}^G W_{ij}^G + 2 \sum_{ij}{}' (W_{ij}^Y)^2 \cdot \sum_{ij}{}' (W_{ij}^G)^2 +
\end{aligned}$$

$$\sum_{ij}{}' W^Y_{ii}W^Y_{jj}\cdot\sum_{ij}{}' W^G_{ii}W^G_{jj}\Big\}$$
$$+(n-3)!\left\{4\sum_{ijk}{}' W^Y_{ij}W^Y_{ik}\cdot\sum_{ijk}{}' W^G_{ij}W^G_{ik}+2\sum_{ijk}{}' W^Y_{ii}W^Y_{jk}\cdot\sum_{ijk}{}' W^G_{ii}W^G_{jk}\right\}$$
$$+(n-4)!\sum_{ijkl}{}' W^Y_{ij}W^Y_{kl}\cdot\sum_{ijkl}{}' W^G_{ij}W^G_{kl}, \tag{3.20}$$

where all the sums over distinct indices ($\sum'$) can be computed from a small number of quantities (see Lemma 7).

Proof.

$$\begin{aligned}
n!\,\mathrm{E}(Q_T^2) &= \sum_{\sigma\in S_n}\mathrm{tr}(\mathbf{W}^{G,\sigma}\mathbf{W}^Y)^2\\
&=\sum_{ijkl}\sum_{\sigma\in S_n} W^{G,\sigma}_{ij}W^Y_{ij}\,W^{G,\sigma}_{kl}W^Y_{kl}\\
&=\sum_{i}\sum_{\sigma\in S_n}(W^Y_{ii})^2(W^{G,\sigma}_{ii})^2\\
&+\sum_{ij}{}'\sum_{\sigma\in S_n}\left\{4W^Y_{ii}W^Y_{ij}W^{G,\sigma}_{ii}W^{G,\sigma}_{ij}+2(W^Y_{ij})^2(W^{G,\sigma}_{ij})^2+W^Y_{ii}W^Y_{jj}W^{G,\sigma}_{ii}W^{G,\sigma}_{jj}\right\}\\
&+\sum_{ijk}{}'\sum_{\sigma\in S_n}\left\{4W^Y_{ij}W^Y_{ik}W^{G,\sigma}_{ij}W^{G,\sigma}_{ik}+2\,W^Y_{ii}W^Y_{jk}W^{G,\sigma}_{ii}W^{G,\sigma}_{jk}\right\}\\
&+\sum_{ijkl}{}'\sum_{\sigma\in S_n}W^Y_{ij}W^Y_{kl}W^{G,\sigma}_{ij}W^{G,\sigma}_{kl}.
\end{aligned}\tag{3.21}$$

Using Lemma 4, the desired result follows. □

Lemma 4. *For an arbitrary symmetric p.s.d. matrix* $\mathbf{A}$*, and up to four distinct indices* i, j, k *and* l*,*

$$\sum_{\sigma\in S_n}(A^\sigma_{ii})^2=(n-1)!\sum_i A^2_{ii},\ \sum_{\sigma\in S_n}A^\sigma_{ii}A^\sigma_{ij}=(n-2)!\sum_{ij}{}' A_{ii}A_{ij},$$
$$\sum_{\sigma\in S_n}(A^\sigma_{ij})^2=(n-2)!\sum_{ij}{}' A^2_{ij},\ \sum_{\sigma\in S_n}A^\sigma_{ii}A^\sigma_{jj}=(n-2)!\sum_{ij}{}' A_{ii}A_{jj},$$

$$\sum_{\sigma\in S_n} A^\sigma_{ij}A^\sigma_{ik} = (n-3)!\sum_{ijk}{}' A_{ij}A_{ik},\ \sum_{\sigma\in S_n} A^\sigma_{ii}A^\sigma_{jk} = (n-3)!\sum_{ijk}{}' A_{ii}A_{jk},$$

$$\sum_{\sigma\in S_n} A^\sigma_{ij}A^\sigma_{kl} = (n-4)!\sum_{ijkl}{}' A_{ij}A_{kl}.$$

Proposition 5. *The third moment of Q_T is given by,*

$$\begin{aligned}
n!\,\mathrm{E}(Q_T^3) &= (n-1)!\sum_i (W^Y_{ii})^3 \sum_i (W^G_{ii})^3 \\
&+ (n-2)!\Bigg\{ 6\sum_{ij}{}' W^Y_{ii}W^Y_{jj}W^Y_{ij}\cdot\sum_{ij}{}' W^G_{ii}W^G_{jj}W^G_{ij} + 12\sum_{ij}{}' W^Y_{ii}(W^Y_{ij})^2\cdot \\
&\sum_{ij}{}' W^G_{ii}(W^G_{ij})^2 + 6\sum_{ij}{}'(W^Y_{ii})^2W^Y_{ij}\cdot\sum_{ij}{}'(W^G_{ii})^2W^G_{ij} + 3\sum_{ij}{}'(W^Y_{ii})^2W^Y_{jj}\cdot \\
&\sum_{ij}{}'(W^G_{ii})^2W^G_{jj} + 4\sum_{ij}{}'(W^Y_{ij})^3\cdot\sum_{ij}{}'(W^G_{ij})^3\Bigg\} \\
&+ (n-3)!\Bigg\{ \sum_{ijk}{}' W^Y_{ii}W^Y_{jj}W^Y_{kk}\cdot\sum_{ijk}{}' W^G_{ii}W^G_{jj}W^G_{kk} + 12\sum_{ijk}{}' W^Y_{ii}W^Y_{jj}W^Y_{ik}\cdot \\
&\sum_{ijk}{}' W^G_{ii}W^G_{jj}W^G_{ik} + 12\sum_{ijk}{}' W^Y_{ii}W^Y_{ij}W^Y_{ik}\cdot\sum_{ijk}{}' W^G_{ii}W^G_{ij}W^G_{ik} + 24\sum_{ijk}{}' W^Y_{ii}W^Y_{ij}W^Y_{jk}\cdot \\
&\sum_{ijk}{}' W^G_{ii}W^G_{ij}W^G_{jk} + 8\sum_{ijk}{}' W^Y_{ij}W^Y_{ik}W^Y_{jk}\cdot\sum_{ijk}{}' W^G_{ij}W^G_{ik}W^G_{jk} + 6\sum_{ijk}{}' W^Y_{ii}(W^Y_{jk})^2\cdot \\
&\sum_{ijk}{}' W^G_{ii}(W^G_{jk})^2 + 3\sum_{ijk}{}'(W^Y_{ii})^2W^Y_{jk}\cdot\sum_{ijk}{}'(W^G_{ii})^2W^G_{jk} + \\
&24\sum_{ijk}{}'(W^Y_{ij})^2W^Y_{ik}\cdot\sum_{ijk}{}'(W^G_{ij})^2W^G_{ik}\Bigg\} \\
&+ (n-4)!\Bigg\{ 3\sum_{ijkl}{}' W^Y_{ii}W^Y_{jj}W^Y_{kl}\cdot\sum_{ijkl}{}' W^G_{ii}W^G_{jj}W^G_{kl} + 12\sum_{ijkl}{}' W^Y_{ii}W^Y_{ij}W^Y_{kl}\cdot \\
&\sum_{ijkl}{}' W^G_{ii}W^G_{ij}W^G_{kl} + 12\sum_{ijkl}{}' W^Y_{ii}W^Y_{jk}W^Y_{jl}\cdot\sum_{ijkl}{}' W^G_{ii}W^G_{jk}W^G_{jl} + 8\sum_{ijkl}{}' W^Y_{ij}W^Y_{ik}W^Y_{il}\cdot \\
&\sum_{ijkl}{}' W^G_{ij}W^G_{ik}W^G_{il} + 24\sum_{ijkl}{}' W^Y_{ij}W^Y_{ik}W^Y_{jl}\cdot\sum_{ijkl}{}' W^G_{ij}W^G_{ik}W^G_{jl} + \\
&6\sum_{ijkl}{}'(W^Y_{ij})^2W^Y_{kl}\cdot\sum_{ijkl}{}'(W^G_{ij})^2W^G_{kl}\Bigg\}
\end{aligned}$$

$$+ (n-5)! \left\{ 12 \sum_{ijklm}{}' W^Y_{ij} W^Y_{ik} W^Y_{lm} \cdot \sum_{ijklm}{}' W^G_{ij} W^G_{ik} W^G_{lm} + \right.$$
$$\left. 3 \sum_{ijklm}{}' W^Y_{ii} W^Y_{jk} W^Y_{lm} \cdot \sum_{ijklm}{}' W^G_{ii} W^G_{jk} W^G_{lm} \right\}$$
$$+ (n-6)! \sum_{ijklmn}{}' W^Y_{ij} W^Y_{kl} W^Y_{mn} \cdot \sum_{ijklmn}{}' W^G_{ij} W^G_{kl} W^G_{mn}, \tag{3.22}$$

where all the sums over the distinct indices ($\sum'$) can be computed from a small number of quantities (see Lemma 7).

Proof.

$$\begin{aligned}
n!\,\mathrm{E}(Q_T^3) &= \sum_{\sigma \in S_n} \mathrm{tr}(\mathbf{W}^{G,\sigma} \mathbf{W}^Y)^3 \\
&= \sum_{ijklmn} \sum_{\sigma \in S_n} W^{G,\sigma}_{ij} W^Y_{ij}\, W^{G,\sigma}_{kl} W^Y_{kl}\, W^{G,\sigma}_{mn} W^Y_{mn} \\
&= \sum_{i} \sum_{\sigma \in S_n} (W^Y_{ii})^3 (W^{G,\sigma}_{ii})^3 \\
&+ \sum_{ij}{}' \sum_{\sigma \in S_n} \big\{ 6 W^Y_{ii} W^Y_{jj} W^Y_{ij} W^{G,\sigma}_{ii} W^{G,\sigma}_{jj} W^{G,\sigma}_{ij} + 12 W^Y_{ii} (W^Y_{ij})^2 W^{G,\sigma}_{ii} (W^{G,\sigma}_{ij})^2 + \\
&6 (W^Y_{ii})^2 W^Y_{ij} (W^{G,\sigma}_{ii})^2 W^{G,\sigma}_{ij} + 3 (W^Y_{ii})^2 W^Y_{jj} (W^{G,\sigma}_{ij})^2 W^{G,\sigma}_{jj} + 4 (W^Y_{ij})^3 (W^{G,\sigma}_{ij})^3 \big\} \\
&+ \sum_{ijk}{}' \sum_{\sigma \in S_n} \big\{ W^Y_{ii} W^Y_{jj} W^Y_{kk} W^{G,\sigma}_{ii} W^{G,\sigma}_{jj} W^{G,\sigma}_{kk} + 12 W^Y_{ii} W^Y_{jj} W^Y_{ik} W^{G,\sigma}_{ii} W^{G,\sigma}_{jj} W^{G,\sigma}_{ik} + \\
&12 W^Y_{ii} W^Y_{ij} W^Y_{ik} W^{G,\sigma}_{ii} W^{G,\sigma}_{ij} W^{G,\sigma}_{ik} + 24 W^Y_{ii} W^Y_{ij} W^Y_{jk} W^{G,\sigma}_{ii} W^{G,\sigma}_{ij} W^{G,\sigma}_{jk} + \\
&8 W^Y_{ij} W^Y_{ik} W^Y_{jk} W^{G,\sigma}_{ij} W^{G,\sigma}_{ik} W^{G,\sigma}_{jk} + 6 W^Y_{ii} (W^Y_{jk})^2 W^{G,\sigma}_{ii} (W^{G,\sigma}_{jk})^2 + \\
&3 (W^Y_{ii})^2 W^Y_{jk} (W^{G,\sigma}_{ii})^2 W^{G,\sigma}_{jk} + 24 (W^Y_{ij})^2 W^Y_{ik} (W^{G,\sigma}_{ij})^2 W^{G,\sigma}_{ik} \big\} \\
&+ \sum_{ijkl}{}' \sum_{\sigma \in S_n} \big\{ 3 W^Y_{ii} W^Y_{jj} W^Y_{kl} W^{G,\sigma}_{ii} W^{G,\sigma}_{jj} W^{G,\sigma}_{kl} + 12 W^Y_{ii} W^Y_{ij} W^Y_{kl} W^{G,\sigma}_{ii} W^{G,\sigma}_{ij} W^{G,\sigma}_{kl} + \\
&12 W^Y_{ii} W^Y_{jk} W^Y_{jl} W^{G,\sigma}_{ii} W^{G,\sigma}_{jk} W^{G,\sigma}_{jl} + 8 W^Y_{ij} W^Y_{ik} W^Y_{il} W^{G,\sigma}_{ij} W^{G,\sigma}_{ik} W^{G,\sigma}_{il} + \\
&24 W^Y_{ij} W^Y_{ik} W^Y_{jl} W^{G,\sigma}_{ij} W^{G,\sigma}_{ik} W^{G,\sigma}_{jl} + 6 (W^Y_{ij})^2 W^Y_{kl} (W^{G,\sigma}_{ij})^2 W^{G,\sigma}_{kl} \big\} \\
&+ \sum_{ijklm}{}' \sum_{\sigma \in S_n} \big\{ 12 W^Y_{ij} W^Y_{ik} W^Y_{lm} W^{G,\sigma}_{ij} W^{G,\sigma}_{ik} W^{G,\sigma}_{lm} + 3 W^Y_{ii} W^Y_{jk} W^Y_{lm} W^{G,\sigma}_{ii} W^{G,\sigma}_{jk} W^{G,\sigma}_{lm} \big\}
\end{aligned}$$

$$+ \sum_{ijklmn}{}' \sum_{\sigma \in S_n} W^Y_{ij} W^Y_{kl} W^Y_{mn} W^{G,\sigma}_{ij} W^{G,\sigma}_{kl} W^{G,\sigma}_{mn}. \tag{3.23}$$

Using Lemma 6, the desired result follows. □

Lemma 6. *For an arbitrary symmetric p.s.d. matrix* $\mathbf{A}$*, and up to six distinct indices* i, j, k, l, m *and* n*,*

$$\sum_{\sigma \in S_n} (A^\sigma_{ii})^3 = (n-1)! \sum_i A^3_{ii}, \sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{jj} A^\sigma_{ij} = (n-2)! \sum_{ij}{}' A_{ii} A_{jj} A_{ij},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ii} (A^\sigma_{ij})^2 = (n-2)! \sum_{ij}{}' A_{ii} A^2_{ij}, \sum_{\sigma \in S_n} (A^\sigma_{ii})^2 A^\sigma_{ij} = (n-2)! \sum_{ij}{}' A^2_{ii} A_{ij},$$

$$\sum_{\sigma \in S_n} (A^\sigma_{ii})^2 A^\sigma_{jj} = (n-2)! \sum_{ij}{}' A^2_{ii} A_{jj}, \sum_{\sigma \in S_n} (A^\sigma_{ij})^3 = (n-2)! \sum_i A^3_{ij},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{jj} A^\sigma_{kk} = (n-3)! \sum_{ijk}{}' A_{ii} A_{jj} A_{kk}, \sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{jj} A^\sigma_{ik} = (n-3)! \sum_{ijk}{}' A_{ii} A_{jj} A_{ik},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{ij} A^\sigma_{ik} = (n-3)! \sum_{ijk}{}' A_{ii} A_{ij} A_{ik}, \sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{ij} A^\sigma_{jk} = (n-3)! \sum_{ijk}{}' A_{ii} A_{ij} A_{jk},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ij} A^\sigma_{ik} A^\sigma_{jk} = (n-3)! \sum_{ijk}{}' A_{ij} A_{ik} A_{jk}, \sum_{\sigma \in S_n} A^\sigma_{ii} (A^\sigma_{jk})^2 = (n-3)! \sum_{ijk}{}' A_{ii} A^2_{jk},$$

$$\sum_{\sigma \in S_n} (A^\sigma_{ii})^2 A^\sigma_{jk} = (n-3)! \sum_{ijk}{}' A^2_{ii} A_{jk}, \sum_{\sigma \in S_n} (A^\sigma_{ij})^2 A^\sigma_{ik} = (n-3)! \sum_{ijk}{}' A^2_{ij} A_{ik},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{jj} A^\sigma_{kl} = (n-4)! \sum_{ijkl}{}' A_{ii} A_{jj} A_{kl}, \sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{ij} A^\sigma_{kl} = (n-4)! \sum_{ijkl}{}' A_{ii} A_{ij} A_{kl},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{jk} A^\sigma_{jl} = (n-4)! \sum_{ijkl}{}' A_{ii} A_{jk} A_{jl}, \sum_{\sigma \in S_n} A^\sigma_{ij} A^\sigma_{ik} A^\sigma_{il} = (n-4)! \sum_{ijkl}{}' A_{ij} A_{ik} A_{il},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ij} A^\sigma_{ik} A^\sigma_{jl} = (n-4)! \sum_{ijkl}{}' A_{ij} A_{ik} A_{jl}, \sum_{\sigma \in S_n} (A^\sigma_{ij})^2 A^\sigma_{kl} = (n-4)! \sum_{ijkl}{}' A^2_{ij} A_{kl},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ij} A^\sigma_{ik} A^\sigma_{lm} = (n-5)! \sum_{ijklm}{}' A_{ij} A_{ik} A_{lm}, \sum_{\sigma \in S_n} A^\sigma_{ii} A^\sigma_{jk} A^\sigma_{lm} = (n-5)! \sum_{ijklm}{}' A_{ii} A_{jk} A_{lm},$$

$$\sum_{\sigma \in S_n} A^\sigma_{ik} A^\sigma_{kl} A^\sigma_{mn} = (n-6)! \sum_{ijklmn}{}' A_{ij} A_{kl} A_{mn}.$$

Lemma 7. *All the sums needed in Propositions 1, 3 and 5 can be expressed in terms of the*

following quantities,

$$T_1^A = \sum_i A_{ii},\ T_2^A = \sum_i A_{ii}^2,\ T_3^A = \sum_i A_{ii}^3,$$
$$S_1^A = \sum_{ij} A_{ij},\ S_2^A = \sum_{ij} A_{ij}^2,\ S_3^A = \sum_{ij} A_{ij}^3,$$
$$R_1^A = \sum_{ij} A_{ii}A_{ij},\ R_2^A = \sum_{ij} A_{ii}A_{ij}^2,\ R_3^A = \sum_{ij} A_{ii}^2A_{ij},\ R_4^A = \sum_{ij} A_{ii}A_{jj}A_{ij},$$
$$C_1^A = \sum_{ijk} A_{ij}A_{ik},\ C_2^A = \sum_{ijk} A_{ij}^2A_{ik},\ C_3^A = \sum_{ijk} A_{ii}A_{ij}A_{ik},\ C_4^A = \sum_{ijk} A_{ii}A_{ij}A_{jk},$$
$$C_5^A = \sum_{ijk} A_{ij}A_{ik}A_{jk},\ D_1^A = \sum_{ijkl} A_{ij}A_{ik}A_{il},\ D_2^A = \sum_{ijkl} A_{ij}A_{ik}A_{jl},$$

where the matrix A is symmetric p.s.d. and can be replaced by $\mathbf{W}^G$ or $\mathbf{W}^Y$ (corresponding to superscript G or Y, respectively).

Proof.

$$\sum_{ij}{}' A_{ij} = S_1^A - T_1^A,$$
$$\sum_{ij}{}' A_{ij}^2 = S_2^A - T_2^A,$$
$$\sum_{ij}{}' A_{ij}^3 = S_3^A - T_3^A,$$
$$\sum_{ij}{}' A_{ii}A_{ij} = R_1^A - T_2^A,$$
$$\sum_{ij}{}' A_{ii}A_{ij}^2 = R_2^A - T_3^A,$$
$$\sum_{ij}{}' A_{ii}^2A_{ij} = R_3^A - T_2^A,$$
$$\sum_{ij}{}' A_{ii}A_{jj} = (T_1^A)^2 - T_2^A,$$
$$\sum_{ij}{}' A_{ii}^2A_{jj} = T_1^AT_2^A - T_3^A,$$

$$\sum_{ij}{}' A_{ii}A_{jj}A_{ij} = R_4^A - T_3^A,$$

$$\sum_{ijk}{}' A_{ii}A_{jk} = T_1^A(S_1^A - T_1^A) + 2(T_2^A - R_1^A),$$

$$\sum_{ijk}{}' A_{ii}A_{jk}^2 = T_1^A(S_2^A - T_2^A) + 2(T_3^A - R_2^A),$$

$$\sum_{ijk}{}' A_{ii}^2A_{jk} = T_2^A(S_1^A - T_1^A) + 2(T_3^A - R_3^A),$$

$$\sum_{ijk}{}' A_{ij}A_{ik} = C_1^A - S_2^A + 2(T_2^A - R_1^A),$$

$$\sum_{ijk}{}' A_{ij}^2A_{ik} = C_2^A - R_2^A - R_3^A - S_3^A + 2T_3^A,$$

$$\sum_{ijk}{}' A_{ii}A_{ij}A_{jk} = C_4^A - R_2^A - R_3^A - R_4^A + 2T_3^A,$$

$$\sum_{ijk}{}' A_{ii}A_{ij}A_{ik} = C_3^A - R_2^A + 2(T_3^A - R_3^A),$$

$$\sum_{ijk}{}' A_{ii}A_{jj}A_{ik} = T_1^A(R_1^A - T_2^A) - R_3^A - R_4^A + 2T_3^A,$$

$$\sum_{ijk}{}' A_{ii}A_{jj}A_{kk} = T_1^A[(T_1^A)^2 - T_2^A] + 2(T_3^A - T_1^AT_2^A),$$

$$\sum_{ijk}{}' A_{ij}A_{ik}A_{jk} = C_5^A - R_2^A + 2(T_3^A - R_2^A),$$

$$\sum_{ijkl}{}' A_{ij}A_{kl} = S_1^A(S_1^A - T_1^A) + 2(R_1^A - C_1^A) - 2\sum_{ijk}{}' A_{ij}A_{ik} - \sum_{ijk}{}' A_{ii}A_{jk},$$

$$\sum_{ijkl}{}' A_{ij}^2A_{kl} = S_2^A(S_1^A - T_1^A) + 2(R_2^A - C_2^A) - 2\sum_{ijk}{}' A_{ij}^2A_{ik} - \sum_{ijk}{}' A_{ii}^2A_{jk},$$

$$\sum_{ijkl}{}' A_{ii}A_{jj}A_{kl} = T_1^A\sum_{ijk}{}' A_{ii}A_{jk} - 2\sum_{ijk}{}' A_{ii}A_{jj}A_{ik} - \sum_{ijk}{}' A_{ii}^2A_{jk},$$

$$\sum_{ijkl}{}' A_{ii}A_{ij}A_{kl} = R_1^A(S_1^A - T_1^A) + 2(R_3^A - C_3^A) - 2\sum_{ijk}{}' A_{ii}A_{ij}A_{jk} - \sum_{ijk}{}' A_{ii}^2A_{jk},$$

$$\sum_{ijkl}{}' A_{ii}A_{jk}A_{jl} = T_1^A\sum_{ijk}{}' A_{ij}A_{ik} - 2\sum_{ijk}{}' A_{ii}A_{ij}A_{jk} - \sum_{ijk}{}' A_{ii}A_{ij}A_{ik},$$

$$\sum_{ijkl}{}' A_{ij}A_{ik}A_{il} = D_1^A - 3C_2^A + 2S_3^A - 3\sum_{ijk}{}' A_{ii}A_{ij}A_{ik},$$

$$\sum_{ijkl}{}' A_{ij}A_{ik}A_{jl} = D_2^A - C_2^A - C_4^A - C_5^A + 2R_2^A - \sum_{ijk}{}' A_{ii}A_{ij}A_{ik}$$
$$- \sum_{ijk}{}' A_{ii}A_{ij}A_{jk} - \sum_{ijk}{}' A_{ij}^2 A_{ik},$$
$$\sum_{ijklm}{}' A_{ii}A_{jk}A_{lm} = T_1^A \sum_{ijkl}{}' A_{ij}A_{kl} - 4\sum_{ijkl}{}' A_{ii}A_{ij}A_{kl},$$
$$\sum_{ijklm}{}' A_{ij}A_{ik}A_{lm} = (C_1^A - S_2^A)(S_1^A - T_1^A) + 4(C_2^A + C_4^A - D_2^A) + 2C_5^A - 6R_2^A$$
$$- 2\sum_{ijkl}{}' A_{ii}A_{ij}A_{kl} - 2\sum_{ijkl}{}' A_{ij}A_{ik}A_{il},$$
$$\sum_{ijklmn}{}' A_{ij}A_{kl}A_{mn} = S_1^A \sum_{ijkl}{}' A_{ij}A_{kl} + 4(R_1^A - C_1^A)(S_1^A - T_1^A) + 8(D_1^A + D_2^A - C_2^A - C_4^A)$$
$$+ 24(R_3^A - C_3^A) - 4\sum_{ijklm}{}' A_{ij}A_{ik}A_{lm} - \sum_{ijklm}{}' A_{ii}A_{jk}A_{lm}.$$

□

REFERENCES

Abney, M., McPeek, M. S., and Ober, C. Estimation of variance components of quantitative traits in inbred populations. *Am J Hum Genet*, 66(2):629–650, 2000.

Abney, M. Permutation testing in the presence of polygenic variation. *Genetic Epidemiology*, 39(4):249–258, 2015.

Abney, M., Ober, C., and McPeek, M. S. Quantitative-Trait Homozygosity and Association Mapping and Empirical Genomewide Significance in Large, Complex Pedigrees: Fasting Serum-Insulin Level in the Hutterites. *The American Journal of Human Genetics*, 70(4): 920–934, 2002.

Alexander, D. H., Novembre, J., and Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009.

Allison, D. B., Thiel, B., St. Jean, P., Elston, R. C., Infante, M. C., and Schork, N. J. Multiple Phenotype Modeling in Gene-Mapping Studies of Quantitative Traits: Power Advantages. *The American Journal of Human Genetics*, 63(4):1190–1201, 1998.

Astle, W. and Balding, D. J. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, pages 451–471, 2009.

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

Bianchi, M., Dahlgren, S., Massey, J., Dietschi, E., Kierczak, M., Lund-Ziener, M., Sundberg, K., Thoresen, S. I., Kämpe, O., Andersson, G., Ollier, W. E., Hedhammar, Å., Leeb, T., Lindblad-Toh, K., Kennedy, L. J., Lingaas, F., and Pielberg, G. R. A multi-breed genome-wide association analysis for canine Hypothyroidism identifies a shared major risk locus on CFA12. *PLoS ONE*, 10(8):e0134720, 2015.

Boyce, A. J. Computation of inbreeding and kinship coefficients on extended pedigrees. *Journal of Heredity*, 74(6):400–404, 1983.

Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A., Bielak, L. F., Zhao, W., Smith, J. A., Peyser, P. A., Kardia, S. L., Ghosh, D., and Epstein, M. P. A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. *American Journal of Human Genetics*, 98(3):525–540, 2016.

Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., Redline, S., Papanicolaou, G. J., Thornton, T. A., Laurie, C. C., Rice, K., and Lin, X. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *American Journal of Human Genetics*, 98(4):653–666, 2016a.

Chen, J., Chen, W., Zhao, N., Wu, M. C., and Schaid, D. J. Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. *Genetic Epidemiology*, 40 (1):5–19, 2016b.

Churchill, G. A. and Doerge, R. W. Naive application of permutation testing leads to inflated type I error rates. *Genetics*, 178(1):609–610, 2008.

Conomos, M. P., Miller, M. B., and Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*, 39(4):276–293, 2015.

Conomos, M. P., Reiner, A. P., Weir, B. S., and Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *American Journal of Human Genetics*, 98(1):127–148, 2016.

Dahl, A., Iotchkova, V., Baud, A., Johansson, Å., Gyllensten, U., Soranzo, N., Mott, R., Kranis, A., and Marchini, J. A multiple-phenotype imputation method for genetic studies. *Nature genetics*, 48(4):466–72, 2016.

Duboscq-Bidot, L., Xu, P., Charron, P., Neyroud, N., Dilanian, G., Millaire, A., Bors, V., Komajda, M., and Villard, E. Mutations in the Z-band protein myopalladin gene and idiopathic dilated cardiomyopathy. *Cardiovascular Research*, 77(1):118–125, 2008.

Dudbridge, F. and Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3):227–234, 2008.

Dufour, C., Capasso, M., Svahn, J., Marrone, A., Haupt, R., Bacigalupo, A., Giordani, L., Longoni, D., Pillon, M., Pistorio, A., Di Michele, P., Iori, A. P., Pongiglione, C., Lanciotti, M., and Iolascon, A. Homozygosis for (12) CA repeats in the first intron of the human IFN-γ gene is significantly associated with the risk of aplastic anaemia in Caucasian population. *British Journal of Haematology*, 126(5):682–685, 2004.

Dutta, D., Scott, L., Boehnke, M., and Lee, S. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genetic Epidemiology*, 2018.

Feinleib, M., Kannel, W. B., Garrison, R. J., McNamara, P. M., and Castelli, W. P. The framingham offspring study. Design and preliminary data. *Preventive Medicine*, 4(4): 518–525, 1975.

Ferreira, M. A., McRae, A. F., Medland, S. E., Nyholt, D. R., Gordon, S. D., Wright, M. J., Henders, A. K., Madden, P. A., Visscher, P. M., Wray, N. R., Heath, A. C., Montgomery, G. W., Duffy, D. L., and Martin, N. G. Association between ORMDL3, IL1RL1 and a deletion on chromosome 17q21 with asthma risk in Australia. *European Journal of Human Genetics*, 19(4):458–464, 2011.

Fisher, R. A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52(02):399–433, 1919.

Fröjmark, A. S., Schuster, J., Sobol, M., Entesarian, M., Kilander, M. B., Gabrikova, D., Nawaz, S., Baig, S. M., Schulte, G., Klar, J., and Dahl, N. Mutations in Frizzled 6 cause isolated autosomal-recessive nail dysplasia. *American Journal of Human Genetics*, 88(6): 852–860, 2011.

Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L., and Vermeulen, S. H. A Comparison of Multivariate Genome-Wide Association Methods. *PLoS ONE*, 9(4):e95923, 2014.

Grotenboer, N. S., Ketelaar, M. E., Koppelman, G. H., and Nawijn, M. C. Decoding asthma: Translating genetic variation in IL33 and IL1RL1 into disease pathophysiology. *Journal of Allergy and Clinical Immunology*, 131(3):856–865.e9, 2013.

Hayward, J. J., Castelhano, M. G., Oliveira, K. C., Corey, E., Balkman, C., Baxter, T. L., Casal, M. L., Center, S. A., Fang, M., Garrison, S. J., Kalla, S. E., Korniliev, P., Kotlikoff, M. I., Moise, N. S., Shannon, L. M., Simpson, K. W., Sutter, N. B., Todhunter, R. J., and Boyko, A. R. Complex disease and phenotype mapping in the domestic dog. *Nature Communications*, 7, 2016a.

Hayward, J. J., Castelhano, M. G., Oliveira, K. C., Corey, E., Balkman, C., Baxter, T. L., Casal, M. L., Center, S. A., Fang, M., Garrison, S. J., Kalla, S. E., Korniliev, P., Kotlikoff, M. I., Moise, N. S., Shannon, L. M., Simpson, K. W., Sutter, N. B., Todhunter, R. J., and Boyko, A. R. Data from: Complex disease and phenotype mapping in the domestic dog, 2016b. URL `https://doi.org/10.5061/dryad.266k4`.

Hodgkinson, C. A., Goldman, D., Jaeger, J., Persaud, S., Kane, J. M., Lipsky, R. H., Malhotra, A. K., Hodgkinson, C. A., Goldman, D., Jaeger, J., Kane, J. M., Lipsky, R. H., and Malhotra, A. K. Disrupted in schizophrenia 1 (DISC1): association with schizophrenia, schizoaffective disorder, and bipolar disorder. *Am J Hum Genet*, 75(5):862–872, 2004.

Hoffman, G. E. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS ONE*, 8(10):e75707, 2013.

Hou, S., Xiao, X., Li, F., Jiang, Z., Kijlstr, A., and Yang, P. Two-stage association study in Chinese Han identifies two independent associations in CCR1/CCR3 locus as candidate for Behçet's disease susceptibility. *Human Genetics*, 131(12):1841–1850, 2012.

Howie, B. N., Donnelly, P., and Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6): e1000529, 2009.

Hua, W. Y. and Ghosh, D. Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics*, 71(3):812–820, 2015.

Iqbal, Z., Vandeweyer, G., Van der voet, M., Waryah, A. M., Zahoor, M. Y., Besseling, J. A., Roca, L. T., Vulto-van silfhout, A. T., Nijhof, B., Kramer, J. M., Van der Aa, N., Ansar, M., Peeters, H., Helsmoortel, C., Gilissen, C., Vissers, L. E., Veltman, J. A., De brouwer, A. P.,

Frank kooy, R., Riazuddin, S., Schenck, A., Van bokhoven, H., and Rooms, L. Homozygous and heterozygous disruptions of ANK3: At the crossroads of neurodevelopmental and psychiatric disorders. *Human Molecular Genetics*, 22(10):1960–1970, 2013.

Jakobsdottir, J. and McPeek, M. S. MASTOR: Mixed-model association mapping of quantitative traits in samples with related individuals. *American Journal of Human Genetics*, 92(5):652–666, 2013.

Jang, W. and Lim, J. A numerical study of PQL estimation biases in generalized linear mixed models under heterogeneity of random effects. *Communications in Statistics: Simulation and Computation*, 38(4):692–702, 2009.

Jiang, D. and Mcpeek, M. S. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genetic Epidemiology*, 38(1):10–20, 2014.

Jiang, D., Zhong, S., and McPeek, M. S. Retrospective Binary-Trait Association Test Elucidates Genetic Architecture of Crohn Disease. *American Journal of Human Genetics*, 98(2):243–255, 2016.

Joehanes, R., Ying, S., Huan, T., Johnson, A. D., Raghavachari, N., Wang, R., Liu, P., Woodhouse, K. A., Sen, S. K., Tanriverdi, K., Courchesne, P., Freedman, J. E., O'Donnell, C. J., Levy, D., and Munson, P. J. Gene expression signatures of coronary heart disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 33(6):1418–1426, 2013.

Joo, J. W. J., Kang, E. Y., Org, E., Furlotte, N., Parks, B., Hormozdiari, F., Lusis, A. J., and Eskin, E. Efficient and accurate multiple-phenotype regression method for high dimensional data considering population structure. *Genetics*, 204(4):1379–1390, 2016.

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

Kanehisa, M. and Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. Efficient Control of Population Stucuture in Model Organism Association Mapping. *Genetics*, 178(3):1709–1723, 2008.

Kazi-Aoual, F., Hitier, S., Sabatier, R., and Lebreton, J. D. Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis*, 20(6):643–656, 1995.

Kennedy, A. E., Ozbek, U., and Dorak, M. T. What has GWAS done for HLA and disease associations? *International Journal of Immunogenetics*, 44(5):195–211, 2017.

Kim, J., Zhang, Y., and Pan, W. Powerful and adaptive testing for multi-trait and Multi-SNP associations with GWAS and sequencing data. *Genetics*, 203(2):715–731, 2016.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., and Lin, X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91(2):224–237, 2012a.

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

Lee, S., Fuchsberger, C., Kim, S., and Scott, L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. *Biostatistics*, 17(1):1–15, 2016.

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

Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):309–326, 1997.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833–835, 2011.

Liu, Y., Nyunoya, T., Leng, S., Belinsky, S. A., Tesfaigzi, Y., and Bruse, S. Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics*, 7(1):1, 2013.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., MacKay, T. F., McCarroll, S. A., and Visscher, P. M. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. The effects of human population structure on large genetic association studies. *Nature Genetics*, 36(5):512–517, 2004.

Mbatchou, J., Abney, M., and McPeek, M. S. Permutation methods for assessing significance in binary trait association mapping with structured samples. *bioRxiv*, page 451377, 2018.

Meguro, A., Bertsias, G., Wood, G. M., Ozyazgan, Y., Satorius, C., Ueda, A., Tugal-Tutkun, I., Kim, Y., Ishigatsubo, Y., Sacli, F. S., Takeno, M., Kastner, D. L., Cakar, A., Inoko, H., Abaci, N., Seyahi, E., Mizuki, N., Ombrello, M. J., Kirino, Y., Gül, A., Emrence, Z., Erer, B., Ustek, D., and Remmers, E. F. Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. *Nature Genetics*, 45(2):202–207, 2013.

Melin, M., Rivera, P., Arendt, M., Elvers, I., Murén, E., Gustafson, U., Starkey, M., Borge, K. S., Lingaas, F., Häggström, J., Saellström, S., Rönnberg, H., and Lindblad-Toh, K. Genome-Wide Analysis Identifies Germ-Line Risk Factors Associated with Canine Mammary Tumours. *PLoS Genetics*, 12(5):e1006029, 2016.

Meyer, T., Ruppert, V., Ackermann, S., Richter, A., Perrot, A., Sperling, S. R., Posch, M. G., Maisch, B., and Pankuweit, S. Novel mutations in the sarcomeric protein myopalladin in patients with dilated cardiomyopathy. *European Journal of Human Genetics*, 21(3): 294–300, 2013.

Minas, C., Curry, E., and Montana, G. A distance-based test of association between paired heterogeneous genomic data. *Bioinformatics*, 29(20):2555–2563, 2013.

Murtagh, F. Complexities of hierarchic clustering algorithms: state of the art. *Computational Statistics Quarterly*, 1(2):101–113, 1984.

Palmer, E. E., Jarrett, K. E., Sachdev, R. K., Zahrani, F. A., Hashem, M. O., Ibrahim, N., Sampaio, H., Kandula, T., Macintosh, R., Gupta, R., Conlon, D. M., Billheimer, J. T., Rader, D. J., Funato, K., Walkey, C. J., Lee, C. S., Loo, C., Brammah, S., Elakis, G., Zhu, Y., Buckley, M., Kirk, E. P., Bye, A., Alkuraya, F. S., Roscioli, T., and Lagor, W. R. Neuronal deficiency of ARV1 causes an autosomal recessive epileptic encephalopathy. *Human Molecular Genetics*, 25(14):3042–3054, 2016.

Pan, W., Kwak, I.-Y., and Wei, P. A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants. *The American Journal of Human Genetics*, 97(1):86–98, 2015.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., and Sunyaev, S. R. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *American Journal of Human Genetics*, 86(6):832–838, 2010a.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010b.

Rabe-hesketh, S., Skrondal, A., and Pickles, A. Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, 2(1):1–21, 2002.

Raudenbush, S. W., Yang, M. L., and Yosef, M. Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics*, 9(1):141–157, 2000.

Rodríguez, G. and Goldman, N. Improved estimation procedures for multilevel models with binary response: A case-study. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 164(2):339–355, 2001.

Safra, N., Bassuk, A. G., Ferguson, P. J., Aguilar, M., Coulson, R. L., Thomas, N., Hitchens, P. L., Dickinson, P. J., Vernau, K. M., Wolf, Z. T., and Bannasch, D. L. Genome-Wide Association Mapping in Dogs Enables Identification of the Homeobox Gene, NKX2-8, as a Genetic Component of Neural Tube Defects in Humans. *PLoS Genetics*, 9(7):e1003646, 2013.

Schaid, D. J., McDonnell, S. K., Sinnwell, J. P., and Thibodeau, S. N. Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. *Genetic Epidemiology*, 37(5):409–418, 2013.

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

Sokal, R. R. A statistical method for evaluating systematic relationship. *University of Kansas science bulletin*, 28:1409–1438, 1958.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T. I., Nudel, R., Lieder, I., Mazor, Y., Kaplan, S., Dahary, D., Warshawsky, D., Guan-Golan, Y., Kohn, A., Rappaport, N., Safran, M., and Lancet, D. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. In *Current Protocols in Bioinformatics*, volume 54, pages 1.30.1–1.30.33. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2016.

Stranger, B. E., Stahl, E. A., and Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383, 2011.

Tang, H., Peng, J., Wang, P., and Risch, N. J. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, 2005.

Tengvall, K., Kierczak, M., Bergvall, K., Olsson, M., Frankowiack, M., Farias, F. H., Pielberg, G., Carlborg, Ö., Leeb, T., Andersson, G., Hammarström, L., Hedhammar, Å., and Lindblad-Toh, K. Genome-Wide Analysis in German Shepherd Dogs Reveals Association of a Locus on CFA 27 with Atopic Dermatitis. *PLoS Genetics*, 9(5):e1003475, 2013.

Thornton, T. and McPeek, M. S. Case-Control Association Testing with Related Individuals: A More Powerful Quasi-Likelihood Score Test. *The American Journal of Human Genetics*, 81(2):321–337, 2007.

Thornton, T. A. Statistical methods for genome-wide and sequencing association studies of complex traits in related samples. *Curr Protoc Hum Genet*, 84(January):1.28.1–1.28.9, 2015.

Vattikuti, S., Guo, J., and Chow, C. C. Heritability and Genetic Correlations Explained by Common SNPs for Metabolic Syndrome Traits. *PLoS Genetics*, 8(3):e1002637, 2012.

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

Wang, Q., Yang, C., Gelernter, J., and Zhao, H. Pervasive pleiotropy between psychiatric disorders and immune disorders revealed by integrative analysis of multiple GWAS. *Human Genetics*, 134(11-12):1195–1209, 2015.

Wang, Z., Sha, Q., Fang, S., Zhang, K., and Zhang, S. Testing an optimally weighted combination of common and/or rare variants with multiple traits. *PLOS ONE*, 13(7): e0201186, 2018.

Wu, B. and Pankow, J. S. Sequence Kernel Association Test of Multiple Continuous Phenotypes. *Genetic Epidemiology*, 40(2):91–100, 2016.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93, 2011.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.

Zhan, X., Plantinga, A., Zhao, N., and Wu, M. C. A fast small-sample kernel independence test for microbiome community-level association analysis. *Biometrics*, 73(4):1453–1463, 2017a.

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

Zhong, S., Jiang, D., and McPeek, M. S. CERAMIC: Case-Control Association Testing in Samples with Related Individuals, Based on Retrospective Mixed Model Analysis with Adjustment for Covariates. *PLoS Genetics*, 12(10):e1006329, 2016.

Zhou, X. and Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, 2012.

Zhou, X. and Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4):407–409, 2014.

Zukosky, K., Meilleur, K., Traynor, B. J., Dastgir, J., Medne, L., Devoto, M., Collins, J., Rooney, J., Zou, Y., Yang, M. L., Gibbs, J. R., Meier, M., Stetefeld, J., Finkel, R. S., Schessl, J., Elman, L., Felice, K., Ferguson, T. A., Ceyhan-Birsoy, O., Beggs, A. H., Tennekoon, G., Johnson, J. O., and Bönnemann, C. G. Association of a novel ACTA1 mutation with a dominant progressive scapuloperoneal myopathy in an extended family. *JAMA Neurology*, 72(6):689–698, 2015.