THE UNIVERSITY OF CHICAGO

GENETIC ASSOCIATION ANALYSIS OF PHENOTYPES JOINTLY INFLUENCED BY A PAIR OF INTERACTING ORGANISMS

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

VASILEIOS KATSIANOS

CHICAGO, ILLINOIS

AUGUST 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This journey has been one of significant personal and professional growth, enriched by the many individuals who have propelled me forward. My development throughout my doctoral studies has been a collective effort, and I wish to acknowledge and thank those who have contributed to it.

First and foremost, I extend my heartfelt gratitude to my advisor, Professor Mary Sara McPeek, whose mentorship over the years has been invaluable. Her dedication, insightful guidance and meticulous attention to detail have been a source of inspiration in my journey as a scholar. I am deeply thankful for the countless hours she devoted to meeting with me and reviewing my writings.

I also want to thank Professors Matthew Stephens and Mark Abney for their invaluable feedback as members of my dissertation committee. My academic growth has been greatly influenced by the faculty members in the department, whose emphasis on rigorous learning has been a cornerstone of my doctoral experience. I am especially grateful to Professor Mei Wang for her continuous words of encouragement. The patience and support of the departmental administrative staff have also been greatly appreciated.

It has been a privilege to learn and grow alongside many fellow students. I am grateful to Huanlin Zhou, Yi Wei, Joonsuk Kang, Sounak Paul, Huy Dang Tran, Wei Kuang, Melissa Adrian, Soumyabrata Kundu, Sean O'Hagan, Annie Xie, Raphael Rossellini, John Hood, Jimmy Lederman, Nathan Waniorek, Peter Laurin, Madhuri Raman and many others for their insightful discussions and camaraderie. To my friends, Panos Andreou and Pavlos Zouboulоglou, thank you for your unwavering support, always believing in me and encouraging me.

Finally, my family has been my anchor throughout this chapter of my life. I am immensely grateful for their constant love and support, especially during moments of self-doubt. I am forever grateful to my father, whose empowerment and encouragement have driven me to strive for excellence and never give up.

ABSTRACT

The virulence of infectious diseases is usually affected by a combination of a host and at least one pathogen organism. Previous experiments have revealed that combining genetic information from different organisms has enabled the identification of more relevant genetic variants than just individually performing an association analysis on each organism. Hence, we are interested in performing a joint association analysis to test for the interaction effect of each possible pair of a host and pathogen genetic variant on the phenotypic trait relating to the infectious disease. Three main issues may arise when performing this joint association analysis.

First, the presence of a non-trivial interaction effect between one of the genetic variants being tested and some unaccounted factor - either observed or unobserved - can lead to heteroscedasticity in the phenotypic trait. Failure to account for this heteroscedasticity may lead to overinflated type I error rates when testing for an interaction effect between this genetic variant and any genetic variant from the other organism. We compare different methods to test and account for the potential heteroscedasticity in the phenotypic trait in the case where the genotype of the pathogen organism is a binary variable.

Secondly, the fact that the phenotypic trait is held fixed while the interacting genotypes vary across different interaction tests in a joint genome-wide association analysis means that the collection of interaction test statistics corresponding to a fixed pathogen genetic variant may often display a tangible departure from the known distribution of the interaction test statistic. Under the global null hypothesis of no interaction, the collection of interaction p-values corresponding to a given pathogen genetic variant might turn out to be consistently smaller than uniform, leading to a phenomenon which has been called the "feast" effect, since we end up with excess false discoveries. Similarly, the collection of interaction p-values corresponding to another fixed pathogen genetic variant might turn out to be consistently larger than uniform, leading to a phenomenon which has been called the "famine" effect, since it limits our ability to make any important discoveries.

This "feast or famine" effect has been shown to result from improper conditioning in the construction of the interaction test statistic in a joint association analysis. The ordinary interaction test statistic conditions on the pair of genetic variants being tested for interaction. Instead, we take the approach

of conditioning on the phenotypic trait and a fixed pathogen genetic variant in order to construct a corrected host-pathogen interaction test statistic which alleviates the feast or famine effect. We focus our efforts on the case of diploid host organisms where an appropriate discrete correction might be required to account for the binomially distributed host genotype. We present a diagnostic tool to predict the prevalence of the feast or famine effect given only the information about a phenotypic trait and a fixed pathogen genetic variant and demonstrate its relationship with the commonly used genomic control inflation factor.

Lastly, accounting for population structure among patients infected with related strains of the same pathogen presents a significant challenge, owing to the presence of genetic variants with differing number of alleles within the pathogen genome. As the number of alleles in a genetic variant increases, some of the alleles may be associated with excessively small observed allele frequencies, which introduce numerical instabilities in the existing methods of constructing a pathogen genetic relatedness matrix (GRM). We build upon previous work to develop a novel pathogen GRM for organisms with multiallelic genetic variants which avoids filtering out genetic variants with exceedingly small observed allele frequencies by introducing an adjusted weighting for rare alleles.

We validate the type I error control and rectification of the feast of famine effect by our correction framework through a host of simulation studies. We demonstrate the applicability of our proposed pathogen GRM and our correction framework by testing for interaction effects between human SNPs and hepatitis C viral genetic variants on pre-treatment viral load in a cohort of HCV infected patients from the BOSON clinical trial.

CHAPTER 1

INTRODUCTION

1.1 Genetics Nomenclature

The **genome** is the collection of all the genetic material of an organism. The genome of most species is divided into multiple **chromosomes**. **Diploid** organisms, such as humans, have 2 copies of each chromosome - one inherited from the father and one inherited from the mother. On the other hand, **haploid** organisms, such as bacteria and viruses, have only 1 copy of each chromosome. Pathogen genomes also typically consist of only a single chromosome, whereas the human genome is organized into 23 pairs of chromosomes - 22 pairs of autosomal chromosomes and one pair of sex chromosomes. Each chromosome is composed of 2 **DNA strands**, which in turn each consist of a sequence of **nucleotides**. A nucleotide contains one of the following 4 **nitrogenous bases**: adenine (A), thymine (T), cytosine (C) or guanine (G). The nitrogenous bases of the 2 DNA strands are bound together according to the following **base pairing rules**: adenine is paired with thymine and cytosine is paired with guanine. The entire human genome comprises over 3 billion nucleotides.

The majority of the human DNA sequence is identical across the population with individuals sharing about 99.5% of their DNA sequence. The remaining segments, which vary among individuals, are called **polymorphic**. Changes in these sequences across individuals in a population are known as **genetic variants**, while the different forms of these sequences are called **alleles**. An individual's **genotype** is the collection of alleles at specific positions in the genome. A common type of genetic variant is the **SNP** (single nucleotide polymorphism): a mutation at a single base position that differentiates individuals. Most SNPs are **biallelic**, meaning they have only two different alleles, which can be arbitrarily labeled as "0" and "1", in the population. For such SNPs, the allele with the lower frequency in the population is called the **minor allele** and its frequency is referred to as the **minor allele frequency** (MAF). Genetic variants are classified as common or rare based on their MAF. Genetic variants are said to be in **linkage disequilibrium** (LD) if certain combinations of alleles occur together more or less often than expected by chance.Give me a list of 10 papers from the last 7 years related to the construction of a genetic relatedness matrix based on genetic variants with more than 2 alleles.

Because diploid organisms have two copies of each autosomal chromosome, possible genotypes at each SNP are "00", "01", "10" or "11". However, since the parental origin is usually not observed, we encode the genotype at each SNP as 0, 1 or 2, depending on whether the individual has 0, 1 or 2 copies of allele "1" at that genetic variant. Genotypes 0 (state "00") and 2 (state "11") are called **homozygous**, meaning they have identical alleles at the SNP, while genotype 1 (state "01" or "10") is called **heterozygous**, indicating different alleles at the SNP. The number of SNPs in an organism can be vast - the human genome consists of roughly 10 million SNPs.

Unlike the human genome, pathogen genomes typically include a **"core"** genome shared by all strains and a **"dispensable"** genome present only in some strains. Accordingly, the genotype at a pathogen SNP can be encoded as "0", "1" or "-" if it has three states with "-" representing the **"deletion"** state. Furthermore, the genome of **single-stranded RNA** viruses consists of just a single sequence of nucleotides containing one of the following 4 **nitrogenous bases**: adenine (A), uracil (U), cytosine (C) or guanine (G). Adding that variation on top of insertion-deletion polymorphisms may lead to viral genetic variants with up to 5 alleles. Finally, a group of 3 consecutive nucleotides in RNA specifies a single **amino acid**, giving rise to the 22 amino acids which are incorporated into proteins. Amino acid genetic variants can display even more than 10 alleles. All things considered, the genotype at a pathogen genetic variant is generally treated as a multilevel categorical variable.

A **phenotypic trait** is a measurable characteristic of an individual, which can be quantitative, e.g. height or weight, or categorical, e.g. presence or absence of a disease. Many phenotypic traits are influenced by both genetic and non-genetic factors. The proportion of phenotypic variance in a population attributable to genetic factors is known as **heritability**. Heritability can be inferred using family-based study designs or mixed-effects regression models.

1.2 Genetic Association Analysis

Genetic association analyses, crucial for uncovering the genetic basis of complex human traits, measure both phenotypes and genotypes (typically SNPs) in a sample of individuals to test for associations between traits and genetic variants. Modern biomedical technologies enable the simul-

taneous examination of millions of genetic variants or all polymorphic sites across the genome. The goal of a genome-wide association study (GWAS) is to identify SNPs in the genome that are, if not directly causal, at least statistically associated with the phenotype of interest. Typically, each SNP is tested independently for association with the phenotype. It has been shown that the power of single-SNP tests using linear or logistic regression increases with the proportion of phenotypic variance explained by the tested SNP [1]. A common approach for testing the association between phenotype (response) and genotype at a SNP (predictor) is regression analysis. Each test generally involves a straightforward statistical method, such as linear, logistic or proportional hazards regression, depending on the type of measurement. These analyses have successfully identified thousands of SNP associations, providing valuable insights into the genetic architecture of complex diseases. Due to the large number of SNPs tested, a stringent significance threshold is used to correct for multiple testing. Additionally, the high interdependence among SNPs and hidden confounding factors among the sampled subjects present significant challenges to traditional statistical methods.

For a typical diploid subject, the genotype at a biallelic SNP is a categorical variable with 3 levels. However, when modeling the effect of a SNP on a trait, additional assumptions are often made to reduce the degrees of freedom. For instance, in an additive trait model, we assume the effect of genotype 2 on the trait is twice that of genotype 1 relative to the baseline effect of genotype 0. This allows the encoded genotype to be treated as a quantitative variable with only 1 additional degree of freedom beyond the intercept in the trait model. Other genotype encodings might be used for recessive or dominant trait models. While different trait models reflect different modes of gene action, the additive trait model is generally preferred because it has reasonably good power even when the true model is dominant or recessive, which is not the case in reverse situations [2].

1.3 Relatedness and Population Structure

In genetic association analyses, a significant statistical challenge arises when sampled individuals are related, leading to dependencies among observations. This relatedness within a sample, termed "population structure", encompasses both known and unknown relationships and is a common confounding factor. When population structure is present, the independence assumption of many standard association testing methods can be violated, compromising their performance and

reliability.

Population structure is prevalent in genetic association data. Many studies include family members with known pedigree relationships, as family-based designs have been popular in traditional genetic research. These family samples are often included in contemporary association analyses because they can increase the power to detect associations by enriching disease-associated SNPs among relatives. Accounting for family structure is crucial to ensure properly controlled type I error rates in association tests and can also improve power when familial correlation is carefully adjusted.

Another source of population structure is latent relatedness among sampled individuals without known family relationships. From an evolutionary perspective, all humans are related to varying degrees through a vast genealogy, though this structure is usually unobserved except in pedigree-based studies. When some individuals in a sample are more closely related than others, this relatedness can confound association tests. Relatedness introduces correlation in observed genotypes and genome-wide variation, potentially producing phenotype correlation. If not accounted for, this can create spurious association signals.

Population stratification is one form of latent relatedness, occurring when a sample includes individuals from different population subgroups. In genetic case-control studies, association tests compare genotype distributions between phenotype groups. If phenotypic and genotypic distributions vary by subpopulation, unassociated SNPs can generate spurious associations without proper correction. Admixed populations, where individuals are genetic mixtures of multiple ancestral populations with varying proportions, can have similar confounding effects.

1.4 A Linear Mixed Model to Account for Population Structure

Linear mixed effects models are particularly effective in addressing various confounding factors, including relatedness and population structure [3]–[6]. Linear mixed models incorporate structure through random effects, whose covariance matrix reflects the dependencies within the sample.

For a sample of n individuals with either known or unknown structure and in the absence of major

genes the trait value for individual i, denoted by Y_i, is modeled as:

$$Y_i = U_i^{\mathrm{T}}\alpha + v_i + \varepsilon_i,$$

where $U_i \in \mathbb{R}^c$ is a vector of covariates including an intercept term, $\alpha \in \mathbb{R}^c$ is a vector of unknown fixed covariate effects, $v = (v_1, v_2, \ldots, v_n)^{\mathrm{T}}$ is a vector of random effects accounting for correlations in phenotype values due to relatedness and ε_i are independent and identically distributed random effects representing environmental influences for $i = 1, 2, \ldots, n$. It is common for linear mixed models to assume that $\mathrm{Var}(v) = \sigma_v^2\Phi$, where $\Phi \in \mathbb{R}^{n\times n}$ is referred to as the kinship matrix.

This model can also be derived from Fisher's polygenic model, which considers the effects of numerous small, equal and additive genetic variants on the phenotype distribution [7]. Specifically, for a set of m independent markers, the phenotype value of individual i is modeled as follows:

$$Y_i = U_i^{\mathrm{T}}\alpha + \sum_{j=1}^{m} w_j \frac{G_{ij} - \mu_j}{\sigma_j} + \varepsilon_i,$$

where G_{ij} is the genotype of individual i at genetic variant j with mean μ_j and standard deviation σ_j and $w_1, w_2, \ldots, w_m$ are independent random effects with $\mathbb{E}(w_j) = 0$ and $\mathrm{Var}(w_j) = \frac{1}{m}\sigma_w^2$ for $j = 1, 2, \ldots, m$. When the number of genetic variants m is large, the distribution of $\sum_{j=1}^{m} w_j \frac{G_{ij}-\mu_j}{\sigma_j}$ can be approximated by a $\mathcal{N}\left(0, \sigma_w^2\Psi\right)$ random variable, where:

$$\Psi = \frac{1}{m}\sum_{j=1}^{m} \frac{(G_j - \mu_j\mathbf{1}_n)(G_j - \mu_j\mathbf{1}_n)^{\mathrm{T}}}{\sigma_j^2}.$$

The matrix Ψ can be thought of as a covariance matrix which measures genetic similarities among individuals in the sample based on whole-genome genotype data.

In practice, the number of independent SNPs is finite, making the exact matrix unknown, so genotype data from numerous genetic variants can be used to empirically estimate the matrix. For example, a commonly used estimate of Ψ based on genome-wide genotype data from a diploid organism is given by:

$$K = \frac{1}{m}\sum_{j=1}^{m} \frac{\left(G_j - 2\widehat{f}_j\mathbf{1}_n\right)\left(G_j - 2\widehat{f}_j\mathbf{1}_n\right)^{\mathrm{T}}}{2\widehat{f}_j\left(1 - \widehat{f}_j\right)},$$

where $\widehat{f}_j = \frac{1}{2}\overline{G}_j$ is an unbiased estimator of the population allele frequency f_j at genetic variant j and we assume that the 2 alleles of individual i are independent draws from the same distribution, so that $\mathrm{Var}\left(G_{ij}\right) = 2f_j\left(1 - f_j\right)$.

Another commonly used estimate of Ψ based on genome-wide genotype data from a haploid organism is given by:

$$K = \frac{1}{m}\sum_{j=1}^{m}\frac{\left(G_j - \widehat{f}_j\mathbf{1}_n\right)\left(G_j - \widehat{f}_j\mathbf{1}_n\right)^{\mathrm{T}}}{\widehat{f}_j\left(1 - \widehat{f}_j\right)},$$

where $\widehat{f}_j = \overline{G}_j$ is an unbiased estimator of the population allele frequency f_j at genetic variant j. Sample dependence can also be modeled using fixed effects, e.g. ancestry-informative covariates, or a combination of fixed and random effects.

CHAPTER 2

CORRECTION FRAMEWORK FOR INTERACTION TEST STATISTIC IN ASSOCIATION ANALYSIS

2.1 Introduction

Previous studies have shown that the effect of a genetic variant on a phenotypic trait may depend on genetic variants of another organism, such as host-pathogen interactions in infectious diseases [8], [9], or on genetic variants of the same organism, i.e. epistatic effects [10]–[12]. The genetic effect can alternatively depend on various environmental factors, such as age, sex, lifestyle or other exposures [13]. Detecting these interaction effects can enhance the identification of genetic effects that might otherwise be diminished or obscured [14]. They are often cited as reasons for the difficulty in replicating results from marginal association studies [15], contribute significantly to missing heritability [16], [17] and improve understanding of the genetic architecture of complex traits and diseases [18], [19].

Despite the widespread popularity of genome-wide association analyses, simultaneous association mapping between interacting species has been infrequent [20]–[23]. Integrating genomes from two organisms into a GWAS could potentially reveal genomic regions indicative of co-evolution between species. Pathosystems, where pathogens and hosts co-evolve to determine disease status, offer pertinent examples [24]–[27]. A lot of existing genome-wide association analysis methods [28], [29] typically focus solely on host genomes [30] or stratify by pathogen strain [31], neglecting the pathogen genome [32]. Other common strategies to examine the relationship between the host and the pathogen genome include mapping one genome onto the other in a GWAS setting [9], [33]–[37], performing a trans-species expression quantitative trait locus (ts-eQTL) analysis [38], [39] or constructing a host-pathogen protein interaction network [40], [41], all of which ignore how associations between host and pathogen genotypes affect any infectious disease trait. As sequencing technology advances, both host and pathogen genomes are increasingly accessible. Identifying genetic associations across both genomes could illuminate the genetic underpinnings of host-pathogen specificity and enhance our understanding of their molecular interactions. Therefore, integrating statistical

methods that encompass genomes from interacting species into genome-wide association analyses stands to yield significant advancements.

Some efforts have been made to develop methods for detecting interactions in genome-wide association analyses, focusing on improving computational efficiency, reducing false positives and increasing power [42]–[47]. Previous studies have shown that replicating interactions in a GWAS can be difficult [48]–[50]. Another critical issue is heteroscedasticity which can manifest in the phenotypic trait if a true interaction effect exists between one of the genetic variants being tested and another genetic variant not accounted for in the model. If not properly addressed, this heteroscedasticity can lead to overinflation of type I error rates [51]–[53]. Several approaches exist to account for heteroscedasticity in the response variable of a linear regression model such as weighted least squares (WLS) regression or the use a heteroscedasticity-consistent (HC) estimator for the covariance matrix of the least squares estimator [54]. We propose the use of an iteratively reweighted least squares (IRLS) method which takes into account the special structure in the heteroscedasticity of the quantitative trait as a function of the binary pathogen genetic variant Z in order to maximize power while maintaining type I error control.

Systematically inflated or deflated p-values have been reported when testing for interactions in GWAS settings based on both real and simulated data sets. Even under simplified conditions, where the quantitative trait is simulated under the global null hypothesis of no interaction and with no unaccounted covariates, relatedness or sample structure, type I error rates and genomic control inflation factors have been shown to be highly variable across different simulation replicates [52], [53], [55]. Under the global null hypothesis of no interaction, the collection of interaction p-values corresponding to a given pathogen genetic variant might turn out to be consistently smaller than uniform, leading to a phenomenon which has been called the "feast" effect, since we end up with excess false discoveries. Similarly, the collection of interaction p-values corresponding to another fixed pathogen genetic variant might turn out to be consistently larger than uniform, leading to a phenomenon which has been called the "famine" effect, since it limits our ability to make any important discoveries. This phenomenon has been shown to arise from the choice of variables to condition on in the construction of the interaction test statistic, and it has been demonstrated that modifying this conditioning choice can potentially resolve this issue [12]. We apply these findings to

develop methods for correcting the feast or famine effect in the context of testing for host-pathogen interaction effects. We present a diagnostic tool to predict the prevalence of the feast or famine effect given only the information about a phenotypic trait and a fixed pathogen genetic variant and demonstrate its relationship with the commonly used genomic control inflation factor.

Integrating information from a pathogen genome into a joint association analysis of an infectious disease phenotypic trait usually entails special considerations with respect to the modeling of the pathogen genetic variants. Contrary to host genetic variants which are commonly biallelic, there are several mechanisms underlying the pathogen genome which can cause genetic variants to display more than 2 alleles. Pathogen genomes often display a large number of insertion-deletion polymorphisms on top of the usual mutation polymorhpisms, which may lead to multiple triallelic pathogen genetic variants. Additionally, single-stranded RNA viral genetic variants consist of a single nucleobase which can be one of adenine (A), uracil (U), cytosine (C) or guanine (G). Adding that variation, on top of insertion-deletion polymorphisms, may lead to viral genetic variants with up to 5 alleles. Finally, a group of 3 consecutive nucleotides in RNA specifies a single amino acid, giving rise to the 22 amino acids which are incorporated into proteins. Amino acid genetic variants can display even more than 10 alleles. Triallelic bacterial genetic variants have previously been handled by including 2 binary indicators for the mutation and deletion polymorphisms into the joint association analysis model and performing score tests for interaction with 2 degrees of freedom [8]. However, handling multiallelic genetic variants with more than 3 alleles in the same way would likely result in test statistics with no substantial power. Our current approach consists of transforming multiallelic pathogen genetic variants into a set of binary allele indicators and testing for interactions between each individual allele indicator and every host genetic variant.

2.2 Notation and Framework for Joint Assocation Analysis

Let $Y \in \mathbb{R}^n$ denote the vector of values for a quantitative trait measured on a sample of n individuals, $X \in \mathbb{R}^n$ the genotype vector of a host genetic variant, $Z \in \mathbb{R}^n$ denote the genotype vector of a pathogen genetic variant and $U \in \mathbb{R}^{n \times c}$ denote a matrix of c covariates - including a column corresponding to the intercept. Similarly, X and Z could denote the genotypes of 2 different genetic variants belonging to the same organism or X could denote the genotype of a genetic

variant and Z could denote the vector of values for an environmental variable. In what follows, we are mostly going to focus on the case where Z is a binary variable, e.g. it represents the genotype of a genetic variant belonging to a haploid organism or an environmental factor such as sex.

For two collections of genotypes $X_1, X_2, \ldots, X_{m_h} \in \mathbb{R}^n$ of m_h host genetic variants and genotypes $Z_1, Z_2, \ldots, Z_{m_p} \in \mathbb{R}^n$ of m_p pathogen genetic variants, let $\mathbb{E}(X_{ij}) = \mu_{X_j}$ and $\mathbb{E}(Z_{ik}) = \mu_{Z_k}$ for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m_h$ and $k = 1, 2, \ldots, m_p$. In a joint association analysis of the quantitative trait Y, we are interested in drawing inferences from the following model:

$$Y_i = U_i^{\mathrm{T}}\alpha + \beta X_{ij} + \gamma Z_{ik} + \delta_{jk}\left(X_{ij} - \mu_{X_j}\right)\left(Z_{ik} - \mu_{Z_k}\right) + \varepsilon_i, \tag{2.1}$$

where $U_i \in \mathbb{R}^c$ is the i-th row of the covariate matrix U and ε_i are independent with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$. After observing the genotype data, our first step is to estimate μ_{X_j} by $\overline{X}_j$ and μ_{Z_k} by $\overline{Z}_k$. Then, we use ordinary least squares (OLS) to fit the following linear regression models:

$$Y = U\alpha + \beta X_j + \gamma Z_k + \delta W_{jk} + \varepsilon, \tag{2.2}$$

where $W_{jk} = \mathrm{diag}(HZ_k)HX_j \in \mathbb{R}^n$ is the interaction term between the genetic variants X_k and Z_k, $H = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ is the projection matrix projecting any vector to its residuals after regressing it on the intercept term and $\mathrm{diag}(HZ_k) \in \mathbb{R}^{n\times n}$ is the diagonal matrix whose diagonal elements are given by the elements of the vector $HZ_k \in \mathbb{R}^n$. Finally, we are interested in performing the hypothesis tests $H_0^{jk} : \delta_{jk} = 0$ vs. $H_1^{jk} : \delta_{jk} \neq 0$, i.e. testing whether there exists any significant interaction effect between each possible pair of a host and pathogen genetic variant on the phenotypic trait. The usual test statistic that one would employ in this scenario of a quantitative trait with no relatedness or population structure among sampled individuals, would be the following interaction t test statistic:

$$t_{jk} = \frac{\widehat{\delta}_{jk}}{S_\varepsilon\sqrt{\left[\left(U_{XZW}^{\mathrm{T}}U_{XZW}\right)^{-1}\right]_{c+3,c+3}}}, \tag{2.3}$$

where $\widehat{\delta}_{jk}$ is the least squares estimator of the interaction coefficient δ_{jk}, $S_\varepsilon^2 = \frac{Y^{\mathrm{T}}P_{XZW}Y}{n-c-3}$ is an unbiased estimator of the residual variance σ_ε^2, $U_{XZW} \in \mathbb{R}^{n\times(c+3)}$ is the design matrix corresponding to model 2.2 and $P_{XZW} = \mathbf{I}_n - U_{XZW}\left(U_{XZW}^{\mathrm{T}}U_{XZW}\right)^{-1}U_{XZW}^{\mathrm{T}}$ is the projection matrix projecting

a vector to its residuals after regressing it on U, X_j, Z_k and W_{jk}. Under the null hypothesis of no interaction effect and the additional assumption that $\varepsilon \sim \mathcal{N}_n\left(0, \sigma_\varepsilon^2 \mathbf{I}_n\right)$, we normally know that $t_{jk} \sim t_{n-c-3}$. More generally, t_{jk} constitutes a Wald test statistic and would asymptotically follow the standard normal distribution under the null hypothesis of no interaction, even without the additional assumption of normality of the error terms. Alternatively, one could choose to perform an asymptotic score or likelihood ratio test for interaction and that test statistic would also be subject to the same phenomena as the ones we are going to discuss in the following sections [12], but we are going to focus on the interaction t test statistic for the purposes of this dissertation. Note that the interaction test statistic would remain unchanged if the interaction term was substituted by the simpler expression $X_{ij}Z_{ik}$, but constructing the interaction term as $(X_{ij} - \mu_X)(Z_{ik} - \mu_Z)$ reduces collinearity and leads to more interpretable parameter estimates.

Without loss of generality, we might choose to conduct a joint GWAS by fixing a pathogen genetic variant Z_1, testing for interactions with every host genetic variant $X_1, X_2, \ldots, X_{m_h}$ and repeating for every other pathogen genetic variant $Z_2, Z_3, \ldots, Z_{m_p}$. In this setting, we might also be interested in performing a suitable global testing procedure [56]–[64] to determine whether there exists at least one host genetic variant interacting with any fixed pathogen genetic variant Z_k for $k = 1, 2, \ldots, m_v$ - in a similar vein to the idea of testing for marginal epistasis [10], [11].

2.3 Heteroscedasticity Due to Latent Interaction Effect

Let $\widetilde{X} \in \mathbb{R}^n$ denote the genotype of an unobserved genetic variant or environmental factor sampled on n individuals with $\mathbb{E}\left(\widetilde{X}_i\right) = \mu_X$ and $\mathrm{Var}\left(\widetilde{X}_i\right) = \sigma_X^2$ for $i = 1, 2, \ldots, n$. Suppose that the true model underlying the quantitative trait $Y \in \mathbb{R}^n$ is the following:

$$Y_i = U_i^{\mathrm{T}}\widetilde{\alpha} + \widetilde{\beta}\widetilde{X}_i + \widetilde{\gamma}Z_i + \widetilde{\delta}\left(\widetilde{X}_i - \mu_X\right)(Z_i - \mu_Z) + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}\left(0, \widetilde{\sigma}_\varepsilon^2\right)$ are independent. Then, we observe that:

$$\begin{aligned} \mathrm{Var}\left(Y_i \mid U_i, Z_i\right) &= \left[\widetilde{\beta} + \widetilde{\delta}\left(Z_i - \mu_Z\right)\right]^2 \sigma_X^2 + \widetilde{\sigma}_\varepsilon^2 \\ &= \left[2\widetilde{\beta} + \widetilde{\delta}\left(1 - 2\mu_Z\right)\right]\widetilde{\delta}\sigma_X^2 Z_i + \left(\widetilde{\beta} - \widetilde{\delta}\mu_Z\right)^2 \sigma_X^2 + \widetilde{\sigma}_\varepsilon^2. \end{aligned} \tag{2.4}$$

In other words, the quantitative trait Y is conditionally heteroscedastic given the binary genetic variant Z. Discounting this heteroscedasticity structure in the response variable when conducting the interaction tests between the pathogen genetic variant Z and all available host genetic variants $X_1, X_2, \ldots, X_{m_h}$ would often lead to excess type I error.

Several methods already exist to tackle the presence of heteroscedasticity in the response variable of a linear regression model. Most notably, the use of heteroscedasticity-consistent (HC) estimators for the covariance matrix of the least squares estimator has been shown to be quite effective in the case of continuous covariates [54], [65]. This approach consists of first fitting the homoscedastic linear regression models 2.2 by ordinary least squares and computing the residuals $\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \ldots, \widehat{\varepsilon}_n$. Let $U_{XZW} \in \mathbb{R}^{n\times(c+3)}$ denote the design matrix corresponding to model 2.2. Then, one makes the assumption that the random error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ in 2.2 are actually heteroscedastic and estimates their covariance structure via one of the following diagonal covariance matrices:

$$\widehat{\Phi}_0 = \operatorname{diag}\left\{\widehat{\varepsilon}_i^2\right\}, \quad \widehat{\Phi}_1 = \frac{n}{n-c-3}\widehat{\Phi}_0, \quad \widehat{\Phi}_2 = \operatorname{diag}\left\{\frac{\widehat{\varepsilon}_i^2}{1-L_{ii}}\right\}, \quad \widehat{\Phi}_3 = \operatorname{diag}\left\{\frac{\widehat{\varepsilon}_i^2}{(1-L_{ii})^2}\right\},$$

where L_{ii} is the leverage of individual i, i.e. the i-th diagonal element of the hat (projection) matrix $L = U_{XZW}\left(U_{XZW}^{\mathrm{T}}U_{XZW}\right)^{-1}U_{XZW}^{\mathrm{T}}$ corresponding to model 2.2. Thus, a heteroscedasticity-consistent estimator for the covariance matrix of the least squares estimator $\widehat{\vartheta}$ of the regression coefficient vector $\vartheta = (\alpha, \beta, \gamma, \delta)^{\mathrm{T}} \in \mathbb{R}^{c+3}$ is given by:

$$\widehat{\operatorname{Var}}\left(\widehat{\vartheta}\right) = \left(U_{XZW}^{\mathrm{T}}U_{XZW}\right)^{-1}U_{XZW}^{\mathrm{T}}\widehat{\Phi}_\ell U_{XZW}\left(U_{XZW}^{\mathrm{T}}U_{XZW}\right)^{-1},$$

where $\ell \in \{0, 1, 2, 3\}$. The $\widehat{\Phi}_3$ estimator has been shown to have better overall performance compared to the other mentioned HC estimators [54]. However, this approach has been shown to perform rather poorly in the case where the covariates in the linear regression model are genotypes, i.e. categorical variables, with fairly small minor allele frequencies [52].

Another popular approach for handling heteroscedasticity in the response variable of a linear regression model is weighted least squares (WLS) regression. However, this approach highly depends on accurate specification of the regression weights and is best suited only for scenarios where some

information about the heteroscedasticity structure in the response variable is known a priori. Luckily enough, in the special case of model 2.1 with binary pathogen genetic variant Z, we have already derived formula 2.4 for the potential conditional variance of the quantitative trait given Z. Hence, it would make more sense to exploit that heteroscedasticity structure instead of using the residuals of the fitted homoscedastic model to estimate it.

Suppose that $Y_i = U_i^{\mathrm{T}}\alpha + \gamma Z_i + \varepsilon_i$ with $\varepsilon_i \mid Z_i \sim \mathcal{N}\left(0, \sigma^2_{Z_i}\right)$ independent for $i = 1, 2, \ldots, n$. Let $n_0 = \sum_{i=1}^{n} \mathbb{1}_{\{z_i=0\}}$, $n_1 = \sum_{i=1}^{n} \mathbb{1}_{\{z_i=1\}}$ and $r_{Y|U,Z}$ denote the vector of residuals of the quantitative trait Y after we regress it on the matrix of covariates U and the binary pathogen genetic variant Z. Then, we define the following conditional variance estimators of Y given that $Z = 0$ and $Z = 1$ respectively:

$$V_0 = \frac{1}{\frac{n-c+1}{n} n_0 - 1} \sum_{i:z_i=0} \left(r^i_{Y|U,Z}\right)^2, \quad V_1 = \frac{1}{\frac{n-c+1}{n} n_1 - 1} \sum_{i:z_i=1} \left(r^i_{Y|U,Z}\right)^2. \tag{2.5}$$

We observe that the total degrees of freedom shared across the two variance estimators are equal to $n - c - 1$, which are exactly the residual degrees of freedom of the fitted linear regression model. However, the $n - c + 1$ degrees of freedom corresponding to the estimation of the coefficients of the $c - 1$ covariates - excluding the intercept and the genetic variant Z - are distributed across the 2 variance estimators proportionally to the number of observations in each of the 2 groups defined by the binary genetic variant Z, since information from both groups of observations is aggregated towards the goal of estimating them. On the other hand, only the first group of observations essentially contributes to the estimation of the intercept coefficient, whereas the second group contributes to the estimation of the coefficient of the genetic variant Z. For $i = 1, 2, \ldots, n$, we define the following regression weights:

$$w_i = \begin{cases} V_0^{-1}, & Z_i = 0 \\ V_1^{-1}, & Z_i = 1 \end{cases}.$$

For $j = 1, 2, \ldots, m_h$, we fit the weighted linear regression models corresponding to 2.2 and calculate new residuals vectors $r_{Y|U,X_1,Z}, r_{Y|U,X_2,Z}, \ldots, r_{Y|U,X_{m_h},Z}$. Since the regression weights are not a priori known and we have used a rough estimate of them based on the observed data, it would be

good practice to iterate this procedure until convergence of the weighted least squares estimator. Hence, we update our conditional variance estimators of Y given U, X_j and Z as follows:

$$V_0^j = \frac{1}{\frac{n-c+1}{n} n_0 - 2} \sum_{i:z_i=0} \left(r^i_{Y|U,X_j,Z} \right)^2, \quad V_1^j = \frac{1}{\frac{n-c+1}{n} n_1 - 2} \sum_{i:z_i=1} \left(r^i_{Y|U,X_j,Z} \right)^2,$$

leading to corresponding updates of our regression weights. Owing to the special covariate and weight structure of this model, it can be proven that the estimates of the regression coefficients are going to remain unchanged after refitting the weighted least squares models using the updated weights. Thus, this iteratively reweighted least squares (IRLS) procedure always converges after just 1 step. We show through multiple simulation studies in Section 2.8 that this IRLS procedure maintains type I error control in the presence of heteroscedasticity in the quantitative trait, consistently higher power than the use of HC covariance matrices for the least squares estimator and does not result in any significant loss in power compared to the ordinary interaction t test.

Even though it is strongly discouraged in the literature to perform heteroscedasticity tests in order to decide whether or not to make use of a suitable heteroscedasticity-consistent method for a specific data set [54], performing heteroscedasticity tests in the context of a joint association analysis might still provide some information about the presence of some interaction effect between the pathogen genetic variant currently being tested and some other observed or unobserved factor. In the case where $U = \mathbf{1}_n$, i.e. in the absence of covariates, a simple F test of equality of variances might be employed to test for the conditional heteroscedasticity of the quantitative trait Y given a pathogen genetic variant Z. In the more general case and under the null hypothesis of homoscedasticity, one might choose to construct the following approximate generalized F test statistic of equality of variances:

$$F' = \frac{V_0}{V_1} \sim F_{\frac{n-c+1}{n} n_0 - 1, \frac{n-c+1}{n} n_1 - 1},$$

where V_0 and V_1 are given by equation 2.5. The validity of this test statistic under the null hypothesis has been verified through simulations and its power is asymptotically comparable to other heteroscedasticity tests such as a likelihood ratio test of equality of variances, the Breusch-Pagan test [66], [67] and the Goldfeld-Quandt test [68].

2.4 The Feast or Famine Effect

Under the null hypothesis of no interaction effect between the host genetic variant X_j and the pathogen genetic variant Z_k, we know that the interaction t test statistic given by equation 2.3 individually follows Student's t distribution with $n-c-3$ degrees of freedom. It has been shown that the collection of interaction t test statistics $t_{1k}, t_{2k}, \ldots, t_{m_h k}$ corresponding to a fixed quantitative trait Y and a fixed pathogen genetic variant Z_k is generally **not** going to be t distributed under the global null hypothesis of no interaction [12]. This happens because this distributional result hinges on the assumption that the response variable Y is random given fixed genetic variants X_j and Z_k, whereas Y and Z_k are held fixed and the host genetic variant X_j is allowed to vary across different interaction tests in this joint association analysis setting. In other words, a set of interaction test statistics between fixed genetic variants X_j and Z_k on a set of quantitative traits $Y_1, Y_2, \ldots, Y_{m_t}$ would naturally be t distributed under the global null hypothesis, but this result breaks down as soon as Y is fixed across different interaction tests instead of X_j.

The ordinary interaction t test statistic is constructed by first subtracting the conditional expectation of $\widehat{\delta}$ given fixed genetic variants X_j and Z_k under the null hypothesis, which is equal to 0 for the test of statistical significance of the interaction effect, so that the test statistic is centered around 0 and is correctly calibrated under the null hypothesis. Then, it is divided by the estimated standard error of $\widehat{\delta}$ given fixed X_j and Z_k under the alternative hypothesis, so that the overall variance of the test statistic is asymptotically equal to 1 and it is easy to compare its observed value against the quantiles of the standardized t distribution in order to draw inferences. As soon as X_j is allowed to vary across different interaction tests in place of Y, it is easy to imagine that the overall sample mean and sample variance of the collection of interaction t test statistics corresponding to fixed pathogen genetic variant Z_k and quantitative trait Y would no longer necessarily be the desired ones, which implies that the collection of test statistics may no longer be assumed to be t distributed.

For a fixed quantitative trait Y and 3 candidate pathogen genetic variants Z_1, Z_2, Z_3, we might potentially observe under the global null hypothesis of no interaction that the sample variance of the collection of interaction t test statistics is significantly larger than 1 for Z_1, significantly

smaller than 1 for Z_2 and approximately equal to 1 for Z_3. We need to highlight that this is not an artifact of potential dependence among the interaction tests being performed [69], [70]. We can keep simulating different sets of independent host genetic variants $X_1, X_2, \ldots, X_{m_h}$ and the sample variances of the collections of interaction test statistics corresponding to Z_1, Z_2, Z_3 are going to consistently turn out in the same direction, even though the test statistics within each collection are independent given fixed Y and Z [12]. As a result, we have an overabundance of small interaction p-values corresponding to Z_1, which we refer to as the "feast" effect, since we end up making many more discoveries than what we would expect under the global null hypothesis, resulting in an overinflated type I error rate. Similarly, we have a severe lack of smaller interaction p-values corresponding to Z_2, which we refer to as the "famine" effect, since our ability to make any important discoveries would be grossly limited, leading to a significant loss in power. On the other hand, the interaction p-values corresponding to Z_3 are indistinguishable from a set of uniformly distributed variates, meaning that our interaction test statistics are actually correctly calibrated in this case. We demonstrate this phenomenon in more detail via multiple simulation studies throughout Section 2.8.

Even though the overall type I error across different collections of interaction test statistics corresponding to different pathogen genetic variants is going to be well controlled for, the variation of type I error rates and genomic control inflation factors within each GWAS is probably going to be staggering - much higher than what one would expect to see for a collection of marginal association analyses. This "feast or famine effect" can possibly lead to a multitude of false discoveries and complete failure to replicate previously identified interaction effects in a joint association analysis, since the feast or famine effect completely depends on the pair of trait vector Y and pathogen genotype vector Z that you happened to have observed in your obtained sample. It has previously been shown that this phenomenon is a result of inappropriate conditioning in the construction of the interaction test statistic and that an appropriate approach for correcting it would be to calculate the conditional expectation and variance of the numerator of the interaction test statistic with respect to a random host genetic variant X given fixed quantitative trait Y and pathogen genetic variant Z, in order to properly standardize the interaction test statistic in a joint association analysis [12]. We refer to this procedure as interaction test statistic correction. We develop this

correction framework specifically in the context of testing for host-pathogen interaction effects and present more details about it throughout the following sections.

It is important to understand why the "feast or famine" effect only appears when testing for interactions in a joint GWAS setting but never in any ordinary marginal GWAS, even though the phenotypic trait is also held fixed while the genetic variant is allowed to vary across different association tests [12]. In order to gain a better understanding of this distinction, it is of vital importance to take a closer look at the expression of the test statistic in both situations. In Section 2.10.1, we explicitly derive the following useful expression for the interaction t test statistic corresponding to the joint association model 2.2:

$$t_{jk} = \frac{W_{jk}^{\mathrm{T}} P_{XZ} Y}{\sqrt{\frac{W_{jk}^{\mathrm{T}} P_{XZ} W_{jk} Y^{\mathrm{T}} P_{XZ} Y - \left(W_{jk}^{\mathrm{T}} P_{XZ} Y\right)^2}{n-c-3}}}, \tag{2.6}$$

where $U_{XZ} = [U, X_j, Z_k] \in \mathbb{R}^{n \times (c+2)}$ is the design matrix corresponding to the covariates and the additive effects of the 2 genetic variants, $P_{XZ} = \mathbf{I}_n - U_{XZ} \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} U_{XZ}^{\mathrm{T}}$ is the projection matrix projecting any vector to its residuals after regressing it on U, X and Z, $H = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}$ is the projection matrix projecting any vector to its residuals after regressing it on the intercept term, $W_{jk} = \mathrm{diag}\left(HZ_k\right) HX_j \in \mathbb{R}^n$ is the interaction term between the 2 genetic variants and $\mathrm{diag}\left(HZ_k\right) \in \mathbb{R}^{n \times n}$ denotes the diagonal matrix whose diagonal elements are given by the elements of the vector $HZ_k \in \mathbb{R}^n$. Now, consider the marginal association analysis model $Y = U\alpha + \beta_j X_j + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n\left(0, \sigma_\varepsilon^2 \mathbf{I}_n\right)$ for $j = 1, 2, \ldots, m_h$. The usual t test statistic for the additive effect of the genetic variant X_j on the phenotypic trait Y can similarly be written as follows:

$$t_j = \frac{X_j^{\mathrm{T}} P Y}{\sqrt{\frac{X_j^{\mathrm{T}} P X_j Y^{\mathrm{T}} P Y - \left(X_j^{\mathrm{T}} P Y\right)^2}{n-c-1}}},$$

where $P = \mathbf{I}_n - U \left(U^{\mathrm{T}} U\right)^{-1} U^{\mathrm{T}}$ is the complementary projection matrix corresponding to just the covariate matrix U. We observe that the association test statistic treats the genetic variant X_j and the quantitative trait Y symmetrically given fixed covariate matrix U, i.e. swapping X_j with Y would yield exactly the same test statistic. Therefore, holding Y and allowing X_j to vary across

marginal association tests, even though the association model assumes that Y is random given fixed X_j, does not cause problems because we could equivalently rewrite the association model as $X_j = U\alpha + \beta_j Y + \varepsilon$ and we would end up with exactly the same test statistic. Unfortunately, this symmetry does not extend to the joint association model, since both the fitted interaction term W_{jk} and the projection matrix P_{XZ} in equation 2.6 explicitly depend on the genetic variants X_j and Z_k, whereas the phenotypic trait Y does not, leading to the conditioning issue we have already discussed.

2.5 Interaction Test Statistic Correction

Let $N = W^{\mathrm{T}} P_{XZ} Y$ denote the numerator of the interaction t test statistic appearing in equation 2.6. Our goal is to construct a new interaction test statistic, which is correctly calibrated given fixed quantitative trait Y and pathogen genetic variant Z, around this numerator. In order to achieve this, we need to be able to calculate the conditional expectation and variance of N with respect to a random host genetic variant X given fixed quantitative trait Y and pathogen genetic variant Z. In order to simplify this procedure, it is important to first take an asymptotic approximation of this numerator given Y and Z.

For any vectors $A, B \in \mathbb{R}^n$, let $S'_{AB} = A^{\mathrm{T}} P B \in \mathbb{R}$, where $P = \mathbf{I}_n - U\left(U^{\mathrm{T}}U\right)^{-1} U^{\mathrm{T}}$ is the projection matrix projecting any vector to its residuals after regressing it on the covariate matrix U. For $i, j, k, \ell = 1, 2, \ldots, n$, we also define the following conditional moments of X given Y and Z:

$$\mu_{X|Y,Z} = \mathbb{E}\left(X \mid Y, Z\right), \quad \Sigma_{X|Y,Z} = \mathrm{Var}\left(X \mid Y, Z\right),$$

$$\gamma^{ijk}_{X|Y,Z} = \mathbb{E}\left[\left(X_i - \mu^i_{X|Y,Z}\right)\left(X_j - \mu^j_{X|Y,Z}\right)\left(X_k - \mu^k_{X|Y,Z}\right)\right],$$
$$K^{ijk\ell}_{X|Y,Z} = \mathbb{E}\left[\left(X_i - \mu^i_{X|Y,Z}\right)\left(X_j - \mu^j_{X|Y,Z}\right)\left(X_k - \mu^k_{X|Y,Z}\right)\left(X_\ell - \mu^\ell_{X|Y,Z}\right)\right].$$

First, we observe that:

$$N = S'_{YW} - \underbrace{\frac{S'_{ZZ}S'_{XY} - S'_{XZ}S'_{YZ}}{S'_{XX}S'_{ZZ} - S'^2_{XZ}}}_{\widehat{\beta}_0} S'_{XW} - \underbrace{\frac{S'_{XX}S'_{YZ} - S'_{XZ}S'_{XY}}{S'_{XX}S'_{ZZ} - S'^2_{XZ}}}_{\widehat{\gamma}_0} S'_{ZW}.$$

Under the null hypothesis of no interaction effect, we consider the following strong law of large numbers (SLLN) approximations of the least squares estimators $\widehat{\beta}_0$ and $\widehat{\gamma}_0$ of the additive effects of the genetic variants X and Z respectively:

$$\beta_* = \frac{S'_{ZZ}S'_{\mu_{X|Y,Z}Y} - S'_{\mu_{X|Y,Z}Z}S'_{YZ}}{\left[S'_{\mu_{X|Y,Z}\mu_{X|Y,Z}} + \operatorname{tr}\left(P\Sigma_{X|Y,Z}\right)\right]S'_{ZZ} - S'^2_{\mu_{X|Y,Z}Z}},$$

$$\gamma_* = \frac{\left[S'_{\mu_{X|Y,Z}\mu_{X|Y,Z}} + \operatorname{tr}\left(P\Sigma_{X|Y,Z}\right)\right]S'_{YZ} - S'_{\mu_{X|Y,Z}Z}S'_{\mu_{X|Y,Z}Y}}{\left[S'_{\mu_{X|Y,Z}\mu_{X|Y,Z}} + \operatorname{tr}\left(P\Sigma_{X|Y,Z}\right)\right]S'_{ZZ} - S'^2_{\mu_{X|Y,Z}Z}}.$$

Note that these asymptotic approximations are taken with respect to X given fixed Y and Z. Lastly, let $r = Y - \gamma_* Z \in \mathbb{R}^n$, $Q = H\operatorname{diag}(HZ)P \in \mathbb{R}^{n\times n}$ and $Q_S = \frac{Q+Q^{\mathrm{T}}}{2}$. In the case where $U = \mathbf{1}_n$, i.e. in the absence of covariates in the joint association model given by 2.1, note that $P = H$, so it follows that $Q = Q^{\mathrm{T}} = Q_S = H\operatorname{diag}(HZ)H$. Then, an asymptotic approximation of the numerator of the interaction t test statistic is given by:

$$N_* = S'_{YW} - \beta_* S'_{XW} - \gamma_* S'_{ZW} = r^{\mathrm{T}}Q^{\mathrm{T}}X - \beta_* X^{\mathrm{T}}Q_S X, \tag{2.7}$$

where $\operatorname{diag}(HZ) \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix whose diagonal elements are given by the elements of the vector $HZ \in \mathbb{R}^n$. We observe that this asymptotic approximation of the numerator of the interaction test statistic can be written as the sum of a linear function and a quadratic form of the random host genetic variant X. By applying known properties for the expectation and the covariance matrix of linear functions and quadratic forms of random vectors, we get the following conditional moment results:

$$\mathbb{E}\left(N_* \mid Y, Z\right) = r^{\mathrm{T}}Q^{\mathrm{T}}\mu_{X|Y,Z} - \beta_*\left[\mu^{\mathrm{T}}_{X|Y,Z}Q\mu_{X|Y,Z} + \operatorname{tr}\left(Q_S\Sigma_{X|Y,Z}\right)\right], \tag{2.8}$$

$$\begin{aligned}\operatorname{Var}\left(N_* \mid Y, Z\right) = {} & r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Qr - 2\beta_*\sum_{i,j,k=1}^{n} Q_S^{ij}\left(Qr\right)^k\gamma^{ijk}_{X|Y,Z} - 4\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} \\ & + \beta_*^2\sum_{i,j,k,\ell=1}^{n} Q_S^{ij}Q_S^{k\ell}\left(K^{ijk\ell}_{X|Y,Z} + 4\gamma^{ijk}_{X|Y,Z}\mu^{\ell}_{X|Y,Z}\right) \\ & + 4\beta_*^2\mu^{\mathrm{T}}_{X|Y,Z}Q_S\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} - \beta_*^2\left[\operatorname{tr}\left(Q_S\Sigma_{X|Y,Z}\right)\right]^2.\end{aligned} \tag{2.9}$$

Therefore, we define the following corrected interaction t test statistic:

$$T_* = \frac{N - \widehat{\mathbb{E}}\left(N_* \mid Y, Z\right)}{\sqrt{\widehat{\text{Var}}\left(N_* \mid Y, Z\right)}}, \tag{2.10}$$

where the estimation of all the unknown parameters in $\mathbb{E}\left(N_* \mid Y, Z\right)$ and $\text{Var}\left(N_* \mid Y, Z\right)$ pertaining to the conditional distribution of X given Y and Z are discussed in the following section. Under the null hypothesis of no interaction effect, this corrected interaction test statistic is going to asymptotically follow the standard normal distribution. Detailed derivations of these asymptotic approximations and conditional moment calculations are presented in Section 2.10.1. We observe that the conditional expectation of the numerator of the interaction t test statistic given Y and Z is just going to be equal to 0 in the special case where the random host genetic variant X is independent of both the quantitative trait Y and the pathogen genetic variant Z. However, it is not necessarily going be to equal to 0 in the general case. On the other hand, the conditional variance of the numerator given Y and Z is going to highly depend on the induced conditional distribution of X given Y and Z.

As is evident from these formulas, the calculation of the corrected interaction t test statistic hinges on the derivation of the required conditional moments of X given Y and Z. These conditional moments can easily be derived through use of Bayes' theorem, but some additional assumption about the conditional distribution of X_i given Z_i for $i = 1, 2 \ldots, n$ is additionally required. Namely, the conditional distribution of X_i given Y_i and Z_i is given by:

$$f_{X_i|Y_i,Z_i}(x \mid y, z) \propto f_{X_i|Z_i}(x \mid z) f_{Y_i|X_i,Z_i}(y \mid x, z),$$

where the conditional density $f_{Y_i|X_i,Z_i}(y \mid x, z)$ is governed by the joint association model given in equation 2.1. In the following subsections, we discuss 2 particularly useful choices for the conditional distribution for X_i given Z_i in the context of joint association analysis.

2.5.1 Gaussian Correction

The Gaussian correction framework [12] makes the assumption that $\left(X_i \mid Z_i = z\right) \sim \mathcal{N}\left(\mu_{X|z}, \sigma_X^2\right)$ are independent for $i = 1, 2, \ldots, n$. Even though X typically never stands for a quantitative variable

in joint association analysis settings, we have found that the Gaussian distribution constitutes a perfectly robust approximation - at least in the case where X represents the genotype of a diploid organism. Under the alternative hypothesis, we infer that $(X_i \mid Y_i = y, Z_i = z) \sim \mathcal{N}\left(\mu^i_{X|Y,Z}, \Sigma^{ii}_{X|Y,Z}\right)$, where:

$$\mu^i_{X|Y,Z} = \frac{\left[y - U_i^{\mathrm{T}}\alpha - \beta\mu_{X|z} - \gamma z - \delta\left(\mu_{X|z} - \mu_X\right)(z - \mu_Z)\right]\left[\beta + \delta(z - \mu_Z)\right]\sigma_X^2}{\left[\beta + \delta(z - \mu_Z)\right]^2\sigma_X^2 + \sigma_\varepsilon^2} + \mu_{X|z},$$

$$\Sigma^{ii}_{X|Y,Z} = \sigma_X^2 - \frac{\left[\beta + \delta(z - \mu_Z)\right]^2\sigma_X^4}{\left[\beta + \delta(z - \mu_Z)\right]^2\sigma_X^2 + \sigma_\varepsilon^2}.$$

Additionally, we know that $K^{ijk\ell}_{X|Y,Z} = \Sigma^{ij}_{X|Y,Z}\Sigma^{k\ell}_{X|Y,Z} + \Sigma^{ik}_{X|Y,Z}\Sigma^{j\ell}_{X|Y,Z} + \Sigma^{i\ell}_{X|Y,Z}\Sigma^{jk}_{X|Y,Z}$ and $\gamma^{ijk}_{X|Y,Z} = 0$ for $i, j, k, \ell = 1, 2, \ldots, n$. Therefore, the formula for the conditional variance of the numerator of the interaction t test statistic given fixed quantitative trait Y and pathogen genetic variant Z is greatly simplified in the context of the Gaussian correction framework with unrelated individuals as follows:

$$\begin{aligned}\mathrm{Var}\left(N_* \mid Y, Z\right) = r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Qr - 4\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} \\ + 2\beta_*^2\mathrm{tr}\left(Q_S\Sigma_{X|Y,Z}Q_S\Sigma_{X|Y,Z}\right) + 4\beta_*^2\mu^{\mathrm{T}}_{X|Y,Z}Q_S\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z}.\end{aligned}$$

2.5.2 Discrete Correction

Suppose that $(X_i \mid Z_i = z)$ are independent and follow some discrete distribution with support $S \subseteq \mathbb{N}$ for $i = 1, 2, \ldots, n$. Then, we know that:

$$p^i(x \mid y, z) = \mathbb{P}\left(X_i = x \mid Y_i = y, Z_i = z\right) \propto \mathbb{P}\left(X_i = x \mid Z_i = z\right) f_{Y_i|X_i,Z_i}\left(y \mid x, z\right),$$

$$\mu^i_r(y, z) = \mathbb{E}\left(X_i^r \mid Y_i = y, Z_i = z\right) = \sum_{x \in S} p^i(x \mid y, z)x^r,$$

$$\mu^i_{X|Y,Z} = \mu^i_1(y, z), \quad \Sigma^{ii}_{X|Y,Z} = \mu^i_2(y, z) - \left[\mu^i_1(y, z)\right]^2,$$

$$\gamma^{iii}_{X|Y,Z} = \mu^i_3(y, z) - 3\mu^i_1(y, z)\mu^i_2(y, z) - \left[\mu^i_1(y, z)\right]^3,$$

$$K^{iiii}_{X|Y,Z} = \mu^i_4(y, z) - 4\mu^i_1(y, z)\mu^i_3(y, z) - 6\left[\mu^i_1(y, z)\right]^2\mu^i_2(y, z) - \left[\mu^i_1(y, z)\right]^4, \quad K^{iijj}_{X|Y,Z} = \Sigma^{ii}_{X|Y,Z}\Sigma^{jj}_{X|Y,Z}.$$

Let $\gamma_{X|Y,Z} = \left[\gamma^{iii}_{X|Y,Z}\right]_{i=1}^{n} \in \mathbb{R}^n$ and $K_{X|Y,Z} = \left[K^{iijj}_{X|Y,Z}\right]_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$. All other conditional central moments of X given Y and Z are equal to 0. Then, we conclude that:

$$\begin{aligned}\mathrm{Var}\left(N_* \mid Y, Z\right) = r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Qr - 2\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\mathrm{Dg}\left(Q_S\right)\gamma_{X|Y,Z} - 4\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} \\ + 2\beta_*^2\mathrm{tr}\left(Q_S\Sigma_{X|Y,Z}Q_S\Sigma_{X|Y,Z}\right) \\ + \beta_*^2\left[\mathrm{diag}\left(Q_S\right)\right]^{\mathrm{T}}\left[\mathrm{Dg}\left(K_{X|Y,Z}\right) - 3\Sigma^2_{X|Y,Z}\right]\mathrm{diag}\left(Q_S\right) \\ + 4\beta_*^2\mu^{\mathrm{T}}_{X|Y,Z}Q_S\mathrm{Dg}\left(Q_S\right)\gamma_{X|Y,Z} + 4\beta_*^2\mu^{\mathrm{T}}_{X|Y,Z}Q_S\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z},\end{aligned}$$

where $\mathrm{Dg}\left(Q_S\right) \in \mathbb{R}^{n\times n}$ is the diagonal matrix whose diagonal elements are given by the diagonal elements of $Q_S \in \mathbb{R}^{n\times n}$ and $\mathrm{diag}\left(Q_S\right) \in \mathbb{R}^n$ is the vector whose elements are given by the diagonal elements of Q_S.

This discrete correction framework is of particular interest in a joint association analysis, where X_i might encode the genotype of a genetic variant on a haploid or diploid organism. In such settings, one reasonable assumption to make would be that $\left(X_i \mid Z_i = z\right) \sim \mathrm{Bernoulli}\left(f_{X|z}\right)$ or $\left(X_i \mid Z_i = z\right) \sim \mathrm{Binomial}\left(2, f_{X|z}\right)$ respectively, where $f_{X|z}$ encodes the allele frequency of X_i given that $Z_i = z$. Then, one might apply this discrete correction framework to calculate a corrected interaction t test statistic.

2.6 Parameter Estimation in the Correction Framework

In order to understand how parameter estimation should be carried out in our proposed correction framework, we first have to take a careful look at the construction of the ordinary t test statistic. As previously discussed, the null conditional expectation of $\widehat{\delta}$ given fixed genetic variants X_j and Z_k is first subtracted, so that the test statistic is centered around 0 under the null hypothesis of no interaction. This ensures that the test statistic has correct type I error control. In the absence of a true interaction effect, estimating the standard error of $\widehat{\delta}$ under the null hypothesis rather than under the alternative hypothesis would make no significant difference, since the interaction effect only explains a negligible portion of the total variance of the quantitative trait in this case. In the presence of a true interaction effect though, estimating the standard error of $\widehat{\delta}$ under the null hypothesis would significantly underestimate the true residual variance of the quantitative trait,

leading to test statistics which are shrunk towards 0 with substantially less power to detect the underlying interaction effect. This is one of the reasons why the use of a Wald test, in which the standard error of $\widehat{\delta}$ is estimated under the alternative hypothesis, might be preferred over the use of a score test, in which the standard error of $\widehat{\delta}$ is estimated under the null hypothesis.

We endeavor to mimic this construction for our corrected interaction t test statistic. Therefore, we use a null model for the quantitative trait Y in order to estimate the conditional moments of X required for the calculation of the conditional expectation of the numerator of the test statistic given Y and Z. On the other hand, we use an alternative model for Y to estimate the conditional moments of X required for the conditional variance of the denominator. The conditional moments of X given Y and Z under the alternative hypothesis for the Gaussian correction framework are given in Section 2.5.1, while the null conditional moments of X can be directly derived by just setting the interaction coefficient δ to be equal to 0. We call this implementation the alternative correction framework.

Even though the alternative correction framework properly mimics the construction of the ordinary t test statistic, albeit with a shift in the conditioning being performed, we have observed that the estimation of the interaction coefficient δ can prove to be highly unreliable in joint association analysis settings, due to the discrete nature of the interacting genetic variants. As the observed minor allele frequency of the pathogen genetic variant Z drifts away from 0.5 and towards 0, we have less and less information available to accurately estimate the interaction effect δ, leading to excess type I errors. We are currently considering some form of regularization or shrinkage to address this behavior. It should be noted though that this estimation issue naturally diminishes as the sample size increases. Therefore, we also propose the use of a null correction framework, where the conditional variance of the numerator of the interaction t test statistic given Y and Z is also estimated under the null hypothesis of no interaction. Even though this null correction framework is naturally more conservative than the alternative one, it manages to maintain better type I error control for observed Z minor allele frequencies below some threshold.

Additionally, we have to be aware of the potential heteroscedasticity in the quantitative trait due to some unaccounted interaction between the pathogen genetic variant currently being tested and

some unobserved factor. As we have previously discussed, this heteroscedasticity in the response variable could lead to overinflated type I error rates if not handled properly. Therefore, we propose combining our correction framework for the feast or famine effect with the IRLS procedure discussed in Section 2.3 to account for heteroscedasticity in the quantitative trait. In this unified framework, the joint association model given by 2.2 is always assumed to be heteroscedastic instead and we elect to fit the resulting heteroscedastic model by using this implementation of the IRLS procedure which always converges after just 1 step in the case of the alternative model. The null and alternative conditional moments of X given Y and Z are finally calculated as previously discussed.

In the Gaussian correction framework, we regress X on the covariate matrix U and the pathogen genetic variant Z to estimate the required parameters for the conditional distribution of X given Z. In the discrete correction framework, we employ naive estimators, which are based on the conditional sample mean of X given Z, for the theoretical allele frequencies $f_{X|z}$ in the conditional distribution of X given that $Z = z$.

2.7 Diagnostic Ratio for the Feast or Famine Effect

For a fixed quantitative trait Y and given pathogen genetic variants $Z_1, Z_2, \ldots, Z_{m_v}$, suppose that we simulate independent host genetic variants $X_1, X_2, \ldots, X_{m_h}$ which are also independent of Y and Z. As discussed in previous sections, we would expect that some pathogen genetic variants are going to display the feast effect, some are going to display the famine effect and some are going to produce roughly uniform interaction p-values. We could potentially perform this Monte Carlo experiment every time we get access to new trait vectors and new pathogen genotype vectors for a joint association analysis, but this procedure could prove to be potentially unreliable for a smaller number of simulated host genetic variants and time consuming for a larger number of simulated host genetic variants. Since we have already ascertained that the feast or famine effect is an intrinsic property of the pair of Y and Z, we would ideally want to calculate an interpretable quantity which only depends on these 2 observed vectors to predict this phenomenon. Then, we could potentially decide to only correct the interaction t test statistics corresponding to the pathogen genetic variants which have been predicted to display the most extreme feast or famine effects using our correction framework.

In order to correct for the feast or famine effect in a collection of interaction t test statistics corresponding to a quantitative trait Y and a pathogen genetic variant Z, our correction framework essentially entails the substitution of the squared denominator of the test statistic by the conditional variance of the numerator of the test statistic given Y and Z. Therefore, we assume that the ratio between these 2 quantities would prove to be a good predictor of the feast or famine effect. Whereas the conditional variance of the numerator of the test statistic given Y and Z only depends on Y and Z, the squared denominator of the test statistic also depends on the random host genetic variant X. Hence, this approach would also require the calculation of the conditional expectation of the squared denominator of the interaction t test statistic given Y and Z.

Suppose that $X \perp\!\!\!\perp (Y, Z)$ with $\mathbb{E}(X) = \mu_X \mathbf{1}_n$ and $\text{Var}(X) = \sigma_X^2 \mathbf{I}_n$. For any vectors $A, B \in \mathbb{R}^n$, we define $S'_{AB} = A^{\text{T}} PB$ and $S'_{ABZZ} = A^{\text{T}} P\text{Dg}\left(HZZ^{\text{T}}H\right) PB \in \mathbb{R}$, where $H = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\text{T}}$, $P = \mathbf{I}_n - U\left(U^{\text{T}}U\right)^{-1}U^{\text{T}}$ and $\text{Dg}\left(HZZ^{\text{T}}H\right) \in \mathbb{R}^{n\times n}$ is the diagonal matrix whose diagonal elements are given by the diagonal elements of $HZZ^{\text{T}}H \in \mathbb{R}^{n\times n}$. In this special case, we calculate that:

$$\text{Var}\left(N_* \mid Y, Z\right) = \sigma_X^2 \frac{S'^2_{ZZ} S'_{YYZZ} - 2S'_{YZ} S'_{ZZ} S'_{YZZZ} + S'^2_{YZ} S'_{ZZZZ}}{S'^2_{ZZ}}.$$

Similarly to the calculations we have previously performed for the numerator of the interaction t test statistic in the context of the correction framework, we first need to take an asymptotic approximation of the denominator of the test statistic given Y and Z. Such an asymptotic approximation is given by:

$$D_*^2 = \frac{S'_{YY} S'_{ZZ} - S'^2_{YZ}}{(n-c-3)\, S'_{ZZ}} \left[S'_{WW} - \frac{\text{tr}\,(Q)}{n-c} S'_{XW}\right] - \frac{N_*^2}{n-c-3},$$

where N_* is the asymptotic approximation of the numerator of the test statistic given by equation 2.7, $Q = H\text{diag}(HZ)P$, $W = \text{diag}\,(HZ)\, HX$ and $\text{diag}\,(HZ) \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix whose diagonal elements are given by the elements of the vector $HZ \in \mathbb{R}^n$. Therefore, we conclude that:

$$\mathbb{E}\left(D_*^2 \mid Y, Z\right) = \frac{\sigma_X^2}{n-c-3} \frac{S'_{YY} S'_{ZZ} - S'^2_{YZ}}{S'_{ZZ}} \frac{(n-c)\text{tr}\left(QQ^{\text{T}}\right) - \left[\text{tr}\,(Q)\right]^2}{n-c} - \frac{\text{Var}\left(N_* \mid Y, Z\right)}{n-c-3}.$$

We define the following ratio between the conditional variance of the numerator and the conditional expectation of the squared denominator of the interaction t test statistic given Y and Z:

$$\begin{aligned} R &= \frac{\mathrm{Var}\left(N_* \mid Y, Z\right)}{\mathbb{E}\left(D_*^2 \mid Y, Z\right)} \\ &\approx \frac{(n-c-3)\,(n-c)}{(n-c)\mathrm{tr}\left(QQ^{\mathrm{T}}\right) - \left[\mathrm{tr}\left(Q\right)\right]^2} \frac{S'^2_{ZZ} S'_{YYZZ} - 2S'_{YZ} S'_{ZZ} S'_{YZZZ} + S'^2_{YZ} S'_{ZZZZ}}{S'_{YY} S'^2_{ZZ} - S'^2_{YZ} S'_{ZZ}}. \end{aligned} \tag{2.11}$$

In the special case where $U = \mathbf{1}_n$, i.e. in the absence of covariates, we note that $\mathrm{tr}(Q) = 0$ and $\mathrm{tr}\left(QQ^{\mathrm{T}}\right) = \frac{n-2}{n} S'_{ZZ}$. Hence, these expressions simplify to the following:

$$\mathbb{E}\left(D_*^2 \mid Y, Z\right) = \frac{\sigma_X^2}{n-4} \frac{n-2}{n} \left(S'_{YY} S'_{ZZ} - S'^2_{YZ}\right) - \frac{\mathrm{Var}\left(N_* \mid Y, Z\right)}{n-4},$$

$$R \approx n \frac{S'^2_{ZZ} S'_{YYZZ} - 2S'_{YZ} S'_{ZZ} S'_{YZZZ} + S'^2_{YZ} S'_{ZZZZ}}{S'_{YY} S'^3_{ZZ} - S'^2_{YZ} S'^2_{ZZ}}. \tag{2.12}$$

Detailed derivations of these asymptotic approximations and conditional moment calculations are presented in Section 2.10.2. In both cases, we observe that the ratio between the conditional variance of the numerator and the conditional expectation of the squared denominator of the interaction t test statistic does not depend on σ_X^2 or X in general but only on Y and Z. If this ratio is larger than 1, then the conditional variance of the numerator given Y and Z is on average larger than the squared denominator, which implies that the interaction test statistics corresponding this specific pathogen genetic variant are systematically larger than what they should normally be under the global null hypothesis of no interaction, i.e. this pathogen genetic variant displays the feast effect. Similarly, if this ratio is smaller than 1, the corresponding pathogen genetic variant displays the famine effect. If the ratio is close to 1, then our proposed correction framework has little effect on the distribution of the interaction t test statistic, which means that the original collection of test statistics corresponding to this pathogen genetic variant is properly calibrated under the global null hypothesis. According to this, we might surmise that this ratio fulfills a similar role to that of the genomic control inflation factor in association analyses - informing us about the potential overinflation or underinflation of a collection of association test statistics due to misspecification of the association analysis model [71]. We have verified through simulations that there indeed exists

a particularly strong linear relationship between this ratio and the genomic control inflation factor of the collection of interaction t test statistics corresponding to the respective pathogen genetic variant. Therefore, our proposed ratio constitutes an appropriate diagnostic tool to predict the feast or famine effect in joint association analysis settings without the need to perform any Monte Carlo experiments.

2.8 Simulation Studies in the Correction Framework

2.8.1 Gaussian Type I Error Study

Before investigating the behavior of the feast or famine effect and assessing the performance of our correction framework with respect to alleviating this phenomenon, it is important to ascertain that our corrected interaction test statistics achieve better type I control and display comparable power to the uncorrected interaction t test statistic. For that reason, we design a couple of simulation studies where we only simulate a few host genetic variants for every simulated quantitative trait and pathogen genetic variant pair. In this setting, we can then evaluate the performance of different interaction test statistics in terms of type I error rate and power across different fixed quantitative traits and pathogen genetic variants, i.e. across a collection of observed interaction test statistics where the feast of famine effect is not applicable, since they do not all correspond to the same fixed quantitative trait and pathogen genetic variant.

We first simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0.1, 0.9]$. Then, we set the sample size n to be $1,000$ individuals and we simulate independent pathogen genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ for $i = 1, 2, \ldots, n$. We set $m_h = 4$ and simulate independent environmental variables $X_{ij} \sim \mathcal{N}(0, 1)$ for $j = 1, 2, \ldots, m_h$. We set the parameter values $\beta = \left(0, \sqrt{0.025}, -\sqrt{0.025}, \sqrt{0.05}\right)$, $\gamma = \sqrt{0.025}$ and $\sigma_\varepsilon^2 = 1 - \|\beta\|^2 - \gamma^2$, so that the total variance of the simulated trait is equal to 1 and the proportion of the total variation explained by each coefficient is prespecified. Lastly, we simulate the following quantitative trait values:

$$Y_i = \sum_{j=1}^{4} \beta_j X_{ij} + \gamma \widetilde{Z}_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ are independent and $\widetilde{Z}_i = \frac{Z_i - f_Z}{\sqrt{f_Z(1-f_Z)}}$ is the standardized pathogen genotype. We observe that there is no true interaction effect between the pathogen genetic variant Z and

any of the simulated environmental variables, meaning that there is no heteroscedasticity in the simulated trait Y. We repeat this simulation $100,000$ times. Type I error rate calculations at significance level 0.05 are shown in Table 2.1. Type I errors which are significantly different from the nominal level of 0.05 at level 0.01 are displayed in bold text.

	Type I Error			
	δ_1	δ_2	δ_3	δ_4
No Correction	0.04918	0.05109	0.04997	0.04964
HC3	0.04947	0.05066	0.05040	0.04984
IRLS	0.04915	0.05168	0.05048	0.04989
Null Gaussian Correction	0.04860	0.05064	0.04971	0.04956
Alternative Gaussian Correction	0.04960	**0.05180**	0.05079	0.05069

TABLE 2.1: Type I Error Rates in the Gaussian Case

We observe that the uncorrected interaction t test statistic maintains correct type I error control in the absence of heteroscedasticity in the quantitative trait. Blanket use of the heteroscedasticity correction methods described in Section 2.3 such as the heteroscedasticity-consistent (HC) estimator corresponding to the covariance matrix $\widehat{\Phi}_3$, our proposed iteratively reweighted least squares (IRLS) procedure and our entire correction framework have no significant impact on that type I error control. It is also important to note that there inherently exists a severe model misspecification in our fitted models for this joint association analysis, since only one X_j is taken into account at a time, ignoring the significant additive effects of the rest on the simulated quantitative trait.

2.8.2 Gaussian Power Study

We first simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0.1, 0.9]$. Then, we set the sample size n to be $1,000$ individuals and we simulate independent pathogen genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ for $i = 1, 2, \ldots, n$. We set $m_h = 4$ and simulate independent environmental variables $X_{ij} \sim \mathcal{N}(0, 1)$ for $j = 1, 2, \ldots, m_h$. We set the parameter values $\beta = \left(0, \sqrt{0.025}, -\sqrt{0.025}, \sqrt{0.05}\right)$, $\gamma = \sqrt{0.025}$, $\delta = \left(0, 0, 0, \sqrt{0.025}\right)$ and $\sigma_\varepsilon^2 = 1 - \|\beta\|^2 - \gamma^2 - \|\delta\|^2$, so that the total variance of the simulated trait is equal to 1 and the proportion of the total variation explained by each coefficient is prespecified.

Lastly, we simulate the following quantitative trait values:

$$Y_i = \sum_{j=1}^{4} \beta_j X_{ij} + \gamma \widetilde{Z}_i + \sum_{j=1}^{4} \delta_j X_{ij} \widetilde{Z}_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ are independent and $\widetilde{Z}_i = \frac{Z_i - f_Z}{\sqrt{f_Z(1-f_Z)}}$ is the standardized pathogen genotype. We observe that there is one true interaction effect between the pathogen genetic variant Z and the environmental variable X_4, meaning that the quantitative trait Y is going to be heteroscedastic when that interaction effect is not accounted for. We repeat this simulation $100,000$ times. Type I error rate calculations at significance level 0.05 and power calculations at significance level 10^{-5} are shown in Table 2.2. Type I errors which are significantly different from the nominal level of 0.05 at level 0.01 are displayed in bold text.

	Type I Error			Power
	δ_1	δ_2	δ_3	δ_4
No Correction	**0.05510**	**0.05650**	**0.05467**	0.78517
HC3	0.05036	**0.05198**	0.05011	0.77222
IRLS	0.05076	**0.05208**	0.05014	0.78514
Null Gaussian Correction	0.04963	0.05108	0.04933	0.74398
Alternative Gaussian Correction	0.05060	**0.05224**	0.05033	0.77951

TABLE 2.2: Type I Error Rates and Power in the Gaussian Case

We observe that the uncorrected interaction t test statistic displays significantly overinflated type I error rates across the board in the presence of heteroscedasticity in the quantitative trait. Blanket use of the heteroscedasticity correction methods described in Section 2.3 manage to attain better type I error control. At the same time, we note that our proposed IRLS procedure achieves slightly higher power than the existing HC3 covariance matrix approach, while also maintaining almost the same power as that of the severely overinflated uncorrected interaction t test statistic. Our proposed null and alternative Gaussian correction frameworks also attain better type I error control, with the alternative Gaussian correction consistently exhibiting slightly more overinflated type I error rates and higher power than the null Gaussian correction, as well as comparable power to the severely overinflated uncorrected interaction test statistic. Again, we note that there inherently exists an even more severe model misspecification in our fitted models for this joint association

analysis, since only one X_j is taken into account at a time, ignoring the significant additive effects of X_2, X_3, X_4 as well as the significant interaction effect between X_4 and Z on the simulated quantitative trait.

2.8.3 Gaussian Feast or Famine Study

We first simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0.1, 0.9]$. Then, we set the sample size n to be $1,000$ individuals and we simulate independent pathogen genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ for $i = 1, 2, \ldots, n$. We set m_h to be equal to $10,000$ and simulate independent environmental variables $X_{ij} \sim \mathcal{N}(0, 1)$ for $j = 1, 2, \ldots, m_h$. Lastly, we simulate independent quantitative trait values $Y_i \sim \mathcal{N}(0, 1)$ under the global null hypothesis of no interaction. We repeat this simulation $m_v = 1,000$ times. At the same time as performing the pertinent joint GWAS testing for interactions between each possible pair of simulated pathogen genetic variants and environmental variables on the quantitative trait, we also perform a marginal association analysis of the simulated environmental variables on the quantitative trait.

For the collection of m_h interaction test statistics corresponding to each simulation replicate, we calculate a list of diagnostic quantities relating to the feast or famine effect. Let F denote the cumulative distribution function of the chi-squared distribution with 1 degree of freedom. Then, we calculate a genomic control inflation factor for the collection of interaction p-values $p_{1k}, p_{2k}, \ldots, p_{m_h k}$ for $k = 1, 2, \ldots, m_v$ as follows:

$$\lambda_k = \frac{\text{median}\left\{F^{-1}(1 - p_{1k}), \ldots, F^{-1}(1 - p_{m_h k})\right\}}{F^{-1}(0.5)},$$

where $F^{-1}(0.5) \approx 0.456$ denotes the theoretical median of the chi-squared distribution with 1 degree of freedom [71]. Under the global null hypothesis of no interaction, we ideally want this quantity to take values close to 1. As previously discussed though, we expect to observe an unusually high variation of genomic control inflation factors in this joint association analysis due to the feast or famine effect. Additionally, we create a Q-Q plot for the collection of interaction p-values $p_{1k}, p_{2k}, \ldots, p_{m_h k}$ with simultaneous confidence based on the equal local levels global test statistic using the qqconf package [59] developed in R. A 2-sided Q-Q plot p-value can then be calculated

as the maximum significance level at which we cannot reject the null hypothesis of uniformity for the set of interaction p-values. We would expect these Q-Q plot p-values to follow the uniform distribution under the global null hypothesis of no interaction. We also repeat these calculations for each set of marginal GWAS p-values as well as each set of corrected interaction p-values. Lastly, we calculate the value of our proposed diagnostic ratio for each pair of a simulated trait and pathogen genetic variant.

First, we look at how the distribution of genomic control inflation factors around 1 based on the joint GWAS compares against the one based on the marginal GWAS, shown in Figure 2.1. We indeed observe a staggering amount of variation in the joint GWAS genomic control inflation factors compared to what one would normally expect in an ordinary marginal GWAS. On the other hand, the distribution of the genomic control inflation factors based on any of our proposed correction frameworks closely matches the one based on the marginal GWAS, as shown in Figure 2.2.

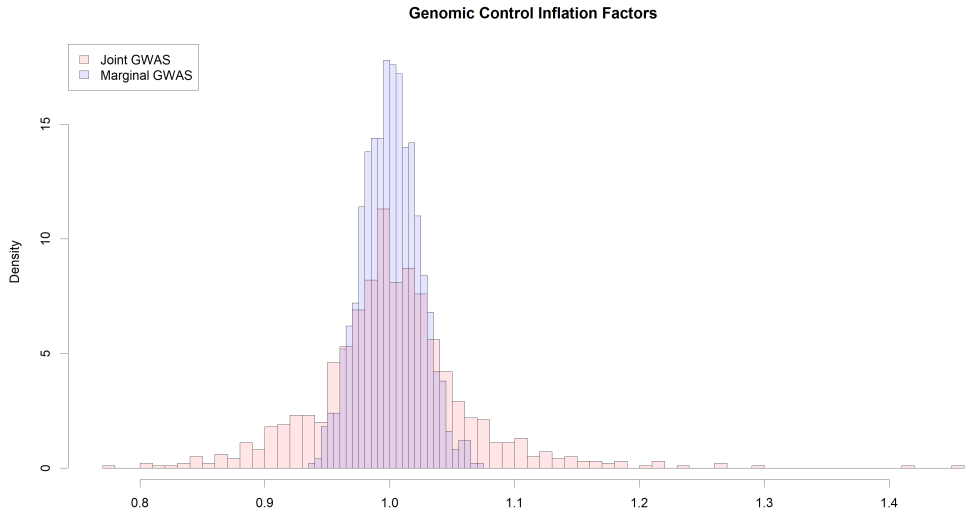

FIGURE 2.1: Histograms of Uncorrected Joint vs. Marginal Genomic Control Inflation Factors in the Gaussian Case

Then, we can take a look at the Q-Q plots of the interaction p-values corresponding to the pathogen genetic variants with the largest and smallest uncorrected genomic control inflation factors, shown in Figure 2.3. The deviation of these collections of interaction p-values from uniformity is astounding, even though this simulation is performed under the global null hypothesis of no interaction. More specifically, the left Q-Q plot represents the potential extremity of the feast effect in joint

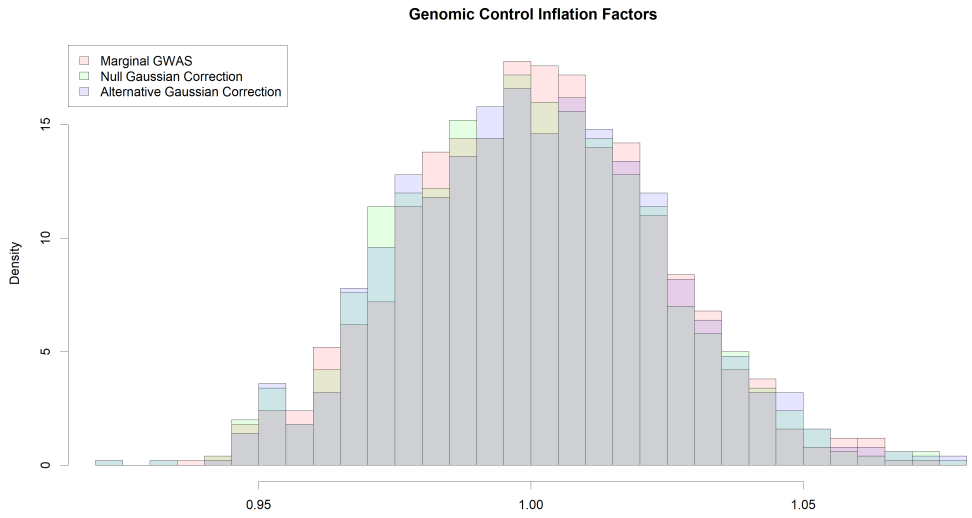

FIGURE 2.2: Histograms of Corrected Joint vs. Marginal Genomic Control Inflation Factors in the Gaussian Case

association analyses, where an unbelievable amount of false discoveries would be committed by using the uncorrected interaction t test statistic, while the right Q-Q plot represents the potential extremity of the famine effect, where there would never be enough power to detect important interaction signals. For reference, we compare these Q-Q plots against the Q-Q plots of the p-values from the marginal association analyses corresponding to the largest and smallest marginal genomic control inflation factors. We note that even the most extreme marginal association p-value distributions are essentially indistinguishable from the uniform distribution, since they do not necessarily coincide with the marginal association p-value distributions with the smallest Q-Q plot p-values. In comparison, the distributions of our corrected interaction p-values are also indistinguishable from uniform in these cases, as shown in Figure 2.4, indicating that our correction frameworks manage to correct for the extreme nature of both the feast as well as the famine effect.

Then, we look at the Q-Q plots of the 2-sided Q-Q plot p-values before and after correction, as well as the the Q-Q plot of the 2-sided Q-Q plot p-values corresponding to the marginal GWAS, all displayed in Figure 2.5. It should be noted that this Q-Q plot is fundamentally different compared to the previously discussed Q-Q plots which were Q-Q plots of actual interaction and association p-values. These Q-Q plots essentially serve as a meta-analysis for our collection of association analyses corresponding to different quantitative trait and pathogen genetic variant pairs, hence we

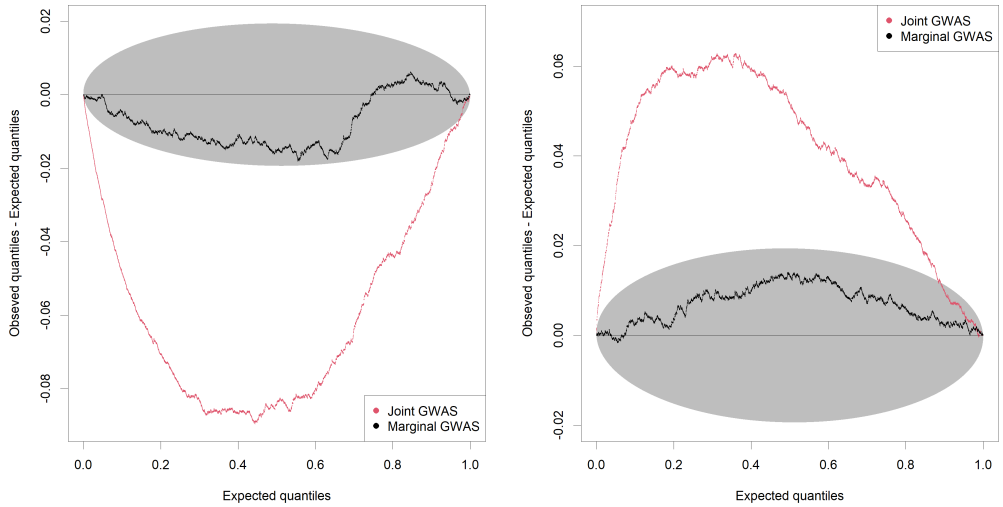

FIGURE 2.3: Q-Q Plots Displaying the Feast or Famine Effect in the Gaussian Case

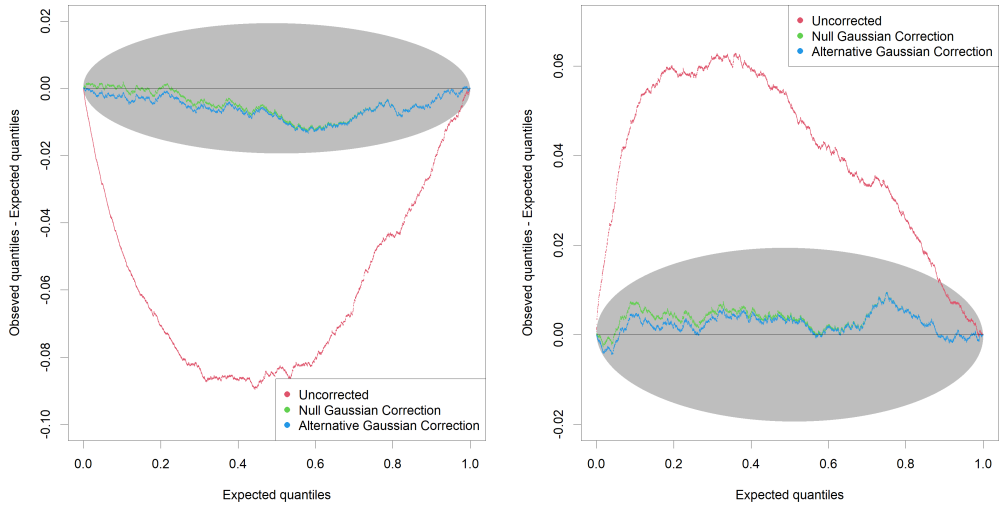

FIGURE 2.4: Q-Q Plots Displaying the Correction of the Feast or Famine Effect in the Gaussian Case

sometimes refer to them as meta Q-Q plots. We observe that the uncorrected Q-Q plot p-values tend to be much smaller than what one would expect under the uniform distribution. On the other hand, the distribution of the Q-Q plot p-values based on our null correction framework are practically indistinguishable from the uniform distribution and closely match those corresponding to the marginal GWAS, which implies that the null correction performs perfectly in terms of correcting the feast or famine effect. The alternative correction framework performs much better than the

uncorrected interaction t test statistic in terms of the feast or famine effect, but consistently displays slightly smaller than uniform Q-Q p-values, mostly owing to the unreliability in the estimation of $\widehat{\delta}$ discussed in Section 2.6.

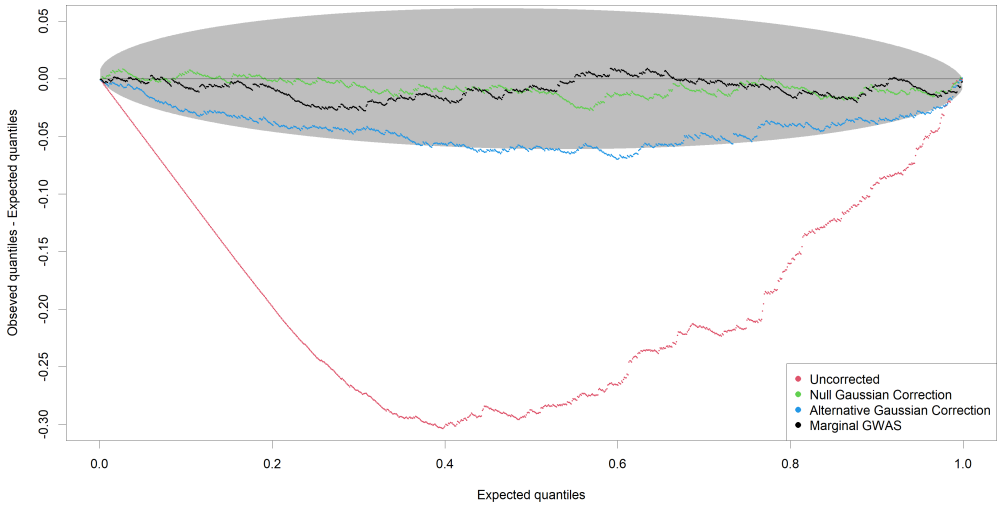

FIGURE 2.5: Comparison of Q-Q Plots of Uncorrected, Corrected and Marginal 2-Sided Q-Q Plot P-Values in the Gaussian Case

Finally, we evaluate the performance of our proposed diagnostic ratio. Plotting histograms of the diagnostic ratio and the uncorrected genomic control inflation factors on top of each other - Figure 2.6 - reveals that the 2 distributions closely match each other, even though the distribution of the diagnostic ratio displays slightly lighter tail behavior. A scatterplot of the uncorrected genomic control inflation factors against the diagnostic ratio, shown in Figure 2.7, verifies the strong linear relationship between them. We note that the sample correlation between these 2 quantities is calculated to be 92.4%.

When performing a GWAS, rather than being interested in the median result, we are particularly interested in the behavior of the smallest p-values. Therefore, we also consider the ability of our diagnostic ratio to predict the tail behavior of the interaction p-values. To do this, we calculate the 5% sample quantiles for each collection of uncorrected interaction p-values $p_{1k}, p_{2k}, \ldots, p_{m_h k}$ and plot them on the $-\log_{10}$ scale against the diagnostic ratio, shown in Figure 2.8. We observe that the diagnostic ratio performs fantastically in predicting the behavior of the smallest uncorrected interaction p-values. As a reference, the sample correlation between these 2 quantities was

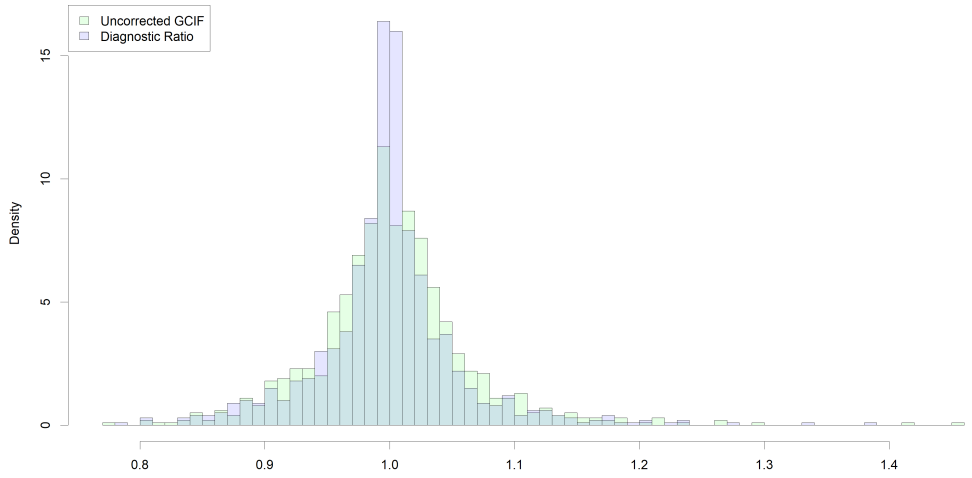

FIGURE 2.6: Histograms of Uncorrected Genomic Control Inflation Factors vs. Diagnostic Ratio in the Gaussian Case

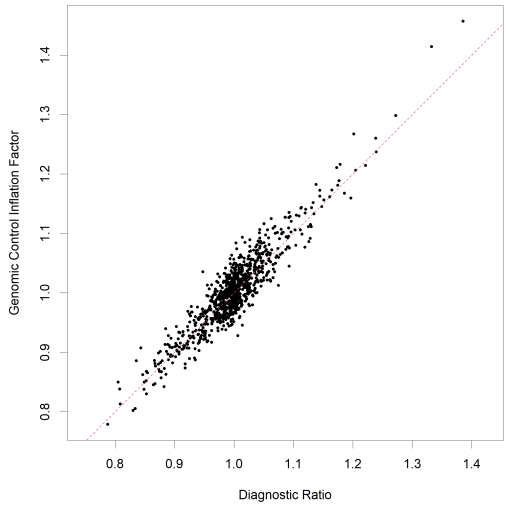

FIGURE 2.7: Scatterplot of Uncorrected Genomic Control Inflation Factors vs. Diagnostic Ratio in the Gaussian Case

calculated to be equal to 94.97%. We note that performance of our proposed diagnostic ratio in predicting the median and tail behavior of the uncorrected interaction p-values remains the same even if the pathogen genetic variant has a significant effect on the simulated quantitative trait, since the relationship between Z and Y is always taken into account in our computations and there are absolutely no assumptions placed upon it.

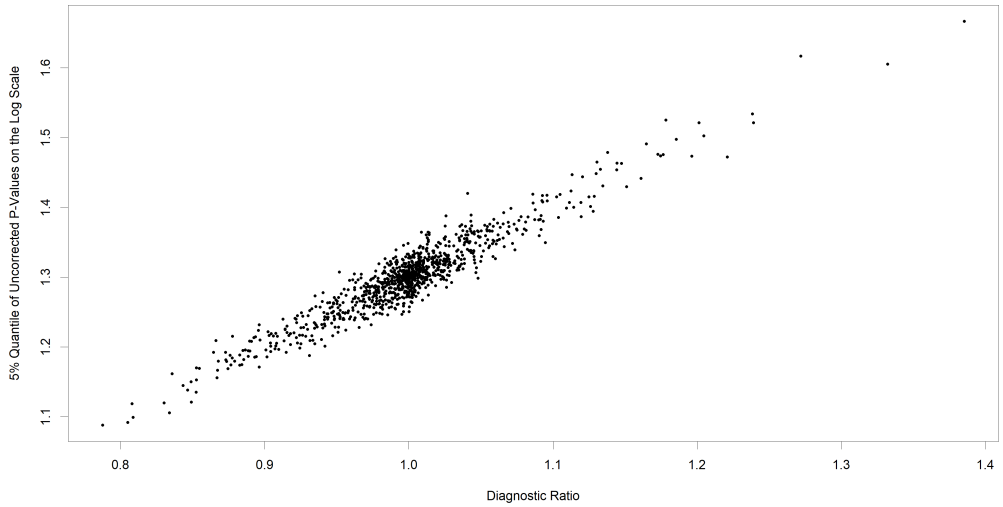

FIGURE 2.8: Scatterplot of 5% Quantiles of Uncorrected P-Values on the Log Scale vs. Diagnostic Ratio in the Gaussian Case

2.8.4 Null Binomial Study

Consider the following 2-by-3 contingency table between a host SNP and a pathogen genetic variant:

		Host		
	Counts	0	1	2
Pathogen	0	a	b	c
	1	d	e	f

We define the minimum cell count of the contingency table and the minor allele count of the host SNP as follows:

$$\text{MCC} = \min\left\{a + \frac{b}{2}, c + \frac{b}{2}, d + \frac{e}{2}, f + \frac{e}{2}\right\},$$

$$\text{MAC} = \min\left\{a + d + \frac{b+e}{2}, c + f + \frac{b+e}{2}\right\}.$$

In other words, each host allele copy counts as 0.5. We first simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0,1]$. Then, we set the sample size n to be $1,000$ individuals and we simulate independent pathogen genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ for $i = 1, 2, \ldots, n$. We set the number of host genetic variants m_h to be equal to 1. We simulate a host allele frequency $f_X \sim \text{Unif}[0,1]$ and independent

host genotypes $X_i \sim \text{Binomial}\,(2, f_X)$. Lastly, we simulate independent quantitative trait values $Y_i \sim \mathcal{N}(0, 1)$ under the global null hypothesis of no interaction. We repeat this simulation $1,000,000$ times.

We are interested in ascertaining the behavior of our corrected interaction test statistics for small allele counts and minimum cell counts. Since we are considering a diploid host organism, the assumption that $(X_i \mid Z_i = z) \sim \text{Binomial}\left(2, f_{X|z}\right)$ makes sense in the context of our discrete correction framework. Hence, we derive the null and alternative binomial correction framework, whose performance we are going to evaluate in the following section on top of that of the established Gaussian correction frameworks. First, we divide our interaction tests into 100 equally sized bins with respect to their Z minor allele count, which ranges from 0 to 500. Plotting the type I error rate of our uncorrected and corrected interaction test statistics as a function of the median Z minor allele count of each bin, shown in Figure 2.9, reveals that the alternative binomial correction becomes severely overinflated as the Z minor allele count approaches 0. Note that 95% Wald confidence bands with multiple testing adjustment are drawn for reference. Therefore, we propose to set a lower threshold of 80 on the Z minor allele count and only consider applying the alternative binomial correction on pathogen genetic variants with a Z minor allele count above that threshold. In contrast, Z minor allele count has no effect on the uncorrected interaction test statistic and the null binomial correction.

Next, we repeat the same with respect to the X minor allele count. Contrary to the behavior of our alternative binomial correction with respect to the Z minor allele count, we observe that the type I error rate of the alternative binomial correction gets severely underinflated as the X minor allele count approaches 0, as shown in Figure 2.10. Hence, we propose to set a threshold of 50 on the X minor allele count and only consider applying the alternative binomial correction on pathogen genetic variants with X minor allele counts above that threshold to avoid substantial loss of power. In contrast, X minor allele count has no effect on the uncorrected interaction test statistic and the null binomial correction.

Finally, we repeat the same with respect to the previously defined notion of a minimum cell count. The behavior of the type I error rates of the alternative binomial correction with respect to minimum

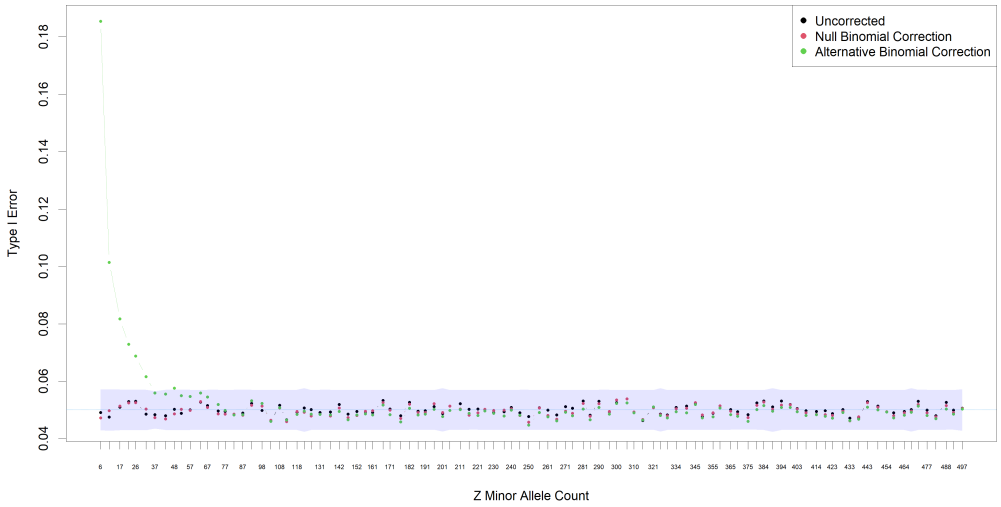

FIGURE 2.9: Uncorrected vs. Corrected Type I Error Rates Aggregated by Z Minor Allele Count in the Null Binomial Case

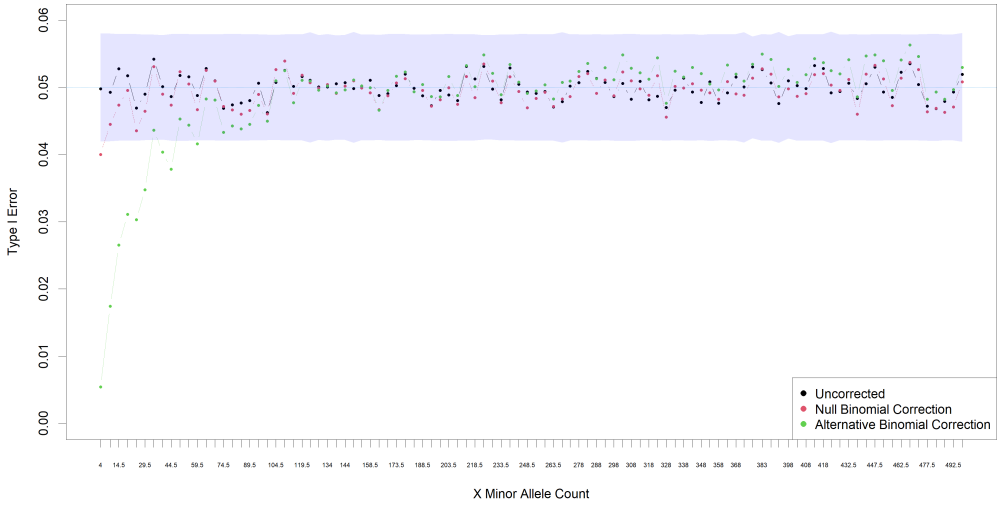

FIGURE 2.10: Uncorrected vs. Corrected Type I Error Rates Aggregated by X Minor Allele Count in the Null Binomial Case

cell count, displayed in Figure 2.11, is hard to disentangle because of the opposite effect that X and Z minor allele counts have on it. However, we can observe that a minimum cell count threshold of 10 should ideally be set on the alternative binomial correction to ensure that there jointly exists enough information between the host and the pathogen genetic variant to accurately estimate all the required parameters for the correction framework. Additionally, we believe that a

minimum cell count threshold of 5 should be set on the null binomial correction and the uncorrected interaction test statistic for the same reason. Further simulation studies, not included as part of this dissertation, reveal that all of these thresholds do not depend on sample size, while also being directly transferable to the null and alternative Gaussian correction frameworks. We are currently considering the use of Fisher's information as a more reliable measure of whether a specific pair of host and pathogen genetic variants contain enough combined information for reliable testing of an interaction effect between them on a quantitative trait of interest.

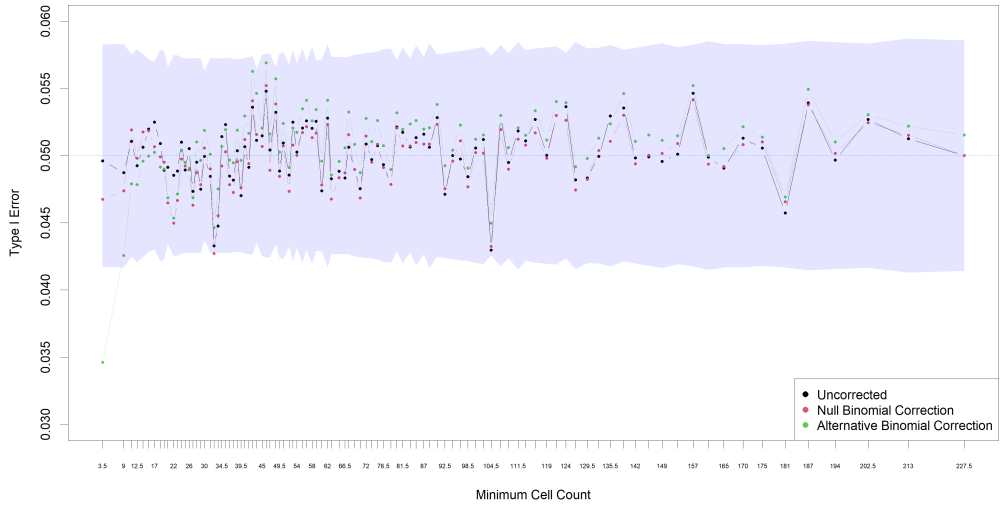

FIGURE 2.11: Uncorrected vs. Corrected Type I Error Rates Aggregated by Minimum Cell Count in the Null Binomial Case

2.8.5 Correlated Binomial Study

We first set the number of host genetic variants m_h to be equal to 1. We simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0.1, 0.5]$, a host allele frequency $f_X \sim \text{Unif}[0.1, 0.5]$ and a correlation level $\rho \sim \text{Unif}[0, 0.2]$. Then, we set the sample size n to be $1,000$ individuals. We simulate independent pathogen genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ and independent host genotypes $X_i \sim \text{Binomial}\,(2, f_X)$ with a correlation of ρ between them for $i = 1, 2, \ldots, n$. Lastly, we simulate independent quantitative trait values $Y_i \sim \mathcal{N}(0, 1)$ under the global null hypothesis of no interaction. We repeat this simulation $1,000,000$ times.

We are interested in ascertaining the behavior of our corrected interaction test statistics for corre-

lated host and pathogen genetic variants. We divide our interaction tests into 100 equally sized bins with respect to their observed correlation level, which roughly ranges from 0 to 0.2. Plotting the type I error rate of our uncorrected and corrected interaction test statistics as a function of the median correlation level of each bin, shown in Figure 2.12, reveals that the type I error rates of our correction frameworks get slightly underinflated as the correlation level increases. Note that 95% Wald confidence bands with multiple testing adjustment are again drawn for reference. This phenomenon can be explained by the fact that as the correlation level between X and Z increases, so does the minimum cell count between them decrease on average. A plot of the type I error rates of our uncorrected and corrected interaction test statistics as a function of the median minimum cell count of each bin, where this underinflation of type I error rates for smaller minimum cell counts is evident, is given in Figure 2.13. Testing for interactions between highly correlated genetic variants is generally hard to interpret in any case, so we propose to set an upper threshold of 0.1 on correlation level and only consider applying the binomial correction framework on pathogen genetic variants with a correlation level below that threshold.

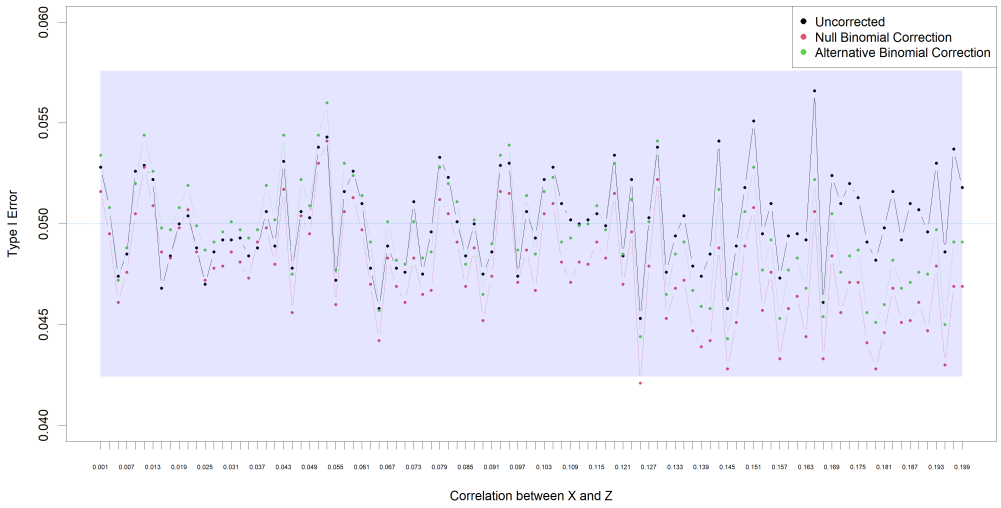

FIGURE 2.12: Uncorrected vs. Corrected Type I Error Rates Aggregated by Correlation between X and Z in the Correlated Binomial Case

2.8.6 Binomial Type I Error Study

We first simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0.1, 0.9]$. Then, we set the sample size n to be $1,000$ individuals and we simulate independent pathogen genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ for

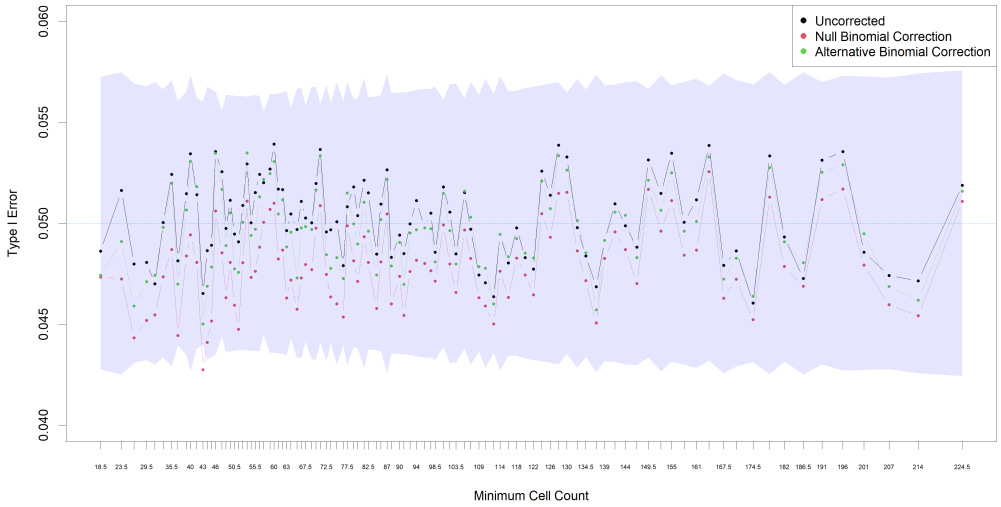

FIGURE 2.13: Uncorrected vs. Corrected Type I Error Rates Aggregated by Minimum Cell Count in the Correlated Binomial Case

$i = 1, 2, \ldots, n$. We set the number of host genetic variants m_h to be equal to 4. We simulate a independent host allele frequencies $f_{X_j} \sim \text{Unif}[0.1, 0.9]$ and host genotypes $X_{ij} \sim \text{Binomial}\left(2, f_{X_j}\right)$ for $j = 1, 2, \ldots, m_h$. We set the parameter values $\beta = \left(0, \sqrt{0.025}, -\sqrt{0.025}, \sqrt{0.05}\right)$, $\gamma = \sqrt{0.025}$ and $\sigma_\varepsilon^2 = 1 - \|\beta\|^2 - \gamma^2$, so that the total variance of the simulated trait is equal to 1 and the proportion of the total variation explained by each coefficient is prespecified. Lastly, we simulate the following quantitative trait values:

$$Y_i = \sum_{j=1}^{4} \beta_j \widetilde{X}_{ij} + \gamma \widetilde{Z}_i + \varepsilon_i, \quad \widetilde{X}_{ij} = \frac{X_{ij} - 2f_{X_j}}{\sqrt{2f_{X_j}\left(1 - f_{X_j}\right)}}, \quad \widetilde{Z}_i = \frac{Z_i - f_Z}{\sqrt{f_Z\left(1 - f_Z\right)}},$$

where $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ are independent. We observe that there is no true interaction effect between the pathogen genetic variant Z and any of the simulated host genetic variants, meaning that there is no heteroscedasticity in the simulated trait Y. We repeat this simulation $100,000$ times. Type I error rate calculations at significance level 0.05 are shown in Table 2.3. Type I errors which are significantly different from the nominal level of 0.05 at level 0.01 are displayed in bold text.

We observe that the uncorrected interaction t test statistic maintains correct type I error control in the absence of heteroscedasticity in the quantitative trait. Blanket use of the heteroscedasticity correction methods described in Section 2.3 such as the heteroscedasticity-consistent (HC) estimator

	Type I Error			
	δ_1	δ_2	δ_3	δ_4
No Correction	0.04930	0.05043	0.04977	0.05031
HC3	0.04946	0.05080	0.05035	0.05121
IRLS	0.04978	0.05100	0.05041	0.05065
Null Gaussian Correction	0.04933	0.04991	0.04958	0.04986
Alternative Gaussian Correction	0.05018	0.05087	0.05057	0.05099
Null Binomial Correction	0.04919	0.05000	0.04972	0.05045
Alternative Binomial Correction	0.05062	0.05139	0.05087	**0.05189**

TABLE 2.3: Type I Error Rates in the Binomial Case

corresponding to the covariance matrix $\widehat{\Phi}_3$, our proposed iteratively reweighted least squares (IRLS) procedure and our entire correction framework have no significant impact on that type I error control. It is also important to note that there inherently exists a severe model misspecification in our fitted models for this joint association analysis, since only one X_j is taken into account at a time, ignoring the significant additive effects of the rest on the simulated quantitative trait.

2.8.7 Binomial Power Study

We first simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0.1, 0.9]$. Then, we set the sample size n to be $1,000$ individuals and we simulate independent pathogen genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ for $i = 1, 2, \ldots, n$. We set the number of host genetic variants m_h to be equal to 4. We simulate a host allele frequency $f_X \sim \text{Unif}[0.1, 0.9]$ and independent host genotypes $X_{ij} \sim \text{Binomial}\,(2, f_X)$ for $j = 1, 2, \ldots, m_h$. We set the parameter values $\beta = \left(0, \sqrt{0.025}, -\sqrt{0.025}, \sqrt{0.05}\right)$, $\gamma = \sqrt{0.025}$, $\delta = \left(0, 0, 0, \sqrt{0.025}\right)$ and $\sigma_\varepsilon^2 = 1 - \|\beta\|^2 - \gamma^2 - \|\delta\|^2$, so that the total variance of the simulated trait is equal to 1 and the proportion of the total variation explained by each coefficient is prespecified. Lastly, we simulate the following quantitative trait values:

$$Y_i = \sum_{j=1}^{4} \beta_j \widetilde{X}_{ij} + \gamma \widetilde{Z}_i + \sum_{j=1}^{4} \delta_j \widetilde{X}_{ij} \widetilde{Z}_i + \varepsilon_i, \quad \widetilde{X}_{ij} = \frac{X_{ij} - 2f_{X_j}}{\sqrt{2f_{X_j}\left(1 - f_{X_j}\right)}}, \quad \widetilde{Z}_i = \frac{Z_i - f_Z}{\sqrt{f_Z\left(1 - f_Z\right)}},$$

where $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ are independent. We observe that there is one true interaction effect between the pathogen genetic variant Z and the host genetic variant X_4, meaning that the quantitative

trait Y is going to be heteroscedastic when that interaction effect is not accounted for. We repeat this simulation $100,000$ times. Type I error rate calculations at significance level 0.05 and power calculations at significance level 10^{-5} are shown in Table 2.4. Type I errors which are significantly different from the nominal level of 0.05 at level 0.01 are displayed in bold text.

	Type I Error			Power
	δ_1	δ_2	δ_3	δ_4
No Correction	**0.05581**	**0.05559**	**0.05469**	0.78594
HC3	0.05162	0.05067	0.05035	0.77295
IRLS	0.05171	0.05123	0.05014	0.78508
Null Gaussian Correction	0.05071	0.05018	0.04917	0.74506
Alternative Gaussian Correction	**0.05181**	0.05134	0.05015	0.78003
Null Binomial Correction	0.05128	0.05054	0.04935	0.74621
Alternative Binomial Correction	**0.05250**	0.05174	0.05081	0.78871

TABLE 2.4: Type I Error Rates and Power in the Binomial Case

We observe that the uncorrected interaction t test statistic displays significantly overinflated type I error rates across the board in the presence of heteroscedasticity in the quantitative trait. Blanket use of the heteroscedasticity correction methods described in Section 2.3 manage to attain correct type I error control. At the same time, we note that our proposed IRLS procedure achieves slightly higher power than the existing HC3 covariance matrix approach, while also maintaining almost the same power as that of the significantly overinflated uncorrected interaction t test statistic. Our proposed null and alternative correction frameworks also attain better type I error control, with the alternative corrections consistently exhibiting slightly more overinflated type I error rates and higher power than the null corrections, as well as comparable power to the severely overinflated uncorrected interaction test statistic. Again, we note that there inherently exists an even more severe model misspecification in our fitted models for this joint association analysis, since only one X_j is taken into account at a time, ignoring the significant additive effects of X_2, X_3, X_4 as well as the significant interaction effect between X_4 and Z on the simulated quantitative trait.

2.8.8 Binomial Feast or Famine Study

We first simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0.1, 0.9]$. Then, we set the sample size n to be $1,000$ individuals and we simulate independent pathogen genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ for $i = 1, 2, \ldots, n$. We set the number of host genetic variants m_h to be equal to $10,000$. We simulate independent host allele frequencies $f_{X_1}, f_{X_2}, \ldots, f_{X_{m_h}} \sim \text{Unif}[0.1, 0.9]$ and independent host genotypes $X_{ij} \sim \text{Binomial}\left(2, f_{X_j}\right)$ for $j = 1, 2, \ldots, m_h$. Lastly, we simulate the following quantitative trait values $Y_i \sim \mathcal{N}(0, 1)$ under the global null hypothesis of no interaction. We repeat this simulation $m_v = 1,000$ times. At the same time as performing the pertinent joint GWAS testing for interactions between each possible pair of simulated pathogen and host genetic variants on the quantitative trait, we also perform a marginal GWAS of the simulated host genetic variants on the quantitative trait.

For the collection of m_h interaction test statistics corresponding to each simulation replicate, we calculate the same list of diagnostic quantities relating to the feast or famine effect. In general, the behavior of the feast or famine effect in the binomial setting is similar to the Gaussian setting in terms of the uncorrected interaction test statistic. However, we can also evaluate the performance of our proposed binomial correction frameworks compared to that of the Gaussian correction frameworks in this setting. First, we look at how the distribution of genomic control inflation factors around 1 based on the joint GWAS compares against the one based on the marginal GWAS, shown in Figure 2.14. We indeed observe a staggering amount of variation in the joint GWAS genomic control inflation factors compared to what one would normally expect in an ordinary marginal GWAS. On the other hand, the distribution of the genomic control inflation factors based on any of our proposed correction frameworks closely matches the one based on the marginal GWAS, as shown in Figure 2.15.

Then, we can take a look at the Q-Q plots of the interaction p-values corresponding to the pathogen genetic variants with the largest and smallest uncorrected genomic control inflation factors, shown in Figure 2.16. The deviation of these collections of interaction p-values from uniformity is astounding, even though this simulation is performed under the global null hypothesis of no interaction. More specifically, the left Q-Q plot represents the potential extremity of the feast effect in joint

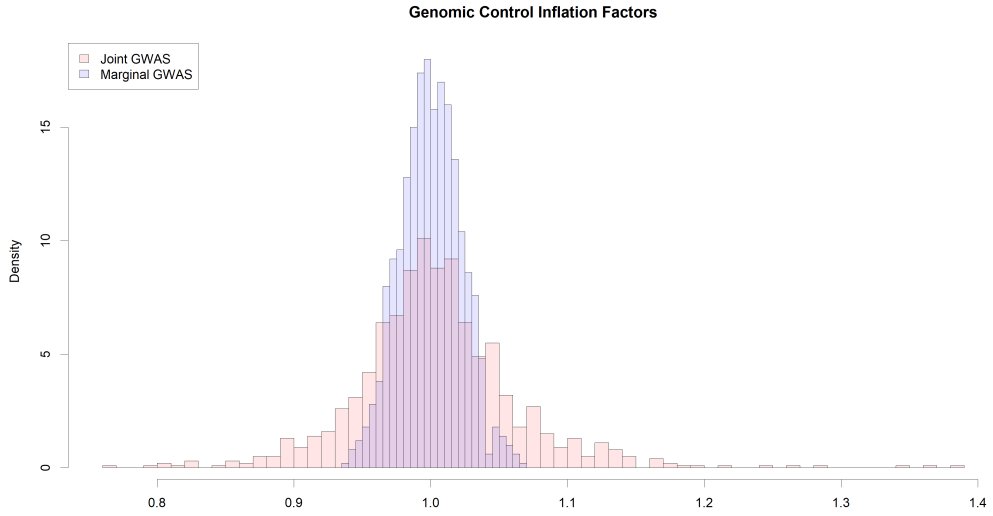

FIGURE 2.14: Histograms of Uncorrected Joint vs. Marginal Genomic Control Inflation Factors in the Binomial Case

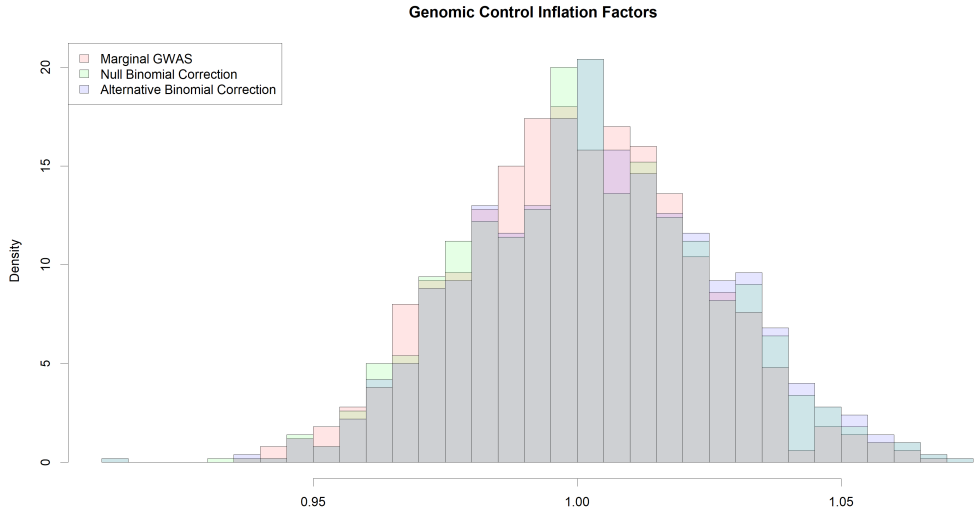

FIGURE 2.15: Histograms of Corrected Joint vs. Marginal Genomic Control Inflation Factors in the Binomial Case

association analyses, where an unbelievable amount of false discoveries would be committed by using the uncorrected interaction t test statistic, while the right Q-Q plot represents the potential extremity of the famine effect, where there would never be enough power to detect important interaction signals. For reference, we compare these Q-Q plots against the Q-Q plots of the p-values from the marginal association analyses corresponding to the largest and smallest marginal genomic

control inflation factors. We note that even the most extreme marginal association p-value distributions are essentially indistinguishable from the uniform distribution, since they do not necessarily coincide with the marginal association p-value distributions with the smallest Q-Q plot p-values. In comparison, the distributions of our corrected interaction p-values are also indistinguishable from uniform in these cases, as shown in Figure 2.17, indicating that our correction frameworks manage to correct for the extreme nature of both the feast as well as the famine effect.

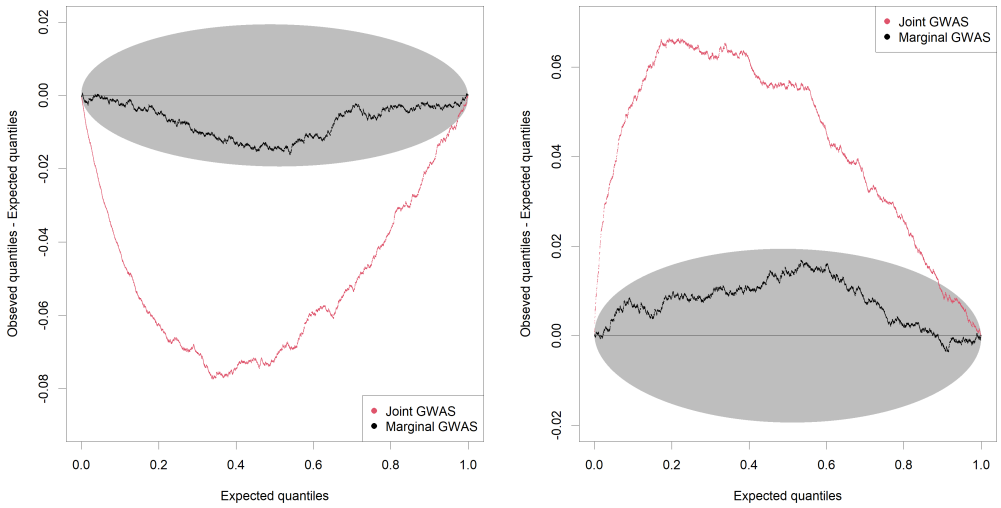

FIGURE 2.16: Q-Q Plots Displaying the Feast or Famine Effect in the Binomial Case

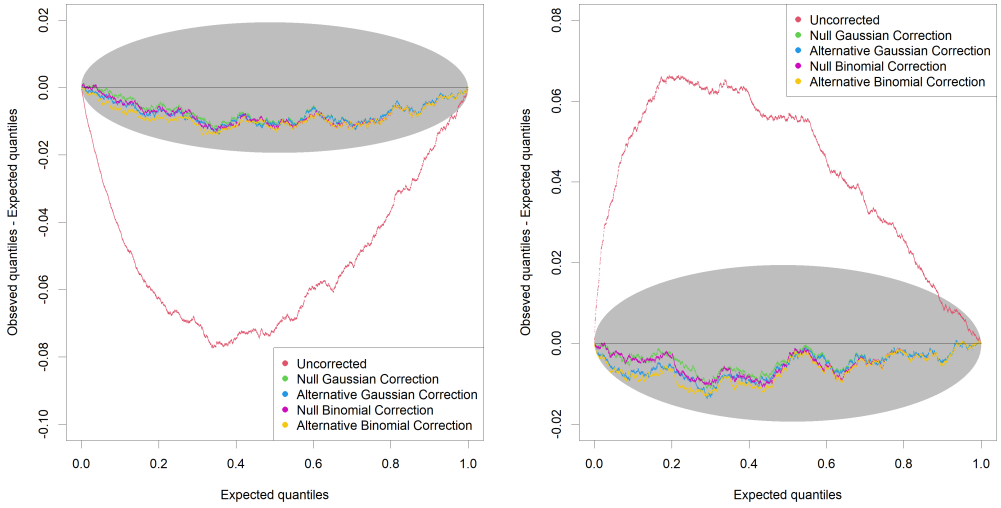

FIGURE 2.17: Q-Q Plots Displaying the Correction of the Feast or Famine Effect in the Binomial Case

Then, we look at the Q-Q plots of the 2-sided Q-Q plot p-values before and after correction, as

well as the the Q-Q plot of the 2-sided Q-Q plot p-values corresponding to the marginal GWAS, all displayed in Figure 2.18. It should be noted that this Q-Q plot is fundamentally different compared to the previously discussed Q-Q plots which were Q-Q plots of actual interaction and association p-values. These Q-Q plots essentially serve as a meta-analysis for our collection of joint association analyses corresponding to different quantitative trait and pathogen genetic variant pairs, hence we sometimes refer to them as meta Q-Q plots. We observe that the uncorrected Q-Q plot p-values tend to be much smaller than what one would expect under the uniform distribution. On the other hand, the distribution of the Q-Q plot p-values based on our null correction frameworks are practically indistinguishable from the uniform distribution and closely match those corresponding to the marginal GWAS, which implies that the null correction performs perfectly in terms of correcting the feast or famine effect. The alternative binomial correction framework performs much better than the uncorrected interaction t test statistic in terms of the feast or famine effect, but consistently displays smaller than uniform Q-Q p-values. The alternative Gaussian correction framework performs much better than both the uncorrected interaction t test statistic and the alternative binomial correction in terms of the feast or famine effect, only displaying slightly smaller than uniform Q-Q plot p-values

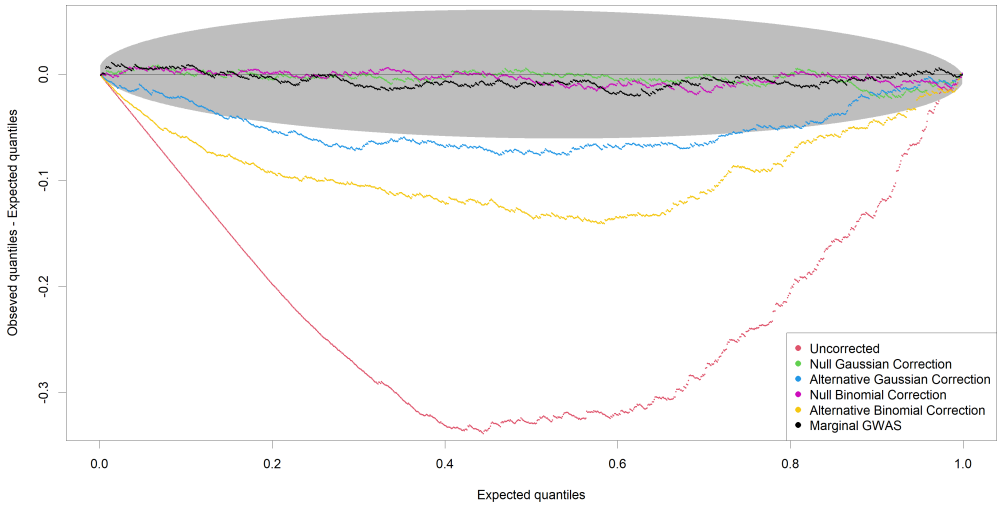

FIGURE 2.18: Comparison of Q-Q Plots of Uncorrected, Corrected and Marginal 2-Sided Q-Q Plot P-Values in the Binomial Case

Finally, we evaluate the performance of our proposed diagnostic ratio. The behavior of the diag-

nostic ratio in the binomial setting is practically identical to that of the Gaussian setting, but we also present the same diagnostic plots in this case, which is of much greater interest in the context of testing for interactions in joint GWAS settings. Plotting histograms of the diagnostic ratio and the uncorrected genomic control inflation factors on top of each other - Figure 2.19 - reveals that the 2 distributions closely match each other, even though the distribution of the diagnostic ratio displays slightly lighter tail behavior. A scatterplot of the uncorrected genomic control inflation factors against the diagnostic ratio, shown in Figure 2.20, verifies the strong linear relationship between them. We note that the sample correlation between these 2 quantities is calculated to be 92.03%.

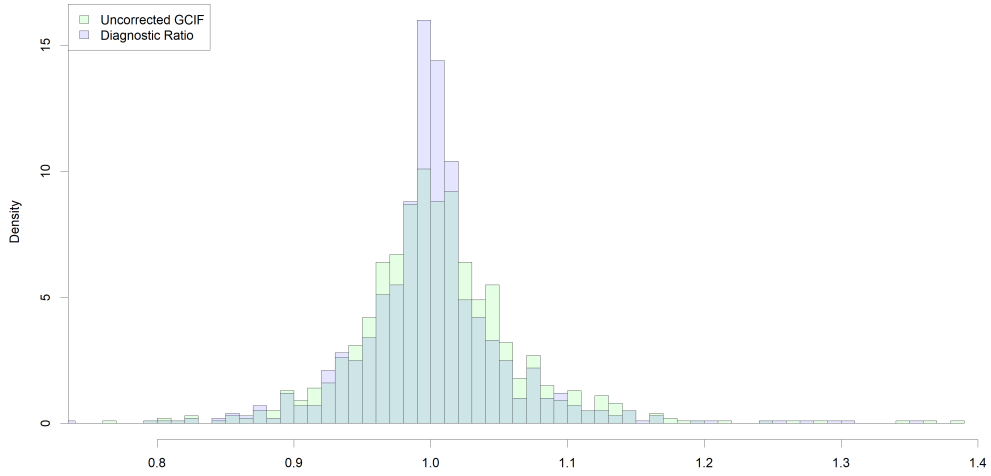

FIGURE 2.19: Histograms of Uncorrected Genomic Control Inflation Factors vs. Diagnostic Ratio in the Binomial Case

When performing a GWAS, rather than being interested in the median result, we are particularly interested in the behavior of the smallest p-values. Therefore, we also consider the ability of our diagnostic ratio to predict the tail behavior of the p-values. To do this, we calculate the 5% sample quantiles for each collection of uncorrected interaction p-values $p_{1k}, p_{2k}, \ldots, p_{m_h k}$ and plot them on the $-\log_{10}$ scale against the diagnostic ratio, shown in Figure 2.21. We observe that the diagnostic ratio performs fantastically in predicting the behavior of the smallest uncorrected interaction p-values. As a reference, the sample correlation between these 2 quantities was calculated to be equal to 94.51%.

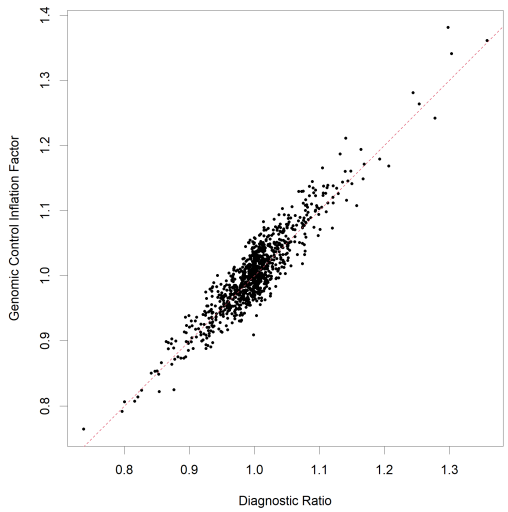

FIGURE 2.20: Scatterplot of Uncorrected Genomic Control Inflation Factors vs. Diagnostic Ratio in the Binomial Case

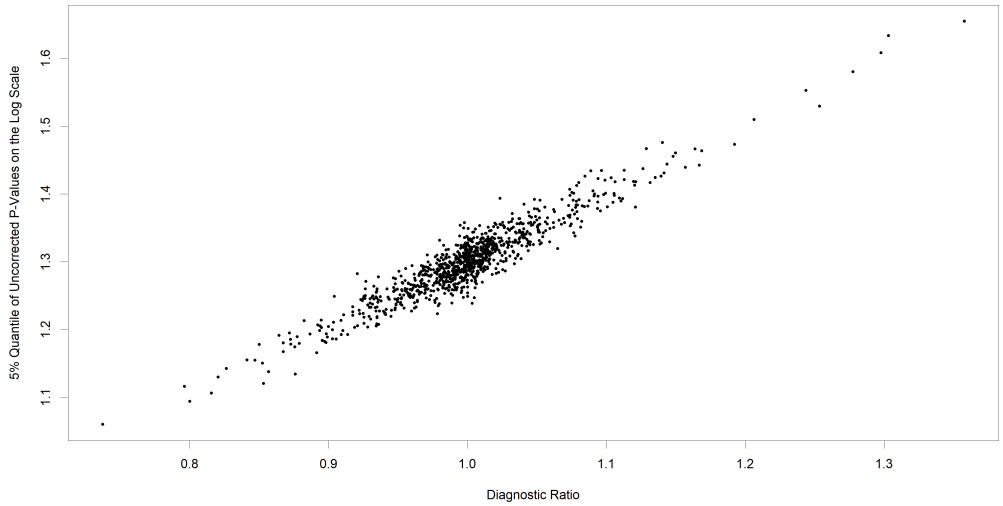

FIGURE 2.21: Scatterplot of 5% Quantiles of Uncorrected P-Values on the Log Scale vs. Diagnostic Ratio in the Binomial Case

2.8.9 Correlated Binomial Feast or Famine Study

We first set the number of host genetic variants m_h to be equal to $10,000$. We simulate a pathogen allele frequency $f_Z \sim \text{Unif}[0.1, 0.5]$, independent host allele frequencies $f_{X_1}, \ldots, f_{X_{m_h}} \sim \text{Unif}[0.1, 0.5]$ and independent correlation levels $\rho_1, \rho_1, \ldots, \rho_{100} \sim \text{Unif}[0, 0.1]$. Then, we set the sample size n to be $1,000$ individuals and $\rho_j = 0$ for $j = 101, 102, \ldots, m_h$. We simulate independent pathogen

genotypes $Z_i \sim \text{Bernoulli}\,(f_Z)$ and independent host genotypes $X_{ij} \sim \text{Binomial}\,(2, f_X)$ with a correlation of ρ_j between them for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_h$. Lastly, we simulate independent quantitative trait values $Y_i \sim \mathcal{N}(0, 1)$ under the global null hypothesis of no interaction. We repeat this simulation $1,000$ times.

In general, the behavior of the feast or famine effect in the correlated binomial setting is similar to the uncorrelated binomial setting in terms of the uncorrected interaction test statistic. The genomic control inflation factors based on any of our proposed correction frameworks are again much more tightly concentrated around 1 compared to those based on the uncorrected interaction test statistic, as shown in Figure 2.22. Looking at the Q-Q plots of the 2-sided Q-Q plot p-values before and after correction, displayed in Figure 2.23, we notice that our alternative correction frameworks perform much better in terms of ameliorating the feast or famine effect compared to the uncorrelated binomial setting. This happens because the tendency of the alternative correction frameworks to produce smaller than uniform Q-Q plot p-values is counteracted by the previously discussed underinflation in the type I error rates of the corrected interaction test statistic in the presence of tangible correlation between host and pathogen genetic variants.

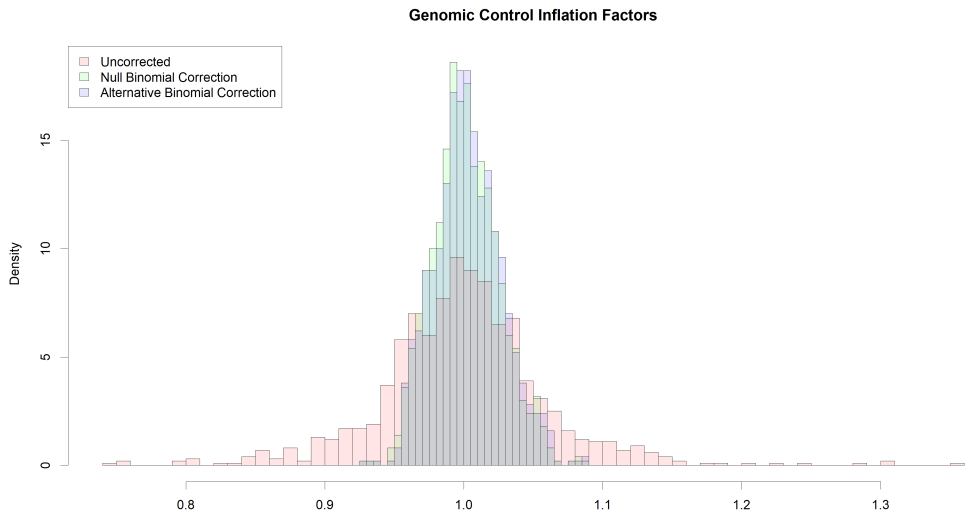

FIGURE 2.22: Histograms of Uncorrected vs. Corrected Genomic Control Inflation Factors in the Correlated Binomial Case

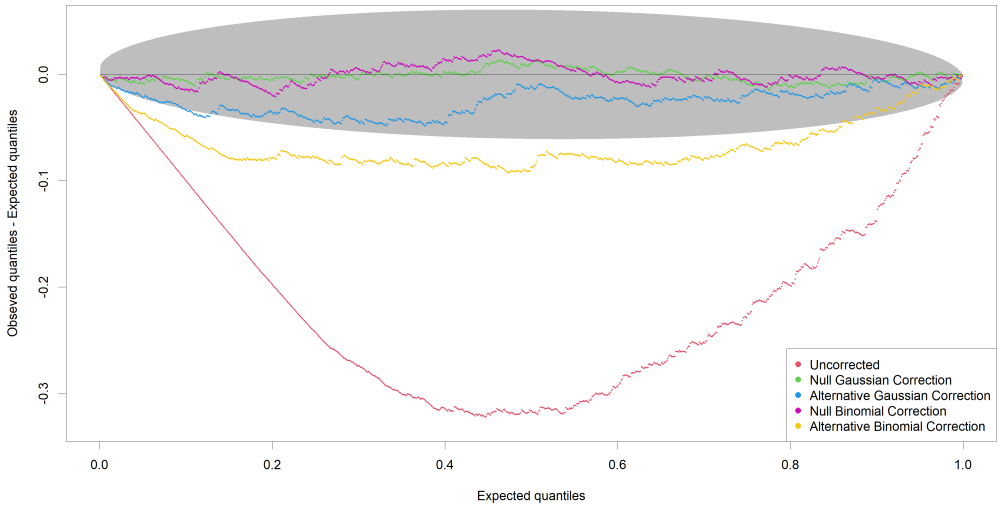

FIGURE 2.23: Comparison of Q-Q Plots of Uncorrected and Corrected 2-Sided Q-Q Plot P-Values in the Correlated Binomial Case

2.9 Discussion and Future Work

Identifying gene-gene and gene-environment interactions provides valuable insights into the genetic architecture on complex traits and underlying biological mechanisms. Integrating genomes from different organisms into a GWAS has the potential of revealing genomic regions indicative of co-evolution between species, especially in the case of pathosystems, where pathogens and hosts co-evolve to determine disease status. In the context of testing for interaction in GWAS settings, we have verified the existence of the "feast or famine" effect, which drives collections of interaction p-values corresponding to a fixed host or pathogen genetic variant to exhibit fundamentally different null distributions from the known null distribution of the employed interaction test statistic [12]. This effect applies to all kinds of host-pathogen genetic variants or environmental variables and affects standard testing methods such as the t test, the Wald test, the likelihood ratio test or the score test for interaction. We clarify that this phenomenon specifically impacts interaction testing in joint GWAS settings and not ordinary association testing in a marginal GWAS of just a single organism.

When considering simulated quantitative traits without any interaction effect between host and pathogen genetic variants, hence without any trait heteroscedasticity, standard interaction testing

methods maintain correct type I error rates overall. However, certain fixed pathogen genetic variants tend to consistently produce excess false discoveries, displaying the "feast" effect, while other fixed pathogen genetic variants tend to consistently produce false negative results, displaying the "famine" effect. The "feast or famine" effect invariably leads to excessive type I error rates, reduced power and association results which are impossible to replicate [12].

This feast or famine effect is an intrinsic property of the fixed quantitative trait and the fixed pathogen genetic variant in the conduction of interaction tests with all available host genetic variants and can be corrected by properly conditioning the interaction test statistic on the fixed quantitative trait and pathogen genetic variant pair instead of conditioning on the pair of genetic variants which are being tested for interaction [12]. This idea of changing the standard conditioning of the interaction test statistic in a joint GWAS led us to the development of a correction framework for the ordinary t test statistic in the context of testing for host-pathogen interaction effects. This framework appropriately incorporates important covariates and accounts for heteroscedasticity arising from latent interaction effects between one of the genetic variants being tested for interaction and some unaccounted factor. Multiple simulation studies demonstrate that the correction framework significantly ameliorates the feast or famine effect while preserving correct overall type I error and comparable power to the improperly calibrated interaction t test statistic.

Finally, we developed a diagnostic ratio which accurately predicts the prevalence of the feast or famine effect given a fixed quantitative trait and pathogen genetic variant pair. This diagnostic ratio is a natural byproduct of the procedure we utilize to adjust interaction t test statistic in the presence of the feast or famine effect, depends only on the fixed quantitative trait and pathogen genetic variant pair and is closely associated with the notion of the genomic control inflation factor for a collection of interaction t test statistics. Therefore, our proposed ratio constitutes a fast and interpretable diagnostic tool to predict the feast or famine effect in joint GWAS settings without the need to perform any arduous Monte Carlo experiments.

Our efforts so far have revolved around fixing a binary variable, which might represent a pathogen genetic variant or an environmental factor, and performing interaction tests between it and all other available genetic variants. However, our correction framework would also be broadly applicable in

studies for the detection of epistasis in diploid organisms, where all available genetic variants would be binomially distributed instead. The induced heteroscedasticity structure in the quantitative trait under the presence of latent interaction effects between genetic variants would be significantly more complex in a setting such as this. It is entirely possible that our current approaches for correcting heteroscedasticity in the trait might be too rudimentary in the presence of more complex heteroscedasticity structures, requiring the development of more sophisticated heteroscedasticity correction methods. More specifically, we think that construction of the correction framework on the basis of an interaction test statistic which already accounts for heteroscedasticity might be required, rather than solely addressing potential heteroscedasticity in the context of model fitting and parameter estimation.

Our proposed alternative correction frameworks are still not entirely correcting the feast or famine effect, which hinders attempts of employing a global testing procedure, such as ADELLE [56], to test whether there exists at least one significant interaction between a fixed pathogen genetic variant and any available host genetic variant. Since our correction framework would theoretically perform perfectly at correcting the feast or famine effect in the absence of the need for parameter estimation and it does normally perform better and better as the sample size increases, our understanding is that this imperfect correction of the feast or famine effect is mostly due to unreliable estimation of the interaction effect between genetic variants with fairly small minor allele frequencies by conventional fitting procedures. We are currently considering the incorporation of some form of regularization, such as Lasso regression, or some sort of shrinkage for the conditional probabilities of X_i given Y_i and Z_i towards the corresponding conditional probabilities given by the null correction framework. Although hard to interpret or generalize, we believe that some form of regularization or shrinkage might be able to address the parameter estimation issues of the alternative correction frameworks.

Testing for interactions between different organisms belonging to the same species, e.g. interactions between different strains of the same pathogen co-infecting the same host organism, or different genetic variants belonging to the same organism, i.e. epistasis detection, also invites other considerations. Electing to fix a pathogen genetic variant and test for interactions between it and all available host genetic variants makes sense in the context of host-pathogen GWAS settings for

reasons of computational efficiency and design of global testing procedures. However, selecting which variable to condition on is not at all obvious when the 2 collections of genetic variants to be tested for interactions are symmetric. Choosing which of the 2 genetic variants to condition on will probably highly depend on their observed minor allele frequencies, but further investigation is required to determine an appropriate approach. It might also be worth considering to not condition on any of them, instead of simply having to choose one of them to be conditioned on, giving rise to a completely new correction framework.

Testing for interaction between 2 genetic variants presents challenges of its own, even discounting the joint GWAS setting where the feast or famine effect is prevalent, because of the joint distribution of allele counts. We have focused on minor allele counts and the notion of a minimum cell count for the contingency table of the allele counts in order to establish some required thresholds to determine when a meaningful test of interaction can be performed between 2 genetic variants. However, we are currently considering the derivation of a more formal measure based on Fisher's information to determine when a pair of genetic variants contains sufficient information for the purposes of interaction testing.

2.10 Appendices

2.10.1 Appendix A

Let $Y = U\alpha + \beta X + \gamma Z + \delta W + \varepsilon$ with $W = \text{diag}(HZ)HX \in \mathbb{R}^n$ and $\varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 \mathbf{I}_n\right)$, where $H = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ and $\text{diag}(HZ) \in \mathbb{R}^{n\times n}$ is the diagonal matrix whose diagonal elements are given by the elements of the vector $HZ \in \mathbb{R}^n$. We define the following design matrices and corresponding projection matrices:

$$U \in \mathbb{R}^{n\times c}, \quad P = \mathbf{I}_n - U\left(U^{\mathrm{T}}U\right)^{-1}U^{\mathrm{T}},$$

$$U_X = \begin{bmatrix} U & X \end{bmatrix} \in \mathbb{R}^{n\times(c+1)}, \quad P_X = \mathbf{I}_n - U_X\left(U_X^{\mathrm{T}}U_X\right)^{-1}U_X^{\mathrm{T}},$$

$$U_{XZ} = \begin{bmatrix} U & X & Z \end{bmatrix} \in \mathbb{R}^{n\times(c+2)}, \quad P_{XZ} = \mathbf{I}_n - U_{XZ}\left(U_{XZ}^{\mathrm{T}}U_{XZ}\right)^{-1}U_{XZ}^{\mathrm{T}},$$

$$U_{XZW} = \begin{bmatrix} U & X & Z & W \end{bmatrix} \in \mathbb{R}^{n\times(c+3)}, \quad P_{XZW} = \mathbf{I}_n - U_{XZW}\left(U_{XZW}^{\mathrm{T}}U_{XZW}\right)^{-1}U_{XZW}^{\mathrm{T}}.$$

Then, the least squares estimator of $\vartheta = (\alpha, \beta, \gamma, \delta)^{\mathrm{T}} \in \mathbb{R}^{c+3}$ is given by:

$$\begin{aligned}
\widehat{\vartheta} &= \left(U_{XZW}^{\mathrm{T}} U_{XZW}\right)^{-1} U_{XZW}^{\mathrm{T}} Y = \begin{bmatrix} U_{XZ}^{\mathrm{T}} U_{XZ} & U_{XZ}^{\mathrm{T}} W \\ W^{\mathrm{T}} U_{XZ} & W^{\mathrm{T}} W \end{bmatrix}^{-1} \begin{bmatrix} U_{XZ}^{\mathrm{T}} \\ W^{\mathrm{T}} \end{bmatrix} Y \\
&= \frac{1}{W^{\mathrm{T}} P_{XZ} W} \begin{bmatrix} W^{\mathrm{T}} P_{XZ} W \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} + \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} U_{XZ}^{\mathrm{T}} W W^{\mathrm{T}} U_{XZ} \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} & -\left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} U_{XZ}^{\mathrm{T}} W \\ -W^{\mathrm{T}} U_{XZ} \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} & 1 \end{bmatrix} \begin{bmatrix} U_{XZ}^{\mathrm{T}} \\ W^{\mathrm{T}} \end{bmatrix} Y \\
&= \frac{1}{W^{\mathrm{T}} P_{XZ} W} \begin{bmatrix} W^{\mathrm{T}} P_{XZ} W \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} U_{XZ}^{\mathrm{T}} Y - W^{\mathrm{T}} P_{XZ} Y \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} U_{XZ}^{\mathrm{T}} W \\ W^{\mathrm{T}} Y - W^{\mathrm{T}} U_{XZ} \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} U_{XZ}^{\mathrm{T}} Y \end{bmatrix}.
\end{aligned}$$

Thus, the least squares estimator of the interaction effect δ is given by:

$$\widehat{\delta} = \frac{W^{\mathrm{T}} P_{XZ} Y}{W^{\mathrm{T}} P_{XZ} W}.$$

Next, we observe that:

$$\begin{aligned}
P_{XZW} &= \mathbf{I}_n - U_{XZ}^{\mathrm{T}} \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} U_{XZ}^{\mathrm{T}} + \frac{U_{XZ}^{\mathrm{T}} \left(U_{XZ}^{\mathrm{T}} U_{XZ}\right)^{-1} U_{XZ}^{\mathrm{T}} W W^{\mathrm{T}} P_{XZ} - W W^{\mathrm{T}} P_{XZ}}{W^{\mathrm{T}} P_{XZ} W} \\
&= P_{XZ} - \frac{P_{XZ} W W^{\mathrm{T}} P_{XZ}}{W^{\mathrm{T}} P_{XZ} W}.
\end{aligned}$$

Similarly, we infer that:

$$P_{XZ} = P_X - \frac{P_X Z Z^{\mathrm{T}} P_X}{Z^{\mathrm{T}} P_X Z}, \quad P_X = P - \frac{P X X^{\mathrm{T}} P}{X^{\mathrm{T}} P X}.$$

Hence, an unbiased estimator of the residual variance σ_ε^2 is given by:

$$S_\varepsilon^2 = \frac{Y^{\mathrm{T}} P_{XZW} Y}{n - c - 3} = \frac{W^{\mathrm{T}} P_{XZ} W Y^{\mathrm{T}} P_{XZ} Y - \left(W^{\mathrm{T}} P_{XZ} Y\right)^2}{(n - c - 3)\, W^{\mathrm{T}} P_{XZ} W}.$$

Therefore, the interaction t test statistic is given by:

$$T = \frac{\widehat{\delta}}{\frac{S_\varepsilon}{\sqrt{W^{\mathrm{T}} P_{XZ} W}}} = \frac{W^{\mathrm{T}} P_{XZ} Y}{\sqrt{\frac{W^{\mathrm{T}} P_{XZ} W Y^{\mathrm{T}} P_{XZ} Y - \left(W^{\mathrm{T}} P_{XZ} Y\right)^2}{n - c - 3}}}.$$

For any vectors $A, B \in \mathbb{R}^n$, we define $S'_{AB} = A^{\mathrm{T}} P B \in \mathbb{R}$. Then, we calculate that:

$$\begin{aligned} P_{XZ} &= P - \frac{PXX^{\mathrm{T}}P}{X^{\mathrm{T}}PX} - \frac{\left(P - \frac{PXX^{\mathrm{T}}P}{X^{\mathrm{T}}PX}\right) ZZ^{\mathrm{T}} \left(P - \frac{PXX^{\mathrm{T}}P}{X^{\mathrm{T}}PX}\right)}{Z^{\mathrm{T}} \left(P - \frac{PXX^{\mathrm{T}}P}{X^{\mathrm{T}}PX}\right) Z} \\ &= P - \frac{PXX^{\mathrm{T}}P}{S'_{XX}} - \frac{S'^2_{XX} PZZ^{\mathrm{T}}P - S'_{XX}S'_{XZ}P\left(XZ^{\mathrm{T}} + ZX^{\mathrm{T}}\right)P + S'^2_{XZ}PXX^{\mathrm{T}}P}{S'_{XX}\left(S'_{XX}S'_{ZZ} - S'^2_{XZ}\right)} \\ &= P - \frac{S'_{ZZ}PXX^{\mathrm{T}}P + S'_{XX}PZZ^{\mathrm{T}}P - S'_{XZ}P\left(XZ^{\mathrm{T}} + ZX^{\mathrm{T}}\right)P}{S'_{XX}S'_{ZZ} - S'^2_{XZ}}. \end{aligned}$$

Let $N = W^{\mathrm{T}} P_{XZ} Y$ denote the numerator of the interaction t test statistic. Then, we observe that:

$$\begin{aligned} N &= S'_{YW} - \frac{S'_{ZZ}S'_{XW}S'_{XY} + S'_{XX}S'_{ZW}S'_{YZ} - S'_{XZ}\left(S'_{XW}S'_{YZ} + S'_{ZW}S'_{XY}\right)}{S'_{XX}S'_{ZZ} - S'^2_{XZ}} \\ &= S'_{YW} - \frac{S'_{ZZ}S'_{XY} - S'_{XZ}S'_{YZ}}{S'_{XX}S'_{ZZ} - S'^2_{XZ}} S'_{XW} - \frac{S'_{XX}S'_{YZ} - S'_{XZ}S'_{XY}}{S'_{XX}S'_{ZZ} - S'^2_{XZ}} S'_{ZW}. \end{aligned}$$

Under the null hypothesis of no interaction effect, we observe that:

$$\widehat{\gamma}_0 = \frac{Z^{\mathrm{T}}P_X Y}{Z^{\mathrm{T}}P_X Z} = \frac{Z^{\mathrm{T}}PY - \frac{Z^{\mathrm{T}}PXX^{\mathrm{T}}PY}{X^{\mathrm{T}}PX}}{Z^{\mathrm{T}}PZ - \frac{Z^{\mathrm{T}}PXX^{\mathrm{T}}PZ}{X^{\mathrm{T}}PX}} = \frac{S'_{XX}S'_{YZ} - S'_{XZ}S'_{XY}}{S'_{XX}S'_{ZZ} - S'^2_{XZ}}.$$

By symmetry, we also infer that:

$$\widehat{\beta}_0 = \frac{S'_{ZZ}S'_{XY} - S'_{XZ}S'_{YZ}}{S'_{XX}S'_{ZZ} - S'^2_{XZ}}.$$

For $i, j, k, \ell = 1, 2, \ldots, n$, we define the following conditional moments of X given Y and Z:

$$\mu_{X|Y,Z} = \mathbb{E}\left(X \mid Y, Z\right), \quad \Sigma_{X|Y,Z} = \mathrm{Var}\left(X \mid Y, Z\right),$$

$$\gamma^{ijk}_{X|Y,Z} = \mathbb{E}\left[\left(X_i - \mu^i_{X|Y,Z}\right)\left(X_j - \mu^j_{X|Y,Z}\right)\left(X_k - \mu^k_{X|Y,Z}\right)\right],$$

$$K^{ijk\ell}_{X|Y,Z} = \mathbb{E}\left[\left(X_i - \mu^i_{X|Y,Z}\right)\left(X_j - \mu^j_{X|Y,Z}\right)\left(X_k - \mu^k_{X|Y,Z}\right)\left(X_\ell - \mu^\ell_{X|Y,Z}\right)\right].$$

According to the strong law of large numbers, we know that:

$$\frac{1}{n}S'_{XY} \sim \frac{1}{n}Y^{\mathrm{T}}P\mu_{X|Y,Z} = \frac{1}{n}S'_{\mu_{X|Y,Z}Y}, \quad \frac{1}{n}S'_{XZ} \sim \frac{1}{n}Z^{\mathrm{T}}P\mu_{X|Y,Z} = \frac{1}{n}S'_{\mu_{X|Y,Z}Z},$$

$$\frac{1}{n}S'_{XX} \sim \frac{1}{n}\left[\mu_{X|Y,Z}^{\mathrm{T}} P \mu_{X|Y,Z} + \mathrm{tr}\left(P\Sigma_{X|Y,Z}\right)\right] = \frac{1}{n}\left[S'_{\mu_{X|Y,Z}\mu_{X|Y,Z}} + \mathrm{tr}\left(P\Sigma_{X|Y,Z}\right)\right].$$

Hence, we infer that:

$$\widehat{\beta}_0 \sim \beta_* = \frac{S'_{ZZ}S'_{\mu_{X|Y,Z}Y} - S'_{\mu_{X|Y,Z}Z}S'_{YZ}}{\left[S'_{\mu_{X|Y,Z}\mu_{X|Y,Z}} + \mathrm{tr}\left(P\Sigma_{X|Y,Z}\right)\right]S'_{ZZ} - S'^2_{\mu_{X|Y,Z}Z}},$$

$$\widehat{\gamma}_0 \sim \gamma_* = \frac{\left[S'_{\mu_{X|Y,Z}\mu_{X|Y,Z}} + \mathrm{tr}\left(P\Sigma_{X|Y,Z}\right)\right]S'_{YZ} - S'_{\mu_{X|Y,Z}Z}S'_{\mu_{X|Y,Z}Y}}{\left[S'_{\mu_{X|Y,Z}\mu_{X|Y,Z}} + \mathrm{tr}\left(P\Sigma_{X|Y,Z}\right)\right]S'_{ZZ} - S'^2_{\mu_{X|Y,Z}Z}}.$$

Thus, an asymptotic approximation of the numerator of the interaction t test statistic is given by:

$$N_* = S'_{YW} - \beta_* S'_{XW} - \gamma_* S'_{ZW} = (Y - \gamma_* Z)^{\mathrm{T}} P\mathrm{diag}(HZ)HX - \beta_* X^{\mathrm{T}} H\mathrm{diag}(HZ)PX.$$

Let $Q = H\mathrm{diag}(HZ)P \in \mathbb{R}^{n\times n}$ and $r = Y - \gamma_* Z \in \mathbb{R}^n$. Then, we observe that:

$$Q = \frac{Q+Q^{\mathrm{T}}}{2} + \frac{Q-Q^{\mathrm{T}}}{2}, \quad X^{\mathrm{T}}Q^{\mathrm{T}}X = X^{\mathrm{T}}QX \in \mathbb{R}.$$

Hence, we define $Q_S = \frac{Q+Q^{\mathrm{T}}}{2}$ and conclude that:

$$N_* = r^{\mathrm{T}}Q^{\mathrm{T}}X - \beta_* X^{\mathrm{T}}Q_S X.$$

In the case where $U = \mathbf{1}_n$, note that $P = H$, so it follows that $Q = Q^{\mathrm{T}} = Q_S = H\mathrm{diag}\left(HZ\right)H$. Next, we calculate that:

$$\mathbb{E}\left(N_* \mid Y, Z\right) = r^{\mathrm{T}}Q^{\mathrm{T}}\mu_{X|Y,Z} - \beta_*\left[\mu_{X|Y,Z}^{\mathrm{T}} Q \mu_{X|Y,Z} + \mathrm{tr}\left(Q_S\Sigma_{X|Y,Z}\right)\right],$$

$$\mathrm{Var}\left(r^{\mathrm{T}}Q^{\mathrm{T}}X\right) = r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Qr,$$

$$\begin{aligned}\mathrm{Cov}\left(X^{\mathrm{T}}Q_S X, r^{\mathrm{T}}Q^{\mathrm{T}}X\right) &= \sum_{i,j,k=1}^{n} Q_S^{ij}\left(Qr\right)^k\left(\gamma_{X|Y,Z}^{ijk} + \Sigma_{X|Y,Z}^{ik}\mu_{X|Y,Z}^{j} + \Sigma_{X|Y,Z}^{jk}\mu_{X|Y,Z}^{i}\right) \\ &= \sum_{i,j,k=1}^{n} Q_S^{ij}\left(Qr\right)^k\gamma_{X|Y,Z}^{ijk} + 2r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z},\end{aligned}$$

$$\begin{aligned}\operatorname{Var}\left(X^{\mathrm{T}}Q_SX\right) &= \sum_{i,j,k,\ell=1}^{n} Q_S^{ij}Q_S^{k\ell}\left(K_{X|Y,Z}^{ijk\ell} + 4\gamma_{X|Y,Z}^{ijk}\mu_{X|Y,Z}^{\ell} + 4\Sigma_{X|Y,Z}^{ik}\mu_{X|Y,Z}^{j}\mu_{X|Y,Z}^{\ell} - \Sigma_{X|Y,Z}^{ij}\Sigma_{X|Y,Z}^{k\ell}\right) \\ &= \sum_{i,j,k,\ell=1}^{n} Q_S^{ij}Q_S^{k\ell}\left(K_{X|Y,Z}^{ijk\ell} + 4\gamma_{X|Y,Z}^{ijk}\mu_{X|Y,Z}^{\ell}\right) + 4\mu_{X|Y,Z}^{\mathrm{T}}Q_S\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} - \left[\operatorname{tr}\left(Q_S\Sigma_{X|Y,Z}\right)\right]^2.\end{aligned}$$

Therefore, we conclude that:

$$\begin{aligned}\operatorname{Var}\left(N_* \mid Y,Z\right) = r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Qr - 2\beta_* \sum_{i,j,k=1}^{n} Q_S^{ij}\left(Qr\right)^k\gamma_{X|Y,Z}^{ijk} - 4\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} \\ + \beta_*^2 \sum_{i,j,k,\ell=1}^{n} Q_S^{ij}Q_S^{k\ell}\left(K_{X|Y,Z}^{ijk\ell} + 4\gamma_{X|Y,Z}^{ijk}\mu_{X|Y,Z}^{\ell}\right) \\ + 4\beta_*^2\mu_{X|Y,Z}^{\mathrm{T}}Q_S\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} - \beta_*^2\left[\operatorname{tr}\left(Q_S\Sigma_{X|Y,Z}\right)\right]^2.\end{aligned}$$

Gaussian Correction

Suppose that $\left(X_i \mid Z_i = z\right) \sim \mathcal{N}\left(\mu_{X|z}, \sigma_X^2\right)$ are independent for $i = 1, 2, \ldots, n$. Under the null hypothesis of no interaction effect, we assume that $Y_i = U_i^{\mathrm{T}}\alpha + \beta X_i + \gamma Z_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. According to Bayes' theorem, we calculate that:

$$\begin{aligned}f_{X_i|Y_i,Z_i}(x \mid y,z) &\propto f_{X_i|Z_i}(x \mid z) f_{Y_i|X_i,Z_i}(y \mid x,z) \\ &\propto \exp\left\{-\frac{1}{2\sigma_X^2}(x-\mu_{X|z})^2\right\}\exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(y-U_i^{\mathrm{T}}\alpha-\beta x-\gamma z)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma_X^2}x^2 + \frac{\mu_{X|z}}{\sigma_X^2}x - \frac{\beta^2}{2\sigma_\varepsilon^2}x^2 + \frac{(y-U_i^{\mathrm{T}}\alpha-\gamma z)\beta}{\sigma_\varepsilon^2}x\right\} \\ &= \exp\left\{-\frac{\beta^2\sigma_X^2+\sigma_\varepsilon^2}{2\sigma_X^2\sigma_\varepsilon^2}x^2 + \frac{(y-U_i^{\mathrm{T}}\alpha-\gamma z)\beta\sigma_X^2+\mu_{X|z}\sigma_\varepsilon^2}{\sigma_X^2\sigma_\varepsilon^2}x\right\} \\ &= \exp\left\{-\frac{\beta^2\sigma_X^2+\sigma_\varepsilon^2}{2\sigma_X^2\sigma_\varepsilon^2}\left[x^2 - 2\frac{(y-U_i^{\mathrm{T}}\alpha-\beta\mu_{X|z}-\gamma z)\beta\sigma_X^2+\left(\beta^2\sigma_X^2+\sigma_\varepsilon^2\right)\mu_{X|z}}{\beta^2\sigma_X^2+\sigma_\varepsilon^2}x\right]\right\}.\end{aligned}$$

Therefore, we infer that:

$$\left(X_i \mid Y_i = y, Z_i = z\right) \sim \mathcal{N}\left(\frac{y-U_i^{\mathrm{T}}\alpha-\beta\mu_{X|z}-\gamma z}{\beta^2\sigma_X^2+\sigma_\varepsilon^2}\beta\sigma_X^2 + \mu_{X|z}, \sigma_X^2 - \frac{\beta^2\sigma_X^4}{\beta^2\sigma_X^2+\sigma_\varepsilon^2}\right).$$

Under the alternative hypothesis, we assume that $Y_i = U_i^{\mathrm{T}}\alpha + \beta X_i + \gamma Z_i + \delta(X_i - \mu_X)(Z_i - \mu_Z) + \varepsilon_i$ with $\mu_X = \mathbb{E}\left(X_i\right)$, $\mu_Z = \mathbb{E}\left(Z_i\right)$ and $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. According to Bayes' theorem, we calculate

that:

$$f_{X_i|Y_i,Z_i}(x \mid y,z) \propto \exp\left\{-\frac{1}{2\sigma_X^2}x^2 + \frac{\mu_{X|z}}{\sigma_X^2}x - \frac{\left[\beta+\delta(z-\mu_Z)\right]^2}{2\sigma_\varepsilon^2}x^2 + \frac{\left[y-U_i^{\mathrm{T}}\alpha-\gamma z+\delta\mu_X(z-\mu_Z)\right]\left[\beta+\delta(z-\mu_Z)\right]}{\sigma_\varepsilon^2}x\right\}$$
$$= \exp\left\{-\frac{\left[\beta+\delta(z-\mu_Z)\right]^2\sigma_X^2+\sigma_\varepsilon^2}{2\sigma_X^2\sigma_\varepsilon^2}x^2 + \frac{\left[y-U_i^{\mathrm{T}}\alpha-\gamma z+\delta\mu_X(z-\mu_Z)\right]\left[\beta+\delta(z-\mu_Z)\right]\sigma_X^2+\mu_{X|z}\sigma_\varepsilon^2}{\sigma_X^2\sigma_\varepsilon^2}x\right\}.$$

Therefore, we infer that:

$$(X_i \mid Y_i=y, Z_i=z) \sim \mathcal{N}\left(\frac{\left[y-U_i^{\mathrm{T}}\alpha-\beta\mu_{X|z}-\gamma z-\delta\left(\mu_{X|z}-\mu_X\right)(z-\mu_Z)\right]\left[\beta+\delta(z-\mu_Z)\right]\sigma_X^2}{\left[\beta+\delta(z-\mu_Z)\right]^2\sigma_X^2+\sigma_\varepsilon^2}+\mu_{X|z}, \sigma_X^2-\frac{\left[\beta+\delta(z-\mu_Z)\right]^2\sigma_X^4}{\left[\beta+\delta(z-\mu_Z)\right]^2\sigma_X^2+\sigma_\varepsilon^2}\right).$$

For $i,j,k,\ell = 1,2,\ldots,n$, we know that $K^{ijk\ell}_{X|Y,Z} = \Sigma^{ij}_{X|Y,Z}\Sigma^{k\ell}_{X|Y,Z} + \Sigma^{ik}_{X|Y,Z}\Sigma^{j\ell}_{X|Y,Z} + \Sigma^{i\ell}_{X|Y,Z}\Sigma^{jk}_{X|Y,Z}$ and $\gamma^{ijk}_{X|Y,Z} = 0$. Therefore, we conclude that:

$$\begin{aligned}\mathrm{Var}\left(N_* \mid Y,Z\right) &= r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Qr - 4\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} \\ &\quad + \beta_*^2\sum_{i,j,k,\ell=1}^{n} Q_S^{ij}Q_S^{k\ell}\left(\Sigma^{ij}_{X|Y,Z}\Sigma^{k\ell}_{X|Y,Z} + \Sigma^{ik}_{X|Y,Z}\Sigma^{j\ell}_{X|Y,Z} + \Sigma^{i\ell}_{X|Y,Z}\Sigma^{jk}_{X|Y,Z}\right) \\ &\quad + 4\beta_*^2\mu^{\mathrm{T}}_{X|Y,Z}Q_S\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} - \beta_*^2\left[\mathrm{tr}\left(Q_S\Sigma_{X|Y,Z}\right)\right]^2 \\ &= r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Qr - 4\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} \\ &\quad + 2\beta_*^2\mathrm{tr}\left(Q_S\Sigma_{X|Y,Z}Q_S\Sigma_{X|Y,Z}\right) + 4\beta_*^2\mu^{\mathrm{T}}_{X|Y,Z}Q_S\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z}.\end{aligned}$$

Discrete Correction

Suppose that $(X_i \mid Z_i = z)$ are independent and follow some discrete distribution with support $S \subseteq \mathbb{N}$ for $i = 1,2,\ldots,n$. Then, we know that:

$$p^i(x \mid y,z) = \mathbb{P}\left(X_i = x \mid Y_i = y, Z_i = z\right) \propto \mathbb{P}\left(X_i = x \mid Z_i = z\right) f_{Y_i|X_i,Z_i}\left(y \mid x,z\right),$$

$$\mu_r^i(y,z) = \mathbb{E}\left(X_i^r \mid Y_i = y, Z_i = z\right) = \sum_{x\in S} p^i(x \mid y,z)x^r,$$

$$\mu^i_{X|Y,Z} = \mu_1^i(y,z), \quad \Sigma^{ii}_{X|Y,Z} = \mu_2^i(y,z) - \left[\mu_1^i(y,z)\right]^2,$$

$$\gamma^{iii}_{X|Y,Z} = \mu_3^i(y,z) - 3\mu_1^i(y,z)\mu_2^i(y,z) - \left[\mu_1^i(y,z)\right]^3,$$

$$K^{iiii}_{X|Y,Z} = \mu^i_4(y,z) - 4\mu^i_1(y,z)\mu^i_3(y,z) - 6\left[\mu^i_1(y,z)\right]^2\mu^i_2(y,z) - \left[\mu^i_1(y,z)\right]^4, \quad K^{iijj}_{X|Y,Z} = \Sigma^{ii}_{X|Y,Z}\Sigma^{jj}_{X|Y,Z}.$$

Let $\gamma_{X|Y,Z} = \left[\gamma^{iii}_{X|Y,Z}\right]_{i=1}^n \in \mathbb{R}^n$ and $K_{X|Y,Z} = \left[K^{iijj}_{X|Y,Z}\right]_{i,j=1}^n \in \mathbb{R}^{n\times n}$. All other conditional central moments of X given Y and Z are equal to 0. Then, we conclude that:

$$\begin{aligned}\mathrm{Var}\left(N_* \mid Y,Z\right) = {} & r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Qr - 2\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\mathrm{Dg}\left(Q_S\right)\gamma_{X|Y,Z} - 4\beta_* r^{\mathrm{T}}Q^{\mathrm{T}}\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z} \\ & + 2\beta_*^2\mathrm{tr}\left(Q_S\Sigma_{X|Y,Z}Q_S\Sigma_{X|Y,Z}\right) \\ & + \beta_*^2\left[\mathrm{diag}\left(Q_S\right)\right]^{\mathrm{T}}\left[\mathrm{Dg}\left(K_{X|Y,Z}\right) - 3\Sigma^2_{X|Y,Z}\right]\mathrm{diag}\left(Q_S\right) \\ & + 4\beta_*^2\mu^{\mathrm{T}}_{X|Y,Z}Q_S\mathrm{Dg}\left(Q_S\right)\gamma_{X|Y,Z} + 4\beta_*^2\mu^{\mathrm{T}}_{X|Y,Z}Q_S\Sigma_{X|Y,Z}Q_S\mu_{X|Y,Z},\end{aligned}$$

where $\mathrm{Dg}\left(Q_S\right) \in \mathbb{R}^{n\times n}$ is the diagonal matrix whose diagonal elements are given by the diagonal elements of $Q_S \in \mathbb{R}^{n\times n}$ and $\mathrm{diag}\left(Q_S\right) \in \mathbb{R}^n$ is the vector whose elements are given by the diagonal elements of Q_S.

2.10.2 Appendix B

Suppose that $X \perp\!\!\!\perp (Y,Z)$. Then, we observe that $\mu_{X|Y,Z} = \mu_X\mathbf{1}_n$ and $\Sigma_{X|Y,Z} = \sigma^2_X\mathbf{I}_n$, which implies that $S'_{\mu_{X|Y,Z}A} = \mu_X\mathbf{1}^{\mathrm{T}}_nPA = 0$ for any vector $A \in \mathbb{R}^n$. Hence, we infer that $\beta_* = 0$ and $\gamma_* = \frac{S'_{YZ}}{S'_{ZZ}}$. For any vectors $A, B \in \mathbb{R}^n$, we define $S'_{ABZZ} = A^{\mathrm{T}}P\mathrm{Dg}\left(HZZ^{\mathrm{T}}H\right)PB \in \mathbb{R}$. Since $HP = PH = P$, we observe that:

$$P\mathrm{diag}\left(HZ\right)H\mathrm{diag}\left(HZ\right)P = P\left[\mathrm{Dg}\left(HZZ^{\mathrm{T}}H\right) - \frac{1}{n}ZZ^{\mathrm{T}}\right]P.$$

Therefore, we conclude that:

$$\begin{aligned}\mathrm{Var}\left(N_* \mid Y,Z\right) &= \sigma^2_X r^{\mathrm{T}}Q^{\mathrm{T}}Qr \\ &= \sigma^2_X\left[\left(S'_{YYZZ} - \frac{1}{n}S'^2_{YZ}\right) - 2\frac{S'_{YZ}}{S'_{ZZ}}\left(S'_{YZZZ} - \frac{1}{n}S'_{YZ}S'_{ZZ}\right) + \frac{S'^2_{YZ}}{S'^2_{ZZ}}\left(S'_{ZZZZ} - \frac{1}{n}S'^2_{ZZ}\right)\right] \\ &= \sigma^2_X\frac{S'^2_{ZZ}S'_{YYZZ} - 2S'_{YZ}S'_{ZZ}S'_{YZZZ} + S'^2_{YZ}S'_{ZZZZ}}{S'^2_{ZZ}}.\end{aligned}$$

Additionally, we observe that:

$$Y^{\mathrm{T}} P_{XZ} Y = S'_{YY} - \widehat{\beta}_0 S'_{XY} - \widehat{\gamma}_0 S'_{YZ} \sim S'_{YY} - \frac{S'_{YZ}}{S'_{ZZ}} S'_{YZ} = \frac{S'_{YY} S'_{ZZ} - S'^2_{YZ}}{S'_{ZZ}},$$

$$W^{\mathrm{T}} P_{XZ} W = S'_{WW} - \frac{S'_{ZZ} S'_{XW} - S'_{XZ} S'_{ZW}}{S'_{XX} S'_{ZZ} - S'^2_{XZ}} S'_{XW} - \frac{S'_{XX} S'_{ZW} - S'_{XZ} S'_{XW}}{S'_{XX} S'_{ZZ} - S'^2_{XZ}} S'_{ZW}.$$

According to the strong law of large numbers, we know that:

$$\frac{1}{n} S'_{XZ} \sim \frac{1}{n} \mu_X \mathbf{1}_n^{\mathrm{T}} P Z = 0, \quad \frac{1}{n} S'_{ZW} \sim \frac{1}{n} \mu_X \mathbf{1}_n^{\mathrm{T}} Q Z = 0,$$

$$\frac{1}{n} S'_{XX} \sim \frac{1}{n} \left[\mu_X^2 \mathbf{1}_n^{\mathrm{T}} P \mathbf{1}_n + \sigma_X^2 \mathrm{tr}\left(P\right) \right] = \frac{n-c}{n} \sigma_X^2,$$
$$\frac{1}{n} S'_{XW} \sim \frac{1}{n} \left[\mu_X^2 \mathbf{1}_n^{\mathrm{T}} Q \mathbf{1}_n + \sigma_X^2 \mathrm{tr}\left(Q\right) \right] = \frac{1}{n} \sigma_X^2 \mathrm{tr}\left(Q\right).$$

Thus, we infer that:

$$\frac{S'_{ZZ} S'_{XW} - S'_{XZ} S'_{ZW}}{S'_{XX} S'_{ZZ} - S'^2_{XZ}} \sim \frac{\mathrm{tr}\left(Q\right)}{n-c}, \quad \frac{S'_{XX} S'_{ZW} - S'_{XZ} S'_{XW}}{S'_{XX} S'_{ZZ} - S'^2_{XZ}} \sim 0,$$

$$W^{\mathrm{T}} P_{XZ} W \sim S'_{WW} - \frac{\mathrm{tr}\left(Q\right)}{n-c} S'_{XW}.$$

Hence, an asymptotic approximation of the squared denominator of the interaction t test statistic is given by:

$$D_*^2 = \frac{S'_{YY} S'_{ZZ} - S'^2_{YZ}}{\left(n-c-3\right) S'_{ZZ}} \left[S'_{WW} - \frac{\mathrm{tr}\left(Q\right)}{n-c} S'_{XW} \right] - \frac{N_*^2}{n-c-3}.$$

Furthermore, we calculate that:

$$\mathbb{E}\left(S'_{WW} \mid Y, Z\right) = \mu_X^2 \mathbf{1}_n^{\mathrm{T}} Q Q^{\mathrm{T}} \mathbf{1}_n + \sigma_X^2 \mathrm{tr}\left(Q Q^{\mathrm{T}}\right) = \sigma_X^2 \mathrm{tr}\left(Q Q^{\mathrm{T}}\right),$$

$$\mathbb{E}\left(N_* \mid Y, Z\right) = \mu_X \mathbf{1}_n^{\mathrm{T}} Q^{\mathrm{T}} r = 0.$$

Therefore, we conclude that:

$$\mathbb{E}\left(D_*^2 \mid Y, Z\right) = \frac{\sigma_X^2}{n-c-3} \frac{S'_{YY} S'_{ZZ} - S'^2_{YZ}}{S'_{ZZ}} \frac{\left(n-c\right) \mathrm{tr}\left(Q Q^{\mathrm{T}}\right) - \left[\mathrm{tr}\left(Q\right)\right]^2}{n-c} - \frac{\mathrm{Var}\left(N_* \mid Y, Z\right)}{n-c-3}.$$

We define the following ratio between the conditional variance of the numerator of the interaction t test statistic given Y and Z and the conditional expectation of the squared denominator of the interaction t test statistic given Y and Z:

$$R = \frac{\mathrm{Var}\left(N_* \mid Y, Z\right)}{\mathbb{E}\left(D_*^2 \mid Y, Z\right)} \approx \frac{(n-c-3)\,(n-c)}{(n-c)\mathrm{tr}\left(QQ^{\mathrm{T}}\right) - \left[\mathrm{tr}\left(Q\right)\right]^2} \frac{S'^2_{ZZ} S'_{YYZZ} - 2S'_{YZ} S'_{ZZ} S'_{YZZZ} + S'^2_{YZ} S'_{ZZZZ}}{S'_{YY} S'^2_{ZZ} - S'^2_{YZ} S'_{ZZ}},$$

where we drop any terms of order $O\left(\frac{1}{n}\right)$.

In the case where $U = \mathbf{1}_n$, we observe that $\mathrm{tr}\left(Q\right) = 0$ and $\mathrm{tr}\left(QQ^{\mathrm{T}}\right) = \frac{n-2}{n} S'_{ZZ}$, so we infer that:

$$\mathbb{E}\left(D_*^2 \mid Y, Z\right) = \frac{\sigma_X^2}{n-4} \frac{n-2}{n} \left(S'_{YY} S'_{ZZ} - S'^2_{YZ}\right) - \frac{\mathrm{Var}\left(N_* \mid Y, Z\right)}{n-4},$$

$$R \approx n \frac{S'^2_{ZZ} S'_{YYZZ} - 2S'_{YZ} S'_{ZZ} S'_{YZZZ} + S'^2_{YZ} S'_{ZZZZ}}{S'_{YY} S'^3_{ZZ} - S'^2_{YZ} S'^2_{ZZ}}.$$

CHAPTER 3

PATHOGEN GENETIC RELATEDNESS MATRIX

3.1 Introduction

Accounting for population structure among subjects infected with related strains of the same pathogen presents a significant challenge, owing to the presence of genetic variants with differing number of alleles within the pathogen genome. Furthermore, as the number of alleles in a genetic variant increases, some of the alleles may be associated with excessively small observed allele frequencies, which introduce numerical instabilities in the existing methods of constructing a pathogen genetic relatedness matrix (GRM). Pathogen GRMs have previously been constructed by first converting multiallelic genetic variants into binary allele indicators, treating all resulting allele indicators as independent genetic variants and computing a regular GRM for a haploid organism on the basis of these allele indicators [9]. Any binary allele indicators with observed allele frequencies above or below some threshold are filtered out to preserve numerical stability. However, treating the binary allele indicators which correspond to the same multiallelic genetic variant as independent genetic variants does not seem like a compelling strategy to adopt.

A previously proposed weighted pathogen GRM for organisms with multiallelic genetic variants makes the assumption that the random effects due to mutation and deletion polymorphisms on the phenotypic trait are independent and integrates information from pathogen genetic variants with differing number of alleles [8]. Nevertheless, construction of this pathogen GRM requires that any genetic variant with at least one observed allele frequency below some prespecified threshold be discarded in order to preserve numerical stability, resulting in highly unreliable sample structure estimates in the case of extremely mutable viral genomes with a large number of rare alleles. We build upon this work to develop a novel pathogen GRM for organisms with multiallelic genetic variants which avoids filtering out genetic variants with exceedingly small observed allele frequencies by introducing an adjusted weighting for rare alleles. This allows the genetic variants to which the rare alleles correspond to still contribute to the estimation of the genetic relationship between different pathogen strains.

3.2 Genetic Relatedness Matrix Based on Multiallelic Genetic Variants

For a sample of n subjects with unknown population structure due to infection with different strains of the same pathogen, suppose that the infectious disease trait value Y_i for individual i is modeled as follows:

$$Y_i = U_i^{\mathrm{T}}\alpha + v_i + \varepsilon_i, \tag{3.1}$$

where $U_i \in \mathbb{R}^c$ is a vector of covariates including an intercept term, $\alpha \in \mathbb{R}^c$ is a vector of unknown fixed covariate effects, $v = (v_1, v_2, \ldots, v_n)^{\mathrm{T}}$ is a vector of random effects accounting for correlations in trait values due to relatedness between pathogen strains with $\mathrm{Var}(v) = \sigma_v^2\Phi$, ε_i are independent and identically distributed random errors representing environmental influences for $i = 1, 2, \ldots, n$ and $\Phi \in \mathbb{R}^{n\times n}$ is the pathogen kinship matrix. Our goal is to develop a framework for the estimation of the unknown pathogen kinship matrix Φ on the basis of pathogen genome-wide data, i.e. on the basis of multiallelic pathogen genetic variants.

For the purposes of association mapping, we elect to estimate the unknown pathogen kinship matrix Φ using a pathogen GRM $K \in \mathbb{R}^{n\times n}$ which satisfies the following properties:

(I) It aggregates information across genetic variants with differing number of alleles;

(II) The contribution of each genetic variant to the overall trait variance is the same, regardless of its number of alleles and their frequencies;

(III) Consider the kinship coefficient ϕ_{ij} between pathogen strains i and j, which quantifies the probability that the genotypes at 2 randomly selected homologous genetic variants from pathogen strains i and j are identical by descent (IBD). Then, it holds that $\mathbb{E}(K_{ij}) = \phi_{ij}$ for $i, j = 1, 2, \ldots, n$, where K_{ij} is the (i, j)-th element of the pathogen GRM K.

Note that property (II) makes more sense in the context of association mapping where a pathogen GRM might be required to control for confounding due to population structure and improve effect size estimates by accounting for the effects of genetic variants other than the one being tested on the phenotypic trait. If the purpose for the computation of a pathogen GRM is to instead reconstruct a phylogenetic tree for the pathogen strains, then a different property which ensures

that the contribution of each genetic variant to the overall trait variance is weighed proportionally to its effective number of alleles might be required.

Suppose that δ_ℓ denotes the number of alleles at pathogen genetic variant ℓ for $\ell = 1, 2, \ldots, m$ and $\eta_\ell = (\eta_{\ell 1}, \eta_{\ell 2}, \ldots, \eta_{\ell \delta_\ell})^{\mathrm{T}} \sim \mathcal{N}_{\delta_\ell}\left(0, \frac{1}{m}\sigma_v^2 V_\ell\right)$ denotes a vector of random effects, where $\eta_{\ell r}$ is the random effect of allele r on the phenotypic trait, $V_\ell \in \mathbb{R}^{\delta_\ell \times \delta_\ell}$ is the covariance matrix between different allelic effects and m is the total number of pathogen genetic variants. Additionally, we assume that the vectors of random effects $\eta_1, \eta_2, \ldots, \eta_m$ corresponding to different genetic variants are independent. Then, it follows that:

$$v_i = \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \eta_{\ell r} \mathbb{1}_{\{G_{i\ell}=r\}}, \tag{3.2}$$

$$\mathrm{Cov}\left(v_i, v_j \mid G\right) = \sigma_v^2 \cdot \underbrace{\frac{1}{m} \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} V_{\ell r s} \mathbb{1}_{\left\{G_{i\ell}=r, G_{j\ell}=s\right\}}}_{K_{ij}}, \tag{3.3}$$

where $G \in \mathbb{R}^{n \times m}$ denotes the pathogen genotype matrix, $G_{i\ell}$ denotes the genotype of pathogen strain i at genetic variant ℓ and $V_{\ell r s}$ denotes the (r, s)-th element of the covariance matrix V_ℓ.

Let $f_\ell = (f_{\ell 1}, f_{\ell 2}, \ldots, f_{\ell \delta_\ell})^{\mathrm{T}}$ denote the vector of allele frequencies at genetic variant ℓ. Then, we observe that:

$$\mathbb{E}\left(K_{ij}\right) = \phi_{ij} \cdot \frac{1}{m} \sum_{\ell=1}^{m} f_\ell^{\mathrm{T}} \mathrm{diag}\left(V_\ell\right) + \left(1 - \phi_{ij}\right) \cdot \frac{1}{m} \sum_{\ell=1}^{m} f_\ell^{\mathrm{T}} V_\ell f_\ell,$$

where $\mathrm{diag}\left(V_\ell\right) \in \mathbb{R}^{\delta_\ell}$ denotes the vector whose elements are given by the diagonal elements of the matrix V_ℓ. More details about these derivations are given in Section 3.6. Note that $\phi_{ii} = 1$ for $i = 1, 2, \ldots, n$ by definition, so $f_\ell^{\mathrm{T}} \mathrm{diag}\left(V_\ell\right)$ constitutes the contribution of genetic variant ℓ to the overall trait variance. In order for property (II) to hold true for the pathogen GRM K, the quantity $f_\ell^{\mathrm{T}} \mathrm{diag}\left(V_\ell\right)$ must be the same for all genetic variants $\ell = 1, 2, \ldots, m$. Combining that with the requirement of property (III), which states that $\mathbb{E}\left(K_{ij}\right) = \phi_{ij}$, we deduce that the following 2 constraints must be placed on the covariance matrix V_ℓ:

$$f_\ell^{\mathrm{T}} \mathrm{diag}\left(V_\ell\right) = 1, \quad f_\ell^{\mathrm{T}} V_\ell f_\ell = 0, \quad \ell = 1, 2, \ldots, m. \tag{3.4}$$

3.3 Weighted Pathogen Genetic Relatedness Matrix

In order to specify a suitable covariance matrix V_ℓ for the random effects of different alleles at genetic variant ℓ, we first need to make an additional assumption about these allelic effects. Hence, we require that our desired pathogen GRM K satisfies the following additional property:

(IV) The variance of an allelic effect is inversely proportional to its allele frequency.

In other words, this property ensures that allelic effects carry more uncertainty the rarer they are, while the contribution of a common allele tends towards 0 as its allele frequency tends towards 1. We should note that this specification forces us to discard all genetic variants with at least one allele below some prespecified threshold, in order to ensure that the variance of all included allelic effects does not blow up. This property again makes more sense in the context of association mapping, where rares alleles play a much less important role than common alleles, since one would generally only be interested in the highest level of structure among pathogen strains in this setting. On the other hand, a different property which ensures that all different allelic effects at a genetic variant have the same variance might be more suitable for the reconstruction of a phylogenetic tree for the pathogen strains.

Let $\frac{1}{\sqrt{f_{\ell r}}}\xi_{\ell r}$ denote the random effect of allele r at genetic variant ℓ, where $\xi_{\ell r} \sim \mathcal{N}\left(0, \frac{1}{m(\delta_\ell - 1)}\sigma_v^2\right)$ are independent for $\ell = 1, 2, \ldots, m$ and $r = 1, 2, \ldots, \delta_\ell$. Then, the average allelic effect at genetic variant ℓ conditional on $\xi_\ell = (\xi_{\ell 1}, \xi_{\ell 2}, \ldots, \xi_{\ell \delta_\ell})^{\mathrm{T}}$ is equal to $\sum_{s=1}^{\delta_\ell} \sqrt{f_{\ell s}}\xi_{\ell s}$. Hence, we define the centered random effect of allele r at genetic variant ℓ as follows:

$$\eta_{\ell r} = \frac{1}{\sqrt{f_{\ell r}}}\xi_{\ell r} - \sum_{s=1}^{\delta_\ell} \sqrt{f_{\ell s}}\xi_{\ell s}.$$

This specification of allelic effects leads to the following covariance structure:

$$\mathrm{Cov}\left(\eta_{\ell r}, \eta_{\ell s}\right) = \frac{1}{m}\sigma_v^2 \cdot \underbrace{\frac{1}{\delta_\ell - 1}\left(\frac{1}{f_{\ell r}}\mathbb{1}_{\{r=s\}} - 1\right)}_{V_{\ell rs}}.$$

We can easily verify that this construction of a covariance matrix V_ℓ among the different allelic effects at genetic variant ℓ satisfies the 2 constraints placed on it in the previous section, so the

resultant pathogen GRM K satisfies the 3 properties also set forth in the previous section. The (i, j)-th element of this weighted pathogen GRM is specified as follows:

$$K_{ij} = \frac{1}{m} \sum_{\ell=1}^{m} \frac{1}{\delta_\ell - 1} \left(\sum_{r=1}^{\delta_\ell} \frac{1}{f_{\ell r}} \mathbb{1}_{\{G_{i\ell} = G_{j\ell} = r\}} - 1 \right). \tag{3.5}$$

More details about these derivations are given in Section 3.6. We observe that this weighted pathogen GRM coincides with the previously proposed pathogen GRM constructed on the basis of genetic variants with differing number of alleles [8]. The contribution of a biallelic genetic variant to the estimation of the genetic relatedness between pathogen strains obviously reduces to the contribution of a genetic variant to an ordinary GRM based on a haploid organism, as described in Section 1.4. Our goal is to build upon this derivation to propose a novel weight-adjusted pathogen GRM which takes into account rare alleles by placing an upper bound on the variance of individual allelic effects.

3.4 Weight-Adjusted Pathogen Genetic Relatedness Matrix

In order to avoid discarding all genetic variants with at least one allele frequency below some prespecified threshold, we first need to make some additional assumptions about the random effects of rare alleles. Hence, we require that our novel pathogen GRM K satisfies the following additional properties:

(V) There exists some prespecified allele frequency threshold τ, e.g. $\tau = 0.05$, such that alleles with frequencies below that threshold are called rare and alleles with frequencies above that threshold are called common;

(VI) An upper bound is placed on the variance of the random effects of rare alleles, so that they can contribute to the estimation of the genetic relatedness between pathogen strains without causing numerical instabilities.

In this specification setting, we would only be required to discard genetic variants for which $\delta_\ell > \frac{1}{\tau}$ or $\max_r f_{\ell r} > 1 - \tau$, i.e. only genetic variants with an exceedingly large number of alleles or genetic variants with only one common allele with an exceedingly large observed allele frequency.

At least in the case of highly mutable genetic variants, this specification amounts to discarding a much smaller proportion of genetic variants to preserve numerical stability in the pathogen GRM computation compared to the previously proposed weighted pathogen GRM.

We define the thresholded allele frequencies $f_{\ell r}^{+} = \max(f_{\ell r}, \tau)$ for $\ell = 1, 2, \ldots, m$ and $r = 1, 2, \ldots, \delta_\ell$. In other words, any allele frequency below the prespecified threshold τ is truncated to τ. We let $B_\ell = \sum_{r=1}^{\delta_\ell} \mathbb{1}_{\{f_{\ell r} \geqslant \tau\}}$ represent the number of common alleles at pathogen genetic variant ℓ, $P_\ell = \sum_{r=1}^{\delta_\ell} f_{\ell r} \mathbb{1}_{\{f_{\ell r} < \tau\}}$ represent the sum of all rare allele frequencies at genetic variant ℓ and $q_\ell = \sum_{r=1}^{\delta_\ell} \frac{f_{\ell r}^2}{f_{\ell r}^{+}}$. We define the "effective" number of alleles at genetic variant ℓ as $c_\ell = B_\ell - q_\ell + \frac{1}{\tau} P_\ell$. In the case where there are no rare alleles at genetic variant ℓ, we observe that $B_\ell = \delta_\ell$, $P_\ell = 0$ and $q_\ell = 1$, so c_ℓ reduces to $\delta_\ell - 1$. However, in the presence of a rare allele at a pathogen genetic variant, its contribution to the "effective" number of alleles reduces linearly from 1 to 0 as its allele frequency moves from the threshold value τ towards 0.

Let $\sqrt{\frac{1}{f_{\ell r}^{+}}} \xi_{\ell r}$ denote the random effect of allele r at genetic variant ℓ, where $\xi_{\ell r} \sim \mathcal{N}\left(0, \frac{1}{m c_\ell} \sigma_v^2\right)$ are independent for $\ell = 1, 2, \ldots, m$ and $r = 1, 2, \ldots, \delta_\ell$. We observe that the use of the thresholded allele frequency $f_{\ell r}^{+}$ in the weighting of the allelic effect places an upper bound of $\frac{1}{m c_\ell \tau} \sigma_v^2$ on its variance. Then, the average allelic effect at genetic variant ℓ conditional on $\xi_\ell = (\xi_{\ell 1}, \xi_{\ell 2}, \ldots, \xi_{\ell \delta_\ell})^{\mathrm{T}}$ is equal to $\sum_{s=1}^{\delta_\ell} \sqrt{\frac{1}{f_{\ell s}^{+}}} f_{\ell s} \xi_{\ell s}$. Hence, we define the centered random effect of allele r at genetic variant ℓ as follows:

$$\eta_{\ell r} = \sqrt{\frac{1}{f_{\ell r}^{+}}} \xi_{\ell r} - \sum_{s=1}^{\delta_\ell} \sqrt{\frac{1}{f_{\ell s}^{+}}} f_{\ell s} \xi_{\ell s}.$$

This specification of allelic effects leads to the following covariance structure:

$$\mathrm{Cov}(\eta_{\ell r}, \eta_{\ell s}) = \frac{1}{m} \sigma_v^2 \cdot \underbrace{\frac{1}{c_\ell} \left(\frac{1}{f_{\ell r}^{+}} \mathbb{1}_{\{r=s\}} - \frac{f_{\ell r}}{f_{\ell r}^{+}} - \frac{f_{\ell s}}{f_{\ell s}^{+}} + q_\ell \right)}_{V_{\ell r s}}.$$

We can easily verify that this construction of a covariance matrix V_ℓ among the different allelic effects at genetic variant ℓ satisfies the 2 constraints placed on it in Section 3.2, so the resultant pathogen GRM K satisfies the 4 properties set forth in the previous sections. The (i, j)-th element

of this weight-adjusted pathogen GRM is specified as follows:

$$K_{ij} = \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{c_\ell}\left(\sum_{r=1}^{\delta_\ell}\frac{1}{f_{\ell r}^{+}}\mathbb{1}_{\{G_{i\ell}=G_{j\ell}=r\}} - \frac{f_{\ell G_{i\ell}}}{f_{\ell G_{i\ell}}^{+}} - \frac{f_{\ell G_{j\ell}}}{f_{\ell G_{j\ell}}^{+}} + q_\ell\right). \tag{3.6}$$

More details about these derivations are given in Section 3.6. In the case where there are no rare alleles at genetic variant ℓ, we observe that $c_\ell = \delta_\ell - 1$, $f_{\ell r}^{+} = f_{\ell r}$ and $-\frac{f_{\ell G_{i\ell}}}{f_{\ell G_{i\ell}}^{+}} - \frac{f_{\ell G_{j\ell}}}{f_{\ell G_{j\ell}}^{+}} + q_\ell = -1$, so the contribution of genetic variant ℓ to the weight-adjusted GRM reduces to being the same as that of the previously proposed weighted pathogen GRM. Additionally, the weight adjustment of this novel pathogen GRM ensures that the contribution of rare alleles at a genetic variant reduces to 0 as their allele frequencies tend towards 0, i.e. the limiting model for the genetic variant would be the same as that of a genetic variant without the presence of these rare alleles.

3.5 Discussion and Future Work

The computation of a pathogen GRM for the estimation of the population structure across different pathogen strains presents challenges due to the highly mutable nature of pathogen genomes, leading to large numbers of multiallelic genetic variants with more than 2 alleles and allele frequencies which decrease as the number of alleles in a genetic variant increases. Previous efforts in the calculation of pathogen GRMs have either required the conversion of all multiallelic genetic variants into binary allele indicators, which are then treated as independent genetic variants from a haploid organism, leading to a binary GRM whose validity is hard to theoretically justify, or the discarding of all genetic variants with at least one rare allele, leaving only a tiny proportion of the pathogen genome available for the computation of a weighted pathogen GRM, leading to unreliable estimation of the genetic relationship between 2 pathogen strains.

We proposed the construction of a novel weight-adjusted pathogen GRM which avoids filtering out genetic variants with exceedingly small observed allele frequencies by placing an upper bound on the random effects of rare alleles. This allows the genetic variants to which the rare alleles correspond to still contribute to the estimation of the genetic relationship between different pathogen strains. This newly proposed pathogen GRM is constructed on the basis of a linear mixed model, which properly weights the random effect of all alleles within each genetic variant, and is suitable for

the purposes of association mapping where a pathogen GRM might be required to control for confounding due to population structure and improve effect size estimates by accounting for the effects of genetic variants other than the one being tested on the phenotypic trait. Furthermore, this weight-adjusted pathogen GRM is constructed in such a way that its expectation with respect to the pathogen genotype matrix is equal to the pathogen kinship matrix, which measures the probability that the genotypes at 2 randomly selected homologous genetic variants belonging to 2 different pathogen strains are identical by descent (IBD).

Construction of this weight-adjusted pathogen GRM is based on specific modeling choices relating to the relative contribution of genetic variants with differing number of alleles to the overall trait variance and the variability of allelic effects with varying allele frequencies. Our modeling choices have been tailored to the setting of estimating pathogen population structure for the purposes of association mapping rather than the reconstruction of phylogeny among pathogen strains, which might require modeling assumptions. Further investigation is required into the effect of different modeling choices in the construction of a pathogen GRM as well as the suitability of these modeling choices for different purposes of a GRM computation.

3.6 Appendix

Suppose that δ_ℓ denotes the number of alleles at pathogen genetic variant ℓ for $\ell = 1, 2, \ldots, m$ and $\eta_\ell = (\eta_{\ell 1}, \eta_{\ell 2}, \ldots, \eta_{\ell \delta_\ell})^{\mathrm{T}} \sim \mathcal{N}_{\delta_\ell}\left(0, \frac{1}{m}\sigma_v^2 V_\ell\right)$ denotes a vector of random effects, where $\eta_{\ell r}$ is the random effect of allele r on the phenotypic trait, $V_\ell \in \mathbb{R}^{\delta_\ell \times \delta_\ell}$ is the covariance matrix between different allelic effects and m is the total number of pathogen genetic variants. Additionally, we assume that the vectors of random effects $\eta_1, \eta_2, \ldots, \eta_m$ corresponding to different genetic variants are independent. For a sample of n individuals with unknown structure due to infection with different strains of the same pathogen, suppose that the infectious disease trait value Y_i for individual i is modeled as follows:

$$Y_i = U_i^{\mathrm{T}}\alpha + v_i + \varepsilon_i,$$

where $U_i \in \mathbb{R}^c$ is a vector of covariates including an intercept term, $\alpha \in \mathbb{R}^c$ is a vector of unknown fixed covariate effects, $v = (v_1, v_2, \ldots, v_n)^{\mathrm{T}}$ is a vector of random effects accounting for correlations

in trait values due to relatedness between pathogen strains with $\mathrm{Var}(v) = \sigma_v^2 \Phi$, ε_i are independent and identically distributed random errors representing environmental influences for $i = 1, 2, \ldots, n$ and $\Phi \in \mathbb{R}^{n \times n}$ is the pathogen kinship matrix. Then, it follows that:

$$v_i = \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \eta_{\ell r} \mathbb{1}_{\{G_{i\ell}=r\}},$$

$$\begin{aligned}
\mathrm{Cov}\left(v_i, v_j \mid G\right) &= \mathrm{Cov}\left(\sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \eta_{\ell r} \mathbb{1}_{\{G_{i\ell}=r\}}, \sum_{\ell=1}^{m} \sum_{s=1}^{\delta_\ell} \eta_{\ell s} \mathbb{1}_{\left\{G_{j\ell}=s\right\}} \middle| G \right) \\
&= \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} \mathrm{Cov}\left(\eta_{\ell r}, \eta_{\ell s}\right) \mathbb{1}_{\left\{G_{i\ell}=r, G_{j\ell}=s\right\}} \\
&= \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} \frac{1}{m} \sigma_v^2 V_{\ell r s} \mathbb{1}_{\left\{G_{i\ell}=r, G_{j\ell}=s\right\}} \\
&= \sigma_v^2 \cdot \underbrace{\frac{1}{m} \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} V_{\ell r s} \mathbb{1}_{\left\{G_{i\ell}=r, G_{j\ell}=s\right\}}}_{K_{ij}},
\end{aligned}$$

where $G \in \mathbb{R}^{n \times m}$ denotes the pathogen genotype matrix, $G_{i\ell}$ denotes the genotype of pathogen strain i at genetic variant ℓ and $V_{\ell r s}$ denotes the (r, s)-th element of the covariance matrix V_ℓ.

For $r, s = 1, 2, \ldots, \delta_\ell$, we calculate that:

$$\mathbb{P}\left(G_{i\ell} = r, G_{j\ell} = s\right) = \phi_{ij} f_{\ell r} \mathbb{1}_{\{r=s\}} + \left(1 - \phi_{ij}\right) f_{\ell r} f_{\ell s}.$$

For $i, j = 1, 2, \ldots, n$, we conclude that:

$$\begin{aligned}
\mathbb{E}\left(K_{ij}\right) &= \frac{1}{m} \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} V_{\ell r s} \mathbb{P}\left(G_{i\ell} = r, G_{j\ell} = s\right) \\
&= \frac{1}{m} \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} V_{\ell r s} \left[\phi_{ij} f_{\ell r} \mathbb{1}_{\{r=s\}} + \left(1 - \phi_{ij}\right) f_{\ell r} f_{\ell s}\right] \\
&= \phi_{ij} \cdot \frac{1}{m} \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} V_{\ell r r} f_{\ell r} + \left(1 - \phi_{ij}\right) \cdot \frac{1}{m} \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} V_{\ell r s} f_{\ell r} f_{\ell s} \\
&= \phi_{ij} \cdot \frac{1}{m} \sum_{\ell=1}^{m} \underbrace{f_\ell^{\mathrm{T}} \mathrm{diag}\left(V_\ell\right)}_{1} + \left(1 - \phi_{ij}\right) \cdot \frac{1}{m} \sum_{\ell=1}^{m} \underbrace{f_\ell^{\mathrm{T}} V_\ell f_\ell}_{0} = \phi_{ij},
\end{aligned}$$

where $f_\ell = (f_{\ell 1}, f_{\ell 2}, \ldots, f_{\ell \delta_\ell})^{\mathrm{T}}$ denotes the vector of allele frequencies at genetic variant ℓ.

Weighted Pathogen Genetic Relatedness Matrix

Let $\frac{1}{\sqrt{f_{\ell r}}}\xi_{\ell r}$ denote the random effect of allele r at genetic variant ℓ, where $\xi_{\ell r} \sim \mathcal{N}\left(0, \frac{1}{m(\delta_\ell - 1)}\sigma_v^2\right)$ are independent for $\ell = 1, 2, \ldots, m$ and $r = 1, 2, \ldots, \delta_\ell$. Then, the average allelic effect at genetic variant ℓ conditional on $\xi_\ell = (\xi_{\ell 1}, \xi_{\ell 2}, \ldots, \xi_{\ell \delta_\ell})^{\mathrm{T}}$ is equal to $\sum_{s=1}^{\delta_\ell} \sqrt{f_{\ell s}}\xi_{\ell s}$. Hence, we define the centered random effect of allele r at genetic variant ℓ as follows:

$$\eta_{\ell r} = \frac{1}{\sqrt{f_{\ell r}}}\xi_{\ell r} - \sum_{q=1}^{\delta_\ell} \sqrt{f_{\ell q}}\xi_{\ell q}.$$

This specification of allelic effects leads to the following covariance structure for $r \neq s$:

$$\begin{aligned}
\mathrm{Var}\left(\eta_{\ell r}\right) &= \mathrm{Var}\left(\frac{1}{\sqrt{f_{\ell r}}}\xi_{\ell r} - \sum_{q=1}^{\delta_\ell} \sqrt{f_{\ell q}}\xi_{\ell q}\right) \\
&= \left(\frac{1}{\sqrt{f_{\ell r}}} - \sqrt{f_{\ell r}}\right)^2 \mathrm{Var}\left(\xi_{\ell r}\right) + \sum_{q \neq r} f_{\ell q}\mathrm{Var}\left(\xi_{\ell q}\right) \\
&= \left[\frac{(1 - f_{\ell r})^2}{f_{\ell r}} + 1 - f_{\ell r}\right] \frac{1}{m\left(\delta_\ell - 1\right)}\sigma_v^2 \\
&= \frac{1}{m}\sigma_v^2 \cdot \underbrace{\frac{1}{\delta_\ell - 1}\left(\frac{1}{f_{\ell r}} - 1\right)}_{V_{\ell r r}},
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}\left(\eta_{\ell r}, \eta_{\ell s}\right) &= \mathrm{Cov}\left(\frac{1}{\sqrt{f_{\ell r}}}\xi_{\ell r} - \sum_{q=1}^{\delta_\ell} \sqrt{f_{\ell q}}\xi_{\ell q}, \frac{1}{\sqrt{f_{\ell s}}}\xi_{\ell s} - \sum_{t=1}^{\delta_\ell} \sqrt{f_{\ell t}}\xi_{\ell t}\right) \\
&= -\left(\frac{1}{\sqrt{f_{\ell r}}} - \sqrt{f_{\ell r}}\right)\sqrt{f_{\ell r}}\mathrm{Var}\left(\xi_{\ell r}\right) - \left(\frac{1}{\sqrt{f_{\ell s}}} - \sqrt{f_{\ell s}}\right)\sqrt{f_{\ell s}}\mathrm{Var}\left(\xi_{\ell s}\right) + \sum_{q \neq r,s} f_{\ell q}\mathrm{Var}\left(\xi_{\ell q}\right) \\
&= \frac{f_{\ell r} - 1 + f_{\ell s} - 1 + 1 - f_{\ell r} - f_{\ell s}}{m\left(\delta_\ell - 1\right)}\sigma_v^2 = \frac{1}{m}\sigma_v^2 \cdot \underbrace{\frac{-1}{\delta_\ell - 1}}_{V_{\ell r s}}.
\end{aligned}$$

We verify that this construction of a covariance matrix V_ℓ among the different allelic effects at genetic variant ℓ satisfies the 2 constraints placed on it as follows:

$$f_\ell^{\mathrm{T}}\mathrm{diag}\left(V_\ell\right) = \sum_{r=1}^{\delta_\ell} \frac{f_{\ell r}}{\delta_\ell - 1}\left(\frac{1}{f_{\ell r}} - 1\right) = \frac{1}{\delta_\ell - 1}\sum_{r=1}^{\delta_\ell}\left(1 - f_{\ell r}\right) = 1,$$

$$
\begin{aligned}
f_\ell^{\mathrm{T}} V_\ell f_\ell &= \sum_{r=1}^{\delta_\ell} \frac{f_{\ell r}^2}{\delta_\ell - 1} \left(\frac{1}{f_{\ell r}} - 1 \right) - \sum_{r=1}^{\delta_\ell} \sum_{s \neq r} \frac{f_{\ell r} f_{\ell s}}{\delta_\ell - 1} \\
&= \frac{1}{\delta_\ell - 1} \sum_{r=1}^{\delta_\ell} f_{\ell r} \left(1 - f_{\ell r} + \sum_{s \neq r} f_{\ell s} \right) = 0.
\end{aligned}
$$

The (i, j)-th element of the resulting weighted pathogen GRM is specified as follows:

$$
\begin{aligned}
K_{ij} &= \frac{1}{m} \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} V_{\ell r s} \mathbb{1}_{\{G_{i\ell}=r, G_{j\ell}=s\}} \\
&= \frac{1}{m} \sum_{\ell=1}^{m} \sum_{r=1}^{\delta_\ell} \sum_{s=1}^{\delta_\ell} \frac{1}{\delta_\ell - 1} \left(\frac{1}{f_{\ell r}} \mathbb{1}_{\{r=s\}} - 1 \right) \mathbb{1}_{\{G_{i\ell}=r, G_{j\ell}=s\}} \\
&= \frac{1}{m} \sum_{\ell=1}^{m} \frac{1}{\delta_\ell - 1} \left(\sum_{r=1}^{\delta_\ell} \frac{1}{f_{\ell r}} \mathbb{1}_{\{G_{i\ell}=G_{j\ell}=r\}} - 1 \right).
\end{aligned}
$$

Additionally, we verify that the contribution of a biallelic genetic variant to this weighted pathogen GRM reduces to its contribution to an ordinary GRM based on a haploid organism as follows:

$$
\sum_{r \in \{0,1\}} \frac{1}{f_{\ell r}} \mathbb{1}_{\{G_{i\ell}=G_{j\ell}=r\}} - 1 = \begin{cases} \frac{1-f_{\ell 1}}{f_{\ell 1}}, & G_{i\ell} = G_{j\ell} = 1 \\ -1, & G_{i\ell} \neq G_{j\ell} \\ \frac{f_{\ell 1}}{1-f_{\ell 1}}, & G_{i\ell} = G_{j\ell} = 0 \end{cases} = \frac{(G_{i\ell} - f_{\ell 1})(G_{j\ell} - f_{\ell 1})}{f_{\ell 1}(1 - f_{\ell 1})}.
$$

Finally, we verify that the weighted pathogen GRM satisfies the properties set forth in Section 3.2 as follows:

$$
\begin{aligned}
\mathbb{E}(K_{ij}) &= \frac{1}{m} \sum_{\ell=1}^{m} \frac{1}{\delta_\ell - 1} \left[\sum_{r=1}^{\delta_\ell} \frac{1}{f_{\ell r}} \mathbb{P}(G_{i\ell} = G_{j\ell} = r) - 1 \right] \\
&= \frac{1}{m} \sum_{\ell=1}^{m} \frac{1}{\delta_\ell - 1} \left[\sum_{r=1}^{\delta_\ell} \frac{\phi_{ij} f_{\ell r} + (1 - \phi_{ij}) f_{\ell r}^2}{f_{\ell r}} - 1 \right] \\
&= \frac{1}{m} \sum_{\ell=1}^{m} \frac{\delta_\ell \phi_{ij} + 1 - \phi_{ij} - 1}{\delta_\ell - 1} = \phi_{ij},
\end{aligned}
$$

$$\begin{aligned}
\text{Var}\left(\sum_{r=1}^{\delta_\ell}\eta_{\ell r}\mathbb{1}_{\{G_{i\ell}=r\}}\right) &= \text{Var}\left[\mathbb{E}\left(\sum_{r=1}^{\delta_\ell}\eta_{\ell r}\mathbb{1}_{\{G_{i\ell}=r\}}\middle| G_{i\ell}\right)\right] + \mathbb{E}\left[\text{Var}\left(\sum_{r=1}^{\delta_\ell}\eta_{\ell r}\mathbb{1}_{\{G_{i\ell}=r\}}\middle| G_{i\ell}\right)\right] \\
&= \text{Var}\left[\sum_{r=1}^{\delta_\ell}\underbrace{\mathbb{E}\left(\eta_{\ell r}\mid G_{i\ell}\right)}_{0}\mathbb{1}_{\{G_{i\ell}=r\}}\right] + \mathbb{E}\left[\sum_{r=1}^{\delta_\ell}\text{Var}\left(\eta_{\ell r}\mid G_{i\ell}\right)\mathbb{1}_{\{G_{i\ell}=r\}}\right] \\
&= \sum_{r=1}^{\delta_\ell}\frac{1}{m}\sigma_v^2 V_{\ell rr}\mathbb{P}\left(G_{i\ell}=r\right) = \frac{1}{m}\sigma_v^2\underbrace{f_\ell^{\mathrm{T}}\text{diag}\left(V_\ell\right)}_{1} = \frac{1}{m}\sigma_v^2.
\end{aligned}$$

Weight-Adjusted Pathogen Genetic Relatedness Matrix

We define the thresholded allele frequencies $f_{\ell r}^{+} = \max\left(f_{\ell r}, \tau\right)$ for $\ell = 1, 2, \ldots, m$ and $r = 1, 2, \ldots, \delta_\ell$. We let $B_\ell = \sum_{r=1}^{\delta_\ell}\mathbb{1}_{\{f_{\ell r}\geqslant\tau\}}$ represent the number of common alleles at pathogen genetic variant ℓ, $P_\ell = \sum_{r=1}^{\delta_\ell} f_{\ell r}\mathbb{1}_{\{f_{\ell r}<\tau\}}$ represent the sum of all rare allele frequencies at genetic variant ℓ and $q_\ell = \sum_{r=1}^{\delta_\ell}\frac{f_{\ell r}^2}{f_{\ell r}^{+}}$. We define the "effective" number of alleles at genetic variant ℓ as $c_\ell = B_\ell - q_\ell + \frac{1}{\tau}P_\ell$.

Let $\sqrt{\frac{1}{f_{\ell r}^{+}}}\xi_{\ell r}$ denote the random effect of allele r at genetic variant ℓ, where $\xi_{\ell r} \sim \mathcal{N}\left(0, \frac{1}{mc_\ell}\sigma_v^2\right)$ are independent for $\ell = 1, 2, \ldots, m$ and $r = 1, 2, \ldots, \delta_\ell$. Then, the average allelic effect at genetic variant ℓ conditional on $\xi_\ell = \left(\xi_{\ell 1}, \xi_{\ell 2}, \ldots, \xi_{\ell\delta_\ell}\right)^{\mathrm{T}}$ is equal to $\sum_{s=1}^{\delta_\ell}\sqrt{\frac{1}{f_{\ell s}^{+}}}f_{\ell s}\xi_{\ell s}$. Hence, we define the centered random effect of allele r at genetic variant ℓ as follows:

$$\eta_{\ell r} = \sqrt{\frac{1}{f_{\ell r}^{+}}}\xi_{\ell r} - \sum_{q=1}^{\delta_\ell}\sqrt{\frac{1}{f_{\ell q}^{+}}}f_{\ell q}\xi_{\ell q}.$$

This specification of allelic effects leads to the following covariance structure for $r \neq s$:

$$\begin{aligned}
\text{Var}\left(\eta_{\ell r}\right) &= \text{Var}\left(\sqrt{\frac{1}{f_{\ell r}^{+}}}\xi_{\ell r} - \sum_{q=1}^{\delta_\ell}\sqrt{\frac{1}{f_{\ell q}^{+}}}f_{\ell q}\xi_{\ell q}\right) \\
&= \frac{\left(1-f_{\ell r}\right)^2}{f_{\ell r}^{+}}\text{Var}\left(\xi_{\ell r}\right) + \sum_{q\neq r}\frac{f_{\ell q}^2}{f_{\ell q}^{+}}\text{Var}\left(\xi_{\ell q}\right) \\
&= \left(\frac{1 - 2f_{\ell r} + f_{\ell r}^2}{f_{\ell r}^{+}} + q_\ell - \frac{f_{\ell r}^2}{f_{\ell r}^{+}}\right)\frac{1}{mc_\ell}\sigma_v^2 \\
&= \frac{1}{m}\sigma_v^2\cdot\underbrace{\frac{1}{c_\ell}\left(\frac{1}{f_{\ell r}^{+}} - \frac{2f_{\ell r}}{f_{\ell r}^{+}} + q_\ell\right)}_{V_{\ell rr}},
\end{aligned}$$

$$
\begin{aligned}
\operatorname{Cov}\left(\eta_{\ell r}, \eta_{\ell s}\right) &= \operatorname{Cov}\left(\sqrt{\frac{1}{f_{\ell r}^{+}}}\xi_{\ell r} - \sum_{q=1}^{\delta_\ell}\sqrt{\frac{1}{f_{\ell q}^{+}}}f_{\ell q}\xi_{\ell q}, \sqrt{\frac{1}{f_{\ell s}^{+}}}\xi_{\ell s} - \sum_{t=1}^{\delta_\ell}\sqrt{\frac{1}{f_{\ell t}^{+}}}f_{\ell t}\xi_{\ell t}\right) \\
&= -\frac{\left(1-f_{\ell r}\right)f_{\ell r}}{f_{\ell r}^{+}}\operatorname{Var}\left(\xi_{\ell r}\right) - \frac{\left(1-f_{\ell s}\right)f_{\ell s}}{f_{\ell s}^{+}}\operatorname{Var}\left(\xi_{\ell s}\right) + \sum_{q\neq r,s}\frac{f_{\ell q}^{2}}{f_{\ell q}^{+}}\operatorname{Var}\left(\xi_{\ell q}\right) \\
&= \left(-\frac{f_{\ell r}}{f_{\ell r}^{+}} + \frac{f_{\ell r}^{2}}{f_{\ell r}^{+}} - \frac{f_{\ell s}}{f_{\ell s}^{+}} + \frac{f_{\ell s}^{2}}{f_{\ell s}^{+}} + q_\ell - \frac{f_{\ell r}^{2}}{f_{\ell r}^{+}} - \frac{f_{\ell s}^{2}}{f_{\ell s}^{+}}\right)\frac{1}{mc_\ell}\sigma_v^2 \\
&= \frac{1}{m}\sigma_v^2 \cdot \underbrace{\frac{1}{c_\ell}\left(-\frac{f_{\ell r}}{f_{\ell r}^{+}} - \frac{f_{\ell s}}{f_{\ell s}^{+}} + q_\ell\right)}_{V_{\ell rs}}.
\end{aligned}
$$

We verify that this construction of a covariance matrix V_ℓ among the different allelic effects at genetic variant ℓ satisfies the 2 constraints placed on it as follows:

$$
\begin{aligned}
f_\ell^{\mathrm{T}}\operatorname{diag}\left(V_\ell\right) &= \sum_{r=1}^{\delta_\ell}\frac{f_{\ell r}}{c_\ell}\left(\frac{1}{f_{\ell r}^{+}} - \frac{2f_{\ell r}}{f_{\ell r}^{+}} + q_\ell\right) \\
&= \frac{1}{c_\ell}\left(\sum_{r=1}^{\delta_\ell}\frac{f_{\ell r}}{f_{\ell r}^{+}}\mathbb{1}_{\{f_{\ell r}\geqslant\tau\}} + \sum_{r=1}^{\delta_\ell}\frac{f_{\ell r}}{f_{\ell r}^{+}}\mathbb{1}_{\{f_{\ell r}<\tau\}} - 2q_\ell + q_\ell\right) \\
&= \frac{B_\ell + \frac{1}{\tau}P_\ell - q_\ell}{c_\ell} = 1,
\end{aligned}
$$

$$
\begin{aligned}
f_\ell^{\mathrm{T}}V_\ell f_\ell &= \sum_{r=1}^{\delta_\ell}\frac{f_{\ell r}^{2}}{c_\ell}\left(\frac{1}{f_{\ell r}^{+}} - \frac{2f_{\ell r}}{f_{\ell r}^{+}} + q_\ell\right) + \sum_{r=1}^{\delta_\ell}\sum_{s\neq r}\frac{f_{\ell r}f_{\ell s}}{c_\ell}\left(-\frac{f_{\ell r}}{f_{\ell r}^{+}} - \frac{f_{\ell s}}{f_{\ell s}^{+}} + q_\ell\right) \\
&= \frac{1}{c_\ell}\sum_{r=1}^{\delta_\ell}f_{\ell r}^{2}\left(\frac{1}{f_{\ell r}^{+}} - \frac{2f_{\ell r}}{f_{\ell r}^{+}} + q_\ell\right) + \frac{1}{c_\ell}\sum_{r=1}^{\delta_\ell}f_{\ell r}\left[-\frac{f_{\ell r}\left(1-f_{\ell r}\right)}{f_{\ell r}^{+}} - q_\ell + \frac{f_{\ell r}^{2}}{f_{\ell r}^{+}} + q_\ell\left(1-f_{\ell r}\right)\right] \\
&= \frac{1}{c_\ell}\sum_{r=1}^{\delta_\ell}f_{\ell r}^{2}\left(\frac{1}{f_{\ell r}^{+}} - \frac{2f_{\ell r}}{f_{\ell r}^{+}} + q_\ell - \frac{1}{f_{\ell r}^{+}} + \frac{f_{\ell r}}{f_{\ell r}^{+}} + \frac{f_{\ell r}}{f_{\ell r}^{+}} - q_\ell\right) = 0.
\end{aligned}
$$

The (i,j)-th element of the weight-adjusted pathogen GRM is specified as follows:

$$
\begin{aligned}
K_{ij} &= \frac{1}{m}\sum_{\ell=1}^{m}\sum_{r=1}^{\delta_\ell}\sum_{s=1}^{\delta_\ell}V_{\ell rs}\mathbb{1}_{\{G_{i\ell}=r, G_{j\ell}=s\}} \\
&= \frac{1}{m}\sum_{\ell=1}^{m}\sum_{r=1}^{\delta_\ell}\sum_{s=1}^{\delta_\ell}\frac{1}{c_\ell}\left(\frac{1}{f_{\ell r}^{+}}\mathbb{1}_{\{r=s\}} - \frac{f_{\ell r}}{f_{\ell r}^{+}} - \frac{f_{\ell s}}{f_{\ell s}^{+}} + q_\ell\right)\mathbb{1}_{\{G_{i\ell}=r, G_{j\ell}=s\}} \\
&= \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{c_\ell}\left(\sum_{r=1}^{\delta_\ell}\frac{1}{f_{\ell r}^{+}}\mathbb{1}_{\{G_{i\ell}=G_{j\ell}=r\}} - \frac{f_{\ell G_{i\ell}}}{f_{\ell G_{i\ell}}^{+}} - \frac{f_{\ell G_{j\ell}}}{f_{\ell G_{j\ell}}^{+}} + q_\ell\right).
\end{aligned}
$$

Finally, we verify that the weight-adjusted pathogen GRM satisfies the properties set forth in Section 3.2 as follows:

$$\begin{aligned}
\mathbb{E}\left(K_{ij}\right) &= \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{c_\ell}\left[\sum_{r=1}^{\delta_\ell}\frac{1}{f_{\ell r}^{+}}\mathbb{P}\left(G_{i\ell}=G_{j\ell}=r\right)-\sum_{r=1}^{\delta_\ell}\frac{f_{\ell r}}{f_{\ell r}^{+}}\mathbb{P}\left(G_{i\ell}=r\right)-\sum_{s=1}^{\delta_\ell}\frac{f_{\ell s}}{f_{\ell s}^{+}}\mathbb{P}\left(G_{j\ell}=s\right)+q_\ell\right] \\
&= \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{c_\ell}\left[\sum_{r=1}^{\delta_\ell}\frac{\phi_{ij}f_{\ell r}+\left(1-\phi_{ij}\right)f_{\ell r}^{2}}{f_{\ell r}^{+}}-q_\ell\right] \\
&= \frac{1}{m}\sum_{\ell=1}^{m}\frac{\phi_{ij}\left(B_\ell+\frac{1}{\tau}P_\ell\right)+\left(1-\phi_{ij}\right)q_\ell-q_\ell}{c_\ell}=\phi_{ij}.
\end{aligned}$$

CHAPTER 4

JOINT ASSOCIATION ANALYSIS OF HEPATITIS C PRE-TREATMENT VIRAL LOAD

4.1 Description of the Data Set

To investigate the capability of our proposed correction framework to detect interaction effects between two genetic variants on a phenotypic trait, we carried out a joint association analysis of pre-treatment viral load (PTVL) on HCV-infected patients from the BOSON clinical trial [72]. Complete clinical and genomic information was available for a total of 540 out of 568 HCV-infected patients. Out of these patients, 450 were of self-reported white ethnicity, 485 were infected with HCV genotype 3a, and 409 were of self-reported white ethnicity while also infected with HCV genotype 3a. Since the vast majority of infected patients in the sample belonged to this last category, we chose to focus on analyzing this subset of patients for reasons of population homogeneity.

4.2 Imputation and Alignment of Genome Sequences

Viral nucleotide and amino acid sequences were downloaded from GenBank under accession codes KY620313-KY620880. These nucleotide and amino acid sequences for each patient were initially aligned using MAFFT version 7 with default settings [73]. The sequences were subsequently filtered according to the viral genotype and the self-reported ethnicity of the infected patient, so that only the 409 sequences corresponding to patients of self-reported white ethnicity and infected with HCV genotype 3a were used for any subsequent analysis.

Human genotype data were sequenced using the Affymetrix UK Biobank array and are deposited in the European Genome-Phenome Archive under accession code EGAS00001002324. Access to the human genotype and clinical data was granted to us by the STOP-HCV consortium. The human genotype data set was first divided into chromosomes and transformed to VCF format using PLINK version 1.9 [74]. Missing human genotypes were then imputed using Beagle version 5.4 [75] with the 1000 Genomes Project phase 3 reference panel [76] and the HapMap GrCh38 human genetic map [77].

Human SNPs on chromosome 6 from coordinates 20 Mbp to 40 Mbp - flanking the HLA region - were extracted using PLINK version 1.9. The extracted SNPs were submitted to the Michigan Imputation Server for genotype imputation using Minimac4 [78] with the 1000G Phase 3 v5 reference panel and the Eagle version 2.4 phasing algorithm [79]. The phased and imputed genotypes were converted from vcf.gz format to Oxford HAPS/SAMPLE format using bcftools version 1.17 [80]. The converted genotypes were used to impute HLA class I and II haplotypes using HLA*IMP:03 [81]. Finally, HLA amino acids and SNPs were imputed using the Michigan Imputation Server with the Four-digit Multi-ethnic HLA reference panel [82].

4.3 Estimation of the Population Structure

All of the aligned viral nucleotide sequences were employed to produce a maximum likelihood tree using RAxML version 8.2 with a general time reversible model of nucleotide substitution under the gamma model of rate heterogeneity [83]. The resulting tree, displayed in Figure 4.1, was rooted at the midpoint and colored by HCV genotype. We notice the prevalence of HCV genotype 3a, colored in cyan, as well as the great performance of the phylogenetic tree at distinguishing between different HCV genotypes.

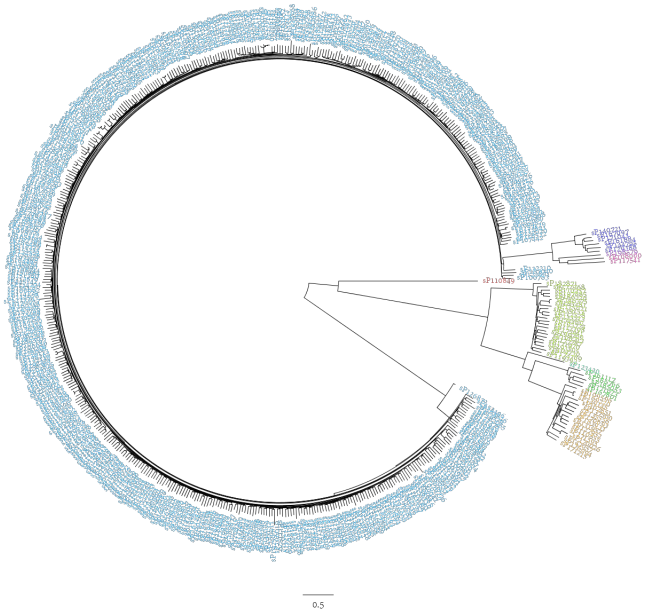

FIGURE 4.1: HCV Phylogenetic Tree

In what follows, we solely focus on the 409 patients of self-reported white ethnicity infected with HCV genotype 3a. Out of the total number of $9,775$ aligned nucleotides corresponding to the HCV coding region, we filtered out 708 nucleotides with a deletion rate above 90%, $4,108$ which displayed no variation within the patient cohort and $2,207$ with major allele frequency above 95%, resulting in a remaining number of $2,752$ nucleotide variants with number of alleles ranging from 2 to 5, based on which we constructed our weight-adjusted pathogen GRM. In contrast, we note that the restriction imposed by the previously proposed weighted pathogen GRM [8] would amount to discarding $4,026$ nucleotide variants with minor allele frequency below 5%, resulting in a remaining number of just 933 nucleotide variants for the computation of the viral GRM. Plotting the scores of the top 2 eigenvectors of our proposed weight-adjusted pathogen GRM, shown in Figure 4.2, unveiled signs of a 3-subpopulation structure within this patient subcohort. The top 2 eigevalues of the weight-adjusted GRM were approximately equal to 7.91 and 5.59, whereas the rest of the eigenvalues were smaller than 3.

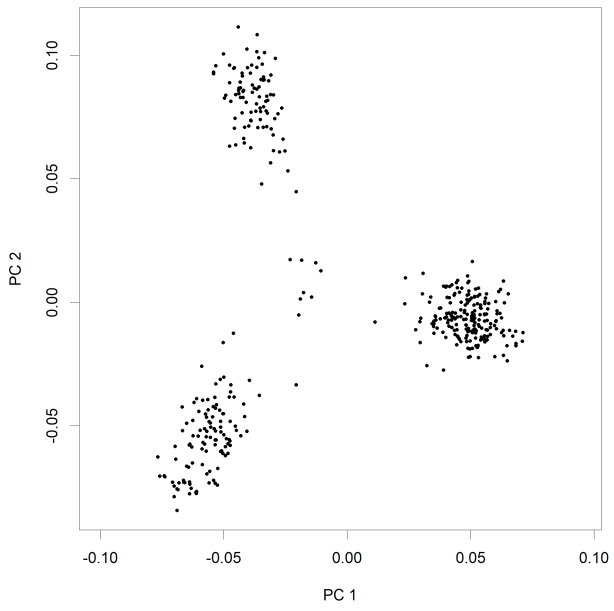

FIGURE 4.2: Top Eigenvector Scores of HCV Weight-Adjusted GRM Based on Patients with Self-Reported White Ethnicity Infected with HCV Genotype 3a

Comparing the diagonal and off-diagonal elements of the weight-adjusted GRM against those of the pathogen GRM constructed on the basis of binary nucleotide allele indicators and the previously

proposed weighted pathogen GRM, shown in Figures 4.3 and 4.4, we observe that the weight-adjusted GRM is in close agreement with the binarized GRM. On the other hand, the previously proposed weighted pathogen GRM displays much higher variation due to having to discard a much larger proportion of nucleotide variants before its computation. For reference, we also do the same GRM calculations based on the entire sample of patients. A plot of the top eigenvectors of the corresponding weight-adjusted GRM is shown in Figure 4.5. Comparisons of the diagonal and off-diagonal elements of the different viral GRMs under consideration are displayed in Figures 4.6 and 4.7.

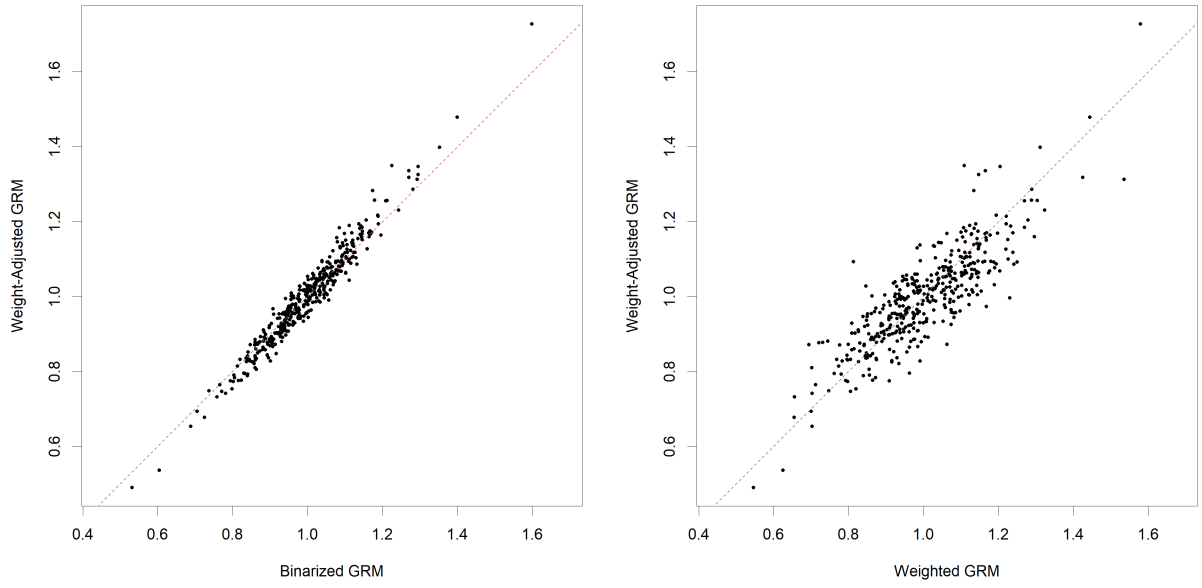

FIGURE 4.3: Comparison of Diagonal Elements of Different HCV GRMs Based on Patients with Self-Reported White Ethnicity Infected with HCV Genotype 3a

Out of the total number of $332,954$ human SNPs, we filtered out $7,616$ SNPs with a minor allele frequency smaller than 5% and 297 SNPs with a chi-square goodness of fit p-value for Hardy-Weinberg equilibrium smaller than $5 \cdot 10^{-8}$. Then, we performed LD pruning on the human genome using PLINK version 1.9, in order to obtain a subset of SNPs with pairwise correlation coefficients smaller than 0.5. This procedure resulted in $184,937$ SNPs being pruned, so we constructed a human GRM based on the remaining $140,104$ SNPs. Plotting the scores of the top 2 eigenvectors of the human GRM, displayed in Figure 4.8, unveiled some heterogeneity within the patient cohort

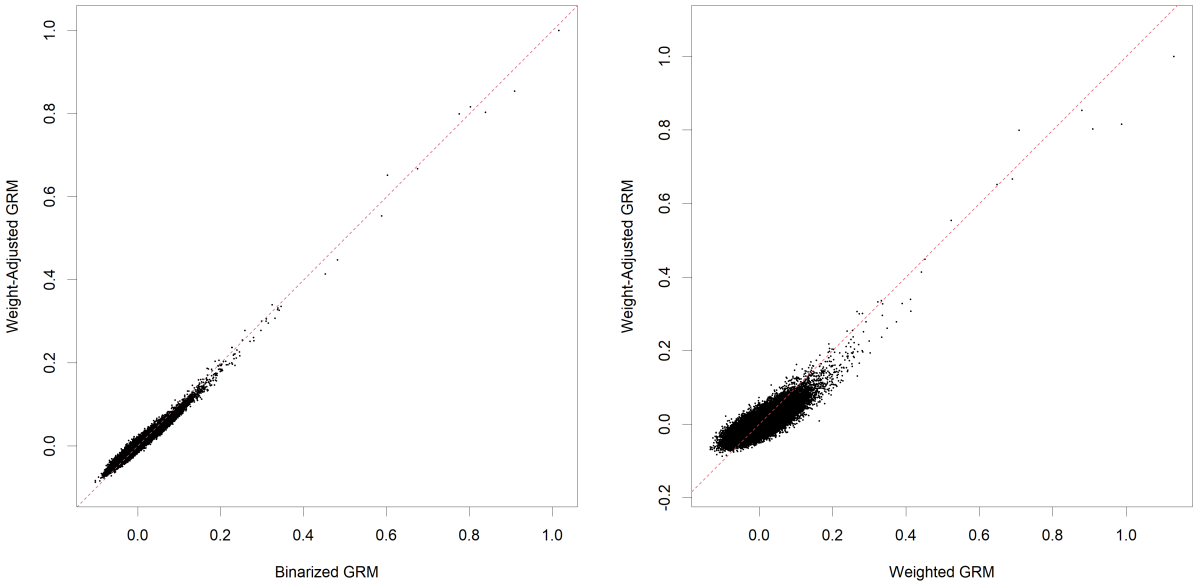

FIGURE 4.4: Comparison of Off-Diagonal Elements of Different HCV GRMs Based on Patients with Self-Reported White Ethnicity Infected with HCV Genotype 3a

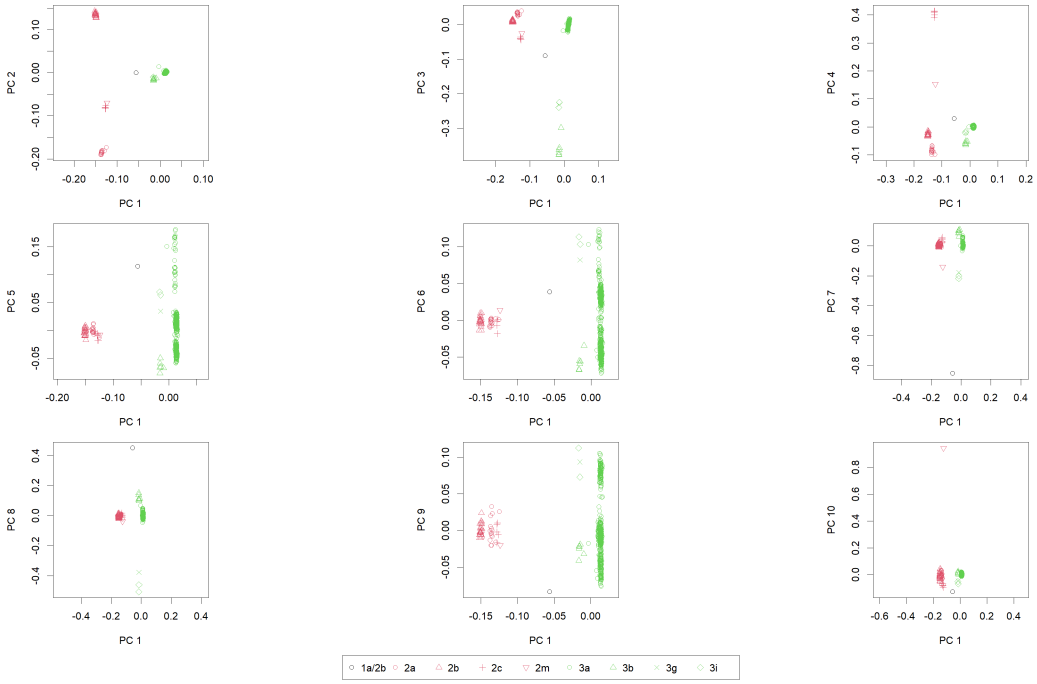

FIGURE 4.5: Top Eigenvector Scores of HCV Weight-Adjusted GRM Based on Entire Sample of Patients

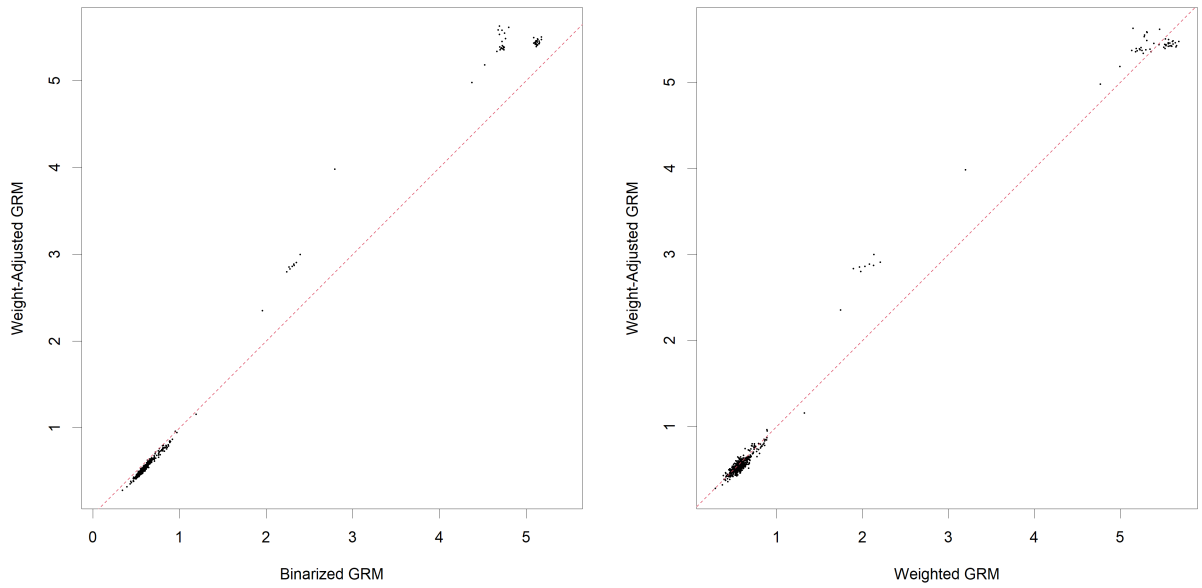

FIGURE 4.6: Comparison of Diagonal Elements of Different HCV GRMs Based on Entire Sample of Patients

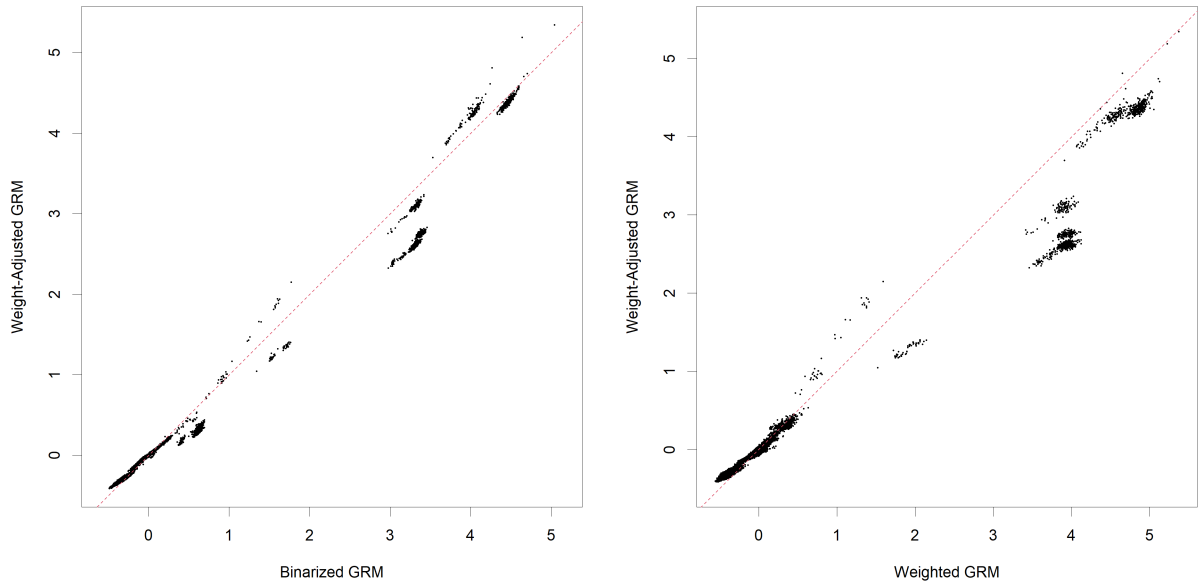

FIGURE 4.7: Comparison of Off-Diagonal Elements of Different HCV GRMs Based on Entire Sample of Patients

of self-reported white ethnicity. Nevertheless, the top eigenvectors of the human GRM explained a negligible portion of the total variation in the trait and the marginal GWAS on the human genome was well-calibrated with a genomic control inflation factor of 1.003 despite lack of adjustment for population structure. For reference, we also provide a plot of the top eigenvectors of the corresponding human GRM based on the entire sample of patients, shown in Figure 4.9.

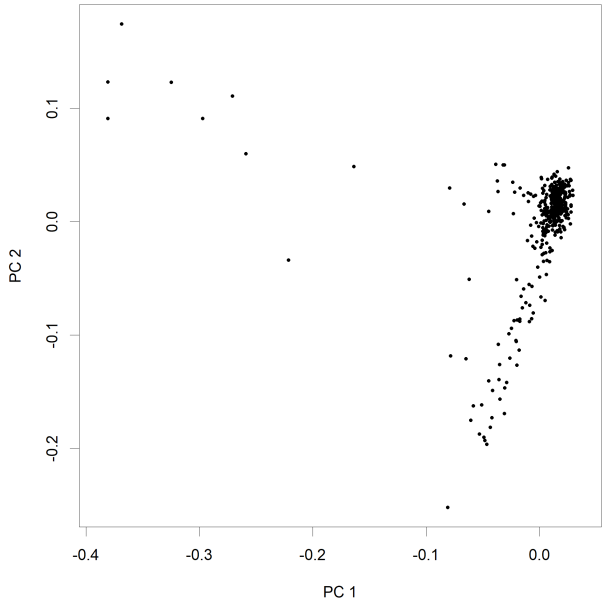

FIGURE 4.8: Top Eigenvector Scores of Human GRM Based on Patients with Self-Reported White Ethnicity Infected with HCV Genotype 3a

4.4 Marginal Association Analyses

Since the distribution of PTVL was heavily skewed and displayed extremely high variation, we chose to log-transform it and use log-PTVL as our phenotypic trait. Additionally, the patients' age and sex had a negligible effect on the trait, so they were ignored. Out of the total number of $332,954$ human SNPs, we filtered out 77 SNPs whose genotype was partially missing after imputation of the human genome, 1 SNP with a minor allele frequency smaller than 1% and 309 SNPs with a chi-square goodness of fit p-value for Hardy-Weinberg equilibrium smaller than $5 \cdot 10^{-8}$, resulting in a remaining number of $332,567$ human SNPs. A Manhattan plot of the human association analysis p-values is displayed in Figure 4.10. We identified 3 human SNPs in chromosome 19 with a genome-wide significant association to log-PTVL: rs8103142 (position 39735106) in the IFNL3

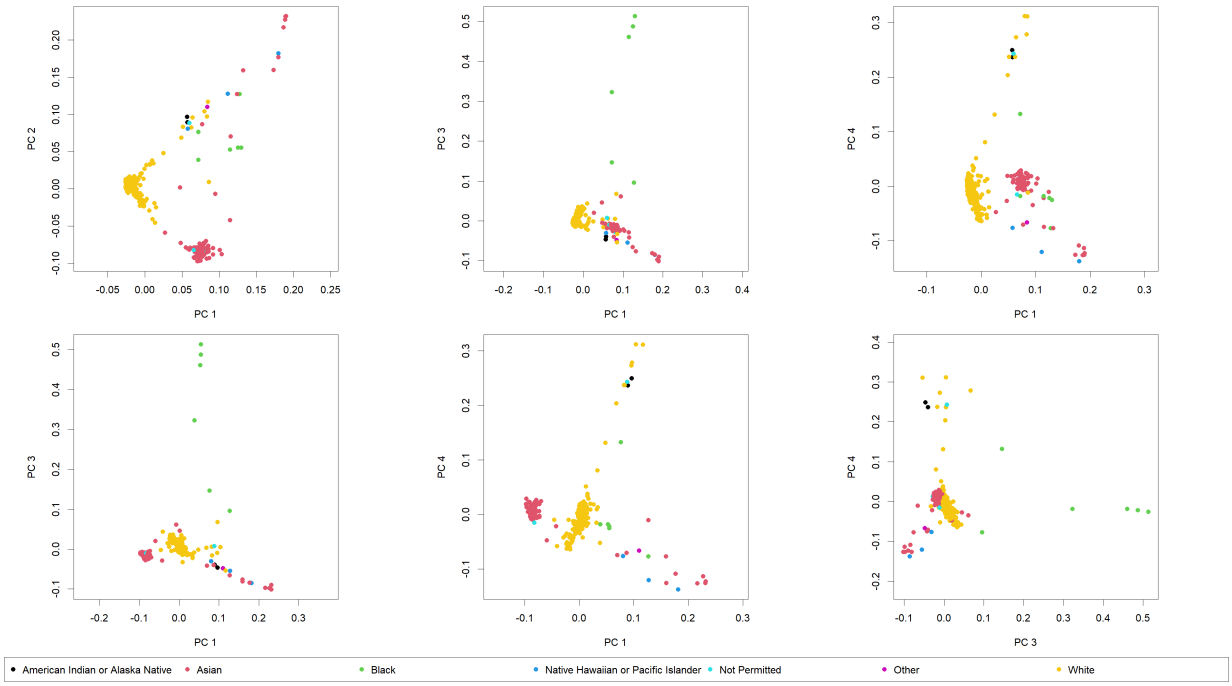

FIGURE 4.9: Top Eigenvector Scores of Human GRM Based on Entire Sample of Patients

gene $\left(\text{p-value } 2.11 \cdot 10^{-8}\right)$, rs12979860 (position 39738787) in the IFNL4 gene $\left(\text{p-value } 1.85 \cdot 10^{-8}\right)$ and rs8099917 (position 39743165) close to the IFNL4 gene $\left(\text{p-value } 1.44 \cdot 10^{-8}\right)$. The first two SNPs displayed a correlation of 0.99 in our cohort, while the third one displayed correlations of 0.73 and 0.72 with each of the first two SNPs respectively. SNP rs12979860 had previously been strongly associated with sustained virological response in a GWAS [84] of 1,137 chronically infected European-American, African-American and Hispanic individuals with HCV genotype 1 who were part of the IDEAL study [85]. This association had been replicated $\left(\text{p-value } 5.9 \cdot 10^{-10}\right)$ on the infected patients with HCV genotype 3a from the BOSON clinical trial [9].

Out of the total number of 3,281 aligned HCV amino acids, we filtered out 260 amino acids with a deletion rate above 90% - leading to a total of 3,021 amino acids - and 1,378 which displayed no variation within the patient cohort, resulting in a remaining number of 1,643 amino acid variants with number of alleles ranging from 2 to 18. Next, we turned each of these amino acid variants into binary amino acid allele indicators, resulting in 4,843 allele indicators. We filtered out 2,722 allele indicators with a minor allele frequency smaller than 1%, resulting in a remaining number of 2,121 allele indicators. A Manhattan plot of the viral amino acid association analysis p-values is

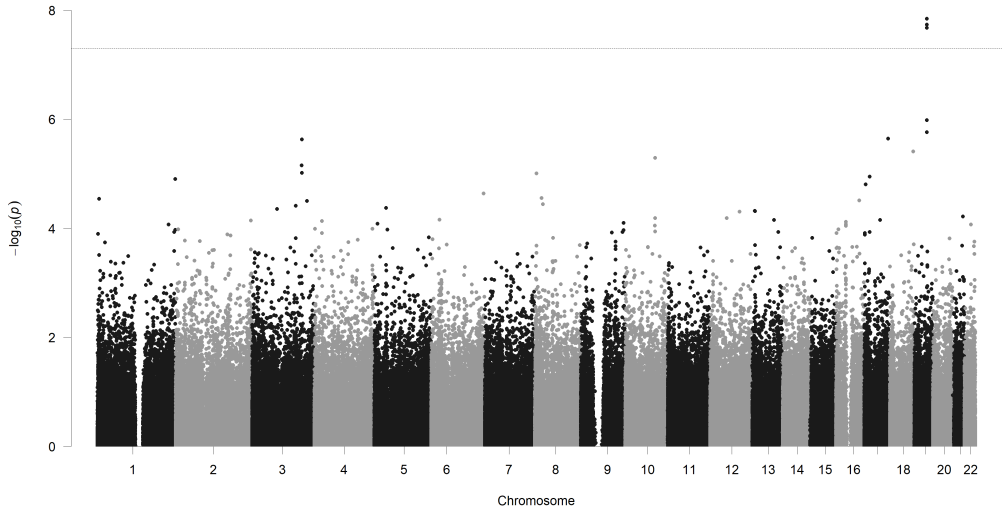

FIGURE 4.10: Manhattan Plot of Human GWAS on log-PTVL

displayed in Figure 4.11. The HCV chromosome is divided into the structural proteins Core, E1, E2 and the non-structural proteins P7, NS2, NS3, NS4A, NS4B, NS5A, NS5B. The marginal GWAS on the amino acid allele indicators, adjusting for the top 2 eigenvectors of the weight-adjusted viral GRM, yielded a few moderate association signals with the top among them being the following: the serine indicator (allele count 320) of the HCV amino acid at position $2,422$ in the NS5A protein $\left(\text{p-value } 1.75 \cdot 10^{-5}\right)$ and the glycine indicator (allele count 26) of the HCV amino acid at position $1,831$ in the NS4B protein $\left(\text{p-value } 7.49 \cdot 10^{-5}\right)$. The NS5A protein is known to contribute to HCV pathogenesis, replication, propagation, modulation of cell signaling pathways and response to interferon treatment [86].

4.5 Feast or Famine Simulation

Before moving on to a joint association analysis of log-PTVL, we first evaluate the prevalence of the feast or famine effect with respect to log-PTVL and the observed HCV amino acid allele indicators. We set the number of simulated host genetic variants m_h to be equal to $10,000$. For each amino acid allele indicator, we simulate independent host allele frequencies $f_{X_1}, f_{X_2}, \ldots, f_{X_{m_h}} \sim \text{Unif}[0.1, 0.9]$ and independent host genotypes $X_{ij} \sim \text{Binomial}\left(2, f_{X_j}\right)$ for $j = 1, 2, \ldots, m_h$.

Focusing on HCV amino acid allele indicators with a minor allele count of at least 80, the aspartate

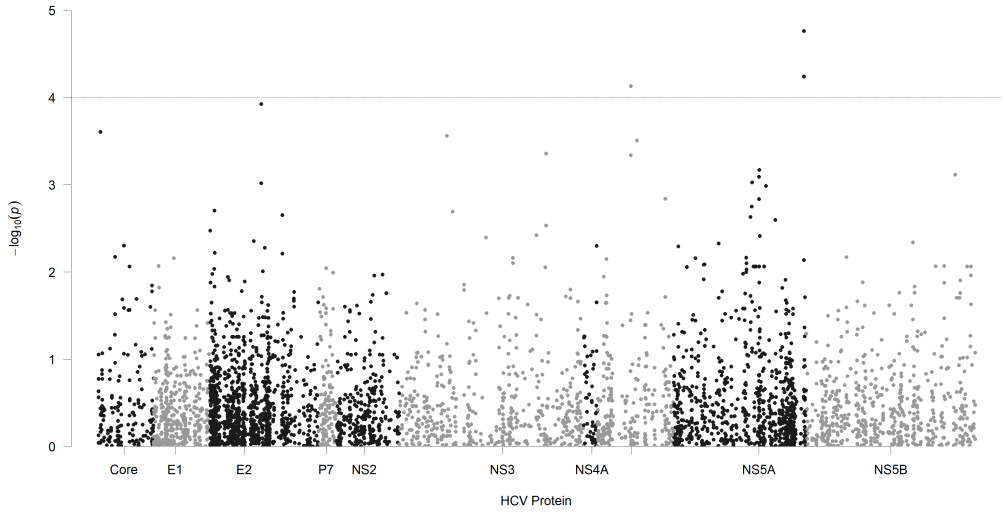

FIGURE 4.11: Manhattan Plot of HCV Amino Acid Allele Indicator GWAS on log-PTVL

indicator (allele count 86) of the HCV amino acid at position 578 in the E2 protein displayed the highest genomic control inflation factor of 1.2, while the phenylalanine indicator (allele count 308) of the HCV amino acid at position 586 in the E2 protein displayed the lowest genomic control inflation factor 0.77. Q-Q plots of the corresponding interaction p-values before and after correction are displayed in Figure 4.12. We observe that all of our proposed correction methods perform similarly in terms of correcting the feast or famine effect in these instances.

Furthermore, we evaluate the performance of our proposed diagnostic ratio on the HCV data set. A scatterplot of the uncorrected genomic control inflation factors against the diagnostic ratio, shown in Figure 4.13, verifies the strong linear relationship between them even based on real HCV amino acid allele indicators. We note that the sample correlation between these 2 quantities is calculated to be 94.78%. This linear relationship does not appear as strong as the one observed based on simulated pathogen genetic variants, but that mostly comes down to the prevalence of smaller Z minor allele counts in the viral genome, leading to the discarding of a substantial proportion of tested pairs with minimum cell counts below 5 and the calculation of genomic control inflation factors on the basis of fewer than $m_h = 10,000$ observed test statistics. We also calculate the 5% sample quantiles for each collection of uncorrected interaction p-values $p_{1k}, p_{2k}, \ldots, p_{m_h k}$ and plot them on the $-\log_{10}$ scale against the diagnostic ratio, shown in Figure 4.14. We observe

that the diagnostic ratio performs even better than previous studies with simulated quantitative traits and pathogen genetic variants in predicting the tail behavior of the uncorrected interaction p-values, partly owing to the prevalence of smaller Z minor allele counts. As a reference, the sample correlation between these 2 quantities was calculated to be equal to 99.29%.

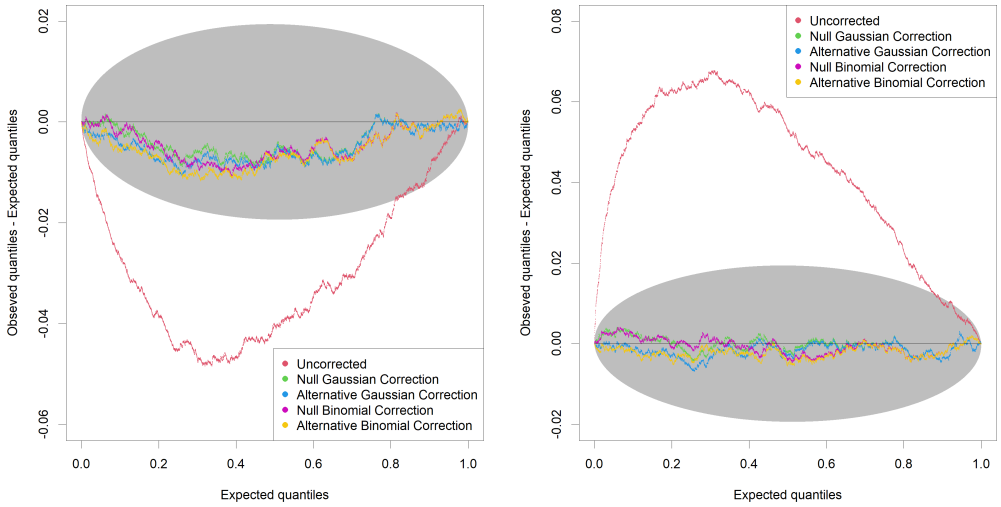

FIGURE 4.12: Q-Q Plots Displaying the Correction of the Feast or Famine Effect Given Real HCV Amino Acid Allele Indicators

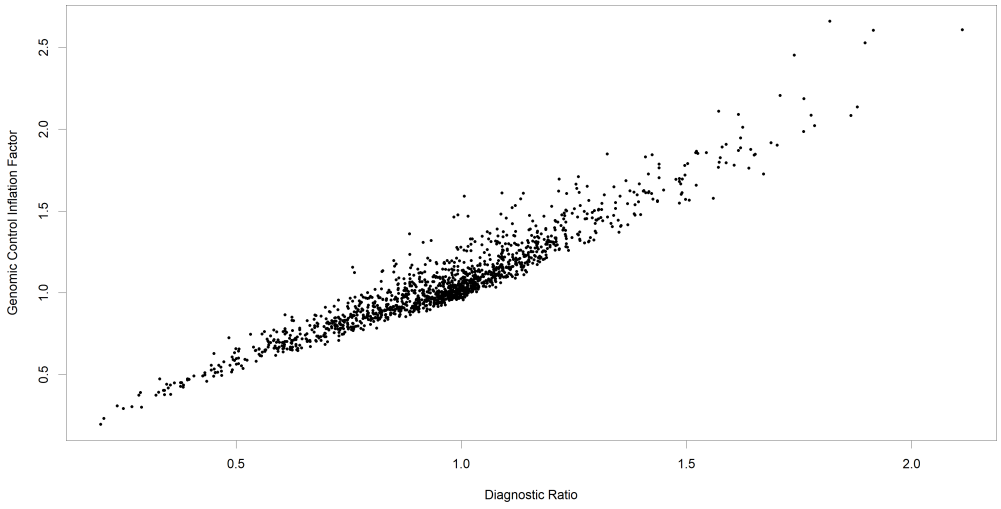

FIGURE 4.13: Scatterplot of Uncorrected Genomic Control Inflation Factors vs. Diagnostic Ratio Given Real HCV Amino Acid Allele Indicators

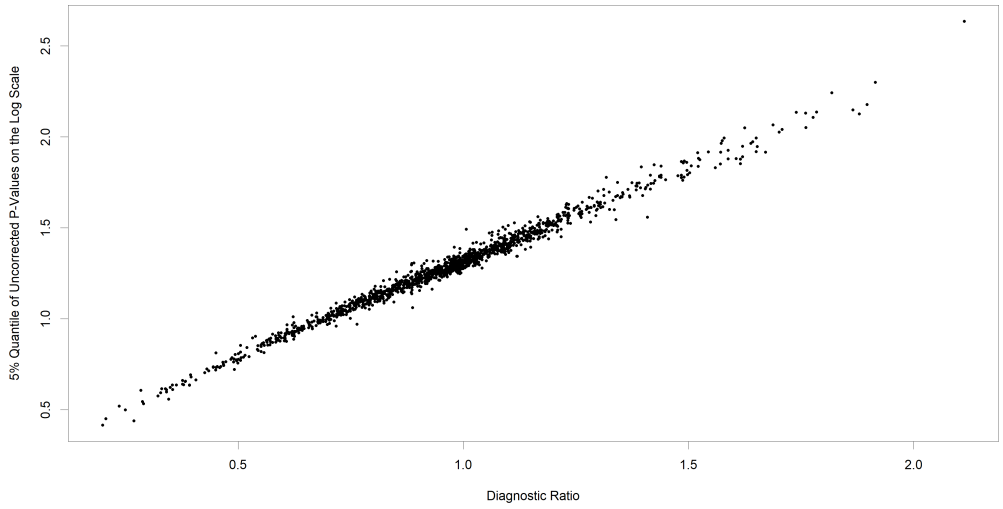

FIGURE 4.14: Scatterplot of 5% Quantiles of Uncorrected P-Values on the Log Scale vs. Diagnostic Ratio Given Real HCV Amino Acid Allele Indicators

4.6 Joint Association Analysis

We filtered out pairs of human SNPs and binary HCV amino acid allele indicators with a minimum cell count smaller than 5, resulting in a remaining number of $257,144,224$ pairs for the purposes of joint association mapping, adjusting for the top 2 eigenvectors of the weight-adjusted viral GRM. For any pair with a human SNP minor allele count larger than or equal to 50, a binary HCV amino acid indicator minor allele count larger than or equal to 80 and a minimum cell count larger than or equal to 10, we calculated a corrected interaction test statistic based on our proposed alternative binomial correction. The number of such pairs was calculated to be equal to $58,259,674$, which accounted for 22.66% of the total number of performed interaction tests. For any other pair, we calculated a corrected interaction test statistic based on our proposed null binomial correction. The top results of the joint GWAS, are displayed in Tables 4.1 through 4.3. The same procedure was also carried out for the 182 HLA haplotype indicators provided by HLA*IMP:03, the $3,209$ HLA amino acid indicators provided by the Michigan Imputation Server and the $10,351$ imputed SNPs flanking the HLA region also provided by the Michigan Imputation Server.

In Tables 4.1 and 4.2, we have listed the top interaction signals between binary HCV amino acid allele indicators and human genetic variants. For each amino acid indicator, we have listed the HCV

protein to which it belongs and its observed alleles in the patient cohort with the corresponding allele counts in parentheses. The alleles to which the identified interaction signal correspond are displayed in bold text. Next to each identified amino acid indicator, we have listed the human genetic variants to which the strongest interaction signals correspond with their chromosomal coordinates in parentheses. For each associated variant, we have listed the closest identified human gene and the corresponding interaction p-values. The uncorrected interaction p-values are displayed in parentheses, while the corrected interaction p-values based on the alternative binomial correction framework displayed in bold text. Lastly, we have listed the minimum cell count for the pair of amino acid indicator and its associated variant with the minor allele count of the associated variant in parentheses.

SNPs in the MICB gene displayed multiple strong to moderate interactions with HCV amino acid indicators - the top among those were that of SNP rs3131638 with the glycine indicator of the HCV amino acid at position 398 in the E2 protein $\left(\text{corrected p-value } 3.83 \cdot 10^{-8}\right)$, that of SNP rs2507971 with the threonine indicator of the HCV amino acid at position 400 in the E2 protein $\left(\text{corrected p-value } 7.87 \cdot 10^{-6}\right)$ and that of SNP rs2507971 with the methionine indicator of the HCV amino acid at position 330 in the E1 protein $\left(\text{corrected p-value } 8.84 \cdot 10^{-6}\right)$. The MICA and MICB genes have proven to be induced on dendritic cells by IFN-α treatment and to be capable of activating natural killer cells in a cohort of Japanese HCV-infected patients and control cases [87]. The MICB gene has been identified to be among 5 genes within the MHC class III - class I boundary which are strongly associated with HCV-related dilated cardiomyopathy in a cohort of Japanese patients with HCV-related dilated cardiomyopathy and control cases [88]. The MICB gene has also been shown to be a strong predictive factor for sustained virological response to pegylated interferon plus ribavirin therapy in a cohort of Japanese HCV-infected patients [89].

Genetic variants corresponding to MHC class I and II displayed multiple moderate interactions with HCV amino acid indicators - the top among those were that of SNP rs562289 in the HLA-DRB1 gene with the asparagine indicator of the HCV amino acid at position 529 in the E2 protein $\left(\text{corrected p-value } 1.11 \cdot 10^{-6}\right)$, that of amino acid 97 in the HLA-A gene with the methionine indicator of the HCV amino acid at position 635 in the E2 protein $\left(\text{corrected p-value } 1.22 \cdot 10^{-6}\right)$ and that of SNP rs3897530 in the HLA-B gene with the serine indicator of the HCV amino acid at

AA Pos.	HCV Prot.	AA Alleles (Counts)	Associated Variants (Position)	Associated Genes	P-Values (Uncorrected)	MCC (MAC)
398	E2	A(15), E(1), F(4), **G(221)**, I(4), K(6), L(3), M(9), N(1), R(31), S(59), T(43), V(10)	rs3131638 (31475127)	MICB	$\mathbf{3.83 \cdot 10^{-8}}$ $\left(1.3 \cdot 10^{-7}\right)$	46 (101.5)
			HLA_B*07:02 (31321650)	HLA-B	$\mathbf{2.89 \cdot 10^{-3}}$ $\left(4.53 \cdot 10^{-3}\right)$	28.5 (63.5)
529	E2	A(2), D(9), E(3), G(3), K(4), **N(100)**, Q(4), R(4), S(21), T(257)	rs562289 (32577046)	HLA-DRB1	$\mathbf{1.11 \cdot 10^{-6}}$ $\left(5.72 \cdot 10^{-6}\right)$	20 (74)
			HLA_DQA1*03:01 (32605197)		$\mathbf{9.11 \cdot 10^{-4}}$ $\left(1.4 \cdot 10^{-3}\right)$	18 (67)
635	E2	A(1), **M(26)**, V(382)	AA_A_97_29911063_exon3	HLA-A	$1.22 \cdot 10^{-6}$ $\left(2.58 \cdot 10^{-7}\right)$	8 (174)
1099	NS3	A(236), D(1), F(2), I(3), T(5), **V(161)**	rs3129771 (32597202)	HLA-DRB1	$\mathbf{4.72 \cdot 10^{-6}}$ $\left(1.29 \cdot 10^{-5}\right)$	59.5 (146)
			HLA_DRB4*01:03 (3830849)	HLA-DRB4	$\mathbf{7.53 \cdot 10^{-6}}$ $\left(2.18 \cdot 10^{-5}\right)$	29.5 (82)
2154	NS5A	A(6), I(3), K(4), L(6), **M(265)**, S(7), T(88), V(35)	rs1264368 (30771612)	HCG20	$\mathbf{4.78 \cdot 10^{-6}}$ $\left(9.31 \cdot 10^{-7}\right)$	39.5 (103.5)
			AA_B_77_31324506_exon2	HLA-B	$1.87 \cdot 10^{-3}$ $\left(1.11 \cdot 10^{-3}\right)$	12.5 (31.5)
541	E2	A(7), D(18), E(183), G(7), H(2), **K(128)**, N(14), P(1), Q(18), R(10), S(1), T(18)	AA_DRB1_37_32552051_exon2	HLA-DRB1	$\mathbf{5.38 \cdot 10^{-6}}$ $\left(1.49 \cdot 10^{-5}\right)$	43.5 (124)
			HLA_DQB1*03:01 (32627257)	HLA-DQB1	$\mathbf{2.17 \cdot 10^{-3}}$ $\left(2.14 \cdot 10^{-3}\right)$	18.5 (63)
384	E2	−(5), A(17), D(25), E(87), G(29), H(14), I(1), K(4), N(39), Q(20), R(4), **S(61)**, T(97), V(1), Y(4)	rs3909115 (30993188)	MUC21	$6.1 \cdot 10^{-6}$ $\left(3.43 \cdot 10^{-5}\right)$	13.5 (94.5)
2076	NS5A	C(4), G(62), **S(343)**	rs3897530 (31323469)	HLA-B	$6.2 \cdot 10^{-6}$ $\left(3.43 \cdot 10^{-5}\right)$	8.5 (56.5)
			HLA_C*04:01	HLA-C	$1.18 \cdot 10^{-3}$ $\left(2.12 \cdot 10^{-3}\right)$	8 (41)
3001	NS5B	H(270), R(1), **Y(138)**	AA_DRB1_37_32552051_exon2	HLA-DRB1	$\mathbf{6.77 \cdot 10^{-6}}$ $\left(1.54 \cdot 10^{-5}\right)$	52.5 (147.5)
			HLA_A*11:01	HLA-A	$5.22 \cdot 10^{-3}$ $\left(5.28 \cdot 10^{-3}\right)$	9.5 (28)
614	E2	I(6), L(91), **M(312)**	AA_B_114_31324150_exon3	HLA-B	$\mathbf{7.36 \cdot 10^{-6}}$ $\left(8.89 \cdot 10^{-6}\right)$	45.5 (195)
400	E2	A(137), F(5), G(1), H(1), K(1), L(3), M(2), N(3), S(24), **T(179)**, V(46), Y(5)	rs2507971 (31461372)	MICB	$\mathbf{7.87 \cdot 10^{-6}}$ $\left(2.96 \cdot 10^{-5}\right)$	66.5 (163)
			HLA_C*07:02 (31236654)	HLA-C	$\mathbf{7.19 \cdot 10^{-3}}$ $\left(1.01 \cdot 10^{-2}\right)$	27.5 (69.5)
879	NS2	**A(88)**, F(4), I(33), T(2), V(282)	rs2261033 (31603591)	PRRC2A	$\mathbf{8.37 \cdot 10^{-6}}$ $\left(7.13 \cdot 10^{-5}\right)$	39.5 (168)
			HLA_DQA1*01:02 (32605186)	HLA-DQA1	$\mathbf{7.56 \cdot 10^{-4}}$ $\left(2.47 \cdot 10^{-3}\right)$	16.5 (92)
330	E1	A(19), I(2), L(8), **M(13)**, V(367)	rs2507971 (31461372)	MICB	$8.84 \cdot 10^{-6}$ $\left(2.71 \cdot 10^{-4}\right)$	5.5 (163)

TABLE 4.1: Top Interaction Signals between HCV Amino Acid Allele Indicators and HLA Variants on log-PTVL.

position 2,076 in the NS5A protein $\left(\text{corrected p-value } 6.2 \cdot 10^{-6}\right)$. Multiple HLA haplotypes have previously been linked with either spontaneous clearance or poor prognosis of HCV infection [90].

SNPs adjacent to the HCG20 gene within the HLA complex displayed a few moderate interactions with HCV amino acid indicators - the top among those was that of SNP rs1264368 with the methionine indicator of the HCV amino acid at position $2,154$ in the NS5A protein $\left(\text{corrected p-value } 4.78 \cdot 10^{-6}\right)$. The HCG20 gene has been identified to be among 8 hub differentially expressed long non-coding RNAs in a cohort of 373 patients with hepatocellular carcinoma and 50 control cases from the Cancer Genome Atlas [91].

SNPs adjacent to the MUC21 gene within the HLA complex displayed a few moderate interactions with HCV amino acid indicators - the top among those was that of SNP rs3909115 with the serine indicator of the HCV amino acid at position 384 in the E2 protein $\left(\text{corrected p-value } 6.1 \cdot 10^{-6}\right)$. Expressions levels of the MUC15, MUC13 and MUC21 genes have been individually associated with survival for digestive cancers in a cohort of patients from the Cancer Genome Atlas [92].

SNPs in the PRRC2A gene within the HLA complex displayed a few moderate interactions with HCV amino acid indicators - the top among those was that of SNP rs2261033 with the alanine indicator of the HCV amino acid at position 879 in the NS2 protein $\left(\text{corrected p-value } 8.37 \cdot 10^{-6}\right)$. High expression of the PRRC2A gene has been associated with poor prognosis in Chinese patients with hepatocellular carcinoma [93], [94].

SNPs adjacent to the NAV2-AS4 gene in chromosome 11 displayed a few strong interactions with HCV amino acid indicators - the top among those was that of SNP rs10082600 with the valine indicator of the HCV amino acid at position $1,651$ in the NS3 protein $\left(\text{corrected p-value } 2.95 \cdot 10^{-8}\right)$. The NAV2-AS4 gene has been found to be significant in predicting overall survival for a cohort of 371 patients with hepatocellular carcinoma from the Cancer Genome Atlas platform [21].

SNPs adjacent to the ZWINT gene in chromosome 10 displayed a few strong interactions with HCV amino acid indicators - the top among those was that of SNP rs7091063 with the alanine indicator of the HCV amino acid at position $1,882$ in the NS4B protein $\left(\text{corrected p-value } 4.13 \cdot 10^{-8}\right)$. The ZWINT gene has been found to be significant in predicting overall survival and making prognostic risk assessments for a cohort of patients with hepatocellular carcinoma from the Cancer Genome Atlas and the International Cancer Genome Consortium [95].

AA Pos.	HCV Prot.	AA Alleles (Counts)	Associated SNPs (Position)	Associated Genes	P-Values (Uncorrected)	MCC (MAC)
628	E2	A(2), F(3), I(58), L(14), Q(1), **V(331)**	rs11246973 (12:132671437)	LINC02361	$1.16 \cdot 10^{-8}$ $\left(1.87 \cdot 10^{-8}\right)$	22 (135)
576	E2	−(2), A(7), D(8), **E(98)**, G(255), H(1), K(7), P(2), R(11), S(4), T(6), V(5)	rs2304389 (9:101068580)	GABBR2	$\mathbf{1.25 \cdot 10^{-8}}$ $\left(1.33 \cdot 10^{-7}\right)$	18 (64)
1435	NS3	**A(181)**, G(1), S(39), T(188)	rs6024051 (20:53974654)	RPL12P4	$\mathbf{1.52 \cdot 10^{-8}}$ $\left(2.29 \cdot 10^{-8}\right)$	23.5 (53.5)
1651	NS3	I(301), L(7), **V(100)**	rs10082600 (11:20087893)	NAV2-AS4	$\mathbf{2.95 \cdot 10^{-8}}$ $\left(2.52 \cdot 10^{-7}\right)$	43.5 (173)
1882	NS4B	**A(235)**, G(12), P(2), S(2), T(158)	rs7091063 (10:57739268)	ZWINT	$\mathbf{4.13 \cdot 10^{-8}}$ $\left(4.79 \cdot 10^{-8}\right)$	23 (52)
433	E2	I(319), **L(81)**, M(1), T(1), V(5)	rs964841 (12:62286036)	TAFA2	$\mathbf{5.84 \cdot 10^{-8}}$ $\left(1.58 \cdot 10^{-6}\right)$	37.5 (186)
2339	NS5A	I(363), L(2), **V(43)**	rs4742488 (9:8371405)	PTPRD	$6.03 \cdot 10^{-8}$ $\left(8.49 \cdot 10^{-7}\right)$	5 (67.5)
2804	NS5B	H(26), I(2), L(131), M(1), Q(38), **R(201)**, V(1), W(9)	rs2662464 (5:115356017)	LVRN	$\mathbf{6.07 \cdot 10^{-8}}$ $\left(1.66 \cdot 10^{-7}\right)$	55 (114)
501	E2	A(70), E(1), I(1), K(6), **L(111)**, P(8), Q(5), R(12), S(180), T(8), V(5)	rs6585043 (10:113010056)	HEAT2	$\mathbf{6.35 \cdot 10^{-8}}$ $\left(1.39 \cdot 10^{-6}\right)$	31 (120.5)
464	E2	A(8), D(7), F(201), H(29), K(1), L(1), N(1), Q(1), **S(128)**, T(1), Y(30)	rs17815047 (12:71920906)	LGR5	$\mathbf{7.11 \cdot 10^{-8}}$ $\left(2.94 \cdot 10^{-7}\right)$	36.5 (115)
2388	NS5A	E(60), **G(328)**, I(1), R(12), S(2), V(3)	rs2242471 (4:76878716)	SDAD1	$7.52 \cdot 10^{-8}$ $\left(6.37 \cdot 10^{-7}\right)$	30.5 (163)
748	E2	A(9), **S(327)**, T(70), V(1)	rs35563441 (3:84603909)	LINC00971	$\mathbf{8.21 \cdot 10^{-8}}$ $\left(2.45 \cdot 10^{-6}\right)$	16 (84.5)
951	NS2	C(26), **F(381)**, L(1)	rs4239261 (17:5980100)	WSCD1	$8.25 \cdot 10^{-8}$ $\left(9.17 \cdot 10^{-7}\right)$	5.5 (72)
208	E1	A(6), H(1), **P(114)**, R(2), S(285)	rs357368 (7:137979927)	RPS17P12	$\mathbf{9.53 \cdot 10^{-8}}$ $\left(3.37 \cdot 10^{-6}\right)$	43.5 (143)
386	E2	−(1), D(3), F(1), G(1), **H(142)**, I(2), L(3), N(1), P(2), Q(3), R(60), T(4), V(2), W(1), Y(182)	rs4284283 (1:17796711)	ARHGEF10L	$\mathbf{9.54 \cdot 10^{-8}}$ $\left(1.9 \cdot 10^{-7}\right)$	46.5 (147.5)

TABLE 4.2: Top Interaction Signals between HCV Amino Acid Allele Indicators and Human SNPs on log-PTVL.

SNPs in the PTPRD gene in chromosome 9 showed a few strong interactions with HCV amino acid indicators - the top among those was that of SNP rs4742488 with the valine indicator of the HCV amino acid at position $2,339$ in the NS5A protein $\left(\text{corrected p-value } 6.03 \cdot 10^{-8}\right)$. The tumor suppressor PTPRD gene has proven to be impaired by HCV infection in hepatocellular

carcinoma lesions. On the other hand, high PTPRD levels in liver tissue adjacent to tumor have been associated with increased survival rate and reduced tumor recurrence in patients undergoing surgical resection at the Gastroenterology and Hepatology outpatient clinic of the Basel University Hospital, Switzerland, the Centre Hospitalier Universitaire de Reims, France and the Hôpitaux Universitaires de Strasbourg, France [96].

SNPs in the LGR5 gene in chromosome 12 displayed a few strong interactions with HCV amino acid indicators - the top among those was that of SNP rs17815047 with the serine indicator of the HCV amino acid at position 464 in the E2 protein $\left(\text{corrected p-value } 7.11 \cdot 10^{-8}\right)$. High protein levels of the LGR5 gene have been associated with poor prognosis in a cohort 66 hepatocellular carcinoma patients who underwent curative surgery at Zhejiang Provincial People's Hospital from 2008 to 2015 and another cohort admitted to the Division of General Surgery, Department of Surgery, Changhua Christian Hospital, Taiwan between November 2013 and September 2017 [97], [98].

SNPs adjacent to the ARHGEF10L gene in chromosome 1 displayed a few strong interactions with HCV amino acid indicators - the top among those was that of SNP rs4284283 with the histidine indicator of the HCV amino acid at position 386 in the E2 protein $\left(\text{corrected p-value } 9.54 \cdot 10^{-8}\right)$. Increased expression of the ARHGEF10L gene has been found to stimulate the tumor cell proliferation and cell migration in a cohort of patients with hepatocellular carcinoma from the Shandong Provincial Qianfoshan Hospital [99].

Amino acid indicators in the E2 protein displayed multiple strong to moderate interactions with human genetic variants - exactly half of the top interaction signals presented in tables 4.1 and 4.2 correspond to this HCV protein. The E2 protein contains two hypervariable regions - HVR1 and HVR2 - which are the most mutable parts of the HCV genome. This heterogeneity can potentially aid the virus in evading the host immune response and developing into chronic infection. The HVR2 region has been shown to contribute to viral receptor binding, while the HVR1 region is the most frequent target for neutralizing antibodies [86].

Lastly, the identified human SNP rs12979860 in the IFNL4 gene from the marginal human GWAS displayed moderate interaction signals with the aspartate allele indicator of the HCV amino acid at position 571 in the E2 protein $\left(\text{corrected p-value } 3.15 \cdot 10^{-4}\right)$ and the leucine allele indicator of the

AA Position	List of Associated Variants
398	rs3131638, rs3095234, rs3130927, rs3134899, rs1065076, rs3130615, rs3132468, rs3131635
529	rs562289, rs9267951, rs1846190
635	rs376646, rs2975043, rs417162, rs2508037, rs9260112
	AA_A_97_29911063_exon3, AA_A_114_29911114_exon3, AA_A_276_29912281_exon5, AA_A_321_29912858_exon6, AA_A_152_29911228_exon3
1099	rs3129771, rs78017592, rs9273138, rs9272990
2076	rs3897530, rs2394986, rs7749442
2388	rs2242471, rs2273

TABLE 4.3: List of Identified Interaction Signals between HCV Amino Acid Allele Indicators and Human Genetic Variants on log-PTVL.

HCV amino acid at position $1,738$ in the NS4B protein $\big($corrected p-value $2.68 \cdot 10^{-4}\big)$. The identified serine allele indicator of the HCV amino acid at position $2,422$ in the NS5A protein displayed a moderate interaction signal with the rs10848105 SNP in the RIMBP2 gene in chromosome 12 $\big($corrected p-value $1.12 \cdot 10^{-7}\big)$. The RIMBP2 gene has been identifited to be among 8 differentially expressed genes for a cohort of 120 patients with advanced hepatocellular carcinoma treated with lenvatinib compared to sorafenib in the Second Hospital of Tianjin Medical University [100].

4.7 Discussion

We performed a joint GWAS testing for interaction effects between human SNPs and hepatitis C viral genetic variants on pre-treatment viral load in a cohort of HCV infected patients from the BOSON clinical trial, focusing on patients of self-reported white ethnicity infected with HCV genotype 3a. We demonstrated that our newly proposed weight-adjusted pathogen GRM performs much better than the previously proposed weighted pathogen GRM [8] in estimating the population structure among different viral strains due to the high mutability of viral genomes leading to a very large number of multiallelic viral genetic variants with rare alleles. We successfully replicated previously identified marginal associations between human SNPs, HCV genetic variants and hepatitis C pre-treatment viral load. We verified through a simulation study the prevalence of the feast or famine effect when testing for interaction effects between certain HCV amino acid allele indicators and simulated human SNPs on pre-treatment viral load. We applied our binomial correction framework to correct for the feast or famine effect and presented a list of the strongest corrected interaction signals between human SNPs and HCV amino acid allele indicators, especially focusing

on the human HLA region which is responsible for the regulation of the immune system. Many of the identified human SNPs belong to human genes which have previously been associated with sustained virological response to interferon therapy and spontaneous clearance or poor prognosis of HCV infection or hepatocellular carcinoma. Many of the identified HCV amino acid allele indicators belong to the E2 protein, which is a highly variable region of the HCV genome, aiding the virus in evading the host immune response and developing into chronic infection, as well as contributing to viral receptor binding.

BIBLIOGRAPHY

[1] J.-H. Park, S. Wacholder, M. H. Gail, *et al.*, "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries," en, *Nat. Genet.*, vol. 42, no. 7, pp. 570–575, Jul. 2010.

[2] E. Silverman, S. Weiss, S. Shapiro, and D. Lomas, *Respiratory Genetics*, en. London, England: Hodder Arnold, Sep. 2005.

[3] H. M. Kang, N. A. Zaitlen, C. M. Wade, *et al.*, "Efficient control of population structure in model organism association mapping," en, *Genetics*, vol. 178, no. 3, pp. 1709–1723, Mar. 2008.

[4] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "FaST linear mixed models for genome-wide association studies," en, *Nat. Methods*, vol. 8, no. 10, pp. 833–835, Sep. 2011.

[5] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," en, *Nat. Rev. Genet.*, vol. 11, no. 7, pp. 459–463, Jul. 2010.

[6] X. Zhou and M. Stephens, "Genome-wide efficient mixed-model analysis for association studies," en, *Nat. Genet.*, vol. 44, no. 7, pp. 821–824, Jun. 2012.

[7] R. A. Fisher, "XV.—the correlation between relatives on the supposition of mendelian inheritance," en, *Trans. R. Soc. Edinb.*, vol. 52, no. 2, pp. 399–433, 1919.

[8] M. Wang, F. Roux, C. Bartoli, *et al.*, "Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 24, E5440–E5449, Jun. 2018.

[9] M. A. Ansari, STOP-HCV Consortium, V. Pedergnana, *et al.*, "Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus," en, *Nat. Genet.*, vol. 49, no. 5, pp. 666–673, May 2017.

[10] L. Crawford, P. Zeng, S. Mukherjee, and X. Zhou, "Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits," en, *PLoS Genet.*, vol. 13, no. 7, e1006869, Jul. 2017.

[11] B. Fu, A. Pazokitoroudi, A. Xue, *et al.*, "A biobank-scale test of marginal epistasis reveals genome-wide signals of polygenic epistasis," en, *bioRxivorg*, Sep. 2023.

[12] H. Zhou and M. S. McPeek, "Overcoming the "feast or famine" effect: Improved interaction testing in genome-wide association studies," en, *bioRxivorg*, Feb. 2024.

[13] A. Dahl, K. Nguyen, N. Cai, M. J. Gandal, J. Flint, and N. Zaitlen, "A robust method uncovers significant context-specific heritability in diverse complex traits," en, *Am. J. Hum. Genet.*, vol. 106, no. 1, pp. 71–91, Jan. 2020.

[14] L. M. Evans, C. H. Arehart, A. D. Grotzinger, *et al.*, "Transcriptome-wide gene-gene interaction associations elucidate pathways and functional enrichment of complex traits," en, *PLoS Genet.*, vol. 19, no. 5, e1010693, May 2023.

[15] F. Vasseur, M. Exposito-Alonso, O. J. Ayala-Garay, *et al.*, "Adaptive diversification of growth allometry in the plant arabidopsis thaliana," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 13, pp. 3416–3421, Mar. 2018.

[16] E. E. Eichler, J. Flint, G. Gibson, *et al.*, "Missing heritability and strategies for finding the underlying causes of complex disease," en, *Nat. Rev. Genet.*, vol. 11, no. 6, pp. 446–450, Jun. 2010.

[17] M. R. Robinson, G. English, G. Moser, *et al.*, "Genotype-covariate interaction effects and the heritability of adult body mass index," en, *Nat. Genet.*, vol. 49, no. 8, pp. 1174–1181, Aug. 2017.

[18] C. Roth, D. Murray, A. Scott, *et al.*, "Pleiotropy and epistasis within and between signaling pathways defines the genetic architecture of fungal virulence," en, *PLoS Genet.*, vol. 17, no. 1, e1009313, Jan. 2021.

[19] T. F. C. Mackay, "Epistasis and quantitative traits: Using model organisms to study gene-gene interactions," en, *Nat. Rev. Genet.*, vol. 15, no. 1, pp. 22–33, Jan. 2014.

[20] C. Bartoli and F. Roux, "Genome-Wide association studies in plant pathosystems: Toward an ecological genomics approach," en, *Front. Plant Sci.*, vol. 8, p. 763, May 2017.

[21] D.-D. Zhang, Y. Shi, J.-B. Liu, *et al.*, "Construction of a myc-associated ceRNA network reveals a prognostic signature in hepatocellular carcinoma," en, *Mol. Ther. Nucleic Acids*, vol. 24, pp. 1033–1050, Jun. 2021.

[22] G. Vogel, G. Giles, K. R. Robbins, M. A. Gore, and C. D. Smart, "Quantitative genetic analysis of interactions in the pepper-phytophthora capsici pathosystem," en, *Mol. Plant. Microbe. Interact.*, vol. 35, no. 11, pp. 1018–1033, Nov. 2022.

[23] P. Krishnan, C. Caseys, N. Soltis, W. Zhang, M. Burow, and D. J. Kliebenstein, "Polygenic pathogen networks influence transcriptional plasticity in the Arabidopsis-Botrytis pathosystem," en, *Genetics*, vol. 224, no. 3, Jul. 2023.

[24] J. D. G. Jones and J. L. Dangl, "The plant immune system," en, *Nature*, vol. 444, no. 7117, pp. 323–329, Nov. 2006.

[25] F. Roux and J. Bergelson, "The genetics underlying natural variation in the biotic interactions of arabidopsis thaliana," in *Genes and Evolution*, ser. Current topics in developmental biology, Elsevier, 2016, pp. 111–156.

[26] N. Aoun, H. Desaint, L. Boyrie, *et al.*, "A complex network of additive and epistatic quantitative trait loci underlies natural variation of arabidopsis thaliana quantitative disease resistance to ralstonia solanacearum under heat stress," en, *Mol. Plant Pathol.*, vol. 21, no. 11, pp. 1405–1420, Nov. 2020.

[27] F. Roux and L. Frachon, "A Genome-Wide association study in arabidopsis thaliana to decipher the adaptive genetics of quantitative disease resistance in a native heterogeneous environment," en, *PLoS One*, vol. 17, no. 10, e0274561, Oct. 2022.

[28] H. M. Kang, J. H. Sul, S. K. Service, *et al.*, "Variance component model to account for sample structure in genome-wide association studies," en, *Nat. Genet.*, vol. 42, no. 4, pp. 348–354, Apr. 2010.

[29] J. Jakobsdottir and M. S. McPeek, "MASTOR: Mixed-model association mapping of quantitative traits in samples with related individuals," en, *Am. J. Hum. Genet.*, vol. 92, no. 5, pp. 652–666, May 2013.

[30] F. Tian, P. J. Bradbury, P. J. Brown, *et al.*, "Genome-wide association study of leaf architecture in the maize nested association mapping population," en, *Nat. Genet.*, vol. 43, no. 2, pp. 159–162, Feb. 2011.

[31] J. A. Corwin, D. Copeland, J. Feusier, *et al.*, "The quantitative basis of the arabidopsis innate immune system to endemic pathogens depends on pathogen genetics," en, *PLoS Genet.*, vol. 12, no. 2, e1005789, Feb. 2016.

[32] R. A. Power, J. Parkhill, and T. de Oliveira, "Microbial genome-wide association studies: Lessons from human GWAS," en, *Nat. Rev. Genet.*, vol. 18, no. 1, pp. 41–50, Jan. 2017.

[33] I. Bartha, J. M. Carlson, C. J. Brumme, *et al.*, "A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control," en, *Elife*, vol. 2, e01123, Oct. 2013.

[34] M. A. Ansari, E. Aranday-Cortes, C. L. C. Ip, *et al.*, "Interferon lambda 4 impacts the genetic diversity of hepatitis C virus," en, *Elife*, vol. 8, Sep. 2019.

[35] W. Correa-Macedo, G. Cambri, and E. Schurr, "The interplay of human and mycobacterium tuberculosis genomic variability," en, *Front. Genet.*, vol. 10, p. 865, Sep. 2019.

[36] Z. M. Xu, O. Naret, M. A. Oumelloul, and J. Fellay, "G2GSnake: A snakemake workflow for host-pathogen genomic association studies," en, *Bioinform. Adv.*, vol. 3, no. 1, vbad142, Oct. 2023.

[37] J. Phelan, P. J. Gomez-Gonzalez, N. Andreu, *et al.*, "Genome-wide host-pathogen analyses reveal genetic interaction points in tuberculosis disease," en, *Nat. Commun.*, vol. 14, no. 1, p. 549, Feb. 2023.

[38] X.-Z. Su, C. Zhang, and D. A. Joy, "Host-malaria parasite interactions and impacts on mutual evolution," en, *Front. Cell. Infect. Microbiol.*, vol. 10, p. 587 933, Oct. 2020.

[39] C. M. Smith, R. E. Baker, M. K. Proulx, *et al.*, "Host-pathogen genetic interactions underlie tuberculosis susceptibility in genetically diverse mice," en, *Elife*, vol. 11, Feb. 2022.

[40] N. Singh, S. Rai, R. Bhatnagar, and S. Bhatnagar, "Network analysis of host-pathogen protein interactions in microbe induced cardiovascular diseases," en, *In Silico Biol.*, vol. 14, no. 3-4, pp. 115–133, 2021.

[41] P. M. Jean Beltran, J. D. Federspiel, X. Sheng, and I. M. Cristea, "Proteomics and integrative omic approaches for understanding host–pathogen interactions and infectious diseases," en, *Mol. Syst. Biol.*, vol. 13, no. 3, p. 922, Mar. 2017.

[42] D. Tang, J. Freudenberg, and A. Dahl, "Factorizing polygenic epistasis improves prediction and uncovers biological pathways in complex traits," en, *Am. J. Hum. Genet.*, vol. 110, no. 11, pp. 1875–1887, Nov. 2023.

[43] B. Sheppard, N. Rappoport, P.-R. Loh, S. J. Sanders, N. Zaitlen, and A. Dahl, "A model and test for coordinated polygenic epistasis in complex traits," en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 15, e1922305118, Apr. 2021.

[44] R. Moore, F. P. Casale, M. Jan Bonder, *et al.*, "A linear mixed-model approach to study multivariate gene-environment interactions," en, *Nat. Genet.*, vol. 51, no. 1, pp. 180–186, Jan. 2019.

[45] C. Lippert, J. Listgarten, R. I. Davidson, *et al.*, "An exhaustive epistatic SNP association analysis on expanded wellcome trust data," en, *Sci. Rep.*, vol. 3, no. 1, p. 1099, Jan. 2013.

[46] C. S. Greene, N. M. Penrod, J. Kiralis, and J. H. Moore, "Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions," en, *BioData Min.*, vol. 2, no. 1, p. 5, Sep. 2009.

[47] M. Emily, T. Mailund, J. Hein, L. Schauser, and M. H. Schierup, "Using biological networks to search for interacting loci in genome-wide association studies," en, *Eur. J. Hum. Genet.*, vol. 17, no. 10, pp. 1231–1240, Oct. 2009.

[48] K. McAllister, L. E. Mechanic, C. Amos, *et al.*, "Current challenges and new opportunities for gene-environment interaction studies of complex diseases," en, *Am. J. Epidemiol.*, vol. 186, no. 7, pp. 753–761, Oct. 2017.

[49] W.-H. Wei, G. Hemani, and C. S. Haley, "Detecting epistasis in human complex traits," en, *Nat. Rev. Genet.*, vol. 15, no. 11, pp. 722–733, Nov. 2014.

[50] S. Ahmad, T. V. Varga, and P. W. Franks, "Gene × environment interactions in obesity: The state of the evidence," en, *Hum. Hered.*, vol. 75, no. 2-4, pp. 106–115, Sep. 2013.

[51] A. Voorman, T. Lumley, B. McKnight, and K. Rice, "Behavior of QQ-plots and genomic control in studies of gene-environment interaction," en, *PLoS One*, vol. 6, no. 5, e19416, May 2011.

[52] T. J. Rao and M. A. Province, "A framework for interpreting type I error rates from a product-term model of interaction applied to quantitative traits," en, *Genet. Epidemiol.*, vol. 40, no. 2, pp. 144–153, Feb. 2016.

[53] T. Zhang and L. Sun, "Beyond the traditional simulation design for evaluating type 1 error control: From the "theoretical" null to "empirical" null," en, *Genet. Epidemiol.*, vol. 43, no. 2, pp. 166–179, Mar. 2019.

[54] J. S. Long and L. H. Ervin, "Using heteroscedasticity consistent standard errors in the linear regression model," *Am. Stat.*, vol. 54, no. 3, p. 217, Aug. 2000.

[55] G. Hemani, J. E. Powell, H. Wang, *et al.*, "Phantom epistasis between unlinked loci," en, *Nature*, vol. 596, no. 7871, E1–E3, Aug. 2021.

[56] T. Akinbiyi, M. S. McPeek, and M. Abney, "ADELLE: A global testing method for Trans-eQTL mapping," en, *bioRxivorg*, Feb. 2024.

[57] R. Sun and X. Lin, "Set-based tests for genetic association using the generalized Berk-Jones statistic," 2017.

[58] I. Barnett, R. Mukherjee, and X. Lin, "The generalized higher criticism for testing SNP-set effects in genetic association studies," en, *J. Am. Stat. Assoc.*, vol. 112, no. 517, pp. 64–76, Jan. 2017.

[59] E. Weine, M. S. McPeek, and M. Abney, "Application of equal local levels to improve Q-Q plot testing bands with R package qqconf," en, *J. Stat. Softw.*, vol. 106, no. 10, Apr. 2023.

[60] D. Donoho and J. Jin, "Higher criticism for large-scale inference, especially for rare and weak effects," *Stat. Sci.*, vol. 30, no. 1, pp. 1–25, Feb. 2015.

[61] V. Gontscharuk, S. Landwehr, and H. Finner, "The intermediates take it all: Asymptotics of higher criticism statistics and a powerful alternative based on equal local levels," en, *Biom. J.*, vol. 57, no. 1, pp. 159–180, Jan. 2015.

[62] V. Gontscharuk, S. Landwehr, and H. Finner, "Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests," *Bernoulli (Andover.)*, vol. 22, no. 3, pp. 1331–1363, Aug. 2016.

[63] V. Gontscharuk and H. Finner, "Asymptotics of goodness-of-fit tests based on minimum p-value statistics," en, *Commun. Stat. Theory Methods*, vol. 46, no. 5, pp. 2332–2342, Mar. 2017.

[64] R. H. Berk and D. H. Jones, "Goodness-of-fit test statistics that dominate the kolmogorov statistics," en, *Z. Wahrscheinlichkeitstheorie verw Gebiete*, vol. 47, no. 1, pp. 47–59, 1979.

[65] P. J. Castaldi, M. H. Cho, L. Liang, *et al.*, "Screening for interaction effects in gene expression data," en, *PLoS One*, vol. 12, no. 3, e0173847, Mar. 2017.

[66] T. S. Breusch and A. R. Pagan, "A simple test for heteroscedasticity and random coefficient variation," *Econometrica*, vol. 47, no. 5, p. 1287, Sep. 1979.

[67] R. Koenker, "A note on studentizing a test for heteroscedasticity," en, *J. Econom.*, vol. 17, no. 1, pp. 107–112, Sep. 1981.

[68] S. M. Goldfeld and R. E. Quandt, "Some tests for homoscedasticity," *J. Am. Stat. Assoc.*, vol. 60, no. 310, p. 539, Jun. 1965.

[69] B. Efron, "Large-scale simultaneous hypothesis testing," *J. Am. Stat. Assoc.*, vol. 99, no. 465, pp. 96–104, Mar. 2004.

[70] L. Sun and M. Stephens, "Solving the empirical bayes normal means problem with correlated noise," 2018.

[71] B. Devlin and K. Roeder, "Genomic control for association studies," en, *Biometrics*, vol. 55, no. 4, pp. 997–1004, Dec. 1999.

[72] G. R. Foster, S. Pianko, A. Brown, *et al.*, "Efficacy of sofosbuvir plus ribavirin with or without peginterferon-alfa in patients with hepatitis C virus genotype 3 infection and treatment-experienced patients with cirrhosis and hepatitis C virus genotype 2 infection," en, *Gastroenterology*, vol. 149, no. 6, pp. 1462–1470, Nov. 2015.

[73] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability," en, *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, Apr. 2013.

[74] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation PLINK: Rising to the challenge of larger and richer datasets," en, *Gigascience*, vol. 4, no. 1, p. 7, Feb. 2015.

[75] B. L. Browning, Y. Zhou, and S. R. Browning, "A one-penny imputed genome from next-generation reference panels," en, *Am. J. Hum. Genet.*, vol. 103, no. 3, pp. 338–348, Sep. 2018.

[76] The 1000 Genomes Project Consortium, A. Auton, G. R. Abecasis, *et al.*, "A global reference for human genetic variation," en, *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015.

[77] V. A. Schneider, T. Graves-Lindsay, K. Howe, *et al.*, "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly," en, *Genome Res.*, vol. 27, no. 5, pp. 849–864, May 2017.

[78] S. Das, L. Forer, S. Schönherr, *et al.*, "Next-generation genotype imputation service and methods," en, *Nat. Genet.*, vol. 48, no. 10, pp. 1284–1287, Oct. 2016.

[79] P.-R. Loh, P. Danecek, P. F. Palamara, *et al.*, "Reference-based phasing using the haplotype reference consortium panel," en, *Nat. Genet.*, vol. 48, no. 11, pp. 1443–1448, Nov. 2016.

[80] P. Danecek, J. K. Bonfield, J. Liddle, *et al.*, "Twelve years of SAMtools and BCFtools," en, *Gigascience*, vol. 10, no. 2, Feb. 2021.

[81] A. Motyer, D. Vukcevic, A. Dilthey, P. Donnelly, G. McVean, and S. Leslie, "Practical use of methods for imputation of HLA alleles from SNP genotype data," Dec. 2016.

[82] Y. Luo, M. Kanai, W. Choi, *et al.*, "A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response," Jul. 2020.

[83] A. Stamatakis, "RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies," en, *Bioinformatics*, vol. 30, no. 9, pp. 1312–1313, May 2014.

[84] D. Ge, J. Fellay, A. J. Thompson, *et al.*, "Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance," en, *Nature*, vol. 461, no. 7262, pp. 399–401, Sep. 2009.

[85] J. G. McHutchison, E. J. Lawitz, M. L. Shiffman, *et al.*, "Peginterferon alfa-2b or alfa-2a with ribavirin for treatment of hepatitis C infection," en, *N. Engl. J. Med.*, vol. 361, no. 6, pp. 580–593, Aug. 2009.

[86] V. A. Morozov and S. Lagaye, "Hepatitis C virus: Morphogenesis, infection and therapy," en, *World J. Hepatol.*, vol. 10, no. 2, pp. 186–212, Feb. 2018.

[87] M. Jinushi, T. Takehara, T. Tatsumi, *et al.*, "Expression and role of MICA and MICB in human hepatocellular carcinomas and their regulation by retinoic acid," en, *Int. J. Cancer*, vol. 104, no. 3, pp. 354–361, Apr. 2003.

[88] D. Shichi, E. F. Kikkawa, M. Ota, *et al.*, "The haplotype block, NFKBIL1-ATP6V1G2-BAT1-MICB-MICA, within the class III-class I boundary region of the human major histocompatibility complex may control susceptibility to hepatitis C virus-associated dilated cardiomyopathy," en, *Tissue Antigens*, vol. 66, no. 3, pp. 200–208, Sep. 2005.

[89] A. Asada, M. Shioya, R. Osaki, *et al.*, "MHC class i-related chain B gene polymorphism is associated with virological response to pegylated interferon plus ribavirin therapy in patients with chronic hepatitis C infection," en, *Biomed. Rep.*, vol. 3, no. 2, pp. 247–253, Mar. 2015.

[90] N. B. Crux and S. Elahi, "Human leukocyte antigen (HLA) and immune regulation: How do classical and non-classical HLA alleles modulate immune response to human immunod-

eficiency virus and hepatitis C virus infections?" en, *Front. Immunol.*, vol. 8, p. 832, Jul. 2017.

[91] F. Qi, X. Du, Z. Zhao, *et al.*, "Tumor mutation burden-associated linc00638/mir-4732-3p/ulbp1 axis promotes immune escape via pd-l1 in hepatocellular carcinoma," en, *Front. Oncol.*, vol. 11, p. 729 340, Sep. 2021.

[92] W. Dai, J. Liu, B. Liu, Q. Li, Q. Sang, and Y.-Y. Li, "Systematical analysis of the cancer genome atlas database reveals EMCN/MUC15 combination as a prognostic signature for gastric cancer," en, *Front. Mol. Biosci.*, vol. 7, p. 19, Feb. 2020.

[93] D.-D. Xiong, Z.-B. Feng, Z.-F. Lai, *et al.*, "High throughput circRNA sequencing analysis reveals novel insights into the mechanism of nitidine chloride against hepatocellular carcinoma," en, *Cell Death Dis.*, vol. 10, no. 9, p. 658, Sep. 2019.

[94] X. Liu, Y. Zhang, Z. Wang, *et al.*, "PRRC2A promotes hepatocellular carcinoma progression and associates with immune infiltration," en, *J. Hepatocell. Carcinoma*, vol. 8, pp. 1495–1511, Dec. 2021.

[95] H.-Y. Zhang, J.-J. Zhu, Z.-M. Liu, Y.-X. Zhang, J.-J. Chen, and K.-D. Chen, "A prognostic four-gene signature and a therapeutic strategy for hepatocellular carcinoma: Construction and analysis of a circRNA-mediated competing endogenous RNA network," en, *Hepatobiliary Pancreat. Dis. Int*, Jun. 2023.

[96] N. Van Renne, A. A. Roca Suarez, F. H. T. Duong, *et al.*, "Mir-135a-5p-mediated downregulation of protein tyrosine phosphatase receptor delta is a candidate driver of HCV-associated hepatocarcinogenesis," en, *Gut*, vol. 67, no. 5, pp. 953–962, May 2018.

[97] W. Chen, Q. Fu, F. Fang, J. Fang, Q. Zhang, and Y. Hong, "Overexpression of leucine-rich repeat-containing g protein-coupled receptor 5 predicts poor prognosis in hepatocellular carcinoma," en, *Saudi J. Biol. Sci.*, vol. 25, no. 5, pp. 904–908, Jul. 2018.

[98] C.-J. Ko, C.-J. Li, M.-Y. Wu, and P.-Y. Chu, "Overexpression of LGR-5 as a predictor of poor outcome in patients with hepatocellular carcinoma," en, *Int. J. Environ. Res. Public Health*, vol. 16, no. 10, p. 1836, May 2019.

[99] J. Tang, C. Liu, B. Xu, D. Wang, Z. Ma, and X. Chang, "ARHGEF10L contributes to liver tumorigenesis through RhoA-ROCK1 signaling and the epithelial-mesenchymal transition," en, *Exp. Cell Res.*, vol. 374, no. 1, pp. 46–68, Jan. 2019.

[100] L. Wang, L. Wang, B. Xiao, M. Cui, and B. Zhang, "Differences between sorafenib and lenvatinib treatment from genetic and clinical perspectives for patients with hepatocellular carcinoma," en, *Med. Sci. Monit.*, vol. 28, e934936, Apr. 2022.