5-2012

Robust Computational Tools for Multiple Testing with Genetic Association Studies

William L. Welbourn Jr.
Utah State University

ROBUST COMPUTATIONAL TOOLS FOR MULTIPLE TESTING WITH GENETIC ASSOCIATION STUDIES

by

William L. Welbourn, Jr.

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

of

DOCTOR OF PHILOSOPHY

in

Mathematical Sciences

Approved:

Dr. Christopher Corcoran Major Professor	Dr. Adele Cutler Committee Member
Dr. Kady Schneiter Committee Member	Dr. John Stevens Committee Member
Dr. Ronald Munger Committee Member	Dr. Mark R. McLellen Vice President for Research and Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2012

ABSTRACT

Robust Computational Tools for Multiple Testing with Genetic Association Studies

by

William L. Welbourn, Jr., Doctor of Philosophy

Utah State University, 2012

Major Professor: Dr. Christopher Corcoran
Department: Mathematics and Statistics

Resolving the interplay of the genetic components of a complex disease is a challenging endeavor. Over the past several years, genome-wide association studies (GWAS) have emerged as a popular approach at locating common genetic variation within the human genome associated with disease risk. Assessing genetic-phenotype associations upon hundreds of thousands of genetic markers using the GWAS approach, introduces the potentially high number of false positive signals and requires statistical correction for multiple hypothesis testing. Permutation tests are considered the gold standard for multiple testing correction in GWAS, because they simultaneously provide unbiased Type I error control and high power. However, they demand heavy computational effort, especially with large-scale data sets of modern GWAS. In recent years, the computational problem has been circumvented by using approximations to permutation tests, but several studies have posed sampling conditions in which these approximations are suggestive to be biased.

We have developed an optimized parallel algorithm for the permutation testing approach to multiple testing correction in GWAS, whose implementation essentially abates the computational problem. When introduced to GWAS data, our algorithm yields rapid, precise, and powerful multiplicity adjustment, many orders of magnitude faster than existing employed GWAS statistical software.

Although GWAS have identified many potentially important genetic associations which will advance our understanding of human disease, the common variants with modest effects on disease risk discovered through this approach likely account for a small proportion of the heritability in complex disease. On the other hand, interactions between genetic and environmental factors could

account for a substantial proportion of the heritability in a complex disease and are overlooked within the GWAS approach.

We have developed an efficient and easily implemented tool for genetic association studies, whose aim is identifying genes involved in a gene-environment interaction. Our approach is amenable to a wide range of association studies and assorted densities in sampled genetic marker panels, and incorporates resampling for multiple testing correction. Within the context of a case-control study design we demonstrate by way of simulation that our proposed method offers greater statistical power to detect gene-environment interaction, when compared to several competing approaches to assess this type of interaction.

(326 pages)

PUBLIC ABSTRACT

Robust Computational Tools for Multiple Testing with Genetic Association Studies

The mapping of the human genome and the completion of the Human HapMap project over the past decade have significantly altered how research is conducted with respect to the genetic epidemiology of human disease. Study designs and analytic approaches have evolved rapidly from investigations involving relatively few targeted candidate genes to hypothesis-free genome-wide association studies, where thousands – and now even millions – of single molecular mutations are simultaneously analyzed to identify regions of the genome that may influence disease. As laboratory techniques continue to improve and costs decrease, the volume of genetic data will inexorably rise, and robust tools for data management, statistical analysis, and computation will likewise need to keep pace.

Multiple hypothesis testing is the core problem in analyzing data from a genome-wide association study (GWAS). A conventional GWAS, focused on genetic risk factors leading to disease incidence, samples some number of disease and non-diseased subjects, genotypes these subjects for a common set of genetic mutations, and then carries out an individual hypothesis test of the association between each marker and disease status. Correction for multiple testing in GWAS typically relies upon the Bonferroni multiple testing procedure. With ever-growing panels of markers (the standard panel currently employs one million markers), this approach engenders numerous problems. First, it is overly conservative, both because of the sheer number of tests as well as the Bonferroni ideal that all tests are mutually independent. The growing density of marker panels results in marker loci that are more physically proximate, yielding hypothesis tests that have some dependence structure. Second, the commonly used corrected significance level on the order of 10^{-8} provides an extreme critical region for which the relative error of asymptotic approximations is large. Third, while approximations can be avoided by using a permutation distribution, such an approach is computationally challenging and has not been widely implemented or used. This is particularly critical in the context of alternative multiple correction procedures that solve the dependence problem, for which permutation distributions are hypothetically available but in practice are seldom used, if ever. Fourth, the distribution of test statistics across the various multiple testing approaches depends on additional features of the data, most prominently on what is referred to as the minor allele frequency (MAF), or the proportion of genetic loci for a given marker within the sampling population that carry the least frequent marker variant.

This research project has led to the development and implementation of a parallel processing algorithm which allows exceptionally rapid computation of the permutation distribution for multiple testing procedures that correct for dependence between tests. This eliminates the need for large-sample approximations, which have been found in prior studies to have poor operating characteristics under some common circumstances. This parallel processing approach relies upon existing hardware and software commonly available in desktop personal computers, allowing for efficient and cost effective computational tools to the research community. In addition, we have leveraged these efficient permutation tools in order to implement MAF-corrected exact tests, to eliminate bias for multiple testing procedures that arise in particular when the MAF is small. We have further extended these tools to other analytic problems in large-scaled genetic association settings, such as tests for gene-environment interactions.

William L. Welbourn, Jr.

This work is dedicated to my parents, William and Linda.

ACKNOWLEDGMENTS

I would first like to thank my parents. Your unconditional love and support have kept my spirits high over the years. I could not have completed this degree without the two of you by my side. I also thank my family and friends for their encouragement and support.

I express sincere gratitude to my advisor, Dr. Christopher Corcoran, for introducing the general problem encompassing each of the respective Chapters 2 and 3 of this manuscript, and for providing encouragement and advice during my course of study at Utah State University.

I thank the members of my dissertation committee, Dr. Adele Cutler, Dr. Kady Schneiter, Dr. John Stevens, and Dr. Ronald Munger, for their commitment and interest in this research project. I appreciate the time you all have set aside on my behalf, and for the encouragement and positive feedback toward my research endeavors.

I am grateful to Dr. Corcoran and Dr. Munger for providing the opportunity to present my research within the USU Center for Epidemiologic Studies Seminars. These invaluable experiences have helped strengthen my interest and knowledge within the biostatistics discipline.

I thank Dr. Peter Zandi of Johns Hopkins University for providing access to the European ancestry portion of the GAIN Bipolar disorder GWAS data set. Your kind gesture helped fuel my interest in acquiring knowledge for the underlying workings of GWAS in general, and provided working data by which to investigate the theses of Chapters 2 and 3 of this manuscript.

To colleagues at the University of Utah: I thank Dr. Martha Slattery for providing access to the colorectal cancer data sets, as these data aided with the development of the methodology governing Chapter 4 of this manuscript. Thanks to Abbie Lundgreen for her help in answering questions I had pertaining to these data sets and help with testing the programming code for GEM. Dr. Slattery, Abbie, and Jennifer Herrick, thank you for the many hours of interesting conversations revolving around these data, which aided with the research of Chapter 4.

I thank the faculty, staff, and students of the USU Department of Mathematics and Statistics for making the past five years a rewarding experience. I thank Cindy Moulton for helping me navigate through the degree requirement process.

To the Department of Biostatistics at the University of Southern California: I thank the faculty for the solid foundation you have provided me within the biostatistics discipline. Dr. Kimberly Siegmund and Dr. W. James Gauderman, thank you both for introducing me to the concept of gene-environment interaction and for working closely with me during the early stages in the development of my research skills.

To the Department of Mathematics at California State University, Fullerton: I thank the faculty for providing me with invaluable knowledge and a rich skill set within the mathematics and statistics disciplines. Dr. James Friel, thank you for taking me under your wing and for introducing me to mathematical probability. Dr. Paul DeLand, thank you for introducing me to mathematical statistics.

William L. Welbourn, Jr.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

CHAPTER 1

INTRODUCTION

Epidemiology is the study of the distribution and determinants of health-related states or events (i.e., health, disease, and health behavior) within human populations. Its aim is discovering disease etiology for prevention in populations. This is accomplished by studying populations which are comprised of healthy and diseased individuals, and identifying environmental and genetic risk factors which serve as intermediates to disease onset.

Genetic epidemiology, a specific focus within the discipline of epidemiology, is the evaluation of the role of inherited genetic causes toward the incidence of disease within families and populations. Its aims lie with the detection of the genetic inheritance pattern for a particular disease, restricting ones' attention to the gene(s) encompassing the disease etiology, and locating positions within the DNA sequence associated with disease risk. This is accomplished by: (1) demonstrating the existence of a genetic association with the disease; (2) reporting the size of the genetic effect, relative to non-genetic contributable factors within the disease etiology; and (3) identification of the gene(s) involved in the disease etiology. For more than 20 years, family-based linkage studies have been a common analytical tool in carrying out the aforementioned three step procedure [1]. However, over the course of the past few years, a new approach has entered the picture to tackle this procedure, called the genome-wide association study (GWAS).

1.1 The Genome-wide Association Study

Linkage studies investigate the genetic inheritance within families (pedigrees), applying the principles of recombination, with the goal of identifying the approximate chromosomal location of a major gene – a major gene is any gene individually associated with pronounced phenotypic effects. More specifically, to identify regions within chromosomes which are shared by family members with the same phenotypic trait, and thus are likely to contain the disease susceptibility genes. This approach has led to the discovery of mutations (mostly rare dominant or recessive) for more than 1,600 diseases, and has been particularly successful in identifying the genetic basis of many human diseases in which the disease penetrance resembles a simple Mendelian model. Examples include Huntington's disease and Cystic Fibrosis [2].

However, many diseases such as cancer and cardiovascular disease are complex, and characterized by a multifactorial etiology. Genes, along with environmental factors, likely interact within a complex causative pathway, ultimately leading to the incidence of disease [3]. The presence of multiple independent and/or interacting disease genes and environmental factors leads to significant problems for genetic linkage analysis. Specifically, linkage studies suffer a loss of statistical power in the presence of such genetic heterogeneity [4]. While the loss of statistical power in the presence of genetic heterogeneity is a considerable limitation of linkage studies, a more substantial limitation in these studies lies with the large resolution of chromosomal regions (often comprising hundreds of genes) shared among family members, in which it can be difficult to narrow the linkage signal sufficiently to identify a disease susceptible gene [5].

Association studies are routinely used by epidemiologists to investigate the relationship between an exposure and a disease. With the completion of the Human Genome Project in 2003 [6] and the development of gene sequencing techniques such as the Polymerase Chain Reaction (PCR), it is now possible to amplify (i.e., clone) specific regions of the human genome, for which these exposures may now include genotypes at one or more susceptibility, candidate, or marker loci. The goals of genetic association studies will differ, depending on ones' knowledge about the given disease. For example, once a susceptibility locus (e.g., *BRCA1* for breast cancer) has been determined and amplified, the goals include estimating the relative risk (RR) and penetrance associated with specific mutations and testing for interaction with environmental exposures or other genes [7]. On the other hand, if a candidate locus has been identified (e.g., the androgen receptor for prostate cancer), the primary goal is testing the null hypothesis of no association between the locus and the disease [8]. Finally, if little is known about specific loci for the disease (e.g., multiple sclerosis), multiple tests of association with finely spaced markers (e.g., the GWAS approach) may be used to screen the genome for candidate regions with the anticipation of detecting linkage disequilibrium (LD)[1] with markers close to one or more disease susceptible loci (DSL) [8].

The [population-based] GWAS approach is revolutionary, insofar as it permits investigation of the entire genome at levels in genetic resolution previously unattainable, among thousands of unrelated individuals, and does not require apriori specification of hypotheses regarding genetic-phenotypic associations (hereinafter, the term genome is assumed synonymous with that of the human nuclear genome). These studies utilize high-throughput genotyping technologies to assay the

[1] LD is the association between the alleles of two SNP loci located near each other on a chromosome, such that they are inherited together more frequently than expected by chance [5].

most common genetic variant, the single nucleotide polymorphism (SNP), and relate these variants to diseases or health-related traits [5]. A SNP is the most common form of genetic variation in the genome, in which a single nucleotide base substitution (mutation) has led to two forms (alleles) of a DNA sequence which differ by a single nucleotide (e.g., the nucleic acid adenine (A) is substituted in lieu of cytosine (C) at the locus). Statistically speaking, a SNP – whose respective alleles are defined from the base pairs adenine and cytosine, say – can be seen as a three level ordinal categorical variable comprised of the genotypes AA, AC, and CC. While significant advances in genotyping technology within recent years has been a key ingredient to providing data for conducting the analysis within a GWAS, the motivation for these studies can be traced back to two papers from 1996.

The articles of [2] and [9] argued that common variants may underlie many common diseases, the variants in which would be more easily determined using population-based association studies rather than family-based linkage analysis, even in the most extreme case of interrogation of every gene within the human genome [2], and that all common variants within the human genome should be identified [9]. These proposals led to the International HapMap Project (IHP), whose aim was indexing common genetic variants within the human genome [10]. The IHP – in conglomeration with advances in genotyping technology – has enabled the GWAS approach to be feasible, leading to discovery of common genetic variants associated with diseases such as coronary heart disease [11, 12, 13] and type II diabetes [14, 15].

The primary aim of the GWAS approach is to locate positions within the genome which pertain to common variants[2] associated with risk to a disease trait [16]. To carry out this task, hundreds of thousands of SNP markers (called tagging SNPs, or tSNPs) are selectively sampled[3] from across the human genome and typed amongst a large random sample of diseased (cases) and healthy (controls) individuals.[4] Allele and genotype frequencies for each of these SNP markers are compared – commonly, by way of the Cochran-Armitage Trend test [17, 18] (see e.g., [19, 20, 21, 22]), assuming the multiplicative risk model [23] (equivalent to the additive – on the log-odds scale – genetic model of inheritance (GMI)) – between cases and controls; an over-representation of alleles or genotypes within one of these groups at a locus is suggestive of a genotype-phenotype association. Aside from

[2]Pertaining to a SNP, a variant (allele) is common if its frequency upon the chromosomes within the study population is at least 5%.

[3]As a result of LD between alleles in close proximity within the genome, not all SNP loci need be typed to capture the majority of common variation within the genome; tSNPs act as proxies to cover the common variation of all SNPs within the genome.

[4]We have described the retrospective case-control GWAS sampling design, by far the most commonly employed GWAS design. As such, henceforth unless otherwise specified, GWAS is assumed to adhere to the case-control study design.

chance, confounding, and data anomalies, a statistically significant genotype-phenotype association at a SNP locus indicates: that the SNP itself carries the risk variant allele (direct association); or the SNP is in LD (i.e., close proximity within the genome) with the SNP carrying the risk variant allele (indirect association). In other words, a direct/indirect association in a GWAS implies the precise/approximate position within the genome pertaining to a DSL.

Due to the fine genetic resolution of the SNP, the GWAS approach is applicable to virtually any complex disease, irrespective of one's knowledge of the underlying genetic components associated with disease risk. For example, as mentioned above *BRCA1* is a disease susceptible locus for breast cancer, but recent GWAS investigations have determined additional novel DSLs for this disease (see e.g., [24, 25]). This makes the GWAS approach a very powerful tool for deciphering the underlying genetic component of complex disease etiology.

1.2 Gene-Environment Interaction

Resolving the interplay of the genetic components of complex diseases is a challenging endeavor, and the architecture of the genetic etiology for these diseases essentially remains a mystery [26]. Although current GWAS have identified many potentially important genetic associations which will advance our understanding of human disease [27], the common variants with modest effects on disease risk discovered through GWAS apparently do not account for all of the heritability in these diseases. In fact, there is a growing consensus that a majority of the heritability in these diseases cannot be assessed by GWAS [26, 28, 29, 30]. Interactions, such as gene-gene (epistasis) or gene-environment (GxE), could account for a significant proportion of the heritability in complex diseases and cannot be detected by the GWAS design of testing for solely genetic main effects.

Many common, complex diseases are believed to be a result of the collective effect of genetic factors, environmental factors, and their interactions [3, 31, 32, 33, 34]. To be clear in discussion, here a genetic factor is broadly defined as any metric which can be used to model genetic variation within the human genome, with the specific aim of associating the variation with risk of disease. For example, the genotypes at a SNP locus can be used to test for a genotype-phenotype association, with the goal of locating specific genotype groupings which are more/less predisposed to disease risk. An environmental factor, on the other hand, is broadly defined as: an exposure, either physical (e.g., temperature, UV radiation from the sun), chemical (e.g., airborne pollutants, such as particulate matter), or biological (e.g., a bacterial infection); a behavioral pattern (e.g., diet, smoking); or, a life event (e.g., injury). Evidence supports the existence of GxE interactions for many complex diseases,

including mental health disorders [35,36], cardiovascular and metabolic disease [37,38,39,40,41,42], infectious disease [43, 44], and trauma and injury [45]. We study GxE interactions for several reasons [46]: they can illuminate fundamental biological mechanisms involved within disease etiology; they can be important for risk prediction and for evaluating the benefit of changes in modifiable environmental exposures; and, failure to adequately account for GxE interaction in a genetic analysis can mask the effects of both genetic and environmental factors [47,48,49,50,51], thereby making it difficult to detect associations using standard genetic or epidemiologic approaches. Studying GxE interactions can lead to a better understanding of the complete etiology of disease, inclusive of both distinct and interacting pathways comprised of genetic and environmental factors. Identifying GxE interactions enables one to target, develop, and prescribe preventative measures to individuals at particularly high risk of disease; these interventions are designed to maximize health and minimize disease.

1.2.1 Types of Interaction

Loosely stated, here the term interaction can imply one of two things: *statistical interaction*, or *biologic interaction* [52, 53]. We have statistical interaction when the relationship between two variables is dependent upon the levels of a third variable. We refer to the third variable as an effect modifier for the relationship of the two variables of interest. Biological interaction, on the other hand, refers to the synergism between/among discrete pathways relating to the maintenance of homeostasis or the expression and progression of a physiological condition [53]. Each of these concepts is central to studying GxE interaction. Gene-environment interaction is essential to biological interaction, insofar as the goal of studying GxE interaction lies with the discovery of novel biological mechanisms – involving both genetic and environmental factors – which are associated with risk of disease. On the other hand, statistical interaction is needed to quantify the presence of biological interaction and is essential to accurately model the true underlying joint effect of genetic and environmental factors in their risk toward disease. Unless otherwise specified, here the term GxE interaction is assumed within the context of statistical interaction.

Gene-environment interactions can exhibit several different patterns of association. To illustrate, we consider a binary disease trait (phenotype), a binary environmental factor, and a three-level genetic factor. The three-levels for the genetic factor correspond to the three genotypes at some SNP locus – measured in terms of the number of copies of the minor allele (the minor/major allele is the less/more frequently occurring allele at the locus within the population) an individual carries

at the locus – whose respective major and minor alleles are denoted by A and a. For clarity in presentation, we consider the additive GMI for the SNP marker. If the environmental factor is an effect modifier for the genotype-phenotype relationship, then [for our setup] essentially three patterns for GxE interaction are tangible: (a) the risk of disease is positively (or, negatively) associated with the number of minor alleles one carries at the locus, where the size of this effect is greater upon the population of exposed individuals. This pattern of GxE interaction is commonly referred to as *complementary* [53, 54]. For example, an allele upon one of the genes related to familial hypercholesterolemia (FH), say the LDL-receptor ($LDLR$) gene [55] (genetic factor), might increase susceptibility to atherosclerosis by increasing the production of LDL cholesterol. Furthermore, a diet enriched in high levels of saturated fat (environmental factor) could contribute to an increased risk of atherosclerosis by increasing blood serum levels of LDL. These genetic and environmental factors complement each other to increase risk for atherosclerosis; (b) the risk of disease is positively (or, negatively) associated with the number of minor alleles one carries at the locus, where the size of this effect is greater upon the population of unexposed individuals. This pattern of GxE interaction is commonly referred to as *antagonistic* [53,54]. For example, mutations upon the β and γ subunits of the epithelial sodium channel ($ENaC$) gene (genetic factor) have been associated with increased risk of resistant hypertension [56]. Whereas, engaging in moderate physical activity for the majority of the days during a given week (environmental factor) is associated with decreasing lifetime risk of hypertension [53]. These genetic and environmental factors antagonize each other in their risk toward hypertension; or, (c) the risk of disease is positively (or, negatively) associated with the number of minor alleles among the population of exposed individuals, and the risk of disease is negatively (or, positively) associated with the number of minor alleles among the population of unexposed individuals. Here, we refer to this pattern of GxE interaction as *cross-interaction* [3], although it has at least two alternative naming conventions in the literature (e.g., called 'flip-flop' interaction by [26], and called *crossover* interaction by [57, 58]). For example, a well-established and replicated GxE interaction for risk of asthma or allergic disease is that resulting from the genotype of a promoter polymorphism (SNP rs2569190; thymine and cytosine allele variants) upon the mononuclear cells (*CD14*) gene (genetic factor) and exposure to microbes (environmental factor) [26]. Several studies (see [26]) found the thymine allele variant to be associated with increased risk for asthma among exposed subjects, whereas the cytosine allele variant was found to be associated with increased risk for asthma among unexposed subjects. These genetic and environmental factors cross-interact in

their effect toward risk of disease. Figure 1.1 illustrates the three patterns of GxE interaction for our setup, where – in terms of the additive GMI – the vertical axis for the plot within each panel could represent, for example, the log odds of disease.

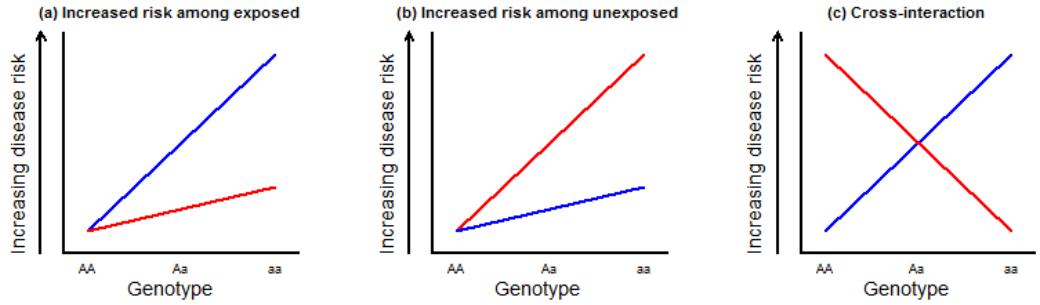

Fig. 1.1: Types of Gene-Environment Interaction upon a Genetic Factor (SNP) with Respective Major and Minor Alleles A and a, and a Binary Environmental Factor. The Blue Line in (a-c) Corresponds to Risk in Exposed Individuals; the Red Line Corresponds to Risk in Unexposed Individuals.

1.3 Approaches for Multiple Hypothesis Testing in Genetic Association Studies

Consider a genetic association study in which several genetic and environmental variables are collected upon a sample of study subjects, where we assume a binary response (phenotype) is recorded for each subject. Such data may arise, for example, from a [cohort] GWAS investigating genetic markers (genetic factors; SNPs) associated with incidence of AIDS among subjects diagnosed with HIV-1 disease [59]: in this case, the response is the indicator (yes or no) for development of AIDS, and the goal is to locate those genetic markers for which the proportion of subjects developing AIDS differs amongst the genotypes upon each of the markers. Restated as a problem in multiple hypothesis testing (MHT): the simultaneous test for each genetic marker of the null hypothesis of no association between the marker and disease (AIDS) incidence. In some circumstances, more specific null hypotheses may be of interest. For example, the null hypothesis of no GxE interaction across the sampled genetic markers and some common environmental factor. Nonetheless, an MHT problem is prevalent amidst such investigations, and unless appropriate measures are taken to account for the multiplicity problem, the chance of committing some Type I errors (i.e., rejecting a particular null hypothesis in favor of the false alternative hypothesis) increases.

To illustrate, consider testing m mutually independent sets of null and alternative hypotheses, each at the prescribed pointwise α_{p} level of significance, some $\alpha_{\mathrm{p}} \in (0, 1)$. Let V be the random variable corresponding to the number of Type I errors committed in testing these m sets of hypothe-

ses. When testing multiple null hypotheses, there are many definitions of the Type I error rate. In terms of the random variable V, [60] describe four most standard definitions for the Type I error rate: the per-comparison error rate (PCER), defined as the expected value of the number of Type I errors divided by the number of hypotheses (i.e., PCER = $E(V)/m$); the per-family error rate (PFER), defined as the expected number of Type I errors (i.e., PFER = $E(V)$); the false discovery rate (FDR) of [61], the expected number of Type I errors among the rejected null hypotheses (i.e., FDR = $E(Q)$, where: $Q = V/R$ if $R > 0$ and 0 if $R = 0$; and R is the random variable corresponding to the number of rejected null hypotheses); and the family-wise error rate (FWER) is defined as the probability of some Type I error, $\Pr(V \geq 1)$. A multiple testing procedure (MTP) is said to control a particular Type I error rate at level α, provided that the error rate is less than or equal to α when the given procedure is applied to produce a list of rejected null hypotheses [60]. For example, the FWER is controlled at level α, provided that the implemented MTP produces a FWER satisfying the inequality FWER $\leq \alpha$. Without loss of generality suppose all m null hypotheses are in fact true – in MHT terminology, we refer to this as the complete null hypothesis [60], the collection of hypotheses in which is denoted $\mathcal{H}_0$. When the tests are independent, it can be shown that

$$\text{FWER} = \Pr(V \geq 1 | \mathcal{H}_0) = 1 - (1 - \alpha_{\text{p}})^m. \tag{1.1}$$

As an example, we consider $m = 100$ and $\alpha_{\text{p}} = 0.05$. It is, FWER = 0.994, for which committing some Type I error is nearly certain. For any fixed α_{p}, we see that (1.1) is increasing in m. Figure 1.2 displays the assumed FWER across m for several choices of α_{p}, under the assumption that each of the m mutually independent null hypotheses are tested at level α_{p} under the complete null hypothesis, where $m = 1, \ldots, 100$.

It is important to note that the expectations and probabilities for the four Type I error rates defined above, are conditional on the true underlying data distribution for the explanatory and response variables involved [60]. In particular, these computations depend upon the specific hypotheses within the collection $\mathcal{H}_0$ which are actually true. Herein, a *partial null hypothesis*, denoted by $\mathcal{H}_0^{\text{p}}$, is defined to be any subset of $\mathcal{H}_0$. Let $\tilde{\mathcal{H}}_0^{\text{p}}$ denote the specific partial null hypothesis, whose elements warehouse the actual true null hypotheses over $\mathcal{H}_0$. Hence, the FWER, for example is given by

$$\text{FWER} = \Pr\left(V \geq 1 | \tilde{\mathcal{H}}_0^{\text{p}}\right). \tag{1.2}$$

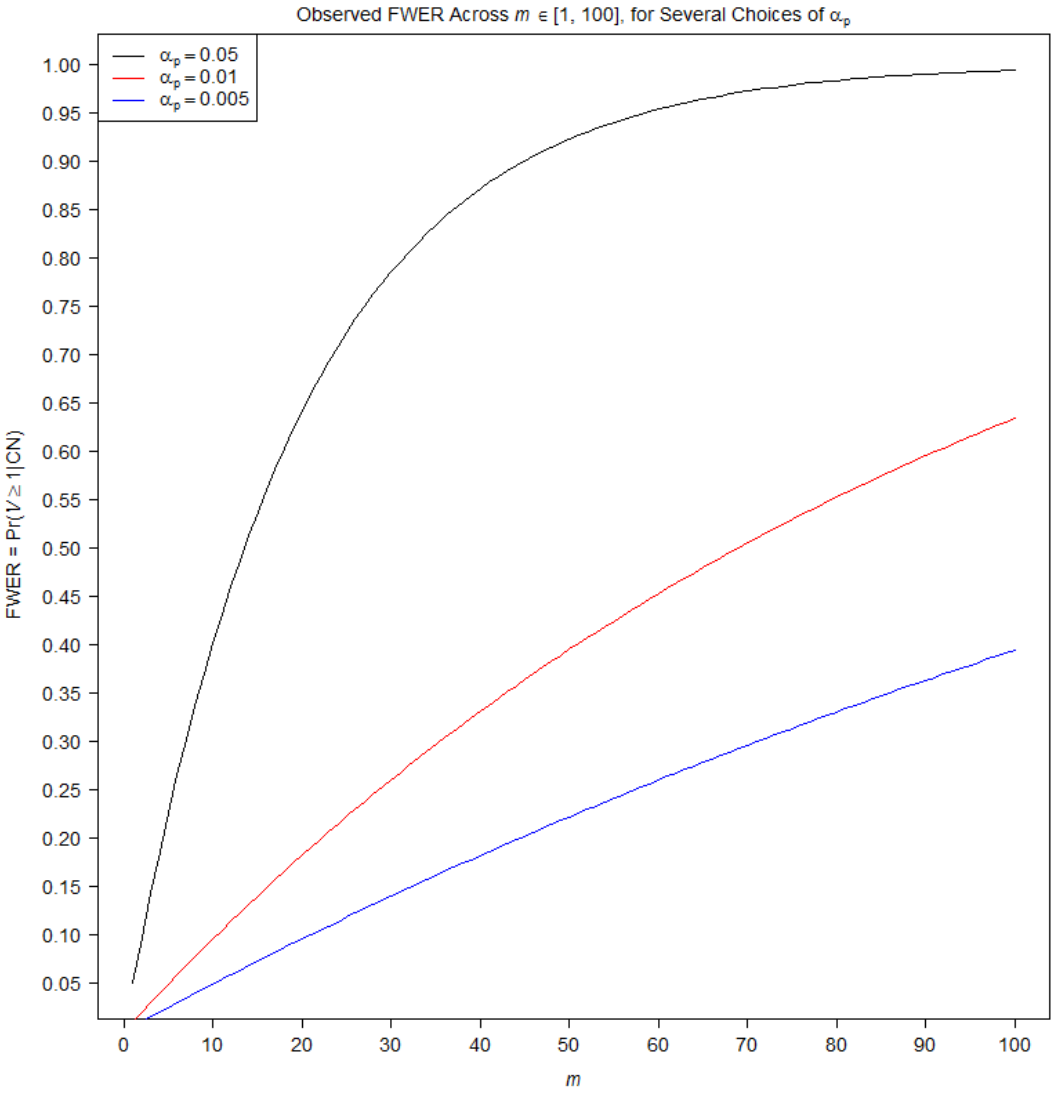

Fig. 1.2: Plot of the Family-wise Type I Error Rate (FWER) Versus the Number of Mutually Independent Tested Null Hypotheses (m), under the Complete Null (CN) Hypothesis, Where Each Null Hypothesis Is Tested at the α_{p} Pointwise Significance Level, $m = 1, \ldots, 100$.

A fundamental, often ignored distinction, is that between strong and weak control of a Type I error rate [60, 62]. An MTP is said to control the Type I error rate (for which it is intended to control) at level α in the *strong* sense (strongly), if the Type I error rate is controlled at said level for every possible partial null hypothesis [60, 62]. For example, the FWER is controlled strongly at level α if

$$\Pr\left(V \geq 1 | \mathcal{H}_0^{\mathrm{p}}\right) \leq \alpha \quad \forall \mathcal{H}_0^{\mathrm{p}} \in \mathcal{P}\left(\mathcal{H}_0\right), \tag{1.3}$$

where $\mathcal{P}(\cdot)$ denotes the power set of the collection $(\cdot)$. On the other hand, we say that an MTP controls the Type I error rate at level α in the *weak* sense (weakly), if it controls the Type I error rate at said level only under the complete null hypothesis [60, 62].

The complete null hypothesis, in general, is not realistic and weak control is unsatisfactory. For example, in a GWAS it is very unlikely that $\tilde{\mathcal{H}}_0^{\mathrm{p}} \equiv \mathcal{H}_0$ (i.e., all null hypotheses are in fact true). That is, among the thousands (or, millions) of tested null hypotheses of no association between genotype and phenotype, it is unreasonable to assume that the genotype for each-and-every SNP locus to not be associated with the disease trait of interest. Realistically, the genotypes for some of the loci will be associated with the disease trait, while genotypes for other loci will not be. However, we do not know which loci fit within $\tilde{\mathcal{H}}_0^{\mathrm{p}}$, nor within $\mathcal{H}_0 \backslash \tilde{\mathcal{H}}_0^{\mathrm{p}}$. This implies that $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ is an unknown proper subset of $\mathcal{H}_0$. Since $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ is unknown, weak control of the Type I error rate at level α provides no assurance that the *actual* Type I error rate is being controlled at level α and is therefore unsatisfactory.

Strong control of the Type I error rate at level α, on the other hand, is a very desirable property, insomuch as it allows one to conveniently compute the Type I error rate assuming the complete null hypothesis to be true, knowing forthright the actual Type I error rate cannot exceed the value α. That is, strong control ensures that the actual Type I error rate (e.g., the actual FWER (1.2)) is controlled at level α, even though the calculation of the Type I error rate is under the assumption that $\mathcal{H}_0$ is true.

Finally, we note that one typically decides among three types of multiple testing procedures: single-step, step-down, and step-up procedures [60]. In single-step procedures, equivalent multiplicity adjustment is carried out for all null hypotheses. That is, $H_0^{(j)}$ is evaluated using a critical value which is independent of the results among the tests of other null hypotheses. For the sake of discussion, we use $H_0^{(j)}$ to denote a test of the null hypothesis of no association between genotype and phenotype for the j^{th} sampled SNP locus within the context of GWAS, for $j = 1, \ldots, m$. Here, j indexes the m GWAS SNP loci. Examples of single-step multiple testing procedures include the Bonferroni, Šidák, and the permutation-based maxT approach implemented within popular GWAS software (e.g., PLINK [63]). Although all multiple testing procedures described henceforth are assumed single-step based, we do mention here that stepwise (i.e., step-down and step-up) procedures can improve statistical power while preserving the Type I error rate, insofar as rejection of $H_0^{(j)}$ is based upon the outcomes of the other tested null hypotheses [60].

1.4 Parallel Computing

To this end, we have introduced several approaches to assessing associations upon complex diseases, and introduced various approaches to tackling their induced MHT problem. The underlying studies of these approaches are typically comprised of exceptionally large samples of data. For example, within a GWAS it is not uncommon to be analyzing hundreds of thousands of SNP markers upon thousands of study subjects [21, 22]. The analysis of these immense data sets, demands both high computational power and appropriate tools for its implementation. Parallel computing is an approach well suited to deliver high computational power for the analysis of such data sets. Within this section, along with its corresponding subsections, we introduce the notion of parallel computing upon the graphical processing unit (GPU) of the personal computer, and outline the requirements (tools) for its implementation.

It is this author's opinion that the dynamic evolution of the personal computer is one of the most intriguing phenomenon occurring within today's research practices. In particular, the recent (mid-late 2006) birth of each of the multi-core central processing unit (CPU) and the programmable manycore GPU. Each of these advancements for the personal computer lends improvement in computational power and reshapes the way one is required to think about solving complex problems. It is through advancements in computing architecture such as these, which allows one to delve into the analysis of ever increasingly more complex subject matter. Whether it be analyzing the tertiary structure of a protein (proteomics) or one's attempt at locating genetic markers which are associated with an increased risk for a disease trait (genetic epidemiology), the field of genetics accurately fits within the domain of analyzing complex subject matter. The demand for computational power in this field is steadily increasing. As genetic technology continues to advance (e.g., through finding more efficient methods to ascertain genetic information; development of methods which allow one to obtain more genetic information) the demand for computational power increases.

1.4.1 The Programming Paradigm for the Future of High Performance Computing upon the Personal Computer

As defined by Almasi and Gottlieb (1989), parallel computing is a form of computation in which many calculations are carried out, where a large problem is broken down into two or more smaller problems, and these smaller problems are simultaneously solved [64]. As opposed to solving the larger problem as it exists (serial computing), the act of simultaneously solving the partitioned

smaller problems can lead to the ascertainment of computational results at a quicker rate. For example, suppose it is desired to compute the sum of the initial four counting numbers. We could solve this problem by: summing the initial two counting numbers together; add the resultant to the third counting number; and add the resultant to the fourth counting number. Note that this solution does not adhere to the Almasi and Gottlieb definition of parallel computing (i.e., this solution is serial computing), the computations in which incorporate a total of three sums, each sum in which entails the storing – say to computer system memory – of the corresponding resultant of said sum. In other words, this serial solution requires a total computational time equal to the aggregation of performing three pairwise sums and the storing of three elements (i.e., positive integers).

On the other hand, noting that addition is commutative, to compute the sum of the initial four counting numbers, we could break this problem down into two disjoint smaller problems (i.e., adhering to the Almasi and Gottlieb definition of parallel computing), each handled by an independent *thread*:[5] the sum of the first two counting numbers, denoting the resultant by s_1; and the sum of the third and fourth counting numbers, denoting the resultant by s_2. These two disjoint problems are simultaneously solved, so that to this end, the computational time is equal to the aggregate of one sum and the storing of one element to system memory. The desired result is obtained by summing the two resultants, s_1 and s_2. Thus, overall the parallel computing solution has required a total computational time equal to the aggregation of two sums and the storing of two elements. All else being equal, the computational time required by the serial solution is 1.5 times that of the parallel solution. Therefore, when compared to serial computing, parallel computing can lead to the ascertainment of computational results at a quicker rate. Note that the actual speedup of the parallel program – over that of the corresponding serial program – is dependent on the proportion of the programming code written in a parallel context. This phenomenon is known as Amdahl's Law [66].

Parallel computing is not a novel notion and has been employed for many years, mainly in high performance computing (e.g., computer clusters and supercomputers), but interest has grown recently at the personal computing level due to physical constraints (e.g., heat dissipation and electricity consumption) of microprocessors (CPUs) [67]. These constraints essentially prevent increases in frequency scaling (a measure of the speed of a microprocessor). In fact, the computer industry has accepted that future performance increases in CPUs must largely come from increasing the number of cores within the CPU, rather than making a single core go faster [67]. Indeed, to circumvent these

[5]Threads are sequential processes that share memory [65].

physical constraints, CPU manufacturers, such as Intel and Advanced Micro Devices (AMD), have recently (mid 2006) developed the multi-core CPU for the personal computer. One can envision each core of a multi-core CPU: analogous to existing as a single-lane upon a multi-lane highway; its assigned computations are performed independently of other cores, which allows for uninterrupted computational flow from-and-to system memory. Thus, all else being equal (e.g., CPU clock speed, memory speed, etc.) the multi-core CPU comprised of c cores is capable of performing c times as many computations per unit time as that of the single-core CPU of yesteryear. The act of unlocking the full capabilities of the multi-core CPU, reduces to parallel computing. That is, the personal computer user streams specially written programming code to the multi-core CPU, thereby activating the cores within said CPU. In brief, the adaptation of parallel computing upon the personal computer consists of two essential components: a multi-core CPU (or, as we will encounter within §1.4.2, manycore GPU); and specialized programming code. Without the latter, the multi-core CPUs of the future are no more useful than the single core CPU of yesteryear. Therefore, parallel computing is indeed the programming paradigm of the future for high performance computing upon the personal computer.

1.4.2 Parallel Computing upon the NVIDIA Manycore GPU

Since many personal computers possess a GPU which is independent of the CPU, there are essentially two competing ways – hardware specific – in which to program in parallel upon the personal computer, either by way of programming specifically to: the multi-core CPU; or, the manycore GPU. Here, we motivate the utility of the manycore GPU over the multi-core CPU as the specific hardware utilized for parallel computing upon the personal computer. In order to do this, let us first briefly outline the required components for parallel computing upon the personal computer:

1. A computer warehousing at least one of a multi-core CPU or manycore GPU;
2. Ability for the user to program within a high-level programming language (e.g., C, C++, FORTRAN);
3. A specialized toolkit – computer hardware (i.e., CPU or GPU) and programming language specific – which provides the user a set of extensions (to harness the parallel computing nature of the hardware) to the high-level programming language; and

4. A compiler capable of compiling the specialized parallel programming code, where parallel programming code is defined as any code written through the collaboration between (2) and (3) above.

Henceforth, any references to CPU and GPU are synonymous with multi-core CPU and [NVIDIA] manycore GPU, respectively.

There is an array of reasons, justifying programming in parallel upon the GPU over that of the CPU. First, whereas the CPU is currently – as of December 2011 – limited to comprise six cores (Intel Westmere/Gulftown processors), the GPU can contain upwards of 1024 cores (NVIDIA GeForce GTX 590). This surplus in core units over the CPU, in-and-of-itself, makes the GPU the more attractable resource for parallel computing upon the personal computer. Moreover, even with hyper-thread – each processing core being able to concurrently process multiple threads – support, the Westmere/Gulftown CPUs are merely capable of processing twelve (12) threads (i.e., operations) concurrently [68]. On the other hand, each of the sixteen (16) multiprocessors upon the NVIDIA GeForce GTX 580 GPU can concurrently process 1536 threads, so that the maximum number of active threads concurrently processed upon this GPU is 24576 [68, 69].

Second, the NVIDIA corporation's – a worldwide leader in graphics card manufacturing – Compute Unified Device Architecture (CUDA) toolkit, designed for parallel computing upon NVIDIA GPUs, is provided free of charge and readily downloadable from the NVIDIA website.[6] Moreover, the CUDA toolkit contains the aforementioned required parallel components for each of (3) (programming language extensions) and (4) (compiler), thereby providing: a consolidated means by which to ascertain said two parallel components; maximum compatibility between the parallel programming code and the compiler utilized to compile said code; and maximum compatibility with its targeted computer hardware. In contrast, obtaining a toolkit for parallel programming upon the CPU is either through a third-party (relative to the CPU manufacturer) – such as Open Multi-Processing (OpenMP) or Open Computing Language (OpenCL) – or, essentially not free of charge. In utilizing a third-party toolkit, one introduces the potential for incompatibility between each of: the parallel programming code; the compiler; and the targeted computer hardware. These ideas hold true since the toolkit is geared toward several possible intended hardware profiles, and the compiler is 'third-party' to the toolkit. As of December 2011, although CPU manufacturer Intel has

[6] Available: `http://www.nvidia.com/object/cuda_get.html`.

released several toolkits (e.g., the Intel Parallel Studio Suite software) there is a fee associated with obtaining the software, of which the minimum MSRP is \$799.[7]

Third, the computational speed of the GPU is substantially greater than that of the CPU. As of May 2011, the computational ability of the fastest NVIDIA GPU (NVIDIA GeForce GTX 580 GPU) was over 1.5 teraflops (one teraflop (TFLOP) = one trillion floating point operations per second) [69]. Whereas, at the same point in history, the computational ability of the fastest CPU (Intel Westmere CPU) was less than 13% of that for this GPU [69]. Fourth, the bandwidth – the quantity of information being able to be moved per unit time – of the memory for the GPU is much greater than that of the CPU. As of May 2011, the memory bandwidth of the GPU ($\sim$ 195 gigabytes per second) was about 450% greater than that of the fastest CPU (NVIDIA GeForce GTX 580 GPU versus the Intel Westmere CPU) [69].

Finally, the computational power of the GPU is readily scalable. Whereas the top-end motherboards for personal computers offer support for a single CPU, many of these motherboards are comprised of multiple graphics card expansion slots. This implies that one can introduce multiple GPUs upon these motherboards, thereby scaling – the factor of which is essentially equal to the number of GPUs warehoused within the personal computer (see §2.6 for an illustration of this notion) – the computational power of the GPU over the CPU. In particular, the ASUS P6T7 WS SuperComputer motherboard[8] supports up to four NVIDIA GeForce GTX 580 GPU graphics cards, providing upwards of six teraflops (1.5 TFLOPs for each GPU) of GPU computing performance.

1.4.3 The NVIDIA CUDA Programming Model

In November 2006, the NVIDIA corporation introduced their Compute Unified Device Architecture (CUDA), "A general purpose parallel computing architecture that leverages the parallel compute engine in NVIDIA GPUs to solve many complex computational problems in a more efficient way than on a CPU" [69, 70]. Here, we interface CUDA with the C programming language, which is called CUDA C programming [69]. CUDA C programming is heterogenous computing, insofar as it involves running code on two different platforms – each embedded within the same personal computer system – concurrently: a *host* system with a CPU; and one or more *devices* (frequently graphics adapter cards) with CUDA-enabled NVIDIA GPUs. This is accomplished by way of the CUDA data processing flow:

[7] Retrieved from `http://software.intel.com/en-us/articles/buy-or-renew/`, December 30, 2011.

[8] Retrieved from `http://usa.asus.com/product.aspx?P_ID=9ca8hJfGz483noLk&templete=2`, December 30, 2011.

1. Copy data from host memory to device (known also as global) memory;

2. Host instructs the device to process data;

3. The device executes in parallel upon its cores; and

4. The results are copied from device memory to host memory.

At its core are three key abstractions – a hierarchy of thread groups, shared memory, and barrier synchronization – which are simply exposed to the programmer as a minimal set of language extensions.

CUDA extends upon the C language by allowing the user to write C [device] functions, known as *kernels*. As opposed to regular C functions being executed once, when invoked kernels are executed N times upon the device in a parallel manner by N different CUDA threads. In other words, a single kernel call of N threads is analogous to simultaneously executing N iterations of a [solely serial based] C function. Threads are organized (i.e., grouped) – at the host level – into a grid of thread blocks. Threads are indexed and identified by the device through the `threadIdx` CUDA resource control variable, while blocks are indexed by way of the `blockIdx` CUDA resource control variable. At the simplest level, this within-blocks thread index is one-dimensional (maximum of three-dimensions), for which threads are identified by the CUDA resource control variable `threadIdx.x` (the '.x' references the first dimension of the `threadIdx` control variable). Similarly, the simplest within-grid block index is one-dimensional (maximum of three-dimensions), for which blocks are identified by the CUDA resource control variable `blockIdx.x`. The number of one-dimensional thread blocks of the CUDA grid – assigned by the user at time of kernel execution at the host level – is referenced within the device by way of the CUDA resource control variable `gridDim.x`; the number of one-dimensional threads per thread block of the CUDA grid – maximum value of 1024 upon the NVIDIA GeForce GTX 470 GPU, the GPU used by this author, assigned by the user at time of kernel execution at the host level – is referenced within the device by way of the CUDA resource control variable `blockDim.x`. Table 1.1 displays a CUDA grid of `gridDim.x` $= B$ one-dimensional thread blocks, each block comprised of `blockDim.x` $= T$ one-dimensional threads.

Thread blocks are required to execute independently – it must be possible to execute them in any order, in parallel or in series. This independence requirement allows thread blocks to be scheduled in any order across any number of cores as depicted by Table 1.1, enabling programmers to write code that scales with the number of cores. Threads within a block can cooperate by sharing

Table 1.1: A CUDA Grid of B Thread Blocks, Each Block Comprised of T Threads.

Grid of `gridDim.x` $\times$ `blockDim.x` $= B \times T$ Threads		
Thread Block 1 (`blockIdx.x` $= 0$)	$\cdots$	Thread Block B (`blockIdx.x` $= B - 1$)
Thread ID (`threadIdx.x`) $0 \quad \cdots \quad T-1$	$\cdots$	Thread ID (`threadIdx.x`) $0 \quad \cdots \quad T-1$

data through a medium called *shared memory*, and the user can place *barrier synchronization* points within the kernel to coordinate memory accesses. More precisely, one can specify synchronization points in the kernel by calling the CUDA `__syncthreads()` intrinsic function; `__syncthreads()` acts as a barrier at which all threads in the block must wait before any is allowed to proceed. Shared memory is expected to be much faster than global device memory – "any opportunity to replace global memory accesses by shared memory accesses should therefore be exploited" [69].

The CUDA architecture is built around a scalable array of multithreaded Streaming Multiprocessors (SMs). When a CUDA program on the host CPU invokes a kernel grid, the blocks of the grid are enumerated and distributed to multiprocessors with available execution capacity. The threads of a thread block execute concurrently on one multiprocessor, and multiple thread blocks can execute concurrently on one multiprocessor. As thread blocks terminate, new blocks are launched on the vacated multiprocessors. A multiprocessor is designed to execute hundreds of threads concurrently. To manage such a large amount of threads, it employs a unique architecture called Single-Instruction, Multiple-Thread (SIMT).

The SIMT within the multiprocessor creates, manages, schedules, and executes threads in groups of thirty-two (32) parallel threads called *warps*. Individual threads composing a warp start together at the same program address, but they have their own instruction address counter and register state and are therefore free to branch and execute independently. The term *warp* originates from weaving, the first parallel thread technology [68]. When a multiprocessor is given one or more thread blocks to execute, it partitions them into warps that get scheduled by a warp scheduler for execution. The way a block is partitioned into warps is always the same; each warp contains threads of consecutive, increasing thread IDs with the first warp containing thread 0. For further details, the reader is encouraged to review the document at `http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf`.

1.4.4 Example

As a simple example of an arithmetic problem which can be solved in a parallel manner within the CUDA C programming environment, consider summing over the elements contained within the SNP profile for the i^{th} study participant, $\mathbf{g}_i$, where $\mathbf{g}_i$ – and its corresponding elements, g_{ji}, $j = 1, \ldots, m$ – are defined within §2.2.1, some $i = 1, \ldots, n$, and for notational clarity and simplicity of explanation we assume that $m = 2^{10} = 1024$. We denote the resultant of this sum by s_i. It is,

$$s_i = \sum_{j=1}^{m} g_{ji}. \tag{1.4}$$

To compute this sum – in serial within a high-level programming language – we could follow the procedure of Algorithm 1.1.

Algorithm 1.1 Serial Sum

$s_i \leftarrow 0$. {Initialize the value of s_i to zero}.

for $j = 1$ **to** m **do**

  $s_i \leftarrow s_i + g_{ji}$. {Increment the value of s_i by that of g_{ji}}.

end for

To carry out the recipe outlined in Algorithm 1.1 within the C programming language environment, we could invoke Code Snippet 1.1. For this code note that: after each of the m iterations of the `for` loop, the resultant `s` (i.e., s_i) is updated with the value `g[j]` (i.e., $g_{\{j+1\}i}$, $j = 0, \ldots, m-1$); the $(k+1)^{\text{st}}$ iteration of the loop does not begin until the k^{th} iteration has completed, $k = 1, \ldots, m-1$. Thus, a total of m arithmetic operations are performed at m distinct points in time. This code is an example of a particular *scan* operation, called sequential scan [71, 72].

Code Snippet 1.1.

```
s = 0;
for(j = 0; j < m; j++)
    s += g[j];□
```

On the other hand, to compute the sum (1.4) in a parallel manner within the CUDA C programming environment, we could follow the procedure of Algorithm 1.2.

Algorithm 1.2 Parallel Sum

1. (Host) Copy the elements g_{ji}, $j = 1, \ldots, m$, from a host memory object to a device memory object, as follows. Suppose the elements g_{ji}, $j = 1, \ldots, m$, reside within the host memory object (vector) `h_data` (of data type, say `unsigned int`). Here, for a given memory object, we use the prefixes `h`, `d`, and `s` to reference host memory, device memory, and shared memory, respectively. Now, a CUDA kernel can only access device or shared memory objects, and cannot directly access the elements within a host memory object (e.g., `h_data`). So, to proceed, we must: create (or, allocate) a device memory object which will warehouse the elements of `h_data`, say `d_data`, and copy the elements of `h_data` to `d_data`. To carry out these respective tasks, we invoke the following two lines of code

```
cudaMalloc((void **) &d_data, m * sizeof(unsigned int));
cudaMemcpy(d_data, h_data, m * sizeof(unsigned int), cudaMemcpyHostToDevice);
```

2. (Host) Invoke a kernel comprised of one block ($B = 1$) and $T = 512$ threads, as follows. We first note that per [69], a kernel is defined using the `__global__` declaration specifier, where the return data type is required to be `void`. Next, note that our kernel requires two parameter specifications: the device object `d_data`, so that the kernel can access and operate upon the corresponding elements within this object; and, a device object, say `d_result`, which will warehouse the value of (1.4). Overall, our kernel declaration – whose name is `SNP_Add` – is

```
__global__ void SNP_Add(unsigned int *d_data, unsigned int *d_result)
```

   Finally, to call this kernel, we must specify the number of blocks (B) and number of threads per block (T). This is carried out by way of the *execution configuration*, `<<< B, T >>>`. The following code is used to call our kernel

```
SNP_Add<<< 1, 512 >>>( (unsigned int *) d_data, (unsigned int *) d_result);
```

3. (Device) Each thread of the kernel – indexed by `threadIdx.x` $= 0, \ldots, T-1$ – loads (copies) two elements from the collection $\{g_{1i}, \ldots, g_{mi}\}$, say $g_{\{\texttt{2threadIdx.x+1}\}i}$ and $g_{\{\texttt{2threadIdx.x+2}\}i}$, to shared memory. Denote these elements by the shared memory object `s_sum`. The following code declares the shared memory object `s_sum`, copies the elements from the device memory object `d_data` to said shared memory object, and synchronizes the threads within the block. The thread synchronization is critical, since more than a single thread warp

is being invoked within our kernel call (in fact, a total of 16 warps (each warp comprised of 32 threads) comprises 512 threads of a block) – we cannot proceed with calculating our sum until all data is loaded into shared memory; threads of different warps cannot communicate with one another.

```
__shared__ unsigned int s_sum[1024];
s_sum[2*threadIdx.x] = d_data[2*threadIdx.x];
s_sum[2*threadIdx.x + 1] = d_data[2*threadIdx.x + 1];
__syncthreads();
```

4. (Device) Perform the sum within the shared memory object `s_sum`:

   for $k = 1$ **to** 10 **do**

   $d \leftarrow 2^{10-k}$. {Set the stride value between paired elements}.

   if `threadIdx.x` $< d$ **then**

   `s_sum[threadIdx.x]` $\leftarrow$ `s_sum[threadIdx.x]`+`s_sum[threadIdx.x + d]`. {Increment the sum}.

   end if

   Synchronize threads. {Wait for each thread to finish its corresponding task before continuing}.

   end for

5. (Device) Note that the element `s_sum[0]` contains the desired sum. If `threadIdx.x` $= 0$, then store said element to the device memory address `d_result[0]`.

6. (Host) Copy the device memory associated with `d_result` to host memory, storing the result within the object `h_result`.

To carry out the recipe outlined within Algorithm 1.2, we could invoke Code Snippet 1.2. The reader should take note that each thread executes the entire code presented within the 'DEVICE KERNEL' portion of said code snippet – this is an important point, which when overlooked, could distort ones' interpretation of the code from the actual interpretation thereof. Note that the loop within this code is comprised of a mere ten (10) iterations, compared to the 1024 iterations within the serial code of Code Snippet 1.1. Hence, all else being equal, obtaining the sum of (1.4) by way of parallel computing is a much more efficient approach than that of serial computing.

Code Snippet 1.2.

```
// DEVICE KERNEL //
__global__ SNP_Add(unsigned int *d_data, unsigned int *d_result)
{
    // INITIALIZE SHARED MEMORY //
    __shared__ unsigned int s_sum[1024];

    // INITIALIZE VARIABLE d (STRIDE FOR ELEMENT REDUCTION) //
     unsigned int d;

    // LOAD PAIRS OF ELEMENTS INTO SHARED MEMORY //
     s_sum[2*threadIdx.x] = d_data[2*threadIdx.x];
     s_sum[2*threadIdx.x + 1] = d_data[2*threadIdx.x + 1];

    // SYNCHRONIZE THREADS //
    __syncthreads();

    // PERFORM SUM W/IN SHARED MEMORY //
    for(d = blockDim.x; d > 0; d >>= 1)
    {
         if(threadIdx.x < d)
            s_sum[threadIdx.x] += s_sum[threadIdx.x + d];
        __syncthreads();
    }

    // WRITE OUT RESULT TO GLOBAL MEMORY //
    if(threadIdx.x == 0) d_result[threadIdx.x] = s_sum[threadIdx.x];
} // END KERNEL //

////////////////////////////////////////

// WITHIN main() SECTION //

// ALLOCATE DEVICE MEMORY FOR OBJECTS d_data AND d_result //
cudaMalloc((void **) &d_data, m * sizeof(unsigned int));
cudaMalloc((void **) &d_result, 1 * sizeof(unsigned int));

// COPY HOST MEMORY TO DEVICE MEMORY //
cudaMemcpy(d_data, h_data, m * sizeof(unsigned int), cudaMemcpyHostToDevice);

// INVOKE KERNEL <<< BLOCKS, THREADS PER BLOCK >>> //
SNP_Add<<< 1, 512 >>> ((unsigned int *) d_data, (unsigned int *) d_result);

// COPY RESULT UPON DEVICE MEMORY TO HOST MEMORY //
cudaMemcpy(h_result, d_result, 1 * sizeof(unsigned int), cudaMemcpyDeviceToHost);□
```

1.5 Outline of the Chapters

Insofar as GWAS is essentially a non-hypothesis driven approach (i.e., it interrogates the genome, by way of the sampled SNP panel, searching for potential DSLs), it unleashes an extraordinary multiple hypothesis testing problem. Within Chapter 2 we argue that control of the FWER is most befitting for the GWAS MHT problem (§2.1.1), describe recently proposed approaches to controlling the FWER in genetic studies, including their limitations when applied to large scaled studies such as GWAS (§2.1.1), introduce our proposed parallel processing approach to abolishing the computational burden when applying the so-called maxT and minP permutation based multiple testing procedures for control of the FWER within a GWAS (§2.1.2), propose data management tools for efficient utilization of the data when processing a permutation null distribution (§2.3), propose a parallel algorithm for exceptionally rapid implementation of the maxT and minP permutation multiple testing procedures upon a GWAS data set (§2.4), outline a methodology for clustering the parallel algorithm (§2.5), demonstrate proof of concept by way of applying our tools against a small proportion of the SNP markers encompassing a GWAS data set (§2.6), benchmark our proposed tools to some of the existing software packages for analyzing GWAS data (§2.7), and provide a summary (§2.8).

GWAS relies upon the common-disease common-variant (CDCV) hypothesis of Lander (1996) [9], which asserts that some of the genetic risk to common diseases is due to several common risk variants; individually the variants alter the risk of disease by a minor amount, but collectively they could increase risk substantially [16]. However, not all genetic effects within complex disease etiology are due to common allele variants, some [moderate-risk] genetic effects could be due to rare allele variants. In the past, these loci were difficult to identify, insomuch as they do not possess a large enough effect to display a clear Mendelian inheritance pattern (i.e., linkage mapping is not an effective tool at finding these loci), and too rare to be efficiently identified by association approaches. Nonetheless, these variants deserve more extensive attention, because only recently are we beginning to identify them in a systematic fashion by way of exome and genome sequencing [16]. We would like to extend upon GWAS and test the null hypothesis of no genotype-phenotype association upon SNPs which are either suggestive (i.e., based upon the sampled data) or known (e.g., by way of the dbSNP database [73]) to possess rare variant alleles. However, to detect such associations, we argue within Chapter 3 that corrections to currently employed GWAS statistical inference techniques are required; we propose a methodology for these corrections (see §3.4).

Within Chapter 3 we introduce the notion that the methodological approaches described within §2.1.1 have lost sight of the central dilemma encompassing multiple testing correction in GWAS (§3.1), highlight the näive approach of reliance upon asymptotic assumptions for the Cochran-Armitage Trend test (CATT) statistic within the scope of the GWAS MHT problem (§3.2), under certain regularity conditions abolish the asymptotic assumptions and correctly identify the test statistics null distribution for the CATT statistic (§3.2.1), illustrate that control of the FWER for the CATT statistic is dependent upon several parameters of a GWAS sample and its underlying population (§3.2.2), illustrate that the asymptotic chi-square assumption for the CATT statistic can lead to improper control of the FWER within a GWAS (§3.2.3). As hinted to above, unless corrected, this notion is particularly detrimental for future genetic association studies whose sampled SNP panels are comprised of some loci possessing rare variant alleles. We propose a methodology which abolishes the asymptotic assumption for the CATT statistic under the null hypothesis of no genotype-phenotype association across SNP loci (§3.4). When implemented in practice, this method yields proper control of the FWER for the CATT statistic within a GWAS. As it turns out, the realized implementation of this methodology in practice, introduces a difficult computational problem. Within §3.5 we propose a methodology to reduce the computational problem. By way of simulation and real GWAS data, within §3.6 we provide proof of concept for the synergistic methodologies proposed within Sections 3.4 and 3.5. Finally, we provide a brief summary (§3.7).

Lastly, a novel tool for detecting gene-environment interaction is proposed within Chapter 4. We begin (§4.1.1) by describing the conventional approach to assessing this type of interaction upon a single genetic marker, introduce the challenges imposed when assessing this type of interaction upon multiple genetic markers, and highlight some alternative approaches to assessing both gene-gene and gene-environment interaction within genetic association studies. As illustrated within Chapter 3, reliance upon an asymptotic approximation to the test statistics null distribution within the realm of multiple hypothesis testing can lead to improper control of the Type I error rate. Within §4.1.2, we sketch our approach to testing for gene-environment interaction, of which addresses the multiple testing problem by embracing the appropriate permutation null distribution for the test statistics null distribution. We outline the implementation of our methodology within §4.2–§4.4, highlight our permutation approach to the multiple testing problem (§4.5), propose an efficient algorithm (Algorithm 4.1) for sampling from the permutation null distribution of the test statistics null distribution (§4.5.3), propose an exact approach to assessing gene-environment interaction upon a single sampled

genetic marker and a binary environmental factor (§4.6), and propose a network algorithm (Algorithm 4.2) for implementing the aforementioned exact approach (§4.6.1). To demonstrate proof of concept, we conduct a simulation analysis (§4.7) and implement our method upon two case-control samples (§4.10). Within §4.8, we delve into the 'special' cross-interaction pattern of GxE interaction; within §4.9 we investigate control of the FWER for our proposed approach at detecting GxE interaction under partial null hypotheses; and, we end the chapter with a brief summary and our future directions (§4.11).

CHAPTER 2

COMPUTATIONAL TOOLS FOR MULTIPLE HYPOTHESIS TESTING IN GENOME-WIDE ASSOCIATION STUDIES

2.1 Introduction

Current statistical inference problems in genome-wide association studies (GWAS) routinely involve the simultaneous test of hundreds of thousands (or, even millions) of null hypotheses. This testing problem entails inference for high-dimensional joint distributions of complex and unknown dependence structures among the sampled genotype and phenotype data. In turn, this leads to complex dependence structures among the test statistics arising from the simultaneous testing of the null hypotheses. Ignoring the dependence structure among the test statistics can lead to a loss in statistical power within a GWAS. The core methodological and computational issue encompassing GWAS is multiple hypothesis testing (MHT). Within this chapter, we discuss approaches to tackling the GWAS multiple hypothesis testing problem, compare and contrast their operating characteristics and computational performance, and develop a parallel programming algorithm to implement the permutation maxT and minP multiple testing procedures (MTPs).

2.1.1 Approaches to Controlling the FWER in Genome-wide Association Studies

Of the four types of Type I error rates defined within §1.3, it seems strong control of the FWER at level α to be most befitting for application within a GWAS. This is due to the fact that MTPs based upon the PFER are generally more conservative (i.e., leads to an increased reporting of Type II errors) than those based upon the FWER [60]; MTPs based upon the PCER are generally less conservative than those which control either the FWER or FDR, but tend to ignore the multiplicity problem altogether [60]. Furthermore, while MTPs based upon the FDR tend to achieve greater statistical power than those based upon the FWER – particularly, when the ratio of false null hypotheses (m_1) to the total number of tested null hypotheses (m) is large (i.e., the ratio m_1/m is large) – in general they can result in a high probability for the occurrence of one or more false positives (i.e., an inflated FWER) [74]. Although it is highly unlikely that all tested null hypotheses in a GWAS are in fact true (i.e., it is unlikely that $m_1 \equiv 0$), it is likely that the ratio m_1/m is exceptionally small, far less than 1%. Under these conditions, control of the FDR is close to weak

control of the FWER [61]; strong control of the FWER is close to the best methods for weak control of the Type I error rate [75, 76]. In light of the above, strong control of the FWER seems most befitting for application within a GWAS, and likely explains why many – not all (see e.g., [77]) – methodological approaches for multiple testing in GWAS have focused upon control of the FWER. As such, all multiple testing procedures discussed within Chapters 2 and 3 of this manuscript are assumed to control, in the strong sense, the FWER at some user specified level α.

As indicated above, there are issues specific to GWAS designs which influence both how investigators control for Type I errors and the decision of which MTP to be most useful for control in the adopted Type I error rate. In a multiple hypothesis testing MHT problem such as a GWAS, the likelihood of committing some Type I errors increases (i.e., the FWER increases), as we have illustrated above through expression (1.1). The goal of the MHT problem is to control some Type I error rate in the strong sense, say the FWER, while simultaneously maximizing statistical power to reject false null hypotheses. To control the FWER at a predefined level, say α, one implements a multiple testing procedure. The choice of implemented MTP is critical – an overly conservative MTP could result in overlooking genetic markers which are truly associated with the disease under investigation (i.e., an excessive Type II error rate); an overly liberal MTP, on the other hand, could result in excessive false positives (i.e., an excessive Type I error rate). The Bonferroni MTP is, by far, the most exploited MTP within a GWAS (for recent articles, see e.g., [78,79,80,81,82,83,84]) for strong control of the FWER at level α, presumably due to its simplicity of application – for a GWAS comprised of m markers, at the FWER α level one rejects a null hypothesis if its corresponding pointwise p-value does not exceed the ratio α/m.

While the Bonferroni MTP is simple to implement, it ignores LD (see footnote 1 within §1.1 for a review of the LD definition) among the sampled SNP markers. As a consequence, in the presence of correlated SNP markers this MTP is overly conservative [20, 21, 22, 85]. So as to maximize the efficiency of a GWAS, SNPs are often selectively sampled to be nearly free of LD (i.e., to avoid ascertaining redundant information, SNPs should be selected to be essentially statistically independent). In spite of this, some degree of correlation typically exists within the sampled genetic data [20]. Permutation-based MTPs, such as the so-called maxT and minP approaches of [62], are widely considered most powerful for strong control of the FWER at level α within a GWAS, insofar as these MTPs account for the correlation structure amongst the sampled data [22]. We outline the maxT and minP MTPs in more detail within §2.2.4, but point out here that they remain largely

unimplemented due high computation effort upon a GWAS data set (see e.g., [20,21,86,87,88]). For example, performing the necessary number of permutations (100K) upon a typical GWAS data set containing 2500 cases and 2500 controls and $m = 500$K SNP markers using standard software (e.g., PLINK [63]) can take upwards of *four CPU years* to complete [21]. To alleviate this computational burden, there have been several recent algorithms proposed to approximate the GWAS permutation-based maxT and minP gold standard.

When correlation exists upon the tested null hypotheses – by way of LD upon the sampled SNP markers – there is less variation among their corresponding test statistics than if the null hypotheses were mutually independent. This decreases the likelihood of extreme test statistics [20]. With correlated tests, we gain information about the plausibility of a particular null hypothesis based upon the tests of other null hypotheses. One alternative approach to permutation MTPs for control of the FWER in GWAS exploits the correlation structure upon the sampled markers. It is based upon estimating the LD within the data. Then, utilizing the LD estimates in turn to estimate the effective number of tested independent null hypotheses (M_{eff}) and modifying the Šidák MTP[1] replacing the value m within said MTP with the less conservative estimate M_{eff}. By exploiting the correlation within the sampled data, this approach results in a less conservative MHT correction than the Šidák MTP [so also the Bonferroni MTP] (i.e., $M_{\text{eff}} < m$); the approach results in a low computational requirement when compared to permutation MTPs. Cheverud (2001) pioneered this approach, and proposed estimating LD from the eigenvalues of the Pearson correlation matrix for the sampled SNP markers [89]. Subsequently, several author's proposed alternative methods for estimating M_{eff} [86, 88, 90]. However, these methods remain conservative when compared to using the actual permutation null distribution [21,22]. Moreover, [91] and [92] illustrated that the effective number of tested independent null hypotheses varies across p-value levels, thereby demonstrating that the M_{eff} approach can be inaccurate.

A second alternative approximation approach to permutation MTPs for control of the FWER is based upon the framework of the multivariate normal distribution (MVN). The joint distribution of the test statistics under the complete null hypothesis for many statistical tests commonly employed within a GWAS – such as the Cochran-Armitage Trend Test – follows an asymptotic MVN [93, 94]. The articles of [93] and [94] proposed simulating replicates of the test statistics from this asymptotic MVN under $\mathcal{H}_0$ (the complete null hypothesis), and ascertaining adjusted p-values by way of comparing the test statistic replicates with those of the observed data. The proposal of [20] increased

[1] For large values of m – as is the case for GWAS – the Šidák and Bonferroni MTPs are nearly equivalent.

the efficiency of this approach, by direct numerical integration over the MVN probability density function (PDF) under $\mathcal{H}_0$. When applied to data sets of the size of candidate gene studies (i.e., a panel of a few hundred SNPs), these methods have been shown to be as accurate as permutation MTPs (less than 1% average error in adjusted p-values) [20]. However, when applied to GWAS data sets, the accuracy of these methods suffer. Utilizing the Wellcome Trust Case Control Consortium (WTCCC) data [13], [21] demonstrated that these MVN methods only remove about two-thirds of the error in the adjusted p-values relative to the Bonferroni MTP. Due to numerical limitations of integrating over high-dimensional MVN PDFs, these methods require the user to partition the data into small LD blocks (of hundreds of markers each) and integrate the MVN PDF within each LD block. Insomuch as inter-block correlation is ignored, these MVN approaches lead to conservative multiplicity correction. To address this problem, [21] proposed a resampling method called SLIDE (a **S**liding-window approach for **L**ocally **I**nter-correlated markers with asymptotic **D**istribution **E**rrors corrected). However, accuracy and computational efficiency for this approach depends on the size of the window: a large window leads to increased accuracy and decreased efficiency, while a small window leads to decreased accuracy and increased efficiency.

Overall, several permutation approximation methods have recently been proposed, with the intent of: (1) controlling the FWER; (2) avoiding the exceptional computational effort of permutation MTPs; and (3) obtaining greater statistical power over the Bonferroni MTP. The accuracy in these methods seems to be increasing, although some concerns linger. First, there is no agreement to a standard alternative method. In fact, there is a lack of consistency in the reported results across the M_{eff} methods. For example, the results of [88] suggest the M_{eff} estimate of [90] to be liberal in controlling the FWER at the 5% level; the results of [95] suggest control of the FWER at the 5% level for the M_{eff} estimate of [86], to vary between 3% and 7%, where the variation is dependent upon LD; and [90] suggest the M_{eff} estimate of [89] is overestimated for some LD structures in the sampled SNP panel. Second, in order to accurately account for the correlation among the sets of tested hypotheses, one must do so utilizing the joint distribution of the test statistics. The Šidák MTP – for which each of the M_{eff} methods make use of in computing their respective pointwise significance level – does not guarantee control of the FWER for arbitrary distributions of the test statistics [20, 60]. These methods fail to account for the distribution in the test statistics, and as such the validity in their respective extension to the Šidák MTP is questionable. Finally, each of the M_{eff} methods, as well as the MVN methods of [20] and [93], cannot cope with missing SNP

data. As such, imputation methods (e.g., the K nearest-neighbor algorithm of [96]) are required to be implemented to fill-in any missing data, which could lead to differential misclassification bias in their reported results.

In contrast, not only is the permutation based MTP approach the GWAS gold standard, it is also robust to patterns of missing data [89] and fully accounts for the correlation structure within the sampled data. The robustness is due to the patterns of missing data being preserved within the permuted data and is thus also included in estimation of the permutation significance thresholds. In addition, as genotyping technology continues to evolve, one is able to sample DNA sequences within the human genome at increasingly finer resolution. This implies future genetic samples will arguably incur increasing presence of correlation among markers within the sample. Thus, continued implementation of the Bonferroni MTP within future genetic association studies, will lead to an increase in the reporting of Type II errors. Therefore, it is imperative that permutation MTPs be implemented within current and future genetic association studies. Over the past three years, significant progress has been made toward resolving this notion. For example, [85] has developed a Java based software called PRESTO, which is markedly faster than PLINK [63]. When performing 1K permutations upon a 450K SNP sample of 2938 controls and 1749 cases of Crohn's disease, PRESTO was approximately eighteen times faster than PLINK at performing this task. More recently, [22] developed a software called PERMORY, which is exceptionally faster than PLINK. For example, when performing 10K permutations upon a simulated balanced (i.e., equal numbers for each of cases and controls) GWAS sample of size 6000 participants and 500K SNP markers, PERMORY completed this task in 1.9 hours. In contrast, extrapolated run times within PLINK were projected by the authors to be 43 days. Based upon this simulated data set, PERMORY is shown to be on the order of approximately 550 times faster than PLINK.

There is however, a significant problem with the PERMORY approach. Namely, it is not clear how to handle missing genotype data with this approach, since the authors fail to include this notion within the description of their algorithm. In fact, within section 2.4 of the article, the author's have miss-stated a critical fact regarding permutation upon the maxT MTP in the presence of missing genotype data. Namely, the author's claim that the permutation of phenotype elements (i.e., the random shuffling of the elements upon the response vector $\mathbf{y}$ – see §2.2.1 for definition of $\mathbf{y}$) does not change the marginal totals of the 2×3 table (Table 2.1; see §2.2.1–§2.2.3 for appropriate definitions of terms) at locus j. However, this notion is not true for the maxT MTP, when some loci are comprised

of missing genotype data. Because the authors fail to handle missing genotype data within their algorithm, the PERMORY approach is essentially incomplete.

2.1.2 An Efficient Approach for Processing the Permutation Null Distribution of the MaxT and MinP Multiple Testing Procedures

We propose an optimized maxT/minP permutation algorithm for conducting multiple hypothesis tests of the null hypothesis of no genotype-phenotype association within large SNP panel genetic association studies entailing a binary disease trait (e.g., a GWAS sample), denoted GPER.[2] Whereas previous maxT permutation algorithm approaches (e.g., PLINK, PRESTO, PERMORY) make use of the central processing unit (CPU) of the personal computer (PC), our approach is novel in that we exploit offloading the computational burden of the permutation procedure to the graphics processing unit (GPU). Not only does this approach abolish the computational problem for the maxT and minP MTPs, it illustrates the utility of the GPU within the framework of a statistical application. This approach incorporates parallel computing, arguably the programming paradigm for the future of high performance computing (HPC) upon the personal computer (see §1.4.1), and is the key ingredient for many of the algorithms developed henceforth within this dissertation. Moreover, we develop an algorithm for clustering GPER upon multiple GPUs – each GPU residing within a single personal computer – of which we demonstrate a linear scaling in the computational power of GPER over the single GPU implementation.

We provide the underlying details of the GPER algorithm (Algorithm 2.1) within §2.4. Here, we proceed with introducing some notation which will be used throughout the remainder of this manuscript (§2.2), and outline two data management techniques for efficient application of GPER (§2.3).

2.2 General Notation

2.2.1 Data Setup for a GWAS

Consider a GWAS in which data is collected upon m genetic markers among n study participants, where a binary response (trait) is recorded for each participant. For example, such a GWAS data set could have arisen from sampling n_0 controls and n_1 cases (where $n_1 = n - n_0$) from some

[2] Named from the acronym GPU and the word permutation, emphasizing the utility of the graphics processing unit (GPU) in the algorithm.

population, whereupon for each participant we obtain – by way of, say, blood samples – genotypes for a collection of m genetic markers. The data can be succinctly represented, utilizing a single vector (warehousing the binary responses) and a single matrix (warehousing the genotypes across participants and SNP loci). Indeed, the data for the i^{th} participant consists of: the binary response y_i, where

$$y_i = \begin{cases} 1, & \text{if the participant is a case (diseased)} \\ 0, & \text{if the participant is a control (healthy, non-diseased);} \end{cases} \tag{2.1}$$

and SNP profile, $\mathbf{g}_i = (g_{1i}, \ldots, g_{mi})'$, where g_{ji} denotes the genotype of the j^{th} SNP locus for participant i. In turn, the genotype at any SNP locus is defined in terms of the number of copies for the minor allele (the less frequently occurring allele at the locus within the population – zero, one, or two) at the locus. That is, for $j = 1, \ldots, m$, and $i = 1, \ldots, n$,

$$g_{ji} = \begin{cases} 2, & \text{if participant } i \text{ carries two copies of the minor allele at SNP locus } j \\ 1, & \text{if participant } i \text{ carries one copy of the minor allele at SNP locus } j \\ 0, & \text{if participant } i \text{ carries no copies of the minor allele at SNP locus } j. \end{cases} \tag{2.2}$$

For notational clarity, we organize the n SNP profiles by the $m \times n$ matrix $\mathbf{G} = (\mathbf{g}_1 \cdots \mathbf{g}_n)$ (referred to as the genotype matrix), whose row and column indices identify SNP loci and participants, respectively; we denote the vector of binary responses for the n study participants by $\mathbf{y} = (y_1, \ldots, y_n)$, referred to as the response vector.

2.2.2 The Genetic Model of Inheritance – Statistical Model

Here, let G_j and Y denote, respectively, the random variables which correspond to the genotype for SNP locus j, $j = 1, \ldots, m$, and binary response. Within a GWAS, we are interested in testing the null hypothesis of no association between Y and G_j, which we denote by $H_0^{(j)}$. There are several ways in which we can define the alternative hypothesis for the existence of an association between G_j and Y. Each of these approaches encompass the notion known as the *genetic model of inheritance*. A genetic model of inheritance (GMI) for a biallelic SNP locus, describes how the risk of disease is expected to change as the number of copies in the minor allele changes. In the circumstance for which we do not know the GMI between G_j and Y – and, rarely do we know the GMI (this notion especially holds true for diseases with little known etiology) – the GMI under the alternative

hypothesis is specified as the general model [97]. On the other hand, if we know the GMI between G_j and Y under the alternative hypothesis, then – in coherence with the literature – it is assumed to lie among one of the three models: (1) additive; (2) recessive; or (3) dominant [63,97]. As mentioned within §1.1, by far the most commonly assumed GMI in GWAS is that of the additive model [23], and for the sake of clarity in discussion is the GMI we assume here.

The additive GMI assumes the change in the log-odds of disease is linear for a one-unit change in the number of copies for the minor allele at SNP locus j; equivalently, a one-unit increase in the number of copies of the minor allele at the locus, leads to an additive change in the log-odds of disease. Mathematically, if $\pi_{jk} = \Pr(Y = 1|G_j = k)$, for $k \in \{0, 1, 2\} = \mathcal{G}$, the additive GMI assumes the behavior in the π_{jk} satisfy the simple logistic regression model

$$\log(\text{Odds}(\pi_{jk})) = \beta_{0j} + \beta_{1j}k \quad \forall k \in \mathcal{G}, \tag{2.3}$$

where β_{0j} and β_{1j} are population parameters. Therefore, in terms of model (2.3), the test of $H_0^{(j)}$ – against the two-sided alternative hypothesis (denoted $H_a^{(j)}$) under the additive GMI – can be expressed by

$$\begin{aligned} H_0^{(j)} &: \beta_{1j} = 0 \\ H_a^{(j)} &: \beta_{1j} \neq 0. \end{aligned} \tag{2.4}$$

2.2.3 The Cochran-Armitage Trend Test

By combining the elements upon the j^{th} row of the genotype matrix with those of the response vector, we can cross-classify the sample of data for G_j and Y, as depicted by a 2×3 contingency table (Table 2.1). To test the null hypothesis of no association between G_j and Y in GWAS, a commonly applied test statistic is based upon the Cochran-Armitage trend test (CATT) [19, 20, 21, 22], which can be expressed by [98]

$$T_j = \frac{n\left(n\sum_{k\in\mathcal{G}} n_{j1k}v_k - n_1\sum_{k\in\mathcal{G}} n_{jk}v_k\right)^2}{(n_0)(n_1)\left(n\sum_{k\in\mathcal{G}} n_{jk}v_k^2 - \left(\sum_{k\in\mathcal{G}} n_{jk}v_k\right)^2\right)}, \tag{2.5}$$

where v_k, $k \in \mathcal{G}$, denotes the score for genotype $G_j = k$ – used to specify the specific tested trend in the π_{jk} under $H_a^{(j)}$ – and n_{j1k} and n_{jk} are the respective genotype counts in cases and the entire sample. Particularly, taking $(v_0, v_1, v_2) = (t, t+1, t+2)$, for some real number t, the CATT statistic

can be used to test $H_0^{(j)}$ against $H_a^{(j)}$ under the additive GMI. Here, the reader may be speculating to the reason(s) for using the CATT in testing $H_0^{(j)}$, and not directly performing inference upon the slope parameter of the simple logistic regression model (2.3) (e.g., conducting a likelihood ratio test (LRT), score test, or Wald-based test under $H_0^{(j)}$ [99]). Indeed, under $H_0^{(j)}$, the CATT statistic (2.5) is equivalent to Rao's Score test statistic in testing the hypotheses given by (2.4) upon said logistic regression model. We provide a formal statement and proof of this notion as Proposition A.1 within Appendix A.

Table 2.1: Cross-classification of Disease Status and Genotype for SNP Locus j.

	Number of Copies of Minor Allele			
	0	1	2	**Totals**
Cases	n_{j10}	n_{j11}	n_{j12}	n_1
Controls	n_{j00}	n_{j01}	n_{j02}	n_0
Totals	n_{j0}	n_{j1}	n_{j2}	n

2.2.4 The MaxT and MinP Multiple Testing Procedures

Let t_j and $p_j = \Pr\left(T_j \geq t_j | H_0^{(j)}\right)$, denote respective realizations of T_j (2.5) and the pointwise p-value in testing $H_0^{(j)}$. Given an MTP, the adjusted p-value in testing $H_0^{(j)}$, denoted $\tilde{p}_j$, is the nominal level of the entire test procedure at which $H_0^{(j)}$ would just be rejected, given the values of all test statistics involved (see e.g., [60, 62]). That is,

$$\tilde{p}_j = \inf\left\{\alpha \in [0,1] : H_0^{(j)} \text{ is rejected at nominal FWER} = \alpha\right\}, \tag{2.6}$$

where the *nominal* FWER is the α level at which the MTP is performed. For control of the FWER, while simultaneously accounting for the joint correlation among the vector of test statistics $(T_1, \ldots, T_m)$, [62] proposed the *single-step minP adjusted p-value* (hereinafter, minP adjusted p-value) for null hypothesis $H_0^{(j)}$, $\tilde{p}_{j(\text{minP})}$, defined by

$$\tilde{p}_{j(\text{minP})} = \Pr\left(\min_{1 \leq k \leq m} P_k \leq p_j | \mathcal{H}_0\right), \tag{2.7}$$

where P_k denotes the random variable for the pointwise p-value in testing null hypothesis $H_0^{(k)}$, $k = 1, \ldots, m$. Alternatively, one may consider multiplicity correction based upon the *single-step maxT adjusted p-values* (hereinafter, maxT adjusted p-value), defined in terms of the test statistics

$(T_1, \ldots, T_m)$ themselves [60, 62]:

$$\tilde{p}_{j(\text{maxT})} = \Pr\left(\max_{1 \leq k \leq m} T_k \geq t_j | \mathcal{H}_0\right). \tag{2.8}$$

It is noted here that the maxT and minP MTPs control the FWER in the weak sense [60], the notion in which is essentially absent within the GWAS literature – particularly, the articles by [85] and [22] fail to make mention of this notion. Strong control of the FWER holds under the property of *subset pivotality* (see pg. 42 of [62]). The distribution of pointwise p-values $(P_1, \ldots, P_m)$ is said to possess subset pivotality, provided that the joint distribution of the random vector $\left\{P_j : H_0^{(j)} \in \mathcal{H}_0^{\text{p}}\right\}$ is identical, for all $\mathcal{H}_0^{\text{p}} \in \mathcal{P}(\mathcal{H}_0)$ [60], where – as previously defined within §1.3 – $\mathcal{H}_0^{\text{p}}$ denotes a partial null hypothesis over $\mathcal{H}_0$ and $\mathcal{P}(\cdot)$ denotes the power set of $(\cdot)$. It turns out that subset pivotality holds among the pointwise p-values for the Cochran-Armitage Trend test statistics (see pg. 157 of [62]), for which we attain strong control of the FWER within the maxT and minP MTPs upon utilizing the Cochran-Armitage Trend test in testing $\mathcal{H}_0$.

When the distributions of $T_{(m)}$ and $P_{(1)}$ are unknown, the maxT and minP adjusted p-values can be estimated by resampling [60, 62], where $T_{(k)}$ and $P_{(k)}$ denote the k^{th} order statistics for the respective vectors $(T_1, \ldots, T_m)$ and $(P_1, \ldots, P_m)$. Here, in accordance with the PERMORY approach, we consider permuting the response vector, $\mathbf{y}$, a total of R times [22]. Then, in accordance with Box 2 of [60], within the r^{th} permutation, $r = 1, \ldots, R$:

1. Randomly shuffle (i.e., permute) the elements of the response vector $\mathbf{y}$. Permuting the elements of $\mathbf{y}$ – while simultaneously preserving the structure of the genotype matrix $\mathbf{G}$ – creates a situation in which $\mathbf{y}$ is independent of $\mathbf{G}$ $\left(\text{i.e., we are simulating } \mathcal{H}_0\right)$ and preserves the correlation structure and distributional properties of the SNP profiles $(\mathbf{g}_i)$ within $\mathbf{G}$.

2. Compute the test statistic for null hypothesis $H_0^{(j)}$, $t_{j,r}$. If implementing the minP MTP, then compute the pointwise p-value corresponding to $t_{j,r}$, $p_{j,r} = \Pr\left(T_j \geq t_{j,r} | H_0^{(j)}\right)$.

3. If implementing the maxT MTP, then locate the maximum of the $t_{j,r}$, denoted by $t_{(m),r}$. If implementing the minP MTP, then locate the minimum of the $p_{j,r}$, denoted by $p_{(1),r}$.

The maxT and minP permutation adjusted p-values are given by

$$\tilde{p}^*_{j(\text{maxT})} = \frac{\sum_{r=1}^{R} I\left(t_{(m),r} \geq t_j\right)}{R}, \tag{2.9}$$

and

$$\tilde{p}^{*}_{j(\text{minP})} = \frac{\sum_{r=1}^{R} I\left(p_{(1),r} \leq p_j\right)}{R}, \tag{2.10}$$

respectively, where t_j and p_j denote the respective realizations of T_j and P_j under $H_0^{(j)}$ for the observed (non-permuted) data and $I(\cdot)$ is the indicator random variable returning the value of one if the argument $(\cdot)$ is true and zero otherwise.

2.3 Data Management Techniques for Efficient Processing of the MaxT and MinP Permutation Null Distributions

Here, we propose two data management strategies for efficient parallel processing of GWAS data upon the maxT and minP permutation null distributions: §2.3.1 outlines a technique for ordering the data (prior to performing any statistical inference), while §2.3.2 develops an approach for optimized CUDA kernel execution.

2.3.1 Strategic Ordering of the Elements upon the Response Vector

Undoubtedly, the development of an efficient parallel algorithm for processing the permutation null distribution of the maxT (or, minP) MTP requires considerable strategic thought. There would seem to be two perspectives to the strategy: (a) locating routines (e.g., arithmetic operations, conditional arguments, looping routines, etc.) upon the programming code, whereby omission of which would enhance efficiency while simultaneously preserving computational integrity; and (b) data management techniques to improve computational efficiency. Within this section, and the subsequent section to follow, we will look into the latter of these two approaches.

Within step 1 of the [60] procedure (§2.2.4), it was stated that the elements upon the response vector $\mathbf{y}$ are to be randomly shuffled (note that this approach is consistent with the proposal of [22]); within step 2 of the GPER pseudocode (see §2.4 to follow), we state that the columns upon a modified version of the genotype matrix $\mathbf{G}$ are to be permuted. On the surface, it appears that the procedures encompassing these two statements are contradictory. As it turns out, the underlying statistical analysis encompassing computation of (2.9) or (2.10) is *invariant* to our choice of permutation – the elements of the response vector ($\mathbf{y}$) or the columns upon the genotype matrix ($\mathbf{G}$). As such, for the maxT and minP MTPs, here we assume the columns upon $\mathbf{G}$ are to be permuted in lieu of the elements upon $\mathbf{y}$. Note that upon each random permutation of the columns upon $\mathbf{G}$, step 3 of

the GPER pseudocode (§2.4), essentially demands that we construct a 2×3 contingency table (as depicted by Table 2.1) for each of the m SNP loci. We will demonstrate that by strategically choosing a specific ordering for $\mathbf{y}$ prior to implementing GPER, one can abate a considerable proportion of the required computations in generating these 2×3 tables.

In our choosing to permute the columns upon the genotype matrix (as opposed to the elements of the response vector) within GPER, this implies that the locations for the elements comprising the response vector $\mathbf{y}$ are fixed throughout the duration of the GPER implementation. In turn, this implies – prior to implementation of GPER – we can choose the ordering of the elements of $\mathbf{y}$ to our liking, and simply rearrange the columns of the genotype matrix in accordance to the ordering we choose for $\mathbf{y}$. That is, suppose we choose to swap elements y_i and $y_{i'}$ within the response vector $\mathbf{y}$, $i \neq i' = 1, \ldots, n$. Note that by also swapping SNP profiles $\mathbf{g}_i$ and $\mathbf{g}_{i'}$ within the genotype matrix, the observed data remains intact. Let $\mathbf{y}^*$ be our designated choice for the ordering of $\mathbf{y}$, defined by any ordering of $\mathbf{y}$ satisfying

$$\mathbf{y}^* = (y_{(1)}, \ldots, y_{(n_0)}, y_{(n_0+1)}, \ldots, y_{(n)}), \tag{2.11}$$

where recall n_0 equals the number of controls within our sample, and where

$$y_{(i)} = 1 - I(i \leq n_0) \quad \forall i = 1, \ldots, n.$$

Thus, the initial n_0 elements of $\mathbf{y}^*$ represent the controls for our random sample of n participants; the final n_1 elements of $\mathbf{y}^*$ represent the cases of our random sample. Let $\mathbf{G}^*$ be the resulting genotype matrix, ascertained by swapping the columns within $\mathbf{G}$ in such a way so that the subscripts for the SNP profiles within $\mathbf{G}^*$ align with those of $\mathbf{y}^*$. Specifically, for every $k = 1, \ldots, n$, there exists a unique $i = 1, \ldots, n$, such that if $y_k \in \mathbf{y}$ and $y_k = y_{(i)} \in \mathbf{y}^*$, then the k^{th} column of $\mathbf{G}$, namely $\mathbf{g}_k$, is the i^{th} column of $\mathbf{G}^*$ (denoted $\mathbf{g}_i^*$).

Next, note that for each $j = 1, \ldots, m$, and each $k \in \mathcal{G}$, the sums

$$n_{jk} = \underbrace{\sum_{i=1}^{n} I(g_{ji} = k)}_{\text{sum over } j^{\text{th}} \text{ row of } \mathbf{G}} = \underbrace{\sum_{i=1}^{n} I(g^*_{ji} = k)}_{\text{sum over } j^{\text{th}} \text{ row of } \mathbf{G}^*}, \tag{2.12}$$

are constant, irrespective of the column permutation for $\mathbf{G}$ or $\mathbf{G}^*$, where g^*_{ji} is the $(j,i)^{\text{th}}$ element

within $\mathbf{G}^*$. That is, the column margin for the 2×3 contingency table – formed by combining the j^{th} row of $\mathbf{G}$ with the vector $\mathbf{y}$ (or, $\mathbf{G}^*$ with $\mathbf{y}^*$ for that matter), whose elements are given within the vector (n_{j0}, n_{j1}, n_{j2}) – is fixed by the column permutation design of $\mathbf{G}$ or $\mathbf{G}^*$, for all $j = 1, \ldots, m$. This implies that the values for the [fixed] column margin of a table, along with the cell counts for a single row of the table, are sufficient to generate a particular table. In fact, if no missing genotype data is present upon $\mathbf{G}$, then both the row and column margins are fixed – in this circumstance, exactly two cell counts within a particular table are sufficient to generate the table.

Now, the "standard" approach to generating a single row upon a particular table would consider the expression

$$n_{jtk} = \sum_{i=1}^{n} I(g_{ji} = k) I(y_i = t),$$

evaluated at each $k \in \mathcal{G}$, some $t \in \{0, 1\}$. In short, the standard approach to generating a single cell upon a row of a table entails an n-fold sum. On the other hand, for any permutation of the columns upon $\mathbf{G}^*$, it holds

$$\underbrace{\sum_{i=n_0+1}^{n} I(g^*_{ji} = k)}_{n_{j1k}} = n_{jk} - \underbrace{\sum_{i=1}^{n_0} I(g^*_{ji} = k)}_{n_{j0k}} \quad \forall j = 1, \ldots, m; k \in \mathcal{G}. \tag{2.13}$$

Thus, given the values upon the column margin of a table, this expression implies that exactly one of the summands (within the expression) over the response vector $\mathbf{y}^*$ – evaluated at each $k \in \mathcal{G}$ – is sufficient to generate the j^{th} 2×3 contingency table, irrespective of the column permutation of $\mathbf{G}^*$. In short, our proposed approach to generating a single cell upon a row of a table – using $\mathbf{y}^*$ and $\mathbf{G}^*$ – entails a $\min\{n_0, n_1\}$-fold sum (because we can choose either summand of (2.13) to evaluate). Hence, the proportion of computations – necessary to generate a particular table upon implementing this approach – is essentially $\min\{n_0, n_1\}/n$ times those upon implementing the standard approach. Note that the maximum of this proportion is $1/2$, occurring upon a balanced (equal numbers of case and controls) GWAS. Therefore, our proposed data management approach can lead to exceptional computational savings in constructing 2×3 tables when compared that of the standard approach, upon generating the permutation null distributions for the maxT (or, minP) MTPs. For the sake of discussion herein we assume that $n_0 = \min\{n_0, n_1\}$.

2.3.2 Data Compression

As with the preceding section, here we propose a data management strategy which should significantly enhance the computational efficiency of GPER. Specifically, we describe a technique which entails *data compression*. When implemented, we conjecture that this approach will provide a considerable boost in computational performance for GPER.

Note that an efficient computational-based program should attempt to minimize the occurrence of (or, time spent upon) program bottlenecks (i.e., lag times between productive computational evaluations). Within the CUDA C programming model, apparently program bottlenecks occur whenever step 3 of the 4-step CUDA data processing flow (see §1.4.3) is not in operation [68]. Hence, an efficient CUDA C program should attempt to minimize its time spent within steps 1, 2, and 4 of the CUDA data processing flow, thereby avoiding [obvious] program bottlenecks. Albeit, this notion set aside, program bottlenecks can also occur at step 3 of the CUDA data processing flow within a CUDA kernel. For example, prior to conducting its computations, kernel threads may need to copy device memory to shared memory. The time required for copying this memory leads to a program bottleneck. Hence, for this example, omission of some memory copies within the CUDA kernel should, in theory, remove program bottlenecks and boost computational performance. In fact, [68] states – referring to kernel optimization strategies – "Kernel access to global memory also should be minimized by maximizing the use of shared memory on the device. Sometimes, the best optimization might even be to avoid any data transfer in the first place by simply recomputing the data whenever it is needed." Here, we employ such a strategy within GPER by way of decompressing (i.e., recomputing (or, recovering)) the *compressed* data which is read-in to the kernel from device memory. We begin describing our data compression technique within the latter of the two paragraphs which follow; we describe our data decompression technique within Algorithm B.4 (see lines 12-31 of the pseudocode upon step 2 therein) of §B.1.2 of Appendix B.

When applying a permutation-based MTP – such as maxT (or, minP) – to correct for the multiple hypothesis testing problem encompassing the CATT statistic (2.5), by far the most computational problem lies with implementation of step 3 of the GPER pseudocode (see §2.4 to follow) – collapsing the randomly ordered columns upon $\mathbf{G}^*$ into a total of m 2×3 contingency tables. This is due to the fact that construction of the control row upon these tables demands the evaluation of $M = m \times \min\{n_0, n_1\} = m \times n_0$ elements over $\mathbf{G}^*$. Now, a CUDA C approach to constructing the control row, for a given permutation upon the columns of $\mathbf{G}^*$, could lie with invoking a kernel

comprised of m blocks of T threads, some T, where each block constructs said row of the 2×3 table for a particular SNP locus. In fact, the kernel we develop within §B.1.2 essentially adheres to this notion. Note that the number of device-to-shared memory copies within the kernel for this approach is M. While this approach is tenable, provided that the magnitude of M is reducible, it is likely not optimal. This is due to the fact that reduction in the value of M, leads to a reduction in device-to-shared memory copies within the kernel; according to [68], this notion adheres with the CUDA C performance optimization strategies.

Now, the value of M can be reduced, provided that the value(s) for at least one of its factors can be reduced. The values for each of these factors can be reduced by way of combining elements within $\mathbf{G}^*$. Specifically, merging any two columns of $\mathbf{G}^*$ would decrease the value of M to $M - m$; merging any two rows of $\mathbf{G}^*$ would decrease the value of M to $M - \min\{n_0, n_1\}$. The permutation of the columns upon $\mathbf{G}^*$ for GPER (see step 2 of the GPER pseudocode §2.4) prohibits the merging of columns within this matrix. Thus, we propose reduction in the value of M, by way of merging (i.e., compressing) rows within the genotype matrix $\mathbf{G}^*$. Here, we consider merging ρ rows upon $\mathbf{G}^*$ (denoted as a row merge operation) to form a single row upon an updated genotype matrix, where without loss of generality it is assumed that the value of m is divisible by that of ρ.[3] If $\mathbf{G}^*_t$ denotes the vector of observations pertaining to row t of $\mathbf{G}^*$, and $m' = m/\rho$ denotes the total number of row merge operations upon $\mathbf{G}^*$, then for $s = 1, \ldots, m'$, we merge the ρ vectors $\mathbf{G}^*_{(s-1)\rho+1}, \ldots, \mathbf{G}^*_{s\rho}$ to form the s^{th} row of the updated genotype matrix $\mathbf{G}^{(*\rho)}$, such that

$$g^{(*\rho)}_{si} = \sum_{j=(s-1)\rho+1}^{s\rho} 4^{j-(s-1)\rho-1} g^*_{ji} \quad \forall i = 1, \ldots, n, \tag{2.14}$$

where $g^{(*\rho)}_{si}$ denotes the $(s,i)^{\text{th}}$ element of $\mathbf{G}^{(*\rho)}$. Assigning missing genotype values to the numerical value of three (3), it can be shown (see Proposition A.2 of Appendix A) that each possible value of $g^{(*\rho)}_{si}$, namely $g^{(*\rho)}_{si} = 0, 1, \ldots, 4^\rho - 1$, corresponds to a unique specification of the vector $\left(g^*_{\{(s-1)\rho+1\}i}, \ldots, g^*_{\{s\rho\}i}\right)$. Hence, for all $s = 1, \ldots, m'$, it follows that the vector $\left(g^{(*\rho)}_{s1}, \ldots, g^{(*\rho)}_{sn}\right)$ is sufficient for the vectors $\mathbf{G}^*_{(s-1)\rho+1}, \ldots, \mathbf{G}^*_{s\rho}$. In turn the $m' \times n$ matrix $\mathbf{G}^{(*\rho)}$ is sufficient for $\mathbf{G}^*$. This implies that the number of device-to-shared memory copies – upon the kernel discussed within the preceding paragraph – is M/ρ for the proposed compressed genotype matrix $\mathbf{G}^{(*\rho)}$, compared

[3]Note: in the circumstance for which m is not divisible by the value of ρ, one can simply augment $\mathbf{G}^*$ so that m (upon the augmented matrix) is divisible by ρ. For example, concatenating $\mathbf{G}^*$ with the [appropriate number of] n-vector(s), each vector equal to say $(2, 2, \ldots, 2)$, will not affect the statistical results from implementation of GPER.

to M memory copies for the genotype matrix $\mathbf{G}^*$. Therefore, in accordance with [68], the utility of $\mathbf{G}^{(*\rho)}$ within the kernel used to construct the 2×3 tables is optimized over that of using $\mathbf{G}^*$.

2.4 The GPER Algorithm

The GPER algorithm is based upon the NVIDIA CUDA [69, 70] for C GPU parallel compute engine[4] (for details see §1.4.3). A simple pseudocode for the GPER algorithm implementation is given by Algorithm 2.1 as follows:

Algorithm 2.1 GPER

1. As the data for $\mathbf{G}^{(*\rho)}$ are being read-in to system memory, compute the column margins for each of the m 2×3 contingency tables – cross-classifying genotype and phenotype across the loci (Table 2.1). Note: this will require the decompression of the data warehousing $\mathbf{G}^{(*\rho)}$ – use the result of Proposition A.3 (see Appendix A) to decompress the data. Initialize r to the value of one.

2. Permute the columns of $\mathbf{G}^{(*\rho)}$.

    a. Generate random numbers by way of a parallel Mersenne Twister pseudorandom number generator [100].

    b. Order these random numbers (i.e., shuffle the columns of $\mathbf{G}^{(*\rho)}$) by way of parallel Bitonic sort [101, 102]. For an example, see Table B.1 within Appendix B.

3. Collapse the randomly ordered data into m 2×3 contingency tables across the rows of $\mathbf{G}^{(*\rho)}$ by way of parallel reduction. For an example of parallel reduction, see Figure B.1 within Appendix B.

    a. Formulate cell counts for, say, the control row upon each of the tables (i.e., second row of Table 2.1) by way of a parallel data reduction routine. The case row for each of the tables can be formulated, by way of subtracting the appropriate control row from its [permutation invariant] column margin across the m loci – in fact, formulation of the case row upon each of the tables is not necessary, as the column margin and the control row of the table are sufficient for calculating (2.5) under the additive GMI (see equivalent form of the CATT statistic under the additive GMI, (A.3)).

[4] See `http://www.nvidia.com/object/cuda_home_new.html`.

4. Compute the test statistics $t_{j,r}$ in a parallel manner. If implementing the minP MTP, then also compute the p-values $p_{j,r}$ in a parallel manner.

5. Find and store the maximum test statistic ($t_{(m),r}$ – maxT) or minimum p-value ($p_{(1),r}$ – minP).

   a. Locate this value by implementing a parallel scan (see Table B.2 within Appendix B for an example). If $r = R$, then proceed to step 6 below; otherwise, increment the value of r by one and proceed to step 2 above.

6. Within a statistical software package, say `R` (version 2.13.1; July 2011) [103], sort the collection $\left\{t_{(m),r}\right\}_{r=1,\ldots,R}$ (maxT) or $\left\{p_{(1),r}\right\}_{r=1,\ldots,R}$ (minP) into increasing order. Denote the k^{th} ordered value of $\left\{t_{(m),r}\right\}_{r=1,\ldots,R}$ and $\left\{p_{(1),r}\right\}_{r=1,\ldots,R}$, by $t_{(k)}^{\text{perm}}$ and $p_{(k)}^{\text{perm}}$, respectively. At the α level in the FWER, reject $H_0^{(j)}$ if $t_j \geq t_{(\lceil(1-\alpha)R\rceil)}^{\text{perm}}$ (maxT) or $p_j \leq p_{(\lfloor\alpha R\rfloor)}^{\text{perm}}$ (minP), where $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$ are the respective ceiling and floor functions for the argument $(\cdot)$. Amongst the rejected null hypotheses, compute the adjusted p-values by way of expression (2.9) or (2.10).

For clarity in presentation, we provide details for the implementation of parallel algorithms – to carry out the respective procedures of steps 2-5 of the aforementioned GPER pseudocode – within §B.1.1-§B.1.4 of Appendix B (see Algorithms B.1-B.6).

2.5 Clustering GPER

One elegant feature of a resampling based computational procedure (e.g., the maxT MTP) applied against a large-scaled sample is its natural affinity to a parallel algorithm. This is due to the independent characteristics of one data set resample to the next, and so also to the large number of statistical tests performed. GPER is itself a parallel algorithm. What we describe here, is a methodology to implement this parallel algorithm in a parallel manner upon more than a single NVIDIA GPU. That is, GPU clustering. By implementing GPER upon a cluster of GPUs, one can, in theory, essentially gain a linear increase – the scale in which is dependent upon the number of GPUs comprising the cluster – in computational power over the single GPU implementation.

Consider a desktop computer system comprised of a CPU (host) and G [identical] NVIDIA GPUs (devices), each GPU of which warehouses, say C CUDA cores, where it is assumed that $G > 1$. There would seem to be two approaches to GPU clustering upon such a computer system: (1) implement a single host thread (i.e., single core of the CPU) to communicate with each of the

GPU devices; or (2) implement a total of G host threads, each of which communicates with a distinct GPU device.

Here, we first consider the latter approach. A significant disadvantage to this approach is that it requires multiple host threads to be simultaneously invoked upon the computer system. When compared to a single host thread CUDA C application, a multiple host thread application requires tracking of several host threads, thereby requiring extra overhead of the user. Nonetheless, there are two apparent approaches to implementing multiple host threads upon the computer system: (a) consider a CUDA C program which incorporates a single host thread, and suppose that one has compiled the program into a .exe file. Furthermore, upon execution, suppose the program integrates a data read (e.g., from an ASCII file) which assigns (maps) the host thread to a specific CUDA enabled GPU device. Here, a total of G host threads may be invoked by simply executing the .exe file a total of G times over, where the ASCII file is manipulated prior to each execution, such that each host thread maps to a distinct GPU device. This approach is somewhat tedious, due to the required manipulation of the ASCII file between successive .exe file executions; or, (b) implement OpenCL within the CUDA C program, by way of integrating the NVIDIA Parallel Insight software[5] with that of the Microsoft Visual Studio 2010 software.[6] However, there is a substantial problem with this approach. Namely, whereas the 'Standard' version of the former software is readily obtained free of charge, there is a fee associated with obtaining the latter software. Although each of the approaches outlined within this paragraph could invoke a CUDA cluster, they each have a formidable drawback: the former through overhead in the ASCII file manipulation between .exe file executions; the latter through an associated cost in software attainment.

On the other hand, the approach of (1) above is much simpler due to the fact that it requires the invocation of merely a single host thread, as opposed to multiple host threads. Due to its simplicity, we would like to implement this approach within the CUDA C programming environment. Prior to doing this, a critical issue must be addressed. Namely, we must confirm that the CUDA toolkit possesses the ability for a single host thread to communicate with multiple GPU devices. As it turns out, this approach is essentially not feasible upon historical installments of CUDA toolkits (i.e., prior to the current version 4.0), as no support for this notion is provided upon the applicable toolkit. However, this is no longer the case with version 4.0 of the toolkit, as a single host thread can communicate with all GPU devices within the computer system [69]. In short, a host thread can

[5] see `http://developer.nvidia.com/nvidia-parallel-nsight`.
[6] see `http://www.microsoft.com/visualstudio/en-us/products/2010-editions`.

set (assign) the device it operates on at any time by calling the CUDA `cudaSetDevice()` function, where the parameter for this function is the device number. By assigning a particular device to the host thread, the user is able to allocate device memory and invoke kernel launches upon the device [69]. Figure 2.1 illustrates a single-host thread induced cluster of G GPUs.

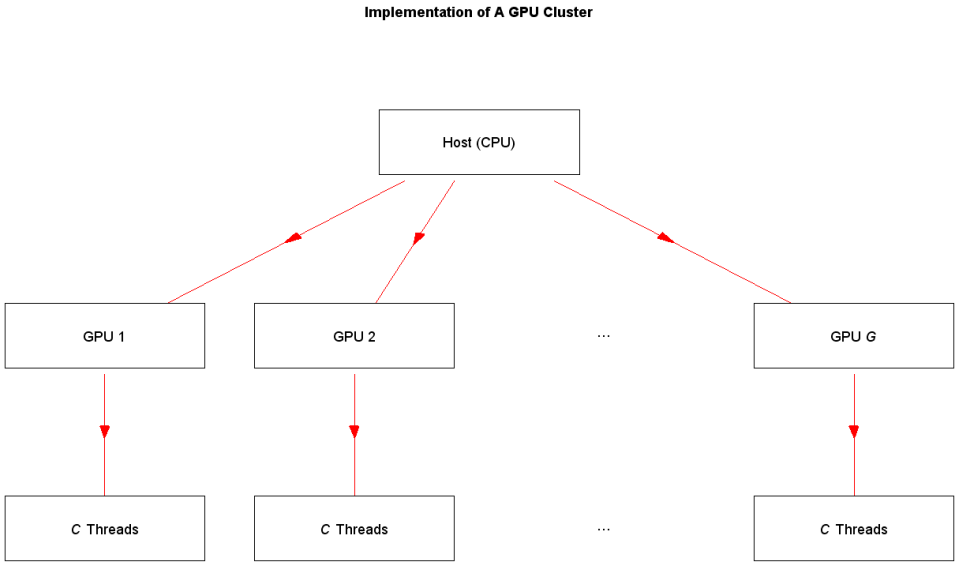

Fig. 2.1: A Single-Host Thread Induced Cluster of G GPUs. Each Arrow Originating from the Host and Terminating upon a Particular GPU, Depicts Control over the GPU by the Host; Each Arrow Originating from a Particular GPU and Terminating upon a Collection of C Threads, Depicts Control of C Simultaneous Operations Invoked upon the GPU. In Theory, a Total of $G \times C$ Computations Can Operate at a Given Point in Time upon the Cluster.

For a given GWAS data set of n subjects and m biallelic SNP markers, suppose R permutations of the column labels upon $\mathbf{G}^{(*\rho)}$ are desired. Without loss of generality, let us presume that $R = k \times G$, for some $k \in \mathbb{N}$. Since we assume the G GPUs are identical (see opening paragraph of this section), it follows that distributing k iterations of steps 2–5 of the GPER algorithm to each GPU, results in a theoretical realized speedup in GPER of $100(G-1)\%$ when compared to a computer system warehousing exactly one of these GPUs. The distribution and implementation of the k iterations of steps 2–5 of the GPER algorithm to each of the G GPUs, precisely describes the procedure for parallelizing (i.e., clustering) GPER.

2.6 Application

To illustrate the computational power of the GPER algorithm upon a live GWAS data set, we applied $R = 20\,480$ maxT permutations[7] of our algorithm against 45 168 SNP markers of chromosome (CHR) 1, for a GWAS investigating Bipolar disorder (BD) amongst individuals of European ancestry, comprised of $n_0 = 1034$ controls, $n_1 = 1001$ cases of BD, and $m = 769\,672$ SNP markers.[8] Details for these data can be found within the articles [105, 106]. For data compression, we chose the value of ρ to be four (4), so that our genotype matrix $\mathbf{G}^{(*4)}$ was of dimension $11\,292 \times 2035$. This choice in ρ is the maximum allowed to accommodate the simplest of C data types for the elements comprising $\mathbf{G}^{(*\rho)}$, namely `unsigned char` (8 bits per element), and maximizes memory management (see Table 2.2).

Table 2.2: Memory Storage Characteristics of the $m' \times n$ Genotype Matrix $\mathbf{G}^{(*\rho)}$ for Select Values of ρ.

ρ	Required† C Data Type for Each of the Elements Within $\mathbf{G}^{(*\rho)}$	Number of Required† Bits of Memory to Warehouse $\mathbf{G}^{(*\rho)}$
1	`unsigned char`	$8 \times m \times n$
2	`unsigned char`	$4 \times m \times n$
4	`unsigned char`	$2 \times m \times n$
8	`unsigned short`	$2 \times m \times n$
16	`unsigned int`	$2 \times m \times n$

†As an absolute minimum.

The asymptotic-based Cochran-Armitage Trend test statistic was used to test the null hypothesis of no genotype-phenotype association at each SNP marker, where the additive genetic model of inheritance was assumed under the two-sided alternative hypothesis across SNP loci. As benchmarking tools for our algorithm, we applied $R = 1000$ maxT permutations within the PLINK software (version 1.07; October 2009) [63] and 20 480 permutations within the PERMORY software (version 1.0; October 2010) [22], against these data.[9] All tests were performed upon the same desktop computer system,[10] whose specifications are listed within Table 2.3.

[7] The GPER algorithm performs blocks of 1024 $(= 2^{10})$ maxT permutations. Here, $R = 20 \times 2^{10}$.

[8] This value of m was obtained after SNP filtering. See e.g., [104] for details to SNP filtering in GWAS.

[9] The following PLINK options were invoked for this test: `--tfile --model-trend --cell 1 --mperm 1000` The `--model-trend` option executes the Cochran-Armitage Trend test for testing the null hypothesis of no genotype-phenotype association at each marker, with the additive genetic model of inheritance assumed under the two-sided alternative hypothesis across SNP loci.

[10] Unless specified otherwise, said computer system is used to conduct all computational analyses of this Dissertation.

Table 2.4 summarizes the results of this benchmarking test, where these data indicate the clustered (upon two GPUs) GPER algorithm is greater than 1500 times faster than PLINK (extrapolated result) and more than 12 times faster than PERMORY, when conducting multiple testing correction by way of the maxT MTP. Equivalently, upon this GWAS subset of data, our clustered GPER algorithm is projected to perform *more than four years worth of PLINK maxT computations in slightly less than one day.*

Table 2.3: Specifications of the Components for the Desktop Computer System Used in the Benchmark Tests.

System Component	Description
CPU	Intel Core i7 920 Quad Core 2.66GHz
System Memory	3GB DDR3 1600MHz
GPUs and RAM	2×NVIDIA GeForce GTX 470 1280MB GDDR5[†]
Operating System	Windows XP Home 32bit
CUDA Toolkit	Version 4.0, May 2011
C Programming Frontend	Microsoft Visual C++ 2010 Express
CUDA C Compiler	nvcc (part of the CUDA Toolkit)

[†]Each GPU is comprised of 448 cores, each core operating at a clock speed of 750MHz. See `http://www.nvidia.com/object/product_geforce_gtx_470_us.html` for further details.

Table 2.4: Summary of the Realized Speedup over PLINK and PERMORY for the GPER Algorithm, upon Implementing $R = 20\,480$ Permutations Within GPER/PERMORY and $R = 1000$ Permutations Within PLINK to 45 168 SNP Markers upon Chromosome 1 of a Bipolar GWAS Dataset Comprised of $n_0 = 1034$ Controls and $n_1 = 1001$ Bipolar Cases.

Number of Active GPUs	Number of Permutations Per Active GPU	Computational Time (minutes)			Speedup of GPER Over	
		GPER	PLINK	PERMORY	PLINK[†]	PERMORY
1	20 480	1.27	47.88	8.4	770x	6x
2	10 240	0.64	47.88	8.4	1530x	12x

[†]Extrapolated estimate.

2.7 Performance Benchmarking

To gain a perspective for the computational performance of GPER applied to varying sampling characteristics of GWAS data sets – particularly, dynamics encompassing sample size and case/control balanced[11] nature of the sample – we simulated subsets of GWAS data sets of varying sample sizes and varying balancing effects upon the underlying cases and controls, where the fixed marker density $m = 40\text{K}$ was used across the simulations. For each data set, 4K SNP loci (of the 40K total) were simulated under assumed Hardy-Weinberg equilibrium (HWE) among population genotype frequencies, upon each of the ten (10) minor allele frequencies (MAF; the frequency within the population of the rarer occurring allele at a particular locus) residing within the collection $\{0.01, 0.02, \ldots, 0.10\}$ (see the simulation setup within §3.2.4.1 for a justification in the use of this collection of values). For each data set, $R = 10\,240$ random permutations were applied within GPER and PERMORY and $R = 1000$ permutations were applied within PLINK.[12]

Table 2.5: Computational Time to Perform $R = 10\,240$ Permutations Within GPER and PERMORY, and $R = 1000$ Permutations Within PLINK.

		Computational Time (minutes)			GPER Speedup Over	
Cases (n_1)	Controls (n_0)	GPER	PLINK	PERMORY	PLINK†	PERMORY
1000	1000	0.6	43.0	5.1	785x	8x
900	1100	0.5	42.3	5.9	830x	10x
800	1200	0.5	42.1	5.4	890x	10x
1500	1500	0.8	64.7	7.3	790x	8x
1350	1650	0.8	63.9	7.2	835x	8x
1200	1800	0.7	62.8	7.1	880x	9x
2000	2000	1.2	89.7	7.9	775x	6x
1800	2200	1.1	88.8	8.3	840x	7x
1600	2400	1.0	87.1	7.6	910x	7x

†Extrapolated estimate.

Table 2.5 summarizes the results from this simulation. In all simulations GPER significantly outperformed each of the PLINK and PERMORY softwares, as demonstrated by the figures depicted within the final two columns of the table. Interestingly, for any fixed balancing characteristic of the sample (i.e., 40%, 45%, or 50% cases within the sample), the relative performance of PERMORY to GPER seems to improve as the sample size increases, as shown by the apparent decreasing trend in the figures upon the final column of the table; exactly the opposite notion seems to hold true for

[11] A balanced/unbalanced GWAS sample is comprised of equal/unequal numbers of cases and controls.

[12] All simulations conducted within this section assume: the value of ρ to be four (4); the asymptotic-based Cochran-Armitage Trend test statistic to be used to test the null hypothesis of no genotype-phenotype association at each SNP marker, where the additive genetic model of inheritance is assumed under the two-sided alternative hypothesis across SNP loci; and, GPER implemented upon a single GPU.

the relative performance of PLINK to GPER for increasing sample size (column 6). Moreover, as expected (per the methodology of §2.3.1), these data suggest that the computational performance of GPER increasingly improves as the sample becomes increasingly unbalanced, as demonstrated by the decreasing trend in computational time for a fixed sample size (column 3). Furthermore, although this notion seems to also be true of PLINK (column 4) – and, PERMORY (column 5) when $n = 3000$ – it is more lucid for GPER. For example, let us consider the samples of $n = 4000$. In comparing the relative timing of the unbalanced sample comprised of 40% cases (row 9) to that of the balanced sample (row 7), we find these values to be: 0.83 for GPER; 0.97 for PLINK; and 0.96 for PERMORY. This suggests that the relative efficiency of GPER to each of PLINK and PERMORY increases as the sample becomes increasingly unbalanced.

To examine the performance of GPER applied against m-size marker panels resembling that of GWAS, we simulated GWAS data sets of varying sample sizes for balanced GWAS samples, assuming marker densities of $m = 500$K and $m = 1$M, under two different scenarios governing the underlying MAFs of the markers. The first (denoted simulation scenario 1), was identical to that given above, where each marker panel was simulated uniformly over the collection of MAFs $\{0.01, \ldots, 0.10\}$. For the second (denoted simulation scenario 2), we noted that, by algorithm design, the computational performance of PERMORY is suggestive to be dependent upon the distribution of MAFs comprising the GWAS sample. Namely, in theory, the computational performance of PERMORY is accelerated upon GWAS samples comprising a large proportion of markers with minute MAF. Thus, when applied against GWAS marker panels comprised of MAF distributions resembling that of the former simulation, the performance of PERMORY could be overstated from its anticipated performance in practice. Hence, to obtain an idea for the relative performance of GPER to PERMORY upon GWAS samples – comprised of marker panels assuming MAFs over the entire domain thereof – we simulated MAFs upon marker panels uniformly over the collection $\{0.01, \ldots, 0.50\}$. Overall, for GPER we anticipated no difference in performance between the two simulation scenarios, since by design, the GPER algorithm does not depend upon the MAF distribution of the markers. However, as previously elucidated to, when compared to the former simulation scenario, we anticipated the computational performance of PERMORY to be lower within the latter scenario.

Table 2.6 summarizes the computational time to perform $R = 10\,240$ maxT permutations within GPER and PERMORY, across the marker panel densities and sample sizes for the two simulation scenarios. In all simulations, GPER significantly outperformed the PERMORY software, as demon-

Table 2.6: Computational Time to Perform $R = 10240$ Permutations Within GPER and PERMORY, Across Several Balanced GWAS Sample Sizes, Marker Densities, and Distribution of SNP Minor Allele Frequencies.

			Computational Time (minutes)	
Marker Density	MAF Range	Sample Size	GPER[†]	PERMORY
$m = 500$K	$0.01 - 0.10$	$n = 2000$	7.0 (8x)	62.6
		$n = 3000$	10.9 (7x)	90.9
		$n = 4000$	15.6 (6x)	110.1
$m = 500$K	$0.01 - 0.50$	$n = 2000$	7.0 (16x)	118.7
		$n = 3000$	10.9 (16x)	180.1
		$n = 4000$	15.6 (13x)	218.7
$m = 1$M	$0.01 - 0.10$	$n = 2000$	14.0 (8x)	120.6
		$n = 3000$	22.0 (7x)	175.8
		$n = 4000$	31.2 (6x)	240.9
$m = 1$M	$0.01 - 0.50$	$n = 2000$	13.9 (20x)	294.1
		$n = 3000$	21.9 (14x)	345.3
		$n = 4000$	30.4 (12x)	394.3

[†]Parenthetic values represent speedup over PERMORY.

strated by the figures presented within the final two columns of the table. In addition, a similar – to that of the simulation conducted above with $m = 40$K – increasing trend in relative computational performance of PERMORY to GPER for increasing sample size is apparent here. Nonetheless, even for $n = 4000$, GPER was at least six (6) times faster than PERMORY. Moreover, as expected, the computational performance of PERMORY appears to depend upon the distribution of MAF amongst the SNP sample. Taking the SNP density $m = 500$K, for example, when compared to simulation scenario 1, PERMORY required essentially twice the time to complete the maxT permutations upon simulation scenario 2. The computational performance of GPER, on the other hand, is impervious to the distribution of MAF upon the SNP sample. Overall, based upon these simulations, GPER appears to be the computational tool of choice for use in the maxT MTP upon GWAS data.

2.8 Conclusions

Multiple hypothesis testing correction is vital within a GWAS, as this ensures the proper reporting of false positive genotype-phenotype associations upon the corresponding sampled SNP panel thereof. When testing multiple null hypotheses, there are many definitions for the Type I error rate. Within a GWAS, it seems control of the family-wise Type I error rate is most befitting. The MHT goal is to control the adopted Type I error rate in the strong sense, while simultaneously maximizing statistical power to reject false null hypotheses. The Bonferroni MTP is a popular approach in

GWAS for strong control of the FWER. However, when implemented upon a sample of correlated data, this approach can suffer a loss in statistical power. Meanwhile, the maxT and minP MTPs – the multiple testing procedures which control the FWER in the strong sense and provide maximum statistical power amongst all MTPs controlling the FWER – are seldom implemented within these studies due to their high computational effort.

There would seem to be two general approaches in addressing the computational problem of the maxT and minP MTPs: accelerate the computational components for these MTPs; or, develop an efficient approximation approach and improve its accuracy. The past decade has seen research primarily focused upon the latter approach. We employed the former approach and have developed GPER, an optimized GPU-based algorithm in conducting multiple tests of association within large-scaled categorical genetic data. Our algorithm presents a significant improvement in computational performance over that of the widely utilized GWAS PLINK software, and is on par with the fastest alternative methods (e.g., PRESTO, PERMORY). However, unlike these methods, our approach is novel insofar as we exploit offloading the computational burden for the maxT and minP MTPs to the GPU of the personal computer. Due to frequency (a measure of the speed for a single processing core) scaling limitations of CPUs, the future of HPC upon the PC is arguably parallel computing. Parallel computing upon the GPU of the PC is a very efficient approach to tackling a computational problem, and has begun to see its interface within the statistics discipline (see e.g., [107, 108]). Our implementation of this approach demonstrates the utility of the GPU in tackling an exceptionally demanding computational problem to a sampled GWAS data set, but its utility is not limited to sampled GWAS data (e.g., the Bipolar data set utilized within §2.6). We utilize GPER within the simulation analysis of the next chapter, in demonstrating two key notions therein: (1) that ones' assumption of an asymptotic null distribution for the Cochran-Armitage trend test statistic under $\mathcal{H}_0$, can lead to the gold standard maxT and minP MTP approach yielding *unbalanced* multiplicity adjustment in a GWAS; and (2) to provide empirical evidence in support of the proposed methodology.

We have developed the GPER algorithm, to address the computational issues of the maxT and minP MTPs within the realm of GWAS. However, by modifying the algorithm, this GPU approach can be extended to include other computationally demanding areas of statistics. In particular, the algorithm can be extended to include other parametric multiple hypothesis testing circumstances in which the maxT or minP MTPs are applicable. For example, our approach can be adapted to mi-

croarray experiments, where the maxT MTP can be utilized to correct for MHT of differential gene expression across probesets of a microarray (i.e., MHT correction for parametric t-tests and F-tests; see [60] for an excellent overview of MHT correction in microarray experiments). Additionally, we have successfully modified/adapted the GPER algorithm in extending the methodology of [109] to include the maxT approach thereof (see pg. 5 of this article for the connection of their methodology to the maxT MTP), for MHT correction when testing for gene-environment interactions. The GPER algorithm can also be extended to controlling, say the k^{th}-level generalized FWER (gFWER(k)), $k = 1, \ldots, n$. Control of the gFWER(k) is a generalization of the FWER, where the maxT and minP MTPs are modified and based upon the respective k^{th} and $(n-k+1)^{\text{st}}$ distributions of the order statistics for the test statistics (maxT) and p-values (minP) (see e.g., pp. 256–257 of [110]) – note: taking $k = n$ recovers the FWER and the respective maxT and minP MTPs. Finally, outside the realm of the maxT and minP MTPs – and extensions to controlling the gFWER(k) thereof – our GPU approach could be adapted to other resampling based MHT procedures, such as SAM (see [111] and [112]).

CHAPTER 3

ENHANCEMENTS TO THE STATISTICAL INFERENCE OF GENOME-WIDE ASSOCIATION STUDIES

3.1 Introduction

There seems to be confusion within the literature regarding the central underlying dilemma encompassing multiplicity correction within a GWAS. Contrary to the focus of recent methodological approaches (i.e., the M_{eff} and MVN approaches described within §2.1.1), this dilemma does not entail the computational problem – a *consequence* – which arises from the implementation of permutation MTPs. Rather, said dilemma is *proper application* of the implemented multiple testing procedure; this notion is essentially lost in the GWAS literature. In conducting statistical inference within GWAS, the asymptotic-based Cochran Armitage Trend test statistic is commonly employed to test the null hypothesis of no genotype-phenotype association on a per-marker basis. Due to the extremely small significance level on a per-marker basis, we have found a discrepancy between the asymptotic chi-square distribution for this test statistic and its true underlying null distribution. Reliance upon asymptotic assumptions for this test statistic in this regard can result in improper control of the FWER within a GWAS.

Herein, we develop a methodology to correct the discrepancy between the chi-square distribution for the asymptotic-based Cochran-Armitage Trend test statistic and its true underlying null distribution. Furthermore, this method embraces the minP MTP, thereby accounting for correlation within the sampled data and achieving unbiased strong control of the FWER in a GWAS. Adaptation of this methodology in practice has several key positive repercussions, including: correcting upon improperly obtained statistical results within historical GWAS; and providing multiple hypothesis testing tools, so that statistical inference is properly conducted within current and future genetic association studies.

3.2 The Test Statistics Null Distribution for the Multiple Hypothesis Testing Problem

As one will recall, the application of an MTP is comprised of several components [110]. Arguably, the most vital component is correct identification of the test statistics null distribution (Q_0) – the distribution which serves as the basis for determining test statistic regions which lead to the rejection of posited null hypotheses. Improper identification of Q_0 could lead to control in the FWER at a level other than that intended (see pg. 255 of [110]). When testing hundreds of thousands of null hypotheses, each against their corresponding two-sided alternative hypothesis, the test statistic rejection regions (as defined by the implemented MTP) for the Cochran-Armitage trend test will call for $H_0^{(j)}$ to be rejected for large [small] realized values in its corresponding test statistic [pointwise p-value]. Reliance upon asymptotic (e.g., MVN, chi-square) assumptions for the distributional properties of the underlying test statistics under $\mathcal{H}_0$ in this regard would seem to be a näive approach – yet, this is common GWAS practice – because in utilizing an asymptotic test statistics null distribution ($\tilde{Q}_0$) for their multiplicity correction, one is assuming the tail region of the accompanying PDF to be representative of that for Q_0. However, the veracity in this assumption is highly speculative, insomuch as derivation of the test statistic rejection regions under $\tilde{Q}_0$ is based upon a *continuous* distribution, whereas the underlying distribution of Q_0 is actually *discrete*. That is, once a case-control sample has been drawn from the population, the margins for the contingency table – cross-classifying genotype and phenotype at the j^{th} SNP locus – are uniquely determined. Conditional on these fixed margins at the locus, there is a finite number of realizations for the cells of the contingency table corresponding to these fixed margin counts. In turn, there is a finite number of possible test statistic realizations comprising the support of Q_0, where it is noted that this notion is invariant to the adopted choice of test statistic. Because of the discrete nature of Q_0, reliance upon $\tilde{Q}_0$ for multiplicity correction when testing hundreds of thousands of null hypotheses, opens the door to errors in the correction.

3.2.1 The Cochran-Armitage Trend Test Statistic

As mentioned within §2.1.1, for many popular statistical tests employed within a GWAS (e.g., the Cochran-Armitage trend test), the vector of test statistics under $\mathcal{H}_0$ asymptotically follows an MVN. If the asymptotic test statistics null distribution ($\tilde{Q}_0$) closely approximates the true test statistics null distribution (Q_0), then one can utilize the corresponding PDF of $\tilde{Q}_0$ to ascertain an accurate multiple testing correction. Recall, [20] demonstrate considerable accuracy of the MVN

approach for small SNP samples (i.e., the size of candidate gene studies; see §2.1.1).

However, due to the extremely small pointwise significance threshold in a GWAS when testing the null hypothesis $H_0^{(j)}$ on a per-marker basis, we observe a discrepancy between $\tilde{Q}_0$ and Q_0 for the asymptotic chi-square Cochran-Armitage trend test (CATT) statistic. Reliance upon $\tilde{Q}_0$ in assessing evidence for/against $H_0^{(j)}$ in this regard can lead to improper multiplicity correction (see §3.2.3 for details). For a GWAS sample of n subjects, it is noted that the discrepancy appears to worsen for an increase in the number of sampled SNP markers (m) and/or a decrease in the minor allele (the less frequently occurring allele at a locus within the population) frequency (MAF) at any SNP locus. To this author's review, the article by [21] is the first to recognize the latter phenomenon (i.e., the discrepancy between Q_0 and $\tilde{Q}_0$, dependent upon the MAF; see pg. 4 therein) – although, we argue within the subsequent paragraph below that the methodology of said article does not correctly identify Q_0. Furthermore, note that discrepancies between $\tilde{Q}_0$ and Q_0 on a per-marker basis (i.e., a discrepancy between the χ_1^2 distribution and Q_0 for the CATT statistic), leads to an incorrect multiplicity adjustment over the MVN joint-marker multiplicity correction approach [21]. In other words, a discrepancy between $\tilde{Q}_0$ and Q_0 for the CATT statistic at *some* SNP loci, leads to incorrect multiplicity correction under $\tilde{Q}_0$ for *all* SNP loci. In order to illustrate the discrepancy between $\tilde{Q}_0$ (hereinafter, unless otherwise specified, $\tilde{Q}_0$ is assumed to denote the χ_1^2 distribution) and Q_0, we must first correctly identify Q_0 for the CATT statistic under $\mathcal{H}_0$.

Regardless of the implemented test statistic for testing $\mathcal{H}_0$, correct identification of Q_0 within a GWAS demands accounting for the case-control sampling design of the study. Indeed, in drawing a case-control GWAS sample, the number of subjects falling within each of the two phenotype strata is *fixed* by design; at each SNP locus, the sampled genotypes (e.g., AA, AC, and CC, for a SNP with adenine (A) and cytosine (C) allele variants) form a multinomial (trinomial) *random* vector within each of the phenotype stratum. That is, each row of the 2×3 contingency table – tabulating the sampled data for an arbitrary SNP locus – forms a trinomial random vector. To illustrate, consider a randomly chosen individual from the study population. At the time of sampling, disease status (case or control) for this individual is fixed and known, insofar as the study subjects are selected according to their disease status and further classified according to their exposure status. However, at the time of sampling, the genotypic information across loci for this individual is blinded to the researcher, until which time a blood (or, buccal) sample has been hybridized to a microarray chip and genotyped within the laboratory. Now, since cases and controls are assumed unrelated, at each

SNP locus we can express the probability of observing the sampled genotypes as the product of the probability mass functions (PMFs) for two independent trinomial random vectors. Thus, once the parameters for these random vectors have been specified under $\mathcal{H}_0$, the resulting probability of observing the sampled genotypes – for any realization thereof – at any locus is determined. Hence, given the values in these parameters under $\mathcal{H}_0$, we can generate the *exact unconditional distribution* for the CATT statistic at each SNP locus. Therefore, taken collectively across the loci, these exact unconditional distributions for the CATT statistic define Q_0 under $\mathcal{H}_0$. In their construction of Q_0, the article of [21] failed to recognize and account for the randomness in the genotype data on a per-marker basis. For this very reason, their construction of Q_0 is not entirely correct.

The parameters for these random vectors are unknown, which presents a problem in completely specifying Q_0. However, under the null hypothesis $H_0^{(j)}$, it can be readily shown (see Proposition A.4 within Appendix A) that the parameter vectors for the random trinomials at locus j are equivalent, leaving us with two nuisance parameters (since the specification of any two parameters for a trinomial random variable determines the third) at the locus. Albeit, we can reduce the nuisance parameters to a single parameter, by noting that under Hardy-Weinberg equilibrium (HWE) – in the absence of migration, mutation, natural selection, and assortative mating – genotype frequencies are a simple function of allele frequencies [113]. The aforementioned cited article notes that the underlying assumptions for HWE appear to hold for most human populations, where deviations from HWE at particular markers may suggest problems with genotyping, population structure (a general problem with population based genetic studies), or an association between the marker and disease if the HWE deviation lies within samples of cases. In fact, the assumption of HWE among population controls is so widely accepted, that part of the data filtering process within a GWAS sample entails excluding SNP loci (from inclusion to $\mathcal{H}_0$) whose genotypes among sampled controls significantly deviate from the HWE assumption (see e.g., [104]). Moreover, the article of [114] suggests that the HWE equation is remarkably robust at providing estimates of genotype frequencies in real populations. Thus, assuming genotype frequencies at SNP locus j adhere to HWE within the population, we can specify the single nuisance parameter under $H_0^{(j)}$ through the population minor allele frequency at the locus. Hence, conditional on this minor allele frequency and the fixed numbers of sampled cases and controls, we can generate the exact unconditional distribution of the CATT statistic for every realization thereof. In this regard, at locus j, Q_0 becomes a function of the assumed population minor allele frequency at the locus (π_j) and the fixed numbers of cases (n_1) and controls (n_0) for the GWAS

sample. Henceforth, we denote this unconditional distribution for the CATT statistic under $H_0^{(j)}$ by $Q_{0j}^{(*H)}(\pi_j, n_0, n_1)$. For notational clarity, we reference this null distribution by $Q_{0j}^{(*H)}$. The only assumption for the derivation of $Q_{0j}^{(*H)}$ is that genotype frequencies within the population at locus j adhere to HWE under $H_0^{(j)}$. This assumption is realistic, per the aforementioned argument presented within this paragraph. Therefore, under HWE among population genotype frequencies at SNP locus j, it follows that $Q_{0j}^{(*H)}$ correctly identifies Q_0 under $H_0^{(j)}$; and collectively, $\left\{Q_{0j}^{(*H)}\right\}_{j=1,\ldots,m}$, correctly identifies Q_0 under $\mathcal{H}_0$.

3.2.2 Control of the FWER Is Dependent upon GWAS Sample Characteristics

Having identified Q_0 for the CATT statistic under $\mathcal{H}_0$ and assumed HWE among population genotype frequencies across loci, we can illustrate the notion of the second paragraph within §3.2.1, namely the discrepancy between $\tilde{Q}_0$ and Q_0 for said statistic under $\mathcal{H}_0$. Figures 3.1 and 3.2 display the Bonferroni corrected exact unconditional probability of Type I error (UPTE) for the CATT statistic (hereinafter, the additive genetic model of inheritance is assumed under the two-sided alternative hypothesis $H_a^{(j)}$) under $Q_{0j}^{(*H)}$ for balanced and unbalanced (two to one ratio of controls to cases) GWAS samples, respectively (sample sizes of $n = 1$K (red curves) and $n = 2$K (blue curves)), across the domain of the minor allele frequency within the population, $\pi_j \in (0, 0.5)$, for the realization of the CATT statistic $F_{\tilde{Q}_0}^{-1}(1 - 0.05/m)$, where $m = 10$K (heavy dashed curves), $m = 100$K (light dashed curves), $m = 500$K (solid curves), and $F_{\tilde{Q}_0}^{-1}(\cdot)$ is the inverse cumulative distribution function (CDF) for $\tilde{Q}_0$ evaluated at $(\cdot)$. The value $F_{\tilde{Q}_0}^{-1}(1 - 0.05/m)$ is the minimum realization of the CATT statistic at locus j under $\mathcal{H}_0$ and $\tilde{Q}_0$, for which the Bonferroni MTP calls for $H_0^{(j)}$ to be rejected at the 5% FWER. If $\tilde{Q}_0$ was to correctly identify $Q_{0j}^{(*H)}$ under $H_0^{(j)}$, the colored curves within each of these figures would lie upon the 5% FWER reference line (heavy black dashed line), across the domain of π_j. However, each figure demonstrates the discrepancy between $\tilde{Q}_0$ and $Q_{0j}^{(*H)}$, since each of the colored curves do not lie upon the reference line. Moreover, we see that the discrepancy between $\tilde{Q}_0$ and $Q_{0j}^{(*H)}$ varies, dependent upon: (1) the minor allele frequency within the population at locus j (π_j). As mentioned within the preceding section, the discrepancy appears to be exacerbated for small values in π_j (less than 0.2 for a balanced GWAS; less than 0.1 for an unbalanced GWAS), as each of the colored curves increasingly separate from the 5% FWER reference line as π_j decreases; (2) the GWAS sample size (n). The discrepancy appears to be exacerbated for smaller sample sizes, since for any fixed m and π_j the corresponding curve for the

sample size $n = 1$K, generally lies further away from the 5% FWER reference line than that for the sample size $n = 2$K; (3) the number of SNP markers for the GWAS sample (m). The discrepancy seems to be exacerbated as the marker density increases, since in general for any fixed n and π_j the corresponding curve for $m = 500$K lies further away from the 5% FWER reference line than that for $m = 100$K. A similar observation holds comparing curves $m = 100$K and $m = 10$K, and the curves $m = 500$K and $m = 10$K; and (4) by way of directly comparing the two figures, the balanced nature of the GWAS sample. Therefore, assuming genotype frequencies within the population adhere to HWE across SNP loci under $\mathcal{H}_0$, control of the FWER for the CATT statistic under Q_0 is dependent upon the magnitude in the values of several factors, including π_j, n, m, and the ratio of controls (n_0) to cases (n_1).

3.2.3 Improper Multiplicity Correction

Insofar as $\tilde{Q}_0$ appears to incorrectly identify $Q_{0j}^{(*H)}$ under $H_0^{(j)}$ and HWE among population genotype frequencies at SNP locus j, it would also seem to fail in correctly identifying Q_0 under $\mathcal{H}_0$ and HWE among population genotype frequencies across SNP loci. Reliance upon $\tilde{Q}_0$ for MHT correction in this regard can lead to improper multiplicity correction within a GWAS for the Cochran-Armitage Trend test. When applied upon a GWAS sample, Figures 3.1 and 3.2, respectively, suggest the Bonferroni – or, because m is assumed large, the Šidák – MTP under $\tilde{Q}_0$ is conservative and liberal/conservative for control of the FWER at the 5% level under $\mathcal{H}_0$. For example, the former figure illustrates that for a balanced GWAS, the Bonferroni MTP under each of $\mathcal{H}_0$, $\tilde{Q}_0$, and HWE among population genotype frequencies, is overly conservative in controlling the FWER at the 5% level amongst a GWAS sample of m mutually independent markers. To illustrate, consider fixed values for each element within the vector (π_j, m, n), where without loss of generality and for clarity, we assume m and n are chosen to be values as depicted within the figure. The point lying upon the appropriate curve – satisfying the fixed values (π_j, m, n) – represents the Bonferroni corrected UPTE for the CATT statistic under Q_0, at realization $F_{\tilde{Q}_0}^{-1}(1 - 0.05/m)$, when $\mathcal{H}_0$ is in fact true and each of the m mutually independent markers is sampled from a HWE population with [common] minor allele frequency π_j. Equivalently, this point represents the actual FWER being controlled by the Bonferroni MTP, for the CATT statistic under Q_0, at the realization $F_{\tilde{Q}_0}^{-1}(1 - 0.05/m)$ (assumes all markers mutually independent and possessing common population MAF π_j). But, at realization $F_{\tilde{Q}_0}^{-1}(1 - 0.05/m)$, the Bonferroni corrected Type I error rate under $\tilde{Q}_0$ (i.e., the assumed FWER being controlled) for the CATT statistic is equal to 5%. Since the assumed FWER under $\tilde{Q}_0$ is larger

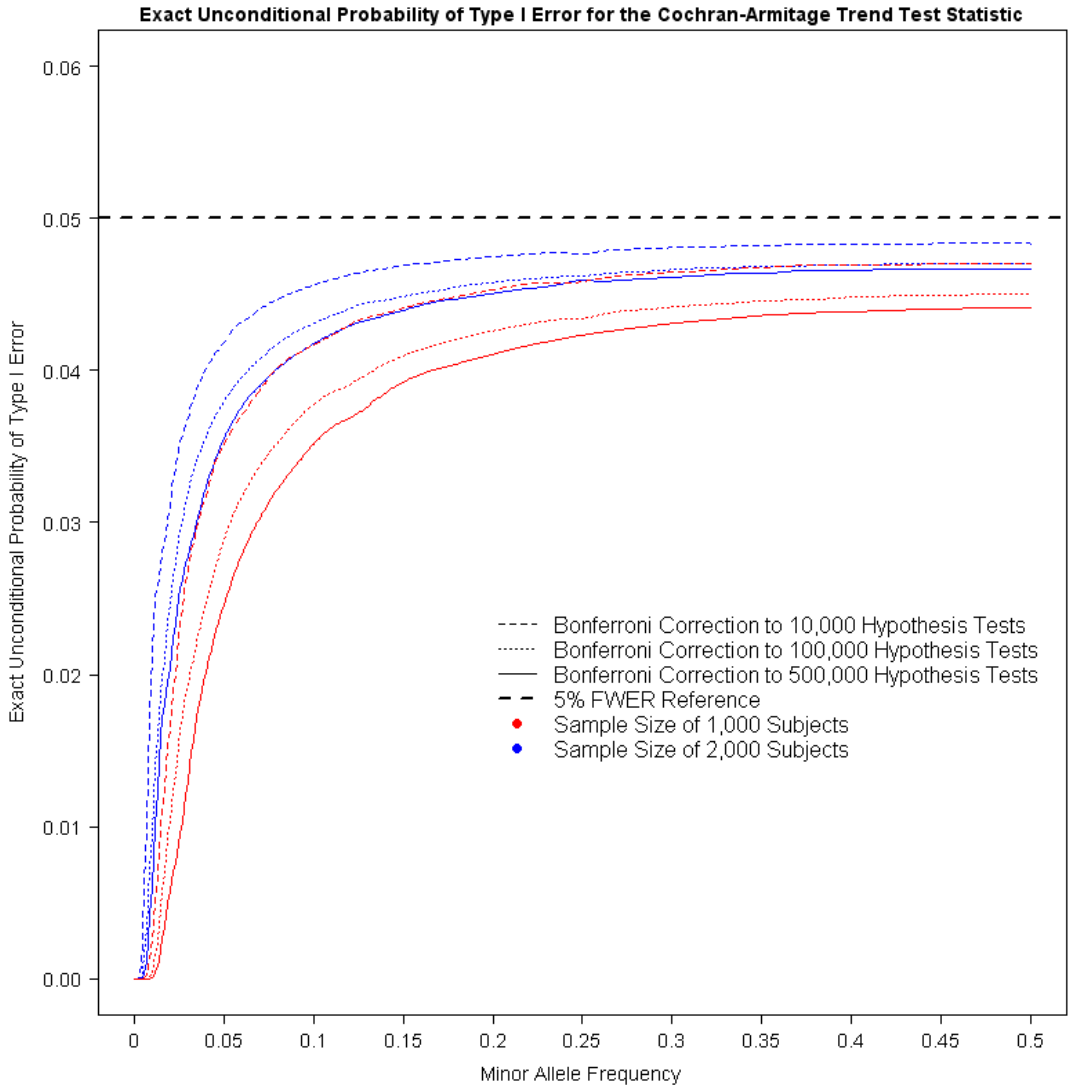

Fig. 3.1: Plot of the Bonferroni Corrected Exact Unconditional Probability of Type I Error for the Cochran-Armitage Trend Test Statistic at the Realization $F^{-1}_{\tilde{Q}_0}(1-0.05/m)$ for the χ^2_1 Distribution ($\tilde{Q}_0$), Across the Population Minor Allele Frequency for a Balanced GWAS, Assuming Population Allele Frequencies Adhere to Hardy-Weinberg Equilibrium. Colored Curves: Heavy Dashed Curves, Light Dashed Curves, and Solid Curves, Assume $m = 10$K, $m = 100$K, and $m = 500$K Tested Null Hypotheses under $\mathcal{H}_0$, Respectively; Red and Blue Curves Assume GWAS Samples of $n = 1$K and $n = 2$K, Respectively. The Assumed FWER under $\tilde{Q}_0$ Is 5% (Heavy Dashed Black Line).

for this realization in the CATT statistic than the actual FWER under Q_0, the Bonferroni correction for the CATT statistic under $\tilde{Q}_0$ is conservative. Mathematically, we can show that the actual Bonferroni corrected FWER for the CATT statistic under Q_0, at realization $F^{-1}_{\tilde{Q}_0}(1-0.05/m)$, can never exceed the supremum of the curve (for the assumed fixed values of m and n) over π_j. Since the supremum for each of the curves within the figure lie below the 5% FWER reference line, the

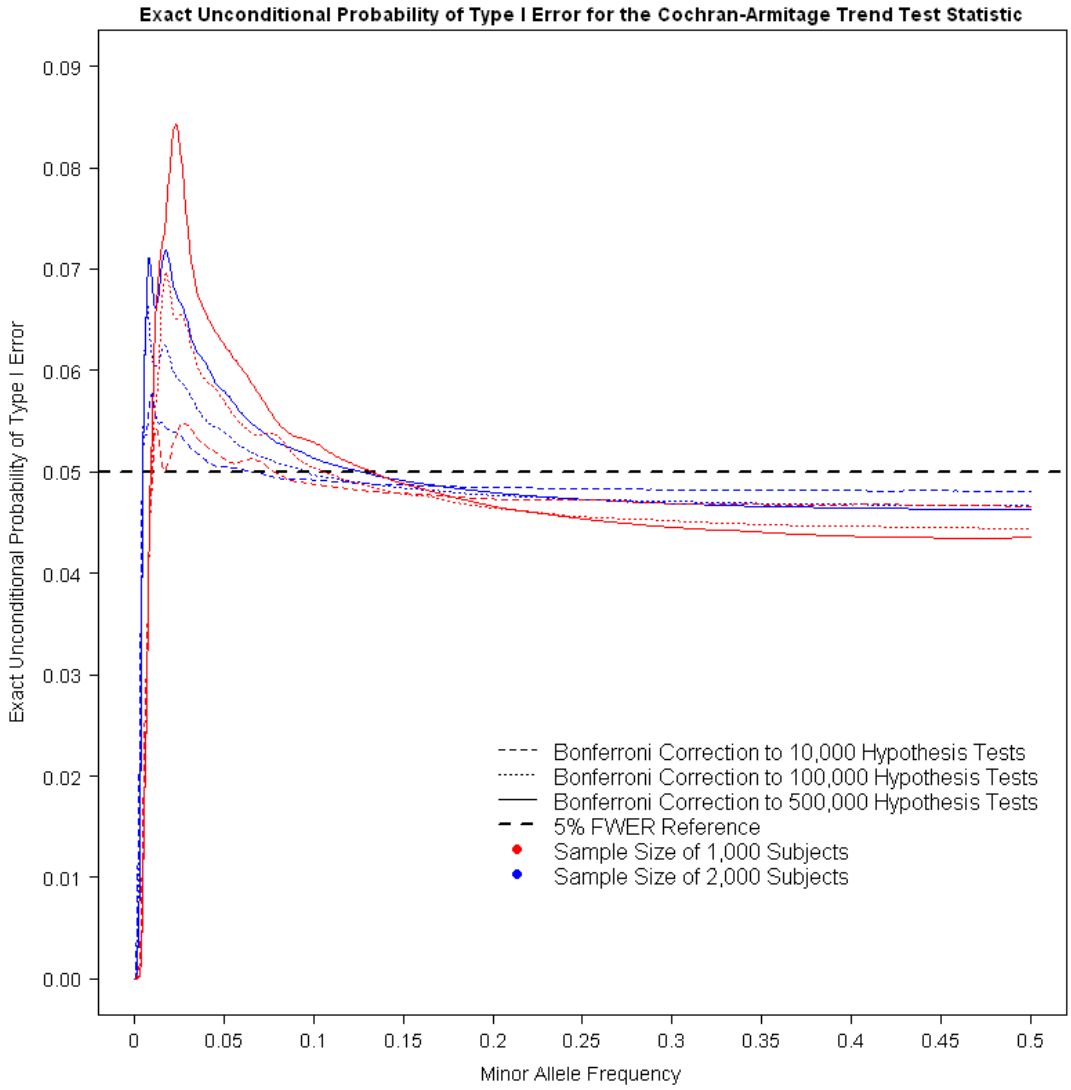

Fig. 3.2: Plot of the Bonferroni Corrected Exact Unconditional Probability of Type I Error for the Cochran-Armitage Trend Test Statistic at the Realization $F^{-1}_{\tilde{Q}_0}(1-0.05/m)$ for the χ^2_1 Distribution ($\tilde{Q}_0$), Across the Population Minor Allele Frequency for an Unbalanced GWAS Comprised of a 2 to 1 Ratio of Controls to Cases, Assuming Population Allele Frequencies Adhere to Hardy-Weinberg Equilibrium. Colored Curves: Heavy Dashed Curves, Light Dashed Curves, and Solid Curves, Assume $m = 10$K, $m = 100$K, and $m = 500$K Tested Null Hypotheses under $\mathcal{H}_0$, Respectively; Red and Blue Curves Assume GWAS Samples of $n = 1$K and $n = 2$K, Respectively. The Assumed FWER under $\tilde{Q}_0$ Is 5% (Heavy Dashed Black Line).

actual FWER for the CATT statistic under Q_0 is strictly less than the assumed 5% level, where this notion holds across π_j. Therefore, the Bonferroni MTP for the CATT statistic under $\tilde{Q}_0$ is overly conservative at controlling the FWER at the 5% level. Adding to the notion of confusion within the literature, the article of [90] – referring to control of the FWER – states, "For independent

tests, the Bonferroni (or Šidák) correction provides a simple and accurate control..."; the article of [20] states, "...Bonferroni and Šidák adjustments are valid in the case of independent tests...." However, the veracity of these statements are contingent upon the application of sound statistical inference tools (i.e., correct identification of Q_0 under $\mathcal{H}_0$), insofar as we have just illustrated that testing independent hypotheses is not sufficient for accurate control of the FWER.

3.2.3.1 Replication of Association Findings in GWAS

Given the major challenge of deciphering the few true-positive associations from the many false-positive associations within a GWAS, an important consideration lies upon replication of significant association findings using independent case-control samples [5, 115]. Criterion for replication of genotype-phenotype associations in GWAS have recently been published, and include: study of the same or similar phenotype and population; exhibition of a similar magnitude/direction of effect – within the same genetic model – upon the same SNP; and similar magnitude of significance upon the same SNP [5, 115].

There are several plausible explanations for the lack of reproducibility of genetic associations in GWAS, such as population stratification and/or genotyping errors [5, 115]. In addition, lack of accounting for gene-gene interactions within the search for susceptibility genes upon complex diseases has been widely suggested to explain difficulties in replicating significant findings in genetic association studies [116]. For example, recent human and animal studies of complex diseases have identified susceptibility genes that marginally contribute to a common trait, to a minor extent at best, but that interact significantly in combined analyses (see e.g., [117,118,119,120,121,122,123,124, 125, 126, 127, 128, 129, 130]). Several studies have found alleles that have opposite effects depending on the genetic background [131, 132] (i.e., cross-interaction), which further raises the likelihood of overlooking epistatic susceptibility genes in single-gene analyses [133]. Finally, lack of accounting for gene-environment interaction can spuriously lead to a non-significant genetic main effect (e.g., cross-interaction – see §1.2) [3, 26, 53, 134, 135, 136] and could explain lack of replication.

The above data anomalies, population characteristics, and analysis approaches set aside, one could also ascertain lack of replication of a genetic association simply due to a chance (i.e., a false-positive) finding within initial GWAS investigations. Indeed, even under strong control of the FWER by way of the Bonferroni MTP, many statistically significant associations within a GWAS have not been replicated, and are believed to be false positives [76]. This disappointing revelation has been attributed to the use of inconsistent thresholds of significance for multiple testing correction [137].

However, in light of the constituents presented within §3.2.3, here we argue by example yet another possible explanation (to this author's review, absent from the literature) for lack of replication of association findings in GWAS. Namely, we argue that lack of replication in these genetic associations could be due to improper selection of Q_0.

Consider an unbalanced GWAS of $n_1 = 400$ cases, $n_0 = 800$ controls, and $m = 500$K SNP markers. Let $p_j^{(AG)}$ (here, A is shorthand for asymptotic and G is shorthand for GWAS) and $p_j^{(UG)}$ (here, U is shorthand for unconditional) denote the respective pointwise p-values in testing $H_0^{(j)}$ with the CATT statistic, under $\tilde{Q}_0$ and $Q_{0j}^{(*H)}(\pi_j, 800, 400)$. Suppose upon a SNP locus with MAF $\pi_j = 0.017$, it is determined that $p_j^{(AG)} = 0.05/m$, so that – upon applying the Bonferroni MTP – at the 5% level in the FWER there appears to be a statistically significantly genotype-phenotype association at said locus. Using the computational tools presented within §3.5, it can be shown that $p_j^{(UG)} = 0.078/m$, so that genotype at this SNP locus is in fact marginally statistically significantly associated with phenotype at the 5% level in the FWER, after MHT correction by way of the Bonferroni MTP. Nonetheless, current GWAS practice (i.e., assuming $\tilde{Q}_0$ for the CATT statistic under $H_0^{(j)}$ and application of the Bonferroni MTP) would flag this marker to be a member of the sampled SNP panel upon a replication study. Mathematically, it is

$$\begin{aligned} & 0.05/m = p_j^{(AG)} = \Pr\left(T_j \geq F_1^{-1}\left(1 - p_j^{(AG)}\right) | \mathcal{H}_0, T_j \sim \tilde{Q}_0\right) \\ \Longrightarrow \quad & 0.078/m = p_j^{(UG)} = \Pr\left(T_j \geq F_1^{-1}\left(1 - p_j^{(AG)}\right) | \mathcal{H}_0, T_j \sim Q_{0j}^{(*H)}(\pi_j, 800, 400)\right), \end{aligned}$$

where F_1^{-1} denotes the inverse cumulative distribution function of the chi-square distribution with one degree-of-freedom.

Now consider a replication case-control study to be comprised of $n_1 = 600$ cases and $n_0 = 600$ controls. Here, in conjunction with the aforementioned replication criteria of 'similar significance' between the two studies, we assume the identical (to that of the initial GWAS) pointwise p-value in testing $H_0^{(j)}$ under $Q_{0j}^{(*H)}(\pi_j, 600, 600)$ within the replication study. Namely, we assume $p_j^{(UR)} = 0.078/m$ (here, R is shorthand for replication study), insofar as the distributions $Q_{0j}^{(*H)}(\pi_j, 800, 400)/Q_{0j}^{(*H)}(\pi_j, 600, 600)$ correctly identify Q_0 upon the initial GWAS/replication study for this SNP locus (i.e., we assume the marginally statistically significant GWAS association is replicated). Here, assuming $\tilde{Q}_0$ to be the distribution of the CATT statistic under $H_0^{(j)}$ (i.e., current GWAS practice), we find – applying the computational tools of §3.5 – $p_j^{(AR)} = 0.37/m$ for the replication study, indicating no statistically significant evidence whatsoever for a genotype-phenotype

association at the locus, after MHT correction by way of the Bonferroni MTP. Mathematically, it is

$$\begin{aligned} & 0.078/m = p_j^{(UR)} = \Pr\left(T_j \geq F_2^{-1}\left(1 - p_j^{(UR)}\right) | \mathcal{H}_0, T_j \sim Q_{0j}^{(*H)}\left(\pi_j, 600, 600\right)\right) \\ \Longrightarrow \quad & 0.37/m = p_j^{(AR)} = \Pr\left(T_j \geq F_2^{-1}\left(1 - p_j^{(UR)}\right) | \mathcal{H}_0, T_j \sim \tilde{Q}_0\right), \end{aligned}$$

where F_2^{-1} denotes the inverse cumulative distribution function of $Q_{0j}^{(*H)}\left(\pi_j, 600, 600\right)$. Thus, after MHT correction by way of the Bonferroni MTP, the assumption of $\tilde{Q}_0$ to be the true underlying distribution of the CATT statistic under $H_0^{(j)}$ upon this SNP locus, has led – at the 5% level in the FWER – to the finding of: a statistically significant genotype-phenotype association within the GWAS ($p_j^{(AG)} = 0.05/m$); no statistically significant evidence to indicate a genotype-phenotype association within the replication study ($p_j^{(AR)} = 0.37/m$). Hence, this example demonstrates that incorrect choice in Q_0 can result in failure to replicate statistically significant genetic associations in GWAS. Therefore, lack of replication in GWAS associations could be attributed to improper selection of Q_0.

3.2.3.2 Unbalanced Multiplicity Adjustment upon the MaxT MTP

Because the underlying asymptotic chi-square assumption for the CATT statistic appears violated – in such a way that the test statistics are not identically distributed under $\mathcal{H}_0$ (i.e., per §3.2.2, the distribution of this statistic seems to depend upon several parameters for the GWAS sample and its underlying population) – the maxT and minP permutation MTPs applied under $\tilde{Q}_0$ for the CATT statistic are inaccurate in their multiplicity adjustment. Effectively, non-identically distributed test statistics under $\mathcal{H}_0$ leads to the maxT multiplicity adjustment being *unbalanced* [60,62]. Due to the one-to-one mapping of the CATT statistic to its pointwise p-value under $\tilde{Q}_0$ (i.e., the chi-square distribution with one degree of freedom), this notion also applies to the minP multiplicity adjustment for said statistic under $\tilde{Q}_0$. To this author's review, no existing GWAS MTP methodological (nor, applied) article recognizes this phenomenon. There are several reasons to abstain from unbalanced multiplicity adjustment in GWAS: (1) there is no compelling reason to systematically favor some null hypotheses over others; (2) assuming HWE holds for population allele frequencies across SNP loci within the human genome, the unbalanced nature of the maxT and minP MTPs will be dependent upon the sampled SNP panel of a GWAS, specifically the distribution of the π_j among the SNP panel; and (3) because the UPTE for the CATT statistic appears to be dependent upon the propor-

tion of sampled cases within the GWAS sample, the unbalanced nature for the maxT adjustment will also depend upon the proportion of cases within the GWAS sample. Hence, all else being equal, said adjustment is likely to not be uniform across balanced and unbalanced GWAS investigations. In this circumstance, the subset of SNPs which are deemed statistically significantly associated with the phenotype upon a balanced GWAS, is likely to be different than that of an unbalanced GWAS, leading to a lack of agreement between the two studies.

Furthermore, the unbalanced adjustment of the maxT MTP under $\tilde{Q}_0$ is especially problematic for investigations extending upon GWAS (e.g., investigating SNP loci with common and rare variants within the same study), because the distortion of $\tilde{Q}_0$ from Q_0 is exacerbated for minute values in π_j – all else being equal, within a [an] balanced [unbalanced] case-control sample, the distortion could lead to an inflated Type II [Type I] error rate among SNPs possessing minute values in π_j. In brief, reliance upon $\tilde{Q}_0$ for the CATT statistic within a GWAS can lead to improper multiplicity correction.

3.2.4 A Simulation Study

By way of simulation we can empirically illustrate: the unbalanced multiplicity adjustment for the maxT MTP under $\tilde{Q}_0$; and violation in the assumption for the CATT statistic being distributed as $\tilde{Q}_0$ under $\mathcal{H}_0$. To see this, we first note: (a) assuming the CATT statistic is truly distributed by $\tilde{Q}_0$ under $\mathcal{H}_0$, it follows that the maxT and minP multiplicity corrections are equivalent [60]; (b) the pointwise p-value in testing $H_0^{(j)}$ under (a) is distributed as $U(0,1)$; and (c) if the pointwise p-values are independent and identically distributed as $U(0,1)$ under $\mathcal{H}_0$, the minP and Šidák multiplicity adjustments are equivalent [60]. Now, consider simulating mutually independent SNP loci under $\mathcal{H}_0$, uniformly across some collection of values for π_j. Under such simulation conditions, non-uniformity in observed Type I errors across the π_j for the maxT MTP is suggestive of said MTP providing unbalanced multiplicity adjustment. Furthermore, an observed discrepancy in the multiplicity correction between the maxT and the Šidák MTPs is suggestive of violation in the assumption for the CATT statistic being distributed as $\tilde{Q}_0$ under $\mathcal{H}_0$.

3.2.4.1 Methods

The illustration of these two notions, requires a sufficient number of observed Type I errors within the simulated data at each of the assumed values in π_j. In turn, this requires the simulation of a great many data sets. For example, at the FWER 5% level – where for the moment we ignore

π_j – we expect to find five random samples exhibiting at least one Type I error amongst 100 random samples. Ideally, we would like to observe hundreds of Type I errors, thereby requiring the simulation of thousands of random samples. Adding π_j into the mix, complicates matters and exacerbates the number of required simulated data sets. For example, at the 5% FWER, we expect to find five data sets exhibiting at least one Type I error for each value in π_j amongst 1K random samples, where the population minor allele frequencies for the SNP loci are assumed uniformly distributed across ten values of π_j. Indeed, we simulated $D = 100$K mutually independent case-control GWAS data sets (samples) under $\mathcal{H}_0$, each data set comprised of size $n = 1200$ and $m = 10$K mutually independent SNP loci, in two different ways: in the first (denoted simulation 1 (S1)), we simulated balanced GWAS samples, each sample comprised of 600 cases and 600 controls; and, in the second (denoted simulation 2 (S2)), we simulated unbalanced GWAS samples, each sample comprised of a 2:1 ratio of controls to cases (i.e., $n_0 = 800$ and $n_1 = 400$).[1] The chosen ratios of cases to controls upon the simulated samples, namely 1:1 and 1:2 for S1 and S2, respectively, were purposefully selected to model those portrayed within the respective Figures 3.1 and 3.2. For each data set, 1K SNP loci (of the 10K total) were simulated – independent of phenotype labeling, ensuring simulation of $\mathcal{H}_0$ – under HWE among population genotype frequencies, upon each of the ten (10) $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$. This collection of values for π_j was purposefully chosen, primarily to empirically illustrate the two notions of the preceding paragraph, but so also to be representative of a large proportion of MAFs likely encountered within a GWAS SNP sample. For example, 29.6% of the 45168 SNP markers (corresponds to 13369 markers) used in the GPER benchmark test (see §2.6) possess observed MAFs not exceeding 0.1; among the four microarray platforms, upon the four corresponding GWAS SNP samples investigated by [104], at least 12.8% of the probes upon three of the arrays corresponded to SNPs whose observed MAFs were less than 0.01. The CATT statistic was used to test each of the null hypotheses $H_0^{(j)}$, where the additive genetic model of inheritance was assumed under the two-sided alternative hypothesis. The maxT and Šidák MTPs, assuming $\tilde{Q}_0$ under $\mathcal{H}_0$, were utilized to control the FWER within each data set. For each data set, $R = 2048$ random shuffles of the phenotype labels were applied within the GPER algorithm (see §2.6) for the maxT MTP. Note that although each of the $D = 100$K simulated data sets (for each of S1 and S2) is comprised of $m = 10$K SNP loci, by randomly aggregating [without replacement] ten/fifty data sets together we obtained 10K/2K mutually independent data sets under $\mathcal{H}_0$, each data set in which was comprised of 100K/500K mutually independent SNP loci (10K/50K loci at each $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$).

[1] Each simulation entailed generating/analyzing approximately four (4) terabytes (TB) of data.

Table 3.1: Number of Data Sets Exhibiting Some Type I Error Cross-Classified by Multiple Testing Procedure (MTP), the Marker Density (m), and Assumed Minor Allele Frequency (MAF; π_j), Within a Population Whose Genotype Frequencies at Each SNP Locus Adhere to Hardy-Weinberg Equilibrium, Assuming the Cochran-Armitage Trend Test Statistic Is Distributed as $\tilde{Q}_0$ under $\mathcal{H}_0$. The True Underlying Family-wise Type I Error Rate (FWER) Is 5%. Assuming Type I Errors Are Independent of MAF, the Expected Number of Type I Errors by MAF Are 500 ($m = 10$K), Fifty ($m = 100$K), and Ten ($m = 500$K). 95% Exact Clopper-Pearson Confidence Intervals (CI) Are for Control in the Overall True Underlying FWER[†].

MTP (m)	Minor Allele Frequency (π_j)										Totals	Observed FWER (95% CI)
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1		
maxT												
(10K)	107	330	416	548	563	565	580	577	636	592	4914	4.91% (4.8%, 5.1%)
(100K)	3	28	30	52	64	49	57	56	65	65	469	4.69% (4.3%, 5.1%)
(500K)	1	5	7	8	10	9	9	10	12	12	83	4.15% (3.3%, 5.1%)
Šidák												
(10K)	61	199	271	356	376	378	381	389	431	394	3236	3.24% (3.1%, 3.3%)
(100K)	2	13	16	27	38	31	29	30	44	41	271	2.71% (2.4%, 3.1%)
(500K)	0	2	4	5	5	4	2	4	5	5	36	1.80% (1.3%, 2.5%)
maxT												
(10K)	565	535	506	506	471	486	493	492	511	497	5062	5.06% (4.9%, 5.2%)
(100K)	54	54	50	50	58	54	48	44	45	53	510	5.10% (4.7%, 5.6%)
(500K)	8	12	14	7	15	10	9	10	12	12	109	5.45% (4.5%, 6.5%)
Šidák												
(10K)	584	544	533	541	493	503	531	505	530	535	5299	5.30% (5.2%, 5.4%)
(100K)	62	62	56	61	61	63	56	50	48	63	582	5.82% (5.4%, 6.3%)
(500K)	13	16	14	10	17	11	11	12	13	14	131	6.55% (5.5%, 7.7%)

[†]Initial/final six rows correspond to simulation 1 (S1)/simulation 2 (S2).

3.2.4.2 Results

Table 3.1 displays the number of data sets exhibiting at least one observed Type I error cross-classified by MTP (maxT or Šidák), the marker density (m), and the assumed minor allele frequency within the population (π_j), at the true underlying 5% FWER, where the initial/final six rows of the table correspond to results obtained under S1/S2. As expected, in ignoring π_j these data support the notion that the maxT MTP controls the FWER at the 5% level, since the six 95% confidence intervals across marker densities and the two simulations (i.e., S1 and S2) cover said level in the FWER.[2] Moreover, these data suggest that the maxT MTP is unbalanced in its control of the FWER at the 5% level, particularly upon S1, since the number of data sets exhibiting some Type I error is not uniform across the π_j, where it is noted that this notion holds irrespective of the marker density. For example, the number of data sets exhibiting a Type I error for MAFs of 0.01, 0.02, and 0.05 upon the $D = 100$K data sets of S1 with marker density $m = 10$K are 107, 330, and 563,

[2]Here, the number of data sets exhibiting at least one Type I error is considered a binomial random variable. Accordingly, each of the 95% confidence intervals, constructed about the true underlying FWER, is a Clopper-Pearson exact confidence interval.

respectively. For this marker density, the observed number of data sets exhibiting a Type I error are 21%, 66%, and 113% relative to that expected (500) for the respective MAFs of 0.01, 0.02, and 0.05.

On the other hand, the data of S2 suggest – for the most part – that the maxT MTP is roughly balanced in its control of the FWER at the 5% level, since the number of data sets exhibiting some Type I error is approximately uniform across the π_j upon each marker density m. However, upon marker density $m = 10$K there is an apparent difference in the balancing nature of the maxT MTP over MAFs 0.01 and 0.05. Upon this marker density, the number of data sets exhibiting a Type I error are 565 and 471 for the respective MAFs 0.01 and 0.05. Equivalently, the observed Type I errors are 113% and 94% relative to that expected (500) for the respective MAFs 0.01 and 0.05. Overall, these data indicate that the balanced nature for the maxT MTP, in the control of the FWER at the 5% level, is different between the two simulation scenarios, S1 and S2.

Furthermore, the data of S1 indicate that the Šidák MTP is overly conservative in its control of the FWER at the 5% level, where the conservatism appears to be positively associated with the marker density. For example, in controlling the true underlying 5% FWER, the Šidák MTP is least conservative at marker density $m = 10$K (observed FWER 3.24%; 95% CI for the true underlying FWER (3.1%, 3.3%)), whereas this MTP reports an exceptionally conservative observed FWER of 1.80% (95% CI for the true underlying FWER (1.3%, 2.5%)) at marker density $m = 500$K. These observations entailing the Šidák MTP – namely, its conservative nature in controlling the true underlying 5% FWER – are in direct coherence with those made within the first paragraph of §3.2.3 for Figure 3.1.

On the other hand, the data of S2 suggest that the Šidák MTP is overly liberal in its control of the FWER at the 5% level, where the magnitude in the liberal nature of the control in the FWER appears to be positively associated with the marker density. For example, in control over the true 5% FWER, this MTP is least liberal at marker density $m = 10$K (observed FWER 5.30%; 95% CI for the true underlying FWER (5.2%, 5.4%)), whereas this MTP reports a very liberal observed FWER of 6.55% (95% CI for the true FWER (5.5%, 7.5%)) upon marker density $m = 500$K. Analogous to S1, these observations entailing the Šidák MTP upon S2 – namely, its liberal nature in controlling the true underlying 5% FWER – are in direct coherence with those made within the first paragraph of §3.2.3 for Figure 3.2.

Finally, these data suggest a discrepancy in the number of observed Type I errors for the maxT

and Šidák MTPs across MAF, particularly upon S1, where this notion holds across the marker density. Upon S1 for example, relative to the Šidák MTP, the observed numbers of Type I errors at the 0.05 MAF for the maxT MTP are 150% ($563/376; m = 10$K), 168% ($64/38; m = 100$K), and 200% ($10/5; m = 500$K). This discrepancy in observed Type I errors between the Šidák and maxT MTPs suggests violation in the assumption for the CATT statistic being distributed by $\tilde{Q}_0$ under $\mathcal{H}_0$.

Figures 3.3 and 3.4 display simultaneous (i.e., corrected for producing the ten independent confidence intervals (CI) across π_j by MTP) exact Clopper-Pearson 95% CIs for control in the FWER across π_j for the maxT and Šidák MTPs, among the $D = 100$K simulated data sets for the respective marker densities $m = 10$K and $m = 500$K (a total of 2K data sets thereof, obtained from randomly aggregating data sets from the $D = 100$K simulated data sets) upon simulation 1 (S1); Figures 3.5 and 3.6 display the analogous exact Clopper-Pearson 95% CIs for control in the FWER across π_j upon the maxT and Šidák MTPs upon simulation 2 (S2). All figures assume the true underlying FWER is 5%, and that the CATT statistic is distributed as $\tilde{Q}_0$ under $\mathcal{H}_0$ – Note: (1) the vertical scale is not consistent across these figures; (2) the observed FWER for the maxT and Šidák MTPs under $\tilde{Q}_0$, across π_j, are displayed by respective circles and squares; and (3) see §3.6.1 for the description/analysis entailing the minP and Šidák MTPs under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ depicted within these figures.

With regard to S1 – Collapsing over the values of π_j, Figures 3.3 and 3.4 illustrate control in the 5% FWER for the maxT MTP, since each of the 95% CIs (gray color) cover the true underlying 5% FWER. However, we do note that the observed FWER for this MTP seems a bit conservative for the marker density $m = 500$K (4.15%). These figures also illustrate the unbalanced control in the FWER at the 5% level for the maxT MTP, since the observed FWER tends to deviate from the 5% expected level across π_j. In particular, these data indicate that the maxT MTP tends to control the 5% FWER at a level lower/higher than that expected for small (0.01-0.04)/large (0.05-0.1) values of π_j. Furthermore, these figures suggest control of the FWER between the maxT and Šidák MTPs to be remarkably different, as seen by the differences in their respective observed FWER across π_j – this notion is particularly true of marker density $m = 10$K, and so also of marker density $m = 500$K for larger values of MAF. In turn, this suggests that $\tilde{Q}_0$ is not the true underlying null distribution for the CATT statistic under $\mathcal{H}_0$. Finally, these figures illustrate the exceptionally conservative control in the 5% FWER for the CATT statistic under $\tilde{Q}_0$ within the Šidák MTP, since the observed FWER

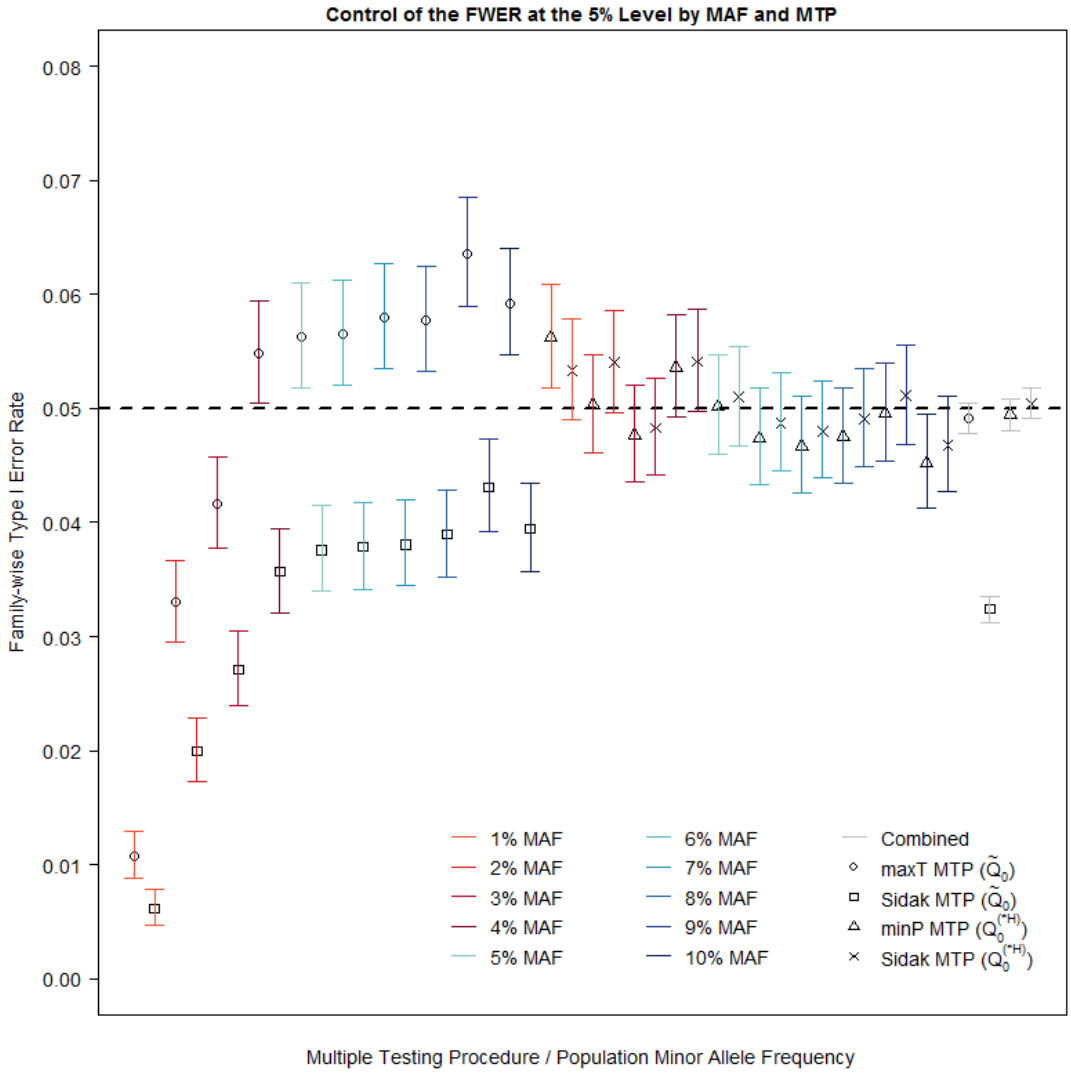

Fig. 3.3: Simultaneous Exact Clopper-Pearson 95% Confidence Intervals (CI) for Control in the Family-wise Type I Error Rate (FWER) for the Cochran-Armitage Trend Test Statistic under $\mathcal{H}_0$, Across Minor Allele Frequencies (MAFs), $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$, Within a Population Whose Genotype Frequencies Adhere to Hardy-Weinberg Equilibrium at Each SNP Locus, Applying Several Multiple Testing Procedures (MTP), Where the True Underlying FWER Is 5% (Heavy Dashed Black Line). This Figure Summarizes the Simulation of $D = 100$K Mutually Independent Data Sets, Each Data Set Comprised of $m = 10$K Mutually Independent SNP Loci Simulated under $\mathcal{H}_0$ and 1K Loci Simulated for Each $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$, upon a Balanced GWAS of Size $n = 1200$. The Symbols (Circle and Square) Depict the Observed Number of Type I Errors for the Respective MaxT and Šidák MTPs under $\tilde{Q}_0$; the Symbols (Triangle and Cross) Depict the Observed Number of Type I Errors for the Respective MinP and Šidák MTPs under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$. The Gray CIs Collapse over All MAFs.

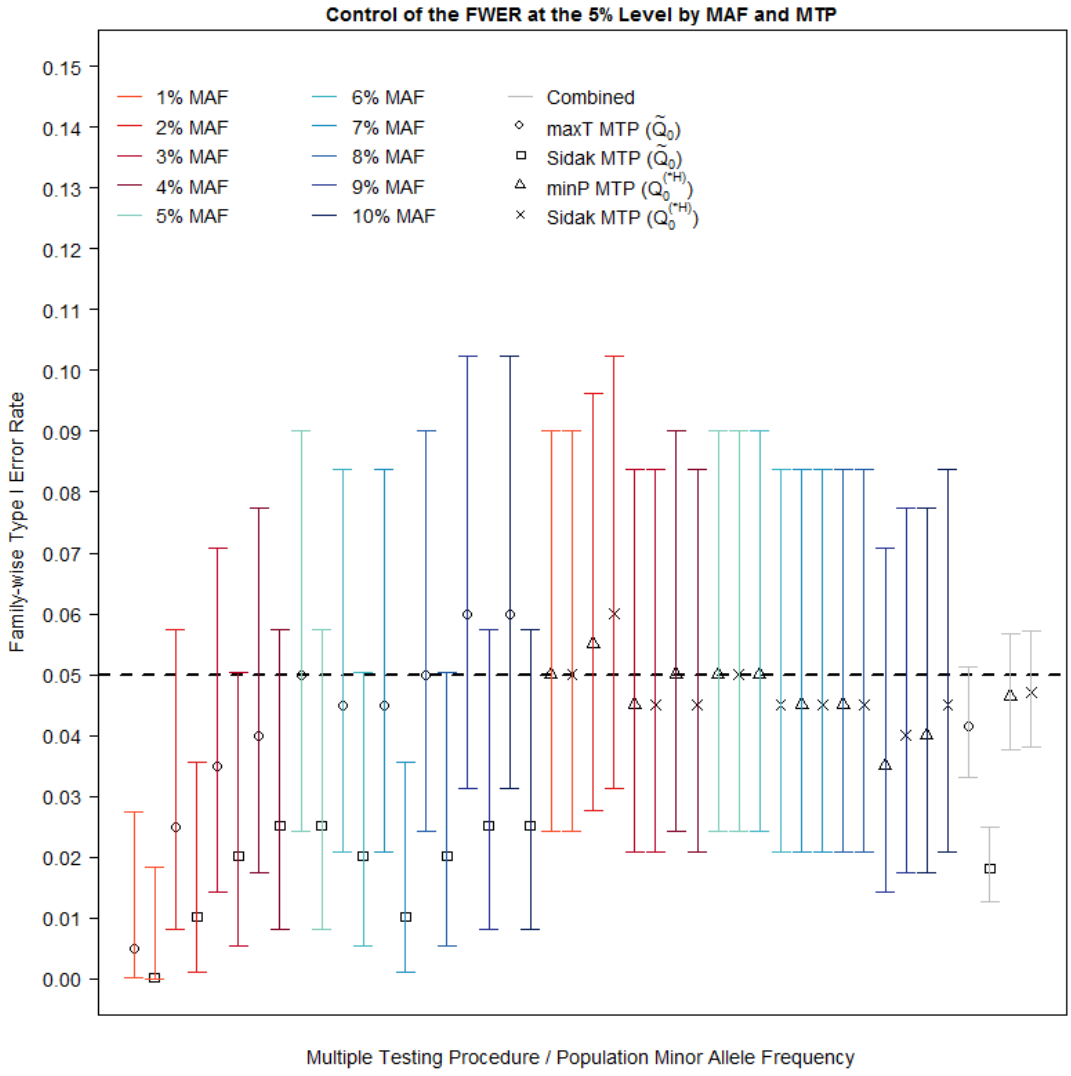

Fig. 3.4: Simultaneous Exact Clopper-Pearson 95% Confidence Intervals (CI) for Control in the Family-wise Type I Error Rate (FWER) for the Cochran-Armitage Trend Test Statistic under $\mathcal{H}_0$, Across Minor Allele Frequencies (MAFs), $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$, Within a Population Whose Genotype Frequencies Adhere to Hardy-Weinberg Equilibrium at Each SNP Locus, Applying Several Multiple Testing Procedures (MTP), Where the True Underlying FWER Is 5% (Heavy Dashed Black Line). This Figure Summarizes the Simulation of 2K Mutually Independent Data Sets, Each Data Set Comprised of $m = 500$K Mutually Independent SNP Loci Simulated under $\mathcal{H}_0$ and 50K Loci Simulated for Each $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$, upon a Balanced GWAS of Size $n = 1200$. The Symbols (Circle and Square) Depict the Observed Number of Type I Errors for the Respective MaxT and Šidák MTPs under $\tilde{Q}_0$; the Symbols (Triangle and Cross) Depict the Observed Number of Type I Errors for the Respective MinP and Šidák MTPs under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$. The Gray CIs Collapse over All MAFs.

is significantly lower than expected (5%) across the two marker densities.

With regard to S2 – Collapsing over the values of π_j, Figures 3.5 and 3.6 illustrate control in the 5% FWER for the maxT MTP, since each of the 95% CIs (gray color) cover the true underlying 5% FWER. Moreover, for marker density $m = 10$K, these data suggest that the maxT MTP is roughly balanced in its control of the 5% FWER across MAF – the exception being at MAF equal to 0.01, where this MTP is suggestive of being slightly liberal in its control of the FWER at the 5% level. The latter of the two figures suggests control of the FWER between the maxT and Šidák MTPs to be remarkably different, as seen by the differences in their respective observed FWER across π_j. This suggests that $\tilde{Q}_0$ is not the true underlying null distribution for the CATT statistic under $\mathcal{H}_0$. Finally, these figures illustrate the liberal control in the 5% FWER for the CATT statistic under $\tilde{Q}_0$ within the Šidák MTP, since the observed FWER is somewhat higher than expected (5%) across the two marker densities.

3.2.4.3 Conclusions

Overall, these simulated data help portray the key notions discussed within §3.2.2-3.2.3. First, based upon application of the Šidák MTP assuming $\tilde{Q}_0$ under $\mathcal{H}_0$, these data suggest that control of the FWER at the 5% level is dependent upon the following characteristics of the case-control sample: the distribution of the minor allele frequency within the sample. This is visually evident within Figures 3.4 and 3.6 upon marker density $m = 500$K, since the observed Type I error rate for this MTP appears to differ across the values in MAF; the marker density m. Here, this is evident empirically within Table 3.1, since there appears to be a positive association between the marker density and the magnitude in which this MTP fails adherence with control in the overall (i.e., ignoring MAF) FWER at the 5% level; and the ratio of controls to cases within the GWAS sample. This can be seen empirically by way of comparing the appropriate rows within Table 3.1 for a given marker density, or visually by way of comparing: Figures 3.3 and 3.5; Figures 3.4 and 3.6. For example, taking MAF equal to 0.01 and $m = 100$K, these data suggest that the observed number of data sets exhibiting some Type I error upon an unbalanced (2:1 ratio of controls to cases) GWAS of size $n = 1200$ is 31 times (62/2) that of the corresponding n-size balanced GWAS.

Second, because the overall (i.e., collapsing over π_j) number of data sets exhibiting some Type I error seems to differ between the Šidák and maxT MTPs, irrespective of marker density and balancing nature of the GWAS sample, these data indicate that $\tilde{Q}_0$ is not the correct distribution for the CATT statistic under $\mathcal{H}_0$. This can have serious negative ramifications in the reporting of

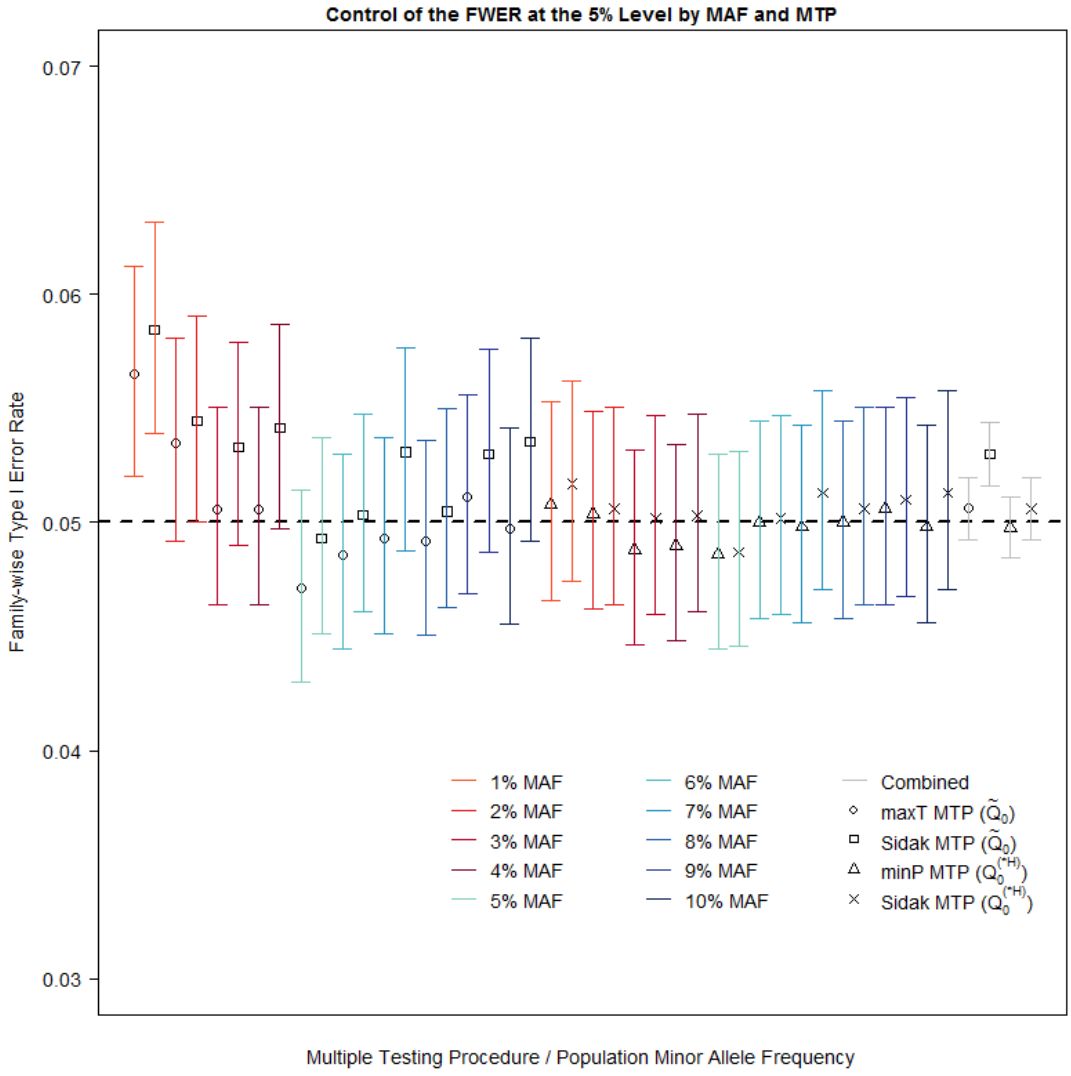

Fig. 3.5: Simultaneous Exact Clopper-Pearson 95% Confidence Intervals (CI) for Control in the Family-wise Type I Error Rate (FWER) for the Cochran-Armitage Trend Test Statistic under $\mathcal{H}_0$, Across Minor Allele Frequencies (MAFs), $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$, Within a Population Whose Genotype Frequencies Adhere to Hardy-Weinberg Equilibrium at Each SNP Locus, Applying Several Multiple Testing Procedures (MTP), Where the True Underlying FWER Is 5% (Heavy Dashed Black Line). This Figure Summarizes the Simulation of $D = 100$K Mutually Independent Data Sets, Each Data Set Comprised of $m = 10$K Mutually Independent SNP Loci Simulated under $\mathcal{H}_0$ and 1K Loci Simulated for Each $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$, upon an Unbalanced GWAS of Size $n = 1200$. The Symbols (Circle and Square) Depict the Observed Number of Type I Errors for the Respective MaxT and Šidák MTPs under $\tilde{Q}_0$; the Symbols (Triangle and Cross) Depict the Observed Number of Type I Errors for the Respective MinP and Šidák MTPs under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$. The Gray CIs Collapse over All MAFs.

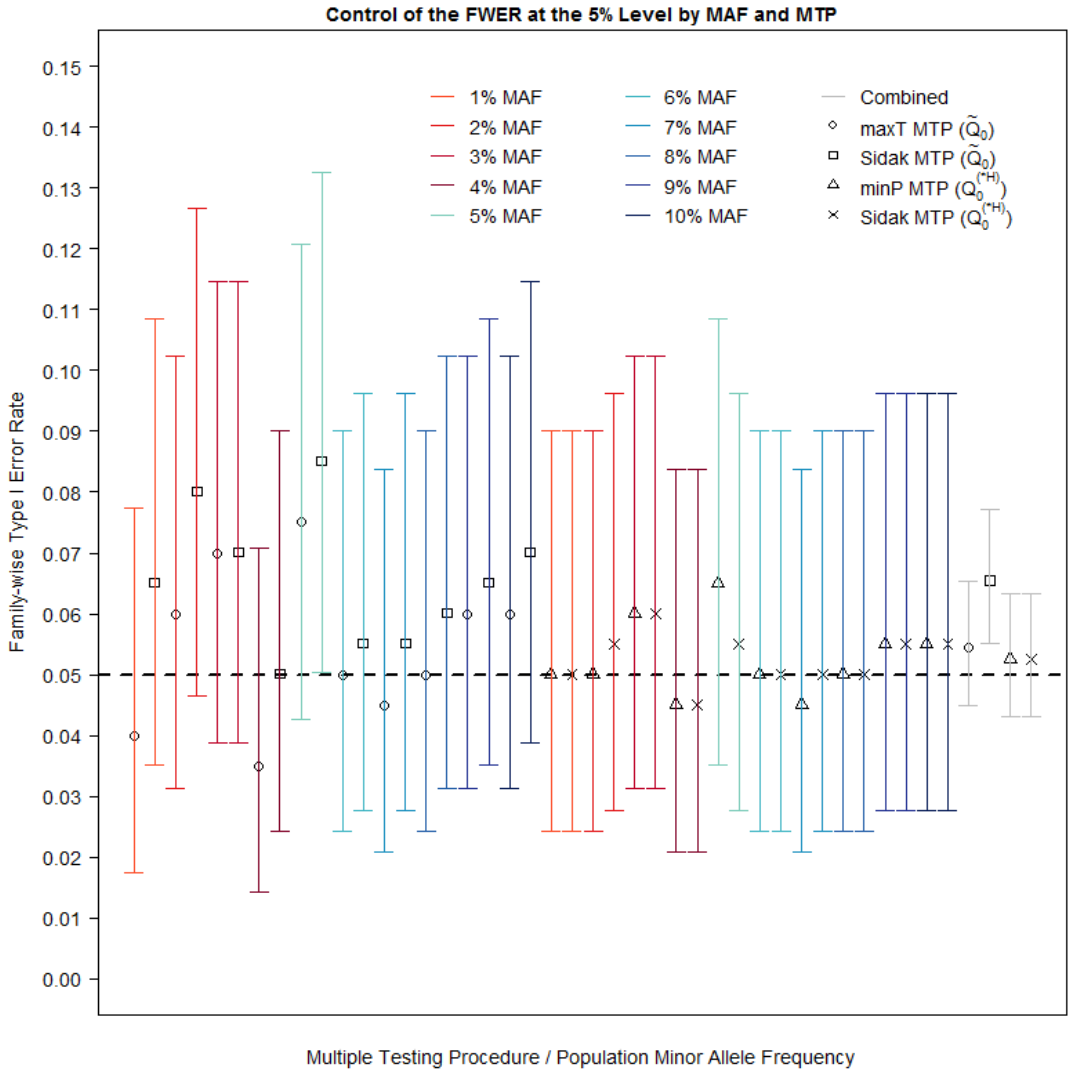

Fig. 3.6: Simultaneous Exact Clopper-Pearson 95% Confidence Intervals (CI) for Control in the Family-wise Type I Error Rate (FWER) for the Cochran-Armitage Trend Test Statistic under $\mathcal{H}_0$, Across Minor Allele Frequencies (MAFs), $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$, Within a Population Whose Genotype Frequencies Adhere to Hardy-Weinberg Equilibrium at Each SNP Locus, Applying Several Multiple Testing Procedures (MTP), Where the True Underlying FWER Is 5% (Heavy Dashed Black Line). This Figure Summarizes the Simulation of 2K Mutually Independent Data Sets, Each Data Set Comprised of $m = 500$K Mutually Independent SNP Loci Simulated under $\mathcal{H}_0$ and 50K Loci Simulated for Each $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$, upon an Unbalanced GWAS of Size $n = 1200$. The Symbols (Circle and Square) Depict the Observed Number of Type I Errors for the Respective MaxT and Šidák MTPs under $\tilde{Q}_0$; the Symbols (Triangle and Cross) Depict the Observed Number of Type I Errors for the Respective MinP and Šidák MTPs under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$. The Gray CIs Collapse over All MAFs.

Type I errors within a GWAS. Taking $m = 500$K for example, these data indicate that the Šidák MTP fails to guard against excessive reporting of Type I errors for an unbalanced GWAS (2:1 controls to cases), upon SNP loci possessing a rare (say, less than 0.05) MAF. In this circumstance, one is likely to be overly reporting false positives. This can lead to increased cost within a replication study, because one is compelled to ascertain genotype data upon SNP marker(s) which should not be included as part of the sampled SNP panel for said study. On the other hand, these data indicate that said MTP is remarkably conservative in its reporting of Type I errors for a balanced GWAS, upon SNP loci possessing a rare MAF. In this circumstance, one is likely to be understating both false positives and true associations. This is problematic, insofar as the Šidák MTP is likely to fail to detect some true genotype-phenotype associations, the associations in which may play an important role within the causative pathway of the disease under study. In short, for fixed values in n and m, based upon these data the Šidák – since m is large, so also the Bonferroni – MTP is suggestive of being inconsistent in its reporting of Type I errors upon SNP loci possessing a rare MAF and this is unacceptable in practice.

Moreover, not only do these data suggest that the maxT MTP is unbalanced in its control of the FWER at the 5% level, particularly upon S1, but also that the unbalanced nature in the control of the FWER is inconsistent between: balanced and unbalanced GWAS samples; and marker densities upon either of S1 or S2. The former of this notions is problematic, because – all else being equal – the reporting of false positives could be different between n-sized unbalanced (2:1 ratio of controls to cases) and balanced GWAS investigations. Taking $m = 500$K for example, these data indicate that the number of Type I errors reported within an unbalanced GWAS upon SNP loci with MAF equal to 0.01, is eight (8/1) times that of a balanced GWAS. The latter of this notions is problematic, because the nature of the reporting of Type I errors for an MTP should not be dependent upon the marker density. Taking MAF of 0.01 for the unbalanced GWAS for example, upon marker density $m = 10$K (Figure 3.5) the maxT MTP is suggestive to be slightly liberal in its reporting of false positive associations, whereas upon marker density $m = 500$K (Figure 3.6) this MTP is suggestive to be quite conservative in its reporting of Type I errors.

3.3 Towards a Resolution: Robustness of the Hardy-Weinberg Equilibrium Assumption

Control of the FWER is of utmost importance within a GWAS, as this ensures the proper reporting of false-positive results. The assumption of $\tilde{Q}_0$ under $\mathcal{H}_0$ for the CATT statistic within GWAS

is not realistic and can lead to improper multiplicity correction. As argued within §3.2.3, this can have serious negative ramifications, including lack of replication of significant genotype-phenotype associations across GWAS samples investigating a common phenotypic trait. On the other hand, under HWE among genotype frequencies within the population at SNP locus j, $Q_{0j}^{(*H)}(\pi_j, n_0, n_1)$ correctly identifies the distribution of the CATT statistic at the locus under $\mathcal{H}_0$. While correct identification of Q_0 under $\mathcal{H}_0$ is a major step towards resolving the GWAS MHT problem, in order to fully resolve the MHT problem we need to implement Q_0 within a MTP.

Indeed, per the above arguments, one might propose multiplicity correction for the CATT statistic under $\mathcal{H}_0$ by exploitation of $\left\{Q_{0j}^{(*H)}\right\}_{j=1,\ldots,m}$. However, there is a rather substantial problem with this approach. Namely, the integrity of pointwise p-values derived under $Q_{0j}^{(*H)}$ is reliant upon the assumption that population genotype frequencies at SNP locus j adhere to HWE. Pursuant to the arguments presented within §3.2.1, while it is true that HWE should in general hold within the population at an arbitrarily sampled SNP locus within the human genome, this notion may not hold true at every SNP locus throughout the human genome. Given π_j, we are interested to know the extent of the robustness in the distribution $Q_{0j}^{(*H)}$ to deviations in the HWE assumption at the locus. That is, we would like to know if deviations in genotype frequencies from HWE at a SNP locus, could lead to different conclusions – relative to genotypes frequencies adhering to HWE – regarding the UPTE under $Q_{0j}^{(*H)}$ for the CATT statistic.

Consider locus j, with respective major (the more frequently occurring allele at the locus within the population) and minor alleles A and a. Let π_j^{aa}, π_j^{Aa}, and π_j^{AA}, denote the respective population frequencies for genotypes aa, Aa, and AA. Under HWE at the locus, it holds that $\pi_j^{aa} = \pi_j^2$, $\pi_j^{Aa} = 2\pi_j(1-\pi_j)$, and $\pi_j^{AA} = (1-\pi_j)^2$. Whenever the assumption of HWE fails at the locus, there are numerous ways in which π_j^{aa}, π_j^{Aa}, and π_j^{AA} can be parameterized. The articles of [113,138,139], for example, each discuss a [common] generalization to the HWE model, of which allows for the over- or under-representation of heterozygotes (genotype Aa; with respect to that under HWE) at SNP locus j by way of the *inbreeding coefficient* (f_j). In terms of the coefficient f_j, these articles express the population frequencies for genotypes aa, Aa, and AA, by $\pi_j^{aa} = \pi_j^2 + \pi_j(1-\pi_j)f_j$, $\pi_j^{Aa} = 2\pi_j(1-\pi_j)(1-f_j)$, and $\pi_j^{AA} = (1-\pi_j)^2 + \pi_j(1-\pi_j)f_j$, respectively, so that $f_j = 0$ recovers the HWE model. By inspection of the formulation for π_j^{Aa}, we see that the difference $(1-f_j)$ is interpreted as the proportion of over- or under-represented heterozygotes (when compared to HWE) at locus j. It can be shown (see Proposition A.5 of Appendix A) that the range of

f_j is $\left[\pi_j\left(\pi_j-1\right)^{-1},1\right]$, which depends on π_j. In a GWAS, $f_j>0$ (i.e., underrepresentation of heterozygotes at the locus, when compared to HWE) could indicate population stratification or inbreeding, while $f_j<0$ (i.e., overrepresentation of heterozygotes at the locus, when compared to HWE) may indicate problems in genotyping [113, 139].

Figure 3.7 displays a contour plot of the Bonferroni corrected UPTE for the CATT statistic under $Q_{0j}^{(*H)}$ (the dimension represented in color within the plot), across the domain in the population minor allele frequency (π_j) and values of the inbreeding coefficient (f_j) within the range $\left(\pi_j\left(\pi_j-1\right)^{-1},0.5\right)$, for a balanced GWAS of $n=1$K subjects and $m=500$K SNP markers, at the realization of the CATT statistic $F_{\tilde{Q}_0}^{-1}\left(1-0.05/m\right)=28.4$ – the minimum value of the CATT statistic under $\tilde{Q}_0$, the chi-square distribution with degrees-of-freedom equal to one, for which the Bonferroni MTP rejects $H_0^{(j)}$ at the 5% FWER level – where the additive genetic model of inheritance is assumed under the two-sided alternative hypothesis $H_a^{(j)}$. This figure is a generalization to that of Figure 3.1, where we note that the solid red curve depicted within the latter figure is shown by the colored contours across π_j upon the dashed black line within the former figure. If $Q_{0j}^{(*H)}$ is truly robust to deviations in the assumption of HWE among the genotype frequencies at SNP locus j, the colored contoured regions within this plot would move in a strictly vertical manner. However, the plot indicates that this may not be the case, particularly for π_j taking values less than 0.2. It appears that the UPTE for the CATT statistic under $Q_{0j}^{(*H)}$ to be under- and over-stated for respective values of $f_j>0$ and $f_j<0$, particularly for π_j assuming values less than 0.2. That is, the utility of $Q_{0j}^{(*H)}$ appears to be conservative/liberal in its control of the FWER at the 5% level, whenever f_j assumes values greater/less than zero (i.e., deviations from the HWE assumption). Moreover, because the width in the colored contour regions appears to be shrinking for decreasing values of π_j (specifically, for $\pi_j\in(0.01,0.20)$), for a given non-zero value in the inbreeding coefficient under Hardy-Weinberg disequilibrium (HWD) we would expect the inaccuracy of the reported UPTE for the CATT statistic under $Q_{0j}^{(*H)}$ to increase as π_j decreases.

Indeed, Table 3.2 summarizes the Bonferroni corrected UPTE for the CATT statistic at realization $F_{\tilde{Q}_0}^{-1}\left(1-0.05/m\right)$ ($\tilde{Q}_0$ is the chi-square distribution with one degree-of-freedom) and 5% FWER, for a balanced GWAS sample of $n=1$K, across several values for each of the population inbreeding coefficient (f_j) and population minor allele frequency (π_j). For notational clarity, let RUPTE denote the observed UPTE for the CATT statistic under HWD relative to that under HWE, for fixed values in π_j, m, and $f_j\neq 0$, where $\mathcal{H}_0$ is assumed true. For fixed values in π_j and

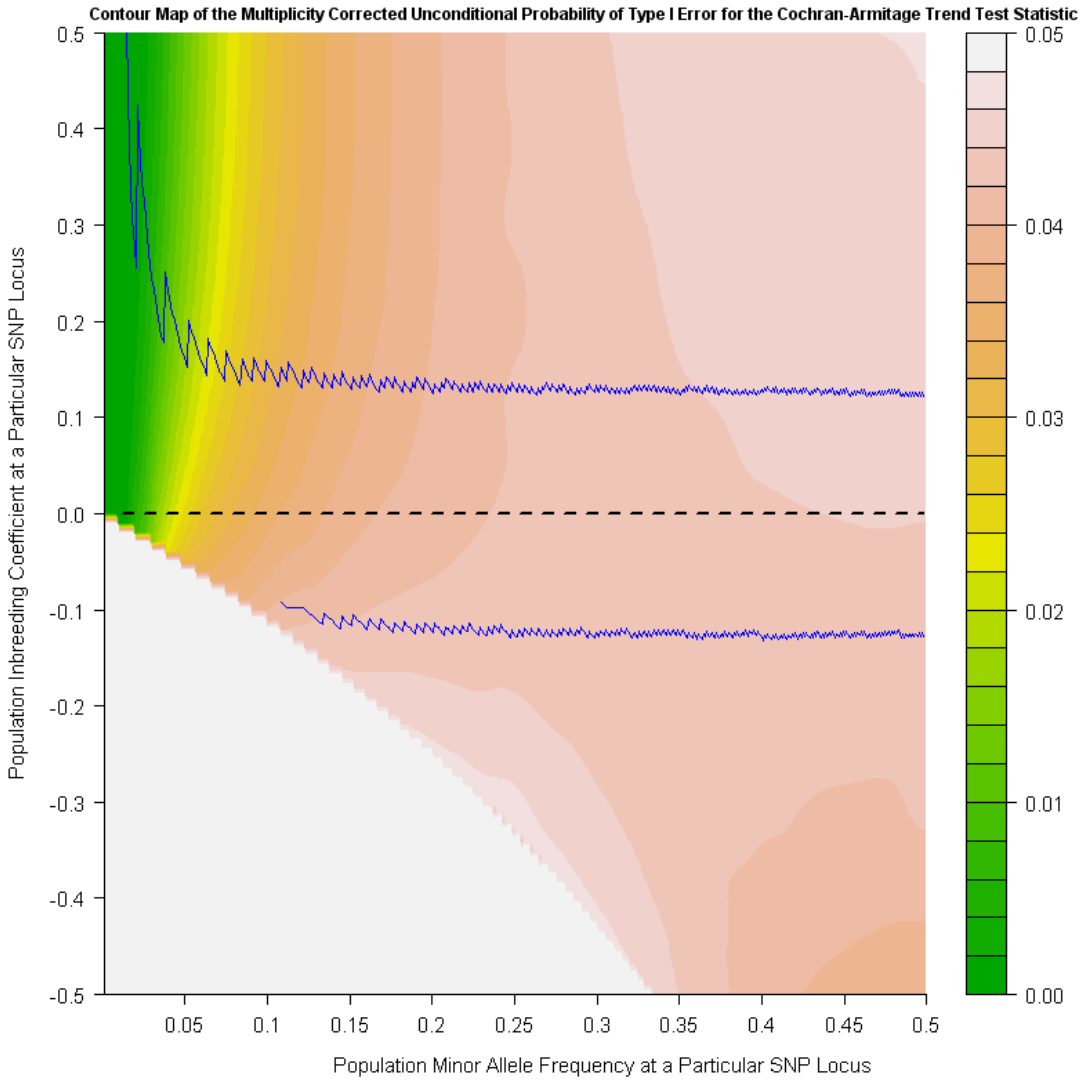

Fig. 3.7: Contour Plot of the Bonferroni Corrected Unconditional Probability of Type I Error for the Cochran-Armitage Trend Test Statistic under $\mathcal{H}_0$ at the Realization $F_{\tilde{Q}_0}^{-1}(1-0.05/m)$ for the χ_1^2 Distribution ($\tilde{Q}_0$) Across the Domain for the Population Minor Allele Frequency ($\pi_j \in (0, 0.5)$) and the Population Inbreeding Coefficient (f_j) Within the Range $\left(\pi_j(\pi_j-1)^{-1}, 0.5\right)$, under a Generalized Model to HWE for Genotype Frequencies for SNP Loci, Against a Balanced GWAS of $n = 1$K and $m = 500$K SNP Loci. The Assumed FWER under $\tilde{Q}_0$ Is 5%. The Heavy Dashed Black Line Indicates HWE; and the Region Bounded Between the Two Blue Curves Indicates the Values of f_j for Which the Exact Test of the Null Hypothesis of HWE among Sampled Controls Possesses Less Than 80% Power to Detect Hardy-Weinberg Disequilibrium at the 5% Pointwise Significance Level.

m, these results indicate a negative association between the RUPTE and f_j. Taking $\pi_j = 0.1$ and $m = 10$K for example, the RUPTE is 0.91 and 1.08 at $f_j = 0.3$ and $f_j = -0.1$, respectively; for fixed values in f_j and m, such that $f_j \neq 0$, the results indicate a positive association between the RUPTE

and π_j. Taking $f_j = 0.3$ and $m = 10$K for example, the RUPTE is 0.98 and 0.91 at $\pi_j = 0.2$ and $\pi_j = 0.1$, respectively; and for fixed values in f_j and π_j, such that $f_j \neq 0$, the results indicate a negative association between the RUPTE and m. Taking $f_j = 0.3$ and $\pi_j = 0.1$ for example, the RUPTE is 0.91 and 0.83 at $m = 10$K and $m = 500$K, respectively. Overall, the results suggest that the RUPTE is a decreasing function for decreasing π_j, and that for any fixed π_j the RUPTE is a decreasing function for increasing m and/or f_j. In other words, taking $\pi_j \in (0, 0.2)$ and $f_j > 0$, these data suggest that $Q_{0j}^{(*H)}$ is conservative in its control of the FWER for the CATT statistic under $H_0^{(j)}$, particularly for $\pi_j \in (0, 0.02)$; taking $\pi_j \in (0, 0.2)$ and $f_j < 0$, these data suggest that $Q_{0j}^{(*H)}$ is liberal in its control of the FWER for the CATT statistic under $H_0^{(j)}$. Note that these observations encompassing the empirical data for this table, are in direct agreement with our visual observations for Figure 3.7.

Table 3.2: Bonferroni Corrected Unconditional Probability of Type I Error for the Cochran-Armitage Trend Test Statistic at the Realization $F_{\tilde{Q}_0}^{-1}(1 - 0.05/m)$ for the χ_1^2 Distribution ($\tilde{Q}_0$) – Where the Two-sided Alternative Hypothesis under the Additive Genetic Model of Inheritance Is Assumed – Assuming a Balanced GWAS of $n = 1$K Subjects and m SNP Markers, Across Several Values for Each of the Population Inbreeding Coefficient (f_j) and Population Minor Allele Frequency (π_j). The Assumed FWER under $\tilde{Q}_0$ Is 5%.

		Number of SNP Markers			Power to Reject the Null
f_j	π_j	10K	100K	500K	of HWE Among Controls[†]
	0.20	0.045 (0.98)[‡]	0.041 (0.97)	0.040 (0.97)	0.99 (0.99)
	0.10	0.038 (0.91)	0.033 (0.86)	0.029 (0.83)	0.98 (0.96)
0.3	0.05	0.027 (0.76)	0.019 (0.65)	0.014 (0.58)	0.92 (0.81)
	0.02	0.007 (0.41)	0.002 (0.18)	< 0.001 (0.13)	0.67 (0.36)
	0.01	< 0.001 (0.09)	< 0.001 (0.01)	< 0.001 (0.01)	0.47 (0.12)
	0.20	0.045 (0.99)	0.042 (0.99)	0.040 (0.98)	0.12 (0.05)
	0.10	0.040 (0.96)	0.035 (0.93)	0.032 (0.91)	0.11 (0.06)
0.1	0.05	0.031 (0.88)	0.024 (0.83)	0.020 (0.80)	0.12 (0.04)
	0.02	0.012 (0.73)	0.005 (0.50)	0.003 (0.47)	0.08 (0.01)
	0.01	0.001 (0.46)	< 0.001 (0.27)	< 0.001 (0.21)	0.07 (0.01)
	0.20	0.045 (1.00)	0.043 (1.00)	0.041 (1.00)	7×10^{-4} (7×10^{-5})
	0.10	0.042 (1.00)	0.038 (1.00)	0.035 (1.00)	4×10^{-4} (9×10^{-5})
0.0	0.05	0.035 (1.00)	0.029 (1.00)	0.025 (1.00)	7×10^{-4} (7×10^{-5})
	0.02	0.017 (1.00)	0.01 (1.00)	0.006 (1.00)	5×10^{-4} (1×10^{-5})
	0.01	0.002 (1.00)	< 0.001 (1.00)	< 0.001 (1.00)	6×10^{-4} (3×10^{-6})
-0.1	0.20	0.046 (1.02)	0.044 (1.02)	0.043 (1.04)	0.12 (0.03)
-0.1	0.10	0.045 (1.08)	0.041 (1.09)	0.040 (1.13)	< 0.001 (< 0.001)

[†] Against the two-sided alternative of HWD for the exact test at pointwise significance levels of 10^{-3} (10^{-4}).

[‡] Unconditional probability of Type I error; parenthetic values are unconditional probabilities of Type I error relative to that under HWE for the given value of π_j.

Table 3.3 summarizes the equivalent information as that of Table 3.2, but for a balanced GWAS of $n = 2$K, as opposed to $n = 1$K for the latter table. The observations regarding the RUPTE for the

latter table also adhere to the former table. When comparing the RUPTE values across the tables for fixed values in π_j, $f_j \neq 0$, and m, we see that deviations in said values from the HWE index of 1.00 are not as extreme for the larger GWAS sample size. Hence, this suggests the magnitude in the conservative/liberal nature of $Q_{0j}^{(*H)}$ – for values of f_j greater/less than zero – in its control of the FWER for the CATT statistic under $H_0^{(j)}$, is decreasing for increasing n. That is, for a balanced GWAS and our adopted model allowing for genotype frequencies to deviate from HWE at SNP locus j under $\mathcal{H}_0$, these results suggest $Q_{0j}^{(*H)}$, in its control of the 5% level of the FWER for the CATT statistic, is asymptotically (in n) robust to HWD.

Table 3.3: Bonferroni Corrected Unconditional Probability of Type I Error for the Cochran-Armitage Trend Test Statistic at the Realization $F_{\tilde{Q}_0}^{-1}(1-0.05/m)$ for the χ_1^2 Distribution ($\tilde{Q}_0$) – Where the Two-sided Alternative Hypothesis under the Additive Genetic Model of Inheritance Is Assumed – Assuming a Balanced GWAS of $n = 2$K Subjects and m SNP Markers, Across Several Values for Each of the Population Inbreeding Coefficient (f_j) and Population Minor Allele Frequency (π_j). The Assumed FWER under $\tilde{Q}_0$ Is 5%.

		Number of SNP Markers			Power to Reject the Null
f_j	π_j	10K	100K	500K	of HWE Among Controls[†]
0.3	0.20	0.047 (0.99)[‡]	0.041 (0.99)	0.040 (0.98)	0.99 (0.99)
	0.10	0.044 (0.96)	0.033 (0.94)	0.029 (0.92)	0.99 (0.99)
	0.05	0.037 (0.89)	0.019 (0.83)	0.014 (0.79)	0.99 (0.99)
	0.02	0.022 (0.70)	0.002 (0.54)	< 0.001 (0.44)	0.94 (0.83)
	0.01	0.006 (0.30)	< 0.001 (0.23)	< 0.001 (0.09)	0.65 (0.65)
0.1	0.20	0.047 (0.99)	0.042 (0.99)	0.040 (0.92)	0.38 (0.20)
	0.10	0.045 (0.98)	0.035 (0.97)	0.032 (0.96)	0.32 (0.16)
	0.05	0.040 (0.95)	0.024 (0.92)	0.020 (0.89)	0.28 (0.16)
	0.02	0.027 (0.86)	0.005 (0.77)	0.003 (0.72)	0.18 (0.06)
	0.01	0.014 (0.65)	< 0.001 (0.62)	< 0.001 (0.45)	0.07 (0.07)
0.0	0.20	0.047 (1.00)	0.043 (1.00)	0.041 (1.00)	7×10^{-4} (7×10^{-5})
	0.10	0.046 (1.00)	0.038 (1.00)	0.035 (1.00)	6×10^{-4} (5×10^{-5})
	0.05	0.042 (1.00)	0.029 (1.00)	0.025 (1.00)	5×10^{-4} (8×10^{-5})
	0.02	0.031 (1.00)	0.01 (1.00)	0.006 (1.00)	4×10^{-4} (2×10^{-5})
	0.01	0.021 (1.00)	< 0.001 (1.00)	< 0.001 (1.00)	7×10^{-5} (7×10^{-5})
-0.1	0.20	0.048 (1.01)	0.044 (1.01)	0.043 (1.02)	0.47 (0.21)
-0.1	0.10	0.047 (1.03)	0.041 (1.05)	0.040 (1.07)	0.74 (0.37)

[†]Against the two-sided alternative of HWD for the exact test at pointwise significance levels of 10^{-3} (10^{-4}).

[‡]Unconditional probability of Type I error; parenthetic values are unconditional probabilities of Type I error relative to that under HWE for the given π_j.

Although the utility of $Q_{0j}^{(*H)}$ – in computing pointwise p-values for the CATT statistic under $H_0^{(j)}$ – does appear to be fairly robust to HWD for a balanced GWAS, particularly for large values of π_j, it seems to be overly conservative in its control of the FWER for the CATT statistic under $H_0^{(j)}$ for minute values (say, not greater than 0.02) of π_j. Furthermore, based upon the empirical evidence presented within Tables 3.2 and 3.3 (by way of the RUPTE), this notion seems to hold

even in the circumstance for which the values of n and f_j are large and small, respectively. Indeed, if one could filter out those SNP loci from the GWAS sample, whose genotype frequencies within the population truly deviate from HWE (particularly those loci which possess minute values of π_j), then exploitation of $\left\{Q_{0j}^{(*H)}\right\}_{j=1,\ldots,m}$ would be a viable approach for multiplicity correction entailing the CATT statistic.

A possible approach to carrying out this SNP filtering process, is to conduct an exact test of the null hypothesis that population genotype frequencies at locus j adhere to HWE (against the two-sided alternative[3]), among the controls within the GWAS sample, for all $j = 1, \ldots, m$. In fact, as mentioned within §3.2.1, this hypothesis testing regimen is recommended as a quality control measure amongst the genotype data of the controls for a GWAS sample [104]. However, applied against a GWAS SNP sample, this test has two problems. First, as demonstrated within each of the articles of [113, 138], the distributional properties for the exact-based test statistic used for testing the null hypothesis of HWE on a per-marker basis, is conservative in its control of the Type I error rate – at pointwise significance levels of 10^{-3} and 10^{-4} this conservatism can also be seen by way of the data presented within the final column for each of the Tables 3.2 and 3.3, taking $f_j = 0$ therein. Although the exact test [by design] is guaranteed to control the Type I error rate at a given significance threshold, due to the discreteness within the genotype (i.e., categorical) data these two articles both demonstrate that the actual assumed Type I error in this test is dependent upon the population MAF at the SNP locus, with an apparent decreasing trend in assumed Type I error for decreasing MAF. All else being equal, this suggests the exact test to possess an inflated Type II error rate for minute values (less than 0.05) in MAF – indeed, at pointwise significance levels of 10^{-3} and 10^{-4} this is precisely the trend in the Type II error rate depicted by the data presented within the final column for each of the Tables 3.2 and 3.3, taking $f_j > 0$ therein; and at the 5% pointwise significance level this trend in decreasing power for decreasing π_j can also be seen within Figure 3.7, by way of the upper blue curve traversing away from the HWE reference line (heavy black dashed line at $f_j = 0$) as π_j decreases. Recall, part of the intent for this Dissertation is to correctly identify Q_0 for the CATT statistic, particularly for rare variant SNPs. Insofar as the exact test of HWE is underpowered for minute values in MAF, it is likely to incorrectly filter out rare variant SNP loci whose genotype frequencies within the population truly deviate from HWE. Second, since the exact test of the null hypothesis of HWE is to be conducted on a per-marker basis, amongst the

[3]The empirical data within Tables 3.2 and 3.3 suggest that a one-sided alternative hypothesis, testing $f_j > 0$ would be practical, since this could elevate the power to screen out rare variant SNP loci which fail adherence to HWE. However, the GWAS convention is to use a two-sided alternative hypothesis (see e.g., the article of [104]).

hundreds of thousands of SNPs within a GWAS sample, multiplicity correction is necessary. The problems encompassing the exact HWE test are exacerbated as the pointwise significance threshold decreases [113].

Figure 3.8 displays the combinations of estimated (at the maximum likelihood estimate (MLE) by marker) population inbreeding coefficients and estimated (at the MLE by marker) population minor allele frequencies among sampled controls, across 45168 SNP loci of CHR 1 for the Bipolar GWAS sample described within §2.6, assuming our adopted model allowing for genotype frequencies to deviate from HWE at an arbitrary SNP locus. In testing the null hypothesis of HWE, each of these markers possesses an exact two-sided pointwise p-value exceeding that of the value 10^{-6}, so that these markers are included within $\mathcal{H}_0$ for the entire GWAS marker sample (recall, $m = 769672$ for the entire GWAS sample). Among the 8082 markers with an MLE for π_j not exceeding the value 0.05, seventy-eight (78) markers – or, about 1% of these markers – possess an MLE for f_j at least equal to 0.1. Based upon the empirical results presented within Table 3.3, along with the magnitude in the value of m for the total GWAS sample ($m = 769672 > 500$K), the integrity of $Q_{0j}^{(*H)}$ in computing pointwise p-values under $\mathcal{H}_0$ for the CATT statistic – so also adjusted p-values within, say, the Bonferroni MTP – among these 78 markers is questionable. Assuming the allele frequencies among these 45168 markers to be representative of all m markers, this implies that the notion of questionable integrity of $Q_{0j}^{(*H)}$ in computing pointwise p-values under $\mathcal{H}_0$ for the CATT statistic would apply to more than 1300 SNP markers within the GWAS sample.

3.4 Proposal for Unbiased Strong Control of the FWER in GWAS

In light of the above potential problems encompassing the HWE assumption, we propose refraining from said assumption altogether and [in doing this] exploiting the resulting exact unconditional distribution for the CATT statistic under $\mathcal{H}_0$ in computing pointwise p-values within a GWAS. The absence of the HWE assumption, leads to the increasing (infinitely many times over) in the generalizability of this distribution, because we make no assumptions about the underlying allele frequency distributions within the population across SNP loci. As a consequence, we incur two nuisance parameters for the random trinomial vectors of sampled cases and controls at each SNP locus (see paragraph four within §3.2.1 for a review to this notion). Nonetheless, given specified values for these parameters and the fixed numbers of sampled cases and controls, we can generate the exact unconditional distribution of the CATT statistic for every realization thereof. In this regard, for SNP locus j, whose respective major and minor alleles are A and a, Q_0 becomes a function of:

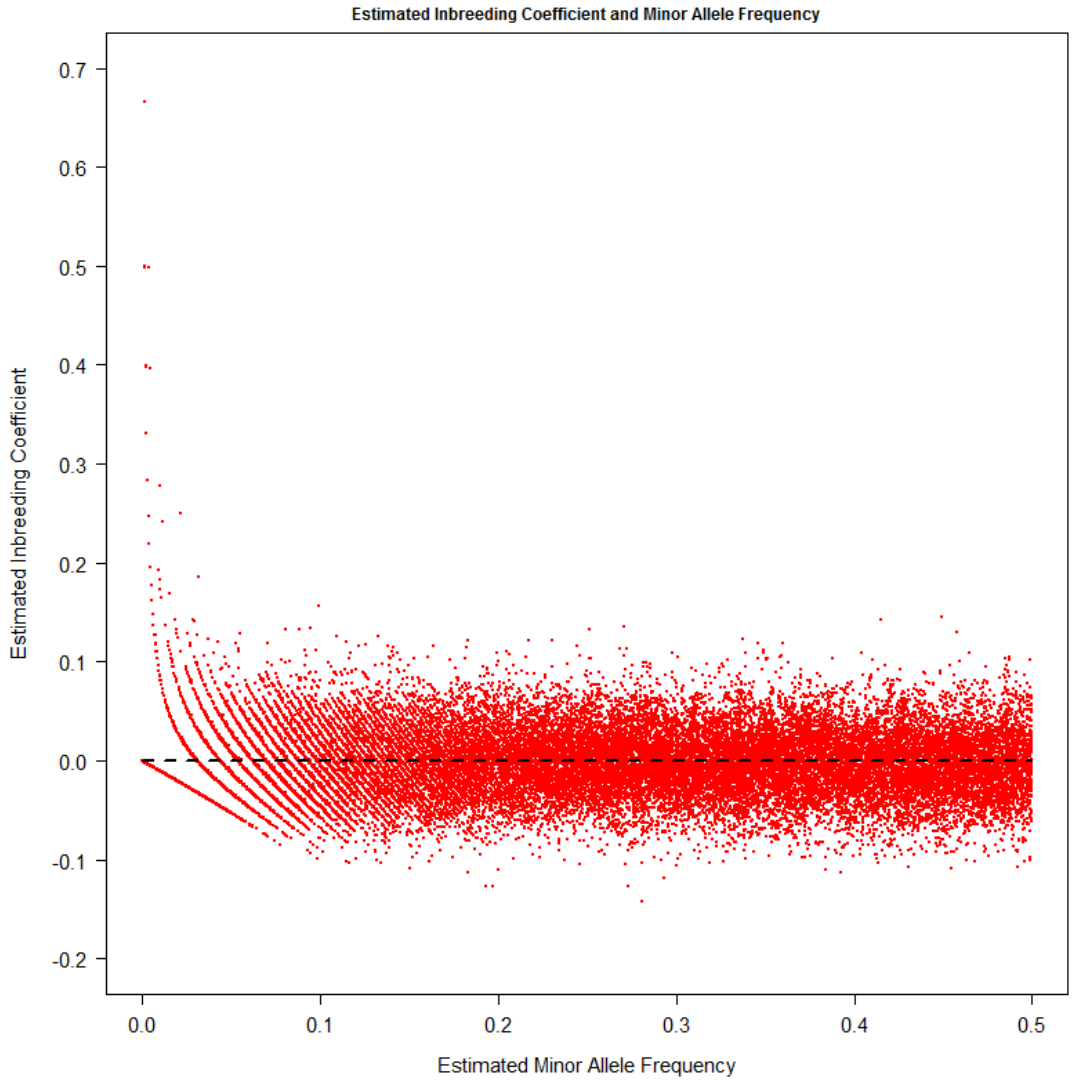

Fig. 3.8: Combinations of Estimated Population Inbreeding Coefficients and Estimated Population Minor Allele Frequencies among Sampled Controls, Across 45168 SNP Loci of Chromosome 1 for a Bipolar GWAS Sample of 1034 Controls and 1001 Cases of Bipolar Disorder. The Heavy Dashed Black Line Indicates Hardy-Weinberg Equilibrium.

the population proportion of homozygotes for the minor allele (subjects within the population possessing two copies of the minor allele at the locus – genotype aa) $\left(\pi_j^{aa}\right)$; the population proportion of heterozygotes at the locus $\left(\pi_j^{Aa}\right)$; and the fixed numbers of cases (n_1) and controls (n_0) for the GWAS sample.[4] Indeed, hereinafter we denote this unconditional distribution for the CATT statistic under $H_0^{(j)}$ by $Q_{0j}^*\left(\pi_j^{aa}, \pi_j^{Aa}, n_0, n_1\right)$ and we denote the vector $\left(\pi_j^{aa}, \pi_j^{Aa}\right)$ by $\boldsymbol{\theta}_j$. For clarity, we

[4]Note that, for SNP locus j, we have illustrated Q_0 to be a function of the two parameters π_j^{aa} and π_j^{Aa}. However, any combination of two elements chosen from the collection $\left\{\pi_j^{aa}, \pi_j^{Aa}, 1-\pi_j^{aa}-\pi_j^{Aa}\right\}$ can be substituted in lieu of π_j^{aa} and π_j^{Aa}.

reference this distribution at locus j by Q_{0j}^*. Therefore, Q_{0j}^* correctly identifies the true underlying test statistics null distribution for the CATT statistic under $H_0^{(j)}$; and collectively, $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ correctly identifies the true underlying test statistics null distribution for the CATT statistic under $\mathcal{H}_0$.

Specifically, for strong control of the FWER in GWAS, we propose exploitation of $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ for the CATT statistic under $\mathcal{H}_0$ within the Bonferroni (or, Šidák) MTP. For a GWAS sample of m mutually independent SNP loci, the implementation of Q_{0j}^* within the Bonferroni MTP will provide nearly exact control of the FWER[5] at level α, all $\alpha \in (0,1)$. Although it is unlikely that a GWAS sample will be comprised of mutually independent SNP loci, the utility of Q_{0j}^* within the Bonferroni MTP is guaranteed to control the FWER at level α for the CATT statistic under $\mathcal{H}_0$, whereas – pursuant to the arguments presented within §3.2.2 and §3.2.3 – no such assurance holds for this MTP under $\tilde{Q}_0$. A simple pseudocode for this implementation is given within Algorithm 3.1.

Algorithm 3.1 The Bonferroni MTP under Q_{0j}^*

1. Select a level in the FWER to control, say α. Compute the realization of the CATT statistic for SNP locus j under $H_0^{(j)}$ and some assumed genetic model of inheritance (e.g., additive model) under the two-sided alternative hypothesis $H_a^{(j)}$, for all $j = 1, \ldots, m$. Denote the realization in said test statistic by t_j.

2. Let t^* be the smallest realization of the CATT statistic which yields a pointwise p-value under $Q_0^*\left(\pi^{aa}, \pi^{Aa}, n_0, n_1\right)$ not exceeding the value α/m, for every $\pi^{aa} \in (0, 0.5)$ and $\pi^{Aa} \in (0,1)$ satisfying the linear inequality $\pi^{Aa} \leq 1 - 2\pi^{aa}$, where $Q_0^*\left(\pi^{aa}, \pi^{Aa}, n_0, n_1\right)$ denotes the unconditional distribution of the CATT statistic under the null hypothesis of no genotype-phenotype association for a locus with respective population proportion of homozygotes for the minor allele and population proportion of heterozygotes given by π^{aa} and π^{Aa}. The value of t^* is the smallest value of the CATT statistic under $\mathcal{H}_0$ which yields an unconditional p-value (over the parameter space for π^{aa} and π^{Aa}) less than the value α/m. That is, t^* is the smallest possible value of t_j for which the Bonferroni MTP rejects null hypothesis $H_0^{(j)}$ at the α level in the FWER under Q_{0j}^*.

[5]The Šidák MTP is exact for mutually independent test statistics under $\mathcal{H}_0$. For large values in m, the Bonferroni and Šidák MTPs are nearly equivalent.

3. For those loci whose values of t_j exceed that of t^*: estimate the values of π_j^{aa} and π_j^{Aa} by their respective maximum likelihood estimators under $H_0^{(j)}$. Denote these MLEs for π_j^{aa} and π_j^{Aa} as $\hat{\pi}_j^{aa}$ and $\hat{\pi}_j^{Aa}$, respectively; utilizing the PMF for $Q_{0j}^* \left(\pi_j^{aa}, \pi_j^{Aa}, n_0, n_1\right)$, evaluated at $\left(\hat{\pi}_j^{aa}, \hat{\pi}_j^{Aa}\right)$ (hereinafter, denoted by $\hat{\boldsymbol{\theta}}_j$), compute pointwise p-values.[6]

4. Reject those null hypotheses whose pointwise p-values do not exceed the value α/m.

Furthermore, to account for correlation within the GWAS sample, we propose implementation of $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ for the CATT statistic under $\mathcal{H}_0$ within the minP MTP. Insofar as the permutation null distribution for the minP MTP is to be derived under $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$, this MTP will provide balanced strong control of the FWER. A simple pseudocode for this implementation – based upon Algorithm 2.5 within [62] – is as follows (we denoted this by Algorithm 3.2):

Algorithm 3.2 The MinP MTP under Q_{0j}^*

---

1. Same as step 1 above for the Bonferroni MTP pseudocode.

2. Assign, without replacement, a label from the collection of counting numbers $\{1, \ldots, n\}$ to each of the n sampled study subjects. The joint distribution of the p-values, for the joint distribution of the CATT statistics under $\mathcal{H}_0$, can be estimated by permuting the labels amongst the subjects (equivalently, we can permute the columns upon the genotype matrix, $\mathbf{G}$, as [essentially] done within GPER). Indeed, randomly permute the labels amongst the subjects, say R times over. That is, in each permutation of the labels, we are randomly reassigning case and control status amongst the subjects. Permuting the labels in this manner ensures the phenotypic trait is independent of the genotype data (i.e., we are simulating $\mathcal{H}_0$), while simultaneously attempts to preserve the correlation structure and distributional properties of the genotype data.

3. For the r^{th} permutation of the labels, $r = 1, \ldots, R$: compute the realization of the CATT statistic for SNP locus j under $H_0^{(j)}$. Note, the genetic model of inheritance assumed under [the two-sided] $H_a^{(j)}$ within step 1 above must also be assumed here. Denote the realization of the CATT statistic for locus j by $t_{j,r}$; utilizing the PMF for $Q_{0j}^* \left(\pi_j^{aa}, \pi_j^{Aa}, n_0, n_1\right)$, evaluated

[6] Insofar as these pointwise p-values are computed under Q_{0j}^* at the MLEs in the nuisance parameters, π_j^{aa} and π_j^{Aa}, the p-values are *approximate*, called bootstrap p-values [140].

at the MLEs for the parameters $\boldsymbol{\theta}_j$ under $H_0^{(j)}$, $\hat{\boldsymbol{\theta}}_j$, compute the pointwise p-value for the realization $t_{j,r}$, which we denote by $p_{j,r}$; and compute the minimum of the $p_{j,r}$, denoted by $p_{(1),r}$. The value of $p_{(1),r}$ is a random sample of size one from the permutation null distribution of the minimum p-value (minP) for the joint distribution of the p-values for the CATT statistic under $\mathcal{H}_0$.

4. Denote the k^{th} ordered value of the collection $\left\{p_{(1),r}\right\}_{r=1,\ldots,R}$ by $p_{(k)}$, $k = 1, \ldots, R$. Let $p^* = p_{(\lfloor \alpha R \rfloor)}$, where $\lfloor \cdot \rfloor$ returns the greatest integer contained within $(\cdot)$. The value of p^* is the maximum observable unconditional pointwise p-value which results in rejection of a null hypothesis at the α level in the FWER. Finally, let t^* be as previously defined within step 2 of the above Bonferroni MTP pseudocode, replacing the fraction α/m with p^* therein.

5. For the observed (non-permuted) data, repeat step 3 of the above Bonferroni MTP pseudocode, omitting the estimation of the parameters comprising $\boldsymbol{\theta}_j$ (i.e., the MLEs for these parameters are permutation invariant and need only be computed once by SNP locus). Reject those null hypotheses whose pointwise p-values do not exceed the value p^*. Note: the footnote within step 3 of the above Bonferroni MTP pseudocode (see Algorithm 3.1) also applies here.

3.5 Computational Tools

3.5.1 Introduction

On the one hand, due to its proper control in the FWER, the unconditional distribution of the CATT statistic under $H_0^{(j)}$, Q_{0j}^*, is exceptionally attractive. Its implementation within large-scaled population based case-control genetic association studies, corrects upon näive asymptotic assumptions entailing the CATT statistic at the SNP locus, and leads to accurate interpretation in the data analysis of these studies. On the other hand, implementation of this distribution in practice, introduces a difficult computational problem.

The computational problem encompassing this distribution originates from the large number of elements making up its support – here, the support of Q_{0j}^* (any $j = 1, \ldots, m$) is called the unconditional reference set and is denoted Γ. In terms of a 2×3 contingency table (e.g., Table 2.1) with fixed row margin values, the unconditional reference set is the collection of all possible tables, such that the cells upon each row of a particular table within this collection must sum to its

corresponding row margin value. Under $\mathcal{H}_0$, each table within this set has an affiliated probability of being realized and a corresponding realization of the CATT statistic. To help illustrate the computational problem here, let Z_{1k} and Z_{0k} represent the respective random numbers of cases and controls carrying k copies of the minor allele at some SNP locus, $k \in \{0,1,2\} = \mathcal{G}$, and let $\mathbf{Z}_1 = (Z_{10}, Z_{11}, Z_{12})$ and $\mathbf{Z}_0 = (Z_{00}, Z_{01}, Z_{02})$ be the random row vectors for a 2×3 contingency table with fixed row margin (n_1, n_0). If the vector $(\mathbf{z}_0, \mathbf{z}_1)$ denotes some table of Γ, note that in computing the pointwise p-value for the realization of the CATT statistic (t_j) under Q_{0j}^*, for all $j = 1, \ldots, m$, one could carry out the following procedure:

for $j = 1$ **to** m **do**

$p_j \leftarrow 0$. {Initialize the p-value to the value of zero upon SNP locus j}.

for each $(\mathbf{z}_0, \mathbf{z}_1) \in \Gamma$ **do**

Compute the realization of the CATT statistic under $H_0^{(j)}$ (against the two-sided alternative assuming the additive GMI) for table $(\mathbf{z}_0, \mathbf{z}_1)$, and denote it by $T(\mathbf{z}_0, \mathbf{z}_1)$.

if $T(\mathbf{z}_0, \mathbf{z}_1) \geq t_j$ **then**

$p_j \leftarrow p_j + \Pr(\text{observing table } (\mathbf{z}_0, \mathbf{z}_1) \,|\boldsymbol{\theta}_j, n_0, n_1)$. {Table contributes to the p-value. Increment the value of p_j by the probability of observing the table under Q_{0j}^*}.

end if

end for

end for

In computing p-value p_j, each $j = 1, \ldots, m$, note that the conditional clause within the aforementioned pseudocode is conducted a total of $n(\Gamma)$ (here, for a set S, $n(S)$ is used to denote the number of elements contained within S – i.e., $n(S)$ denotes the cardinality of S) times over. Hence, the computational problem encompassing utility of Q_{0j}^* in practice is an increasing function in $n(\Gamma)$. For a case-control GWAS of n_0 controls and n_1 cases, it can be shown (see Proposition A.6 within Appendix A) that

$$n(\Gamma) = \binom{n_0 + 2}{2}\binom{n_1 + 2}{2}. \tag{3.1}$$

Expanding the binomial coefficients of expression (3.1), we find that $n(\Gamma)$ is the product of quadratic functions in each of n_0 and n_1. For a balanced GWAS, this observation reduces to $n(\Gamma)$ taking a value on the order of a quartic function in the number of cases (or, controls). This implies that $n(\Gamma)$ can obtain large values, even for small GWAS sample sizes. For example, consider a balanced

GWAS of size $n = 2\text{K}$. According to expression (3.1), we have

$$n(\Gamma) = \binom{1002}{2}^2 = 2.52 \times 10^{11} \approx (0.25)(n_0)^4, \tag{3.2}$$

a very large number indeed. Therefore, the computational problem encompassing the distribution Q_{0j}^* originates from the large number of elements making up its support.

Pursuant to the above argument, it is clear that computing the pointwise p-value for a single realization in the CATT statistic under Q_{0j}^*, presents a large computational problem. However, the magnitude of this computational problem is dwarfed, when compared to that of implementing $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ in computing pointwise p-values within the minP MTP. This can be seen by noting that in carrying out R permutations of the phenotype labels within the minP MTP, the aforementioned pseudocode must be repeated a total of R times over. For example, taking $m = 500\text{K}$ and $R = 100\text{K}$ – realistic values for these variables within a GWAS – it follows that the number of iterations upon the conditional clause within the above pseudocode is $m \times R = 5 \times 10^{10}$. To illustrate the extent of this computational problem, suppose for each $r = 1, \ldots, R$ that the aforementioned pseudocode can compute the pointwise p-values upon the realizations in the CATT statistic $\{t_{j,r}\}_{j=1,\ldots,m}$ (replacing t_j with $t_{j,r}$ therein), in m seconds. Under this [highly suspect] supposition, it would take more than 1580 computational years to compute the $m \times R$ pointwise p-values under $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ within the minP MTP. Nonetheless, performing the aforementioned pseudocode in m seconds, each $r = 1, \ldots, R$, seems rather optimistic. Indeed, we require fast computational tools for the realized implementation of $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ within the minP MTP.

3.5.2 Approach

By recognizing that the computed pointwise p-values between two SNP loci will be similar, whenever their corresponding parameter vectors $\boldsymbol{\theta} = \left(\pi^{aa}, \pi^{Aa}\right)$ (where dropping the subscript j from the vector $\boldsymbol{\theta}_j$ indicates general values of π^{aa} and π^{Aa} within their joint parameter space) and CATT statistic realizations are similar, to resolve the computational problem – entailing the implementation of $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ within the minP MTP – we propose the following five-step procedure (Algorithm 3.3):

Algorithm 3.3 An Efficient Approach for Multiple Testing Correction under Q_{0j}^*

1. Estimate the collection of vectors $\{\boldsymbol{\theta}_j\}_{j=1,\ldots,m}$ at their corresponding MLEs, $\left\{\hat{\boldsymbol{\theta}}_j\right\}_{j=1,\ldots,m}$ under $\mathcal{H}_0$. For example, Figure 3.9 displays this collection of MLE vectors across 45168 SNP loci of CHR 1 for the Bipolar GWAS sample described within §2.6. Formulate a subspace of the parameter space for $\boldsymbol{\theta}$ – the triangular region of the first quadrant within the Cartesian plane (where π^{aa} and π^{Aa} represent the respective x- and y-axis), bounded by each axis within said plane and the downward sloping line $\pi^{Aa} = 1 - 2\pi^{aa}$ – of which captures the collection $\left\{\hat{\boldsymbol{\theta}}_j\right\}_{j=1,\ldots,m}$. Partition this subspace by way of horizontal and vertical line segments, and partition the domain for the CATT statistic into disjoint intervals. The finer the resolution in this partitioning scheme, the better the precision in the resulting estimates of $p_{j,r}$ from this approach (see step 5 below for details in the reasoning here).

2. For each upper interval endpoint within the partition of the CATT statistic domain and each ordered pair – formed from an intersecting horizontal and vertical line segment within the subspace partition – of the parameter vector $\boldsymbol{\theta}$, compute the corresponding pointwise p-value under $Q_0^*(x, y, n_0, n_1)$, where n_0 and n_1 are assumed given and (x, y) corresponds to an ordered pair of a realization in the elements comprising the parameter vector $\boldsymbol{\theta}$ within the subspace and where Q_0^* is as previously defined within step 2 of Algorithm 3.1. For example, Figure 3.10 displays a contour plot of the Bonferroni corrected UPTE for the CATT statistic under Q_0^* (the dimension represented in color) at realization $F_{\tilde{Q}_0}^{-1}(1 - 0.05/m)$ – the minimum value in this statistic for which the Bonferroni MTP under $\tilde{Q}_0$ calls for $H_0^{(j)}$ to be rejected at the 5% FWER – within the parameter space of $\boldsymbol{\theta}$, assuming a balanced GWAS of $n = 1\text{K}$ and $m = 500\text{K}$. This plot essentially portrays the notion of step 2, for a single interval endpoint of the CATT statistic, namely the value $F_{\tilde{Q}_0}^{-1}(1 - 0.05/m)$.

3. Implement parallel algorithms to calculate the pointwise p-values of step 2.

4. Formulate a lookup table from these p-value calculations.

5. Utilize this table within step 3 of the aforementioned minP pseudocode (see Algorithm 3.2 within §3.4). Essentially, this table can serve as a proxy for all possible realizations in the CATT statistic, and all realizations of the vector $\boldsymbol{\theta}$ within the formulated subspace. That is, this table can be used to approximate $p_{j,r}$, all $j = 1, \ldots, m$ and $r = 1, \ldots, R$. Note that

the precision in these approximations improve, as the resolutions in the partitioning scheme – provided within step 1 above – become finer.

3.5.3 Generating the *P*-value Lookup Table

3.5.3.1 Introduction

Note that the general idea encompassing the implementation of Algorithm 3.3, lies with generating a pointwise p-value lookup table (PPT) which can be utilized within the minP MTP pseudoalgorithm (see Algorithm 3.2) to estimate the pointwise p-values therein. In generating this lookup table a priori to application of the minP MTP, we avoid having to directly utilize the joint null distribution $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ in computing the pointwise p-values within the R permutations of the minP MTP. Provided that the number of elements comprising the PPT to be considerably smaller than the value $m \times R$ – the number of pointwise p-values computed within $\left\{Q_{0j}^*\right\}_{j=1,\ldots,m}$ of the minP MTP upon R permutations thereof (i.e., implementation of the minP MTP without use of the PPT) – the utility of the PPT within the minP MTP possesses the potential to substantially reduce the scale of the computational problem outlined within §3.5.1.

Let $\mathcal{O}$ and $\mathcal{T}$, denote the respective collections of ordered pairs $\boldsymbol{\theta} = (\pi^{aa}, \pi^{Aa})$ and interval endpoints for the CATT statistic, formulated as a result of implementing step 1 of Algorithm 3.3. In terms of $\mathcal{O}$ and $\mathcal{T}$, the PPT can be considered as a $n(\mathcal{T}) \times n(\mathcal{O})$ dimensional matrix (denoted $\mathbf{P}$) whose $(u,w)^{\text{th}}$ element, $p_{u,w}^{\text{o}} = [\mathbf{P}]_{u,w}$, is defined by

$$p_{u,w}^{\text{o}} = \Pr\left(T_w \geq \tau_u | Q_0^*\left(\boldsymbol{\theta}_w, n_0, n_1\right)\right), \tag{3.3}$$

where τ_u and $\boldsymbol{\theta}_w$ correspond to the respective u^{th} and w^{th} elements of $\mathcal{T}$ and $\mathcal{O}$; where $T_w \sim Q_0^*\left(\boldsymbol{\theta}_w, n_0, n_1\right)$ under $\mathcal{H}_0$ (Q_0^* is as defined within step 2 of the Bonferroni pseudoalgorithm – see Algorithm 3.1); and where, recall, n_1 and n_0 denote the respective number of cases and controls amongst the sample of n subjects. Here, we use u and w to index the respective rows and columns of $\mathbf{P}$, $u = 1, \ldots, n(\mathcal{T})$ and $w = 1, \ldots, n(\mathcal{O})$. Calculation of $p_{u,w}^{\text{o}}$ is not a trivial task, and demands considerable computational power, each $u = 1, \ldots, n(\mathcal{T})$ and each $w = 1, \ldots, n(\mathcal{O})$. This is particularly due to the number of elements comprising the collection Γ (i.e., the support of Q_0^*), as previously elucidated to within §3.5.1.

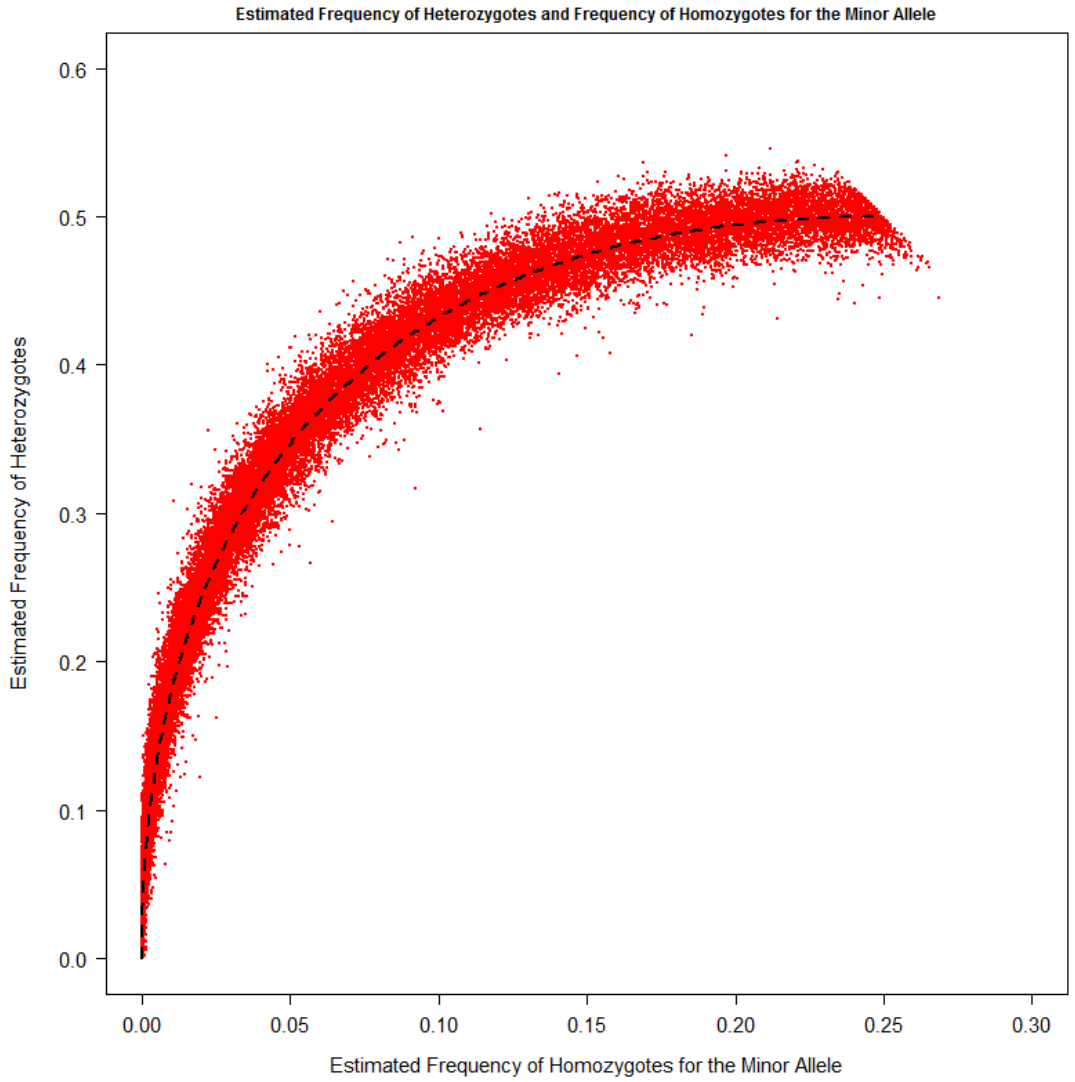

Fig. 3.9: Combinations of Estimated Population Frequencies of Heterozygotes and Estimated Population Frequencies of Homozygotes for the Minor Allele, Across 45168 SNP Loci of Chromosome 1 for a Bipolar GWAS Sample of 1034 Controls and 1001 Cases of Bipolar Disorder. The Heavy Dashed Black Curve Indicates Hardy-Weinberg Equilibrium.

One approach to generating the PPT could entail the implementation of a network algorithm (see e.g., [141, 142]). Given fixed margins of a 2×3 contingency table, a network is depicted as a directed acyclic graph of nodes connected by arcs [143]. This network is constructed in four stages, where a series of calculations are performed upon each of the nodes within each stage. The goal of the network algorithm is to compute the p-value for the observed test statistic under $H_0^{(j)}$, by *implicit* evaluation of the performed calculations upon the nodes of the constructed network (see §4.6.1 for explicit details of a network algorithm). While a network algorithm is perhaps the best approach

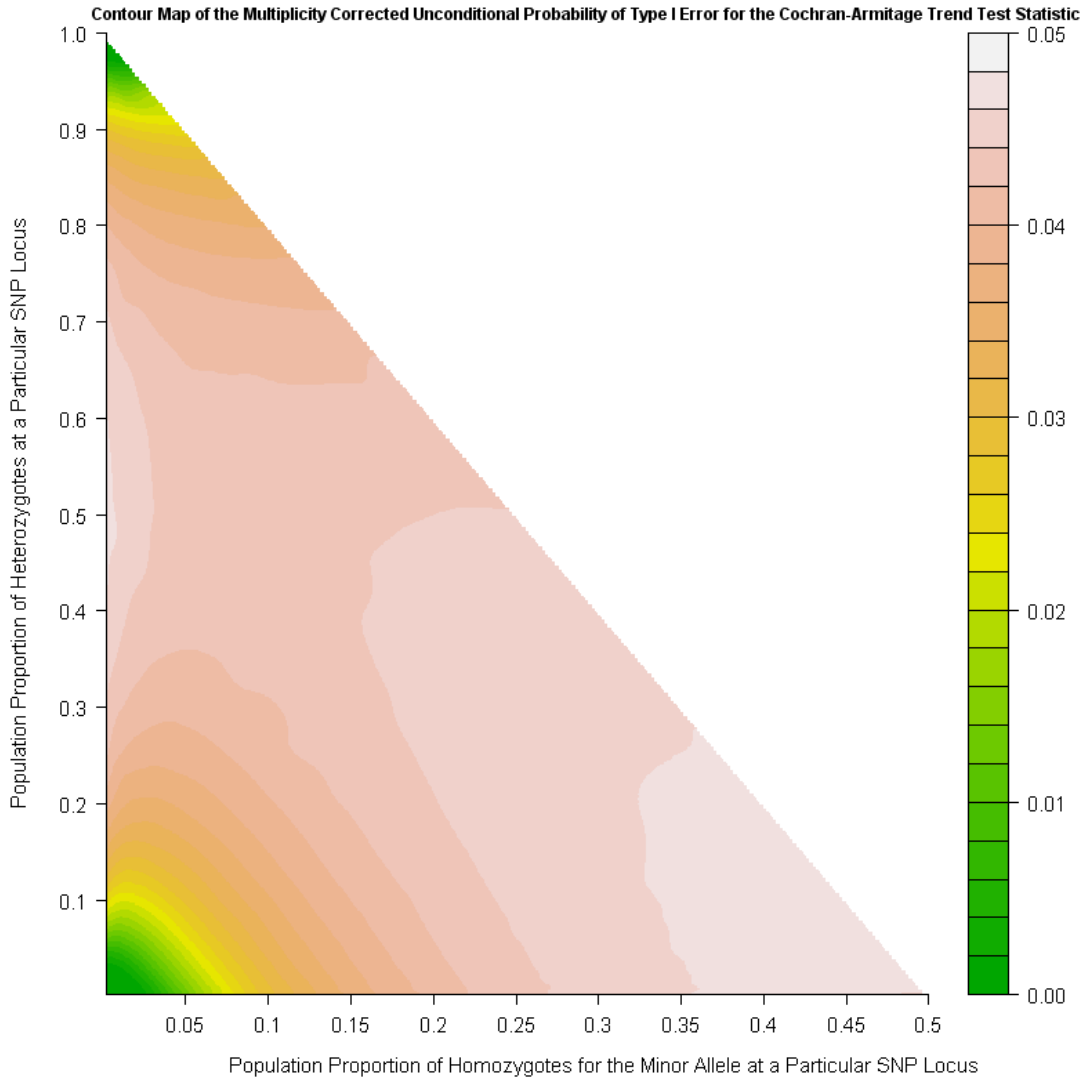

Fig. 3.10: Contour Plot of the Bonferroni Corrected Unconditional Probability of Type I Error for the Cochran-Armitage Trend Test Statistic under $\mathcal{H}_0$ at the Realization $F^{-1}_{\tilde{Q}_0}(1- \ 0.05/m)$ for the χ^2_1 Distribution ($\tilde{Q}_0$), Across the Parameter Space $\boldsymbol{\theta}_j$, Against a Balanced GWAS of $n = 1$K and $m = 500$K SNP Loci. The Assumed FWER under $\tilde{Q}_0$ Is 5%.

in computing the exact *conditional* (i.e., fixed margins for a 2×3 contingency table) p-value for the observed test statistic under $H_0^{(j)}$, there are several problems with this approach in generating $\mathbf{P}$. The first problem lies with the scale in the number of possible observable margins for the 2×3 tables within the *unconditional* reference set (i.e., the support of Q_0^*; equivalent to the collection Γ), which leads to the construction/evaluation of a great many networks. Since the row margin for any table is considered fixed at the time of case-control sampling, the number of observable 2×3 table margins comprising the unconditional reference set is equal to the number of possible combinations

for the column margin upon the n-size random sample of study subjects, denoted n_c. It is,

$$n_c = \binom{n+2}{2}. \tag{3.4}$$

Hence, the computation of $p^{\circ}_{u,w}$ of $\mathbf{P}$, entails evaluation of n_c (3.4) total networks. For example, taking $n_0 = n_1 = 1\text{K}$, the computation of $p^{\circ}_{u,w}$ demands evaluation of more than two million networks, which likely requires an inordinate quantity of time to traverse over in generating $\mathbf{P}$. The second problem with the network algorithm approach in generating the elements of $\mathbf{P}$, lies with the fact that the elements comprising the collection $\mathcal{T}$ are not fixed. This complicates implementation of the network algorithm and could require considerable computational time to generate the PPT. In particular, this approach would require at least a single pass through each of the n_c (3.4) networks. However, within the k^{th} network, $k = 1, \ldots, n_c$, note that the *Forward pass* – for the network algorithm of [142] (see Appendix therein) – is required to be iterated through for each $\tau_u \in \mathcal{T}$. This likely requires considerable computational time to accomplish.

Here, we take a different perspective to tackling the computational problem in generating the elements of $\mathbf{P}$. Namely, for a fixed $\boldsymbol{\theta}_w \in \mathcal{O}$, we recognize that some elements within the unconditional reference set have exceptionally small (nearly zero) probability of being observed. In this regard, for each $w = 1, \ldots, n(\mathcal{O})$, we anticipate accurate estimation of $p^{\circ}_{u,w}$, $u = 1, \ldots, n(\mathcal{T})$, amongst a truncated (i.e., deleted) unconditional reference set. Within §3.5.3.3, we develop a methodology for deleting elements upon the unconditional reference set, while simultaneously preserving the integrity of the estimates amongst the elements comprising the appropriate column of $\mathbf{P}$. Within §3.5.3.4 we sketch an iterative algorithm for obtaining a truncated unconditional reference set, and within §3.5.3.5 we provide examples for the implementation of the iterative algorithm. As it turns out, this approach – generating [an estimated] PPT from truncated unconditional reference sets – lends elegantly to parallel computing. Within §3.5.3.6, we sketch a parallel computing approach to generating an estimated PPT from truncated unconditional reference sets.

3.5.3.2 The Exact Unconditional Probability of Type I Error

Before we can discuss constructing a truncated unconditional reference set, we need to: define, mathematically (through set notation), the unconditional reference set; and define the formula which relates (3.3) to the unconditional reference set. Here, for each $k \in \{0, 1, 2\} = \mathcal{G}$, let Z_{1k} and Z_{0k} be as previously defined, the respective random numbers of cases and controls carrying k copies of the

minor allele at some SNP locus. Under the null hypothesis $H_0^{(j)}$, for each $y \in \{0,1\} = \mathcal{Y}$, it can be shown (see Proposition A.4 within Appendix A) that

$$\mathbf{Z}_y = (Z_{y0}, Z_{y1}, Z_{y2}) \sim \text{Multinomial}\left(n_y, \boldsymbol{\pi}_j = (\pi_{0j}, \pi_{1j}, \pi_{2j})\right),$$

where $\pi_{kj} = \Pr(G_j = k)$ for all $k \in \mathcal{G}$, and $G_j \in \mathcal{G}$ denotes the number of copies for the minor allele at locus j. Hence, for each $y \in \mathcal{Y}$, under $H_0^{(j)}$ the density for the random vector $\mathbf{Z}_y$ is given by

$$h\left(\mathbf{z}_y|\boldsymbol{\pi}_j\right) = \Pr\left(\mathbf{Z}_y = \mathbf{z}_y|\boldsymbol{\pi}_j\right) = \binom{n_y}{z_{y0}, z_{y1}, z_{y2}} \left(\pi_j^{AA}\right)^{z_{y0}} \left(\pi_j^{Aa}\right)^{z_{y1}} \left(\pi_j^{aa}\right)^{z_{y2}}, \tag{3.5}$$

where

$$\binom{n_y}{z_{y0}, z_{y1}, z_{y2}} = \frac{n_y!}{z_{y0}! z_{y1}! z_{y2}!},$$

such that A and a denote the respective major and minor alleles at locus j. Since cases and controls are assumed unrelated, the exact unconditional probability of observing table $(\mathbf{z}_0, \mathbf{z}_1) \in \Gamma$ under $H_0^{(j)}$, is given by

$$\begin{aligned} g\left(\mathbf{z}_0, \mathbf{z}_1|\boldsymbol{\pi}_j\right) &= \Pr\left(\mathbf{Z}_0 = \mathbf{z}_0, \mathbf{Z}_1 = \mathbf{z}_1|\boldsymbol{\pi}_j\right) \\ &= h\left(\mathbf{z}_0|\boldsymbol{\pi}_j\right) h\left(\mathbf{z}_1|\boldsymbol{\pi}_j\right) \\ &= \prod_{y \in \mathcal{Y}} \binom{n_y}{z_{y0}, z_{y1}, z_{y2}} \left(\pi_j^{AA}\right)^{z_{y0}} \left(\pi_j^{Aa}\right)^{z_{y1}} \left(\pi_j^{aa}\right)^{z_{y2}}, \end{aligned} \tag{3.6}$$

where Γ (i.e., the unconditional reference set) is given by

$$\Gamma = \left\{ (\mathbf{z}_0, \mathbf{z}_1) : \sum_{k \in \mathcal{G}} z_{yk} = n_y, \forall y \in \mathcal{Y} \right\}. \tag{3.7}$$

Consider $(\mathbf{z}_0, \mathbf{z}_1) \in \Gamma$ to be an arbitrarily chosen table from the unconditional reference set, and let $T(\mathbf{z}_0, \mathbf{z}_1) > 0$ denote the realization of the CATT statistic corresponding to the table, computed under $H_0^{(j)}$, some j, where the two-sided alternative hypothesis $H_a^{(j)}$ is assumed. Thus, for any other realization in the CATT statistic (computed under $H_0^{(j)}$), say $t > 0$, the critical region of the asymptotic test corresponding with t, denoted $\Gamma_A(t)$, is given by

$$\Gamma_A(t) = \{(\mathbf{z}_0, \mathbf{z}_1) \in \Gamma : T(\mathbf{z}_0, \mathbf{z}_1) \geq t\}. \tag{3.8}$$

Therefore, under $H_0^{(j)}$, the exact unconditional probability of Type I error for the CATT statistic at realization t_j, p_j, is given by

$$p_j = \sum_{(\mathbf{z}_0,\mathbf{z}_1)\in\Gamma_A(t_j)} g\left(\mathbf{z}_0,\mathbf{z}_1|\boldsymbol{\pi}_j\right), \tag{3.9}$$

where $g\left(\cdot|\cdot\right)$ is given by (3.6). Note that by substituting the appropriate parameter vector $\boldsymbol{\theta}_w$ and realization τ_u within this expression, one obtains the value of (3.3).

3.5.3.3 A Truncated Unconditional Reference Set

Here, for a fixed $\boldsymbol{\theta}_w \in \mathcal{O}$ and some user defined precision (denoted, ϵ) in estimating (3.3), our goal is to formulate a truncated unconditional reference set (denoted, $\Gamma\left(\boldsymbol{\theta}_w\right)$), with the intentions of: rapid generation of estimates for the elements $p^{\mathrm{o}}_{u,w}$ over the assembled truncated unconditional reference set, for all $u = 1,\ldots,n(\mathcal{T})$; and maintaining high accuracy upon these estimates for the $p^{\mathrm{o}}_{u,w}$. In other words, given $\boldsymbol{\theta}_w \in \mathcal{O}$ and ϵ, we would like to delete as many elements from Γ as possible (this ensures rapid generation of the estimate for $p^{\mathrm{o}}_{u,w}$, $u = 1,\ldots,n(\mathcal{T})$), without compromising the accuracy in the estimates of the $p^{\mathrm{o}}_{u,w}$.

Let Γ_1 and Γ_0 denote the respective collections of all possible case and control rows for the 2×3 contingency tables upon the elements within Γ. That is, for each $y \in \mathcal{Y}$,

$$\Gamma_y = \left\{ \mathbf{z}_y : \sum_{k\in\mathcal{G}} z_{yk} = n_y \right\}. \tag{3.10}$$

Now, given $\boldsymbol{\theta}_w$ – where, for the moment we set ϵ aside – let $\Gamma_y\left(\boldsymbol{\theta}_w\right)$ denote the elements within Γ_y which are to be preserved (i.e., not deleted) within the truncated unconditional reference set, and denote the complement of $\Gamma_y\left(\boldsymbol{\theta}_w\right)$ by $\Gamma'_y\left(\boldsymbol{\theta}_w\right)$. For a particular specification over each of the collections $\Gamma_0\left(\boldsymbol{\theta}_w\right)$ and $\Gamma_1\left(\boldsymbol{\theta}_w\right)$, it therefore follows that the truncated unconditional reference set is given by

$$\Gamma\left(\boldsymbol{\theta}_w\right) = \left\{(\mathbf{z}_0,\mathbf{z}_1)\in\Gamma : \left(\mathbf{z}_0 \in \Gamma_0\left(\boldsymbol{\theta}_w\right)\right) \cap \left(\mathbf{z}_1 \in \Gamma_1\left(\boldsymbol{\theta}_w\right)\right)\right\}, \tag{3.11}$$

and its complement is given by

$$\Gamma'\left(\boldsymbol{\theta}_w\right) = \left\{(\mathbf{z}_0,\mathbf{z}_1)\in\Gamma : \left(\mathbf{z}_0 \in \Gamma'_0\left(\boldsymbol{\theta}_w\right)\right) \cup \left(\mathbf{z}_1 \in \Gamma'_1\left(\boldsymbol{\theta}_w\right)\right)\right\}. \tag{3.12}$$

Under $\mathcal{H}_0$, (3.9) can be expressed as

$$p^{\text{o}}_{u,w} = \sum_{(\mathbf{z}_0,\mathbf{z}_1)\in(\Gamma(\boldsymbol{\theta}_w)\cap\Gamma_A(\tau_u))} g\left(\mathbf{z}_0,\mathbf{z}_1|\boldsymbol{\theta}_w\right) + \sum_{(\mathbf{z}_0,\mathbf{z}_1)\in(\Gamma'(\boldsymbol{\theta}_w)\cap\Gamma_A(\tau_u))} g\left(\mathbf{z}_0,\mathbf{z}_1|\boldsymbol{\theta}_w\right)$$

$$= p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) + e_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right), \tag{3.13}$$

where τ_u is the u^{th} element of $\mathcal{T}$, $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ is the estimate of $p^{\text{o}}_{u,w}$ under $\Gamma\left(\boldsymbol{\theta}_w\right)$, and $e_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ is the acquired error in estimating $p^{\text{o}}_{u,w}$ with $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$. For a given $\Gamma\left(\boldsymbol{\theta}_w\right)$, we make the observation that the ratio, $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ to $p^{\text{o}}_{u,w}$, must assume values between zero and one, where a value approximately equal to one in said ratio indicates high accuracy in the estimate of $p^{\text{o}}_{u,w}$ – equivalently, the value of $e_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ is small relative to that of $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$, whenever $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ is a very accurate estimate of $p^{\text{o}}_{u,w}$. Hence, we can utilize the ratio, $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ to $p^{\text{o}}_{u,w}$, as a measure of accuracy for the estimate $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$.

Here, for $\boldsymbol{\theta}_w \in \mathcal{O}$, let $\epsilon > 0$ be the smallest value in the ratio $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ to $p^{\text{o}}_{u,w}$ over all possible $\Gamma\left(\boldsymbol{\theta}_w\right) \subset \Gamma$ for which the user is willing to accept. For a given $\Gamma\left(\boldsymbol{\theta}_w\right)$, we note that

$$\begin{aligned} e_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) &= \sum_{(\mathbf{z}_0,\mathbf{z}_1)\in(\Gamma'(\boldsymbol{\theta}_w)\cap\Gamma_A(\tau_u))} g\left(\mathbf{z}_0,\mathbf{z}_1|\boldsymbol{\theta}_w\right) \\ &\leq \sum_{(\mathbf{z}_0,\mathbf{z}_1)\in\Gamma'(\boldsymbol{\theta}_w)} g\left(\mathbf{z}_0,\mathbf{z}_1|\boldsymbol{\theta}_w\right) \\ &< \sum_{\mathbf{z}_0\in\Gamma'_0(\boldsymbol{\theta}_w)} h\left(\mathbf{z}_0|\boldsymbol{\theta}_w\right) + \sum_{\mathbf{z}_1\in\Gamma'_1(\boldsymbol{\theta}_w)} h\left(\mathbf{z}_1|\boldsymbol{\theta}_w\right) \\ &= e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right). \end{aligned} \tag{3.14}$$

Hence, we have

$$\frac{p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)}{p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) + e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)} \geq \epsilon \qquad \Longrightarrow \qquad \frac{p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)}{p^{\text{o}}_{u,w}} > \epsilon, \tag{3.15}$$

so that if one can demonstrate that the inequality of the premise within (3.15) holds for the given choice in $\Gamma\left(\boldsymbol{\theta}_w\right)$, then $\Gamma\left(\boldsymbol{\theta}_w\right)$ has essentially been created (through the selection of ϵ). Since ϵ and – for a given $\Gamma\left(\boldsymbol{\theta}_w\right)$ – $e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ are constants, the inequality of the premise within (3.15) reduces to

$$p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) \geq \left(\frac{\epsilon}{1-\epsilon}\right) e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right). \tag{3.16}$$

Therefore, for the given value of $\epsilon > 0$ and collection $\Gamma(\boldsymbol{\theta}_w)$, if the condition imposed by (3.16) holds for all $u = 1, \ldots, n(\mathcal{T})$, then one has essentially constructed a truncated unconditional reference set, of which can be utilized to estimate the w^{th} column upon $\mathbf{P}$.

3.5.3.4 An Iterative Algorithm for Obtaining a Truncated Unconditional Reference Set

To this end, for a given $\epsilon > 0$ and $\boldsymbol{\theta}_w \in \mathcal{O}$, evaluation of (3.16) assumes that one possesses some collection $\Gamma(\boldsymbol{\theta}_w)$, for all $u = 1, \ldots, n(\mathcal{T})$. Here, given $\boldsymbol{\theta}_w \in \mathcal{O}$ and $\epsilon > 0$, we construct a systematic procedure (algorithm) which generates some $\Gamma(\boldsymbol{\theta}_w) \subset \Gamma$ satisfying the condition imposed by (3.16) for all $u = 1, \ldots, n(\mathcal{T})$. In brief, the algorithm determines a value $\delta_w^* \in (0, 1)$, such that for every table $(\mathbf{z}_0, \mathbf{z}_1) \in \Gamma$, $(\mathbf{z}_0, \mathbf{z}_1) \in \Gamma(\boldsymbol{\theta}_w)$ if and only if $\bigcap_{y \in \mathcal{Y}} \{h(\mathbf{z}_y | \boldsymbol{\theta}_w) > \delta_w^*\}$ implies (3.16) holds for all $u = 1, \ldots, n(\mathcal{T})$. Given an initial approximation to δ_w^*, denoted δ_1, the algorithm evaluates the condition imposed by (3.16), for all $u = 1, \ldots, n(\mathcal{T})$. If said condition fails for some value of u, the approximation of δ_w^* is updated to some [lesser] value, denoted δ_2, and condition (3.16) is evaluated (for all $u = 1, \ldots, n(\mathcal{T})$) upon this updated estimate of δ_w^*. This process of: updating the estimate of δ_w^*; and, evaluation – across the possible values in u – of the condition imposed by (3.16), continues until which time said condition is satisfied for all $u = 1, \ldots, n(\mathcal{T})$.

Prior to stating the algorithm, we outline two strategies for efficient implementation thereof: we refine the above approach, so that evaluation of the condition (3.16) is to occur upon a single [strategically selected] value of u over the collection $\{1, \ldots, n(\mathcal{T})\}$. We demonstrate that adherence to the condition imposed by (3.16) for the selected value in u is sufficient, so that said condition is satisfied for all $u = 1, \ldots, n(\mathcal{T})$; and, to 'prime' the iterative process, we designate a value for δ_1. First, note that for any $\Gamma(\boldsymbol{\theta}_w) \subset \Gamma$, the inequality (3.16) is most sensitive in failing to hold true for small values of $p_{u,w}(\Gamma(\boldsymbol{\theta}_w))$, because ϵ and $e(\Gamma(\boldsymbol{\theta}_w))$ are constants. By construction of $\Gamma_A(\tau_u)$ (3.8), the $p_{u,w}(\Gamma(\boldsymbol{\theta}_w))$ are decreasing for increasing $\tau_u \in \mathcal{T}$. This implies that the inequality (3.16) is most sensitive in failing to hold true for large values of $\tau_u \in \mathcal{T}$. So, let

$$\iota = \{s, s = 1, \ldots, n(\mathcal{T}) : \tau_s = \max\{\tau_u \in \mathcal{T}\}\}.$$

Hence, for any $\Gamma(\boldsymbol{\theta}_w) \subset \Gamma$, if the condition of (3.16) holds taking $u = \iota$, then it holds for all $u = 1, \ldots, n(\mathcal{T})$. Therefore, within our iterative algorithm (Algorithm 3.4), we evaluate the condition (3.16) solely upon $u = \iota$.

Second, without loss of generality, suppose

$$p^{o}_{\iota,w} \doteq \Pr\left(X \geq \tau_{\iota} | \tilde{Q}_0\right),$$

where $X \sim \tilde{Q}_0$ – the asymptotic chi-square distribution with degrees-of-freedom equal to one – under $\mathcal{H}_0$. Thus, upon a properly chosen value for δ^*_w, the left hand side of (3.16) will essentially satisfy

$$p_{\iota,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) \doteq \Pr\left(X \geq \tau_{\iota} | \tilde{Q}_0\right).$$

for which we have

$$e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) \leq \left(\frac{1-\epsilon}{\epsilon}\right) \Pr\left(X \geq \tau_{\iota} | \tilde{Q}_0\right). \tag{3.17}$$

Moreover, pursuant to the opening paragraph within this section, by (3.14) the value of δ^*_w must satisfy

$$e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) \leq \delta^*_w \left\{n\left(\Gamma'_0\left(\boldsymbol{\theta}_w\right)\right) + n\left(\Gamma'_1\left(\boldsymbol{\theta}_w\right)\right)\right\}.$$

Taking $e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ at the upper bound of this inequality and substituting within (3.17), we have

$$\delta^*_w \leq \left(\frac{1-\epsilon}{\epsilon}\right) \frac{\Pr\left(X \geq \tau_{\iota} | \tilde{Q}_0\right)}{n\left(\Gamma'_0\left(\boldsymbol{\theta}_w\right)\right) + n\left(\Gamma'_1\left(\boldsymbol{\theta}_w\right)\right)} < \left(\frac{1-\epsilon}{\epsilon}\right) \Pr\left(X \geq \tau_{\iota} | \tilde{Q}_0\right). \tag{3.18}$$

Since the right hand side of the strict inequality within (3.18) is a constant, we utilize the value in said constant as our designated choice for δ_1. We are now poised to state the algorithm. It is given by Algorithm 3.4.

Algorithm 3.4 An Iterative Algorithm for Generating a Truncated Unconditional Reference Set

1. Initialize the value of s – a counter, indicating the number of "visits" to this step of the algorithm – to one. In conjunction with (3.18), let δ_s be defined by[7]

$$\delta_s = \left(\frac{1-\epsilon}{\epsilon}\right) \Pr\left(X \geq \tau_{\iota} | \tilde{Q}_0\right) I(s=1) + \left(\frac{\delta_{s-1}}{10}\right) I(s>1), \tag{3.19}$$

[7] This definition of δ_s is a suggestion. The user may choose to define δ_s in some other manner, under the condition that $\delta_{s+1} < \delta_s$ for all $s \in \mathbb{N}$.

where $\delta_0 = 0$. For each $y \in \mathcal{Y}$, let the collection $\Gamma'_y(\boldsymbol{\theta}_w)$ (the elements of Γ_y not preserved within the truncated unconditional reference set – defined following (3.10)) be given by

$$\Gamma'_y(\boldsymbol{\theta}_w) = \{\mathbf{z}_y \in \Gamma_y : h(\mathbf{z}_y|\boldsymbol{\theta}_w) \leq \delta_s\}. \tag{3.20}$$

2. Compute the value of $e(\Gamma(\boldsymbol{\theta}_w))$ by way of (3.14).

3. Define the collection $\Gamma(\boldsymbol{\theta}_w)$ by way of (3.11).

4. Compute $p_{\iota,k}(\Gamma(\boldsymbol{\theta}_w))$ by way of (3.13).

5. Evaluate the inequality within (3.16), for the computed value $p_{\iota,w}(\Gamma(\boldsymbol{\theta}_w))$. If this value fails to adhere to said inequality, then $\Gamma(\boldsymbol{\theta}_w)$ is unsatisfactory for the given ϵ – increment the value of s by one and proceed to step 1 above, omitting the initial sentence thereof. On the other hand, if $p_{\iota,w}(\Gamma(\boldsymbol{\theta}_w))$ satisfies the inequality within (3.16), then $\Gamma(\boldsymbol{\theta}_w)$ properly defines a truncated unconditional reference set for the specified value of ϵ – let $\delta^*_w = \delta_s$ and terminate the algorithm. ■

In brief, given $\boldsymbol{\theta}_w$ and ϵ, this algorithm searches for a truncated unconditional reference set, $\Gamma(\boldsymbol{\theta}_w)$, for which the computed value $p_{u,w}(\Gamma(\boldsymbol{\theta}_w))$ satisfies (3.16), for all $u = 1, \ldots, n(\mathcal{T})$. It does this in an iterative manner, by truncating the collection $\Gamma'_y(\boldsymbol{\theta}_w)$ by way of the value of δ_s (3.19). In each iteration through the steps of the algorithm, δ_s becomes smaller by an order of magnitude equal to 0.1 (step 1), decreasing the number of elements comprising $\Gamma'_y(\boldsymbol{\theta}_w)$ (step 1), thereby decreasing the upper bound in the incurred error, $e(\Gamma(\boldsymbol{\theta}_w))$ (step 2). This increases the number of elements contained within the truncated unconditional reference set, $\Gamma(\boldsymbol{\theta}_w)$ (step 3), thereby producing a more precise estimate $p_{\iota,w}(\Gamma(\boldsymbol{\theta}_w))$ (step 4).

3.5.3.5 Examples for the Implementation of Algorithm 3.4

Example 3.1.

Consider a balanced GWAS of $n = 2\text{K}$ study subjects, and $\boldsymbol{\theta}_w = \left(\pi^{aa}, \pi^{Aa}\right) = (0.25, 0.50)$ (this is equivalent to genotypes at a SNP locus adhering to HWE with MAF equal to 0.5). Here, we assign ϵ to be the value, such that $100\epsilon\% = 99.99\%$, where it is assumed that $\tau_\iota = 40$ for $\mathbf{P}$ (see §3.6.1.1 for the motivation in assigning this value of τ_ι). We utilize Algorithm 3.4 to define a

truncated unconditional reference set $\Gamma(\boldsymbol{\theta}_w)$ for the CATT statistic, assuming the additive genetic model of inheritance under $H_a^{(j)}$, some $j = 1, \ldots, m$.

Table 3.4 summarizes the application of Algorithm 3.4 for this example – implementation of the algorithm is carried out by way of the programming code presented within §D.1 of Appendix D. Overall, to attain the desired precision in the estimate of $p^{\circ}_{u,w}$ for the given value of ϵ, all $u = 1, \ldots, \mathcal{T}$, we find that the collection $\Gamma_y(\boldsymbol{\theta}_w)$, each $y \in \mathcal{Y}$, comprises only about 7.3% (36705/501501) of the total elements for the collection Γ_y. Moreover, the number of elements comprising the truncated unconditional reference set is about 0.5% $(36705/501501)^2$ of that for Γ, and the integrity of the PPT remains intact. The total time needed to generate the estimate $p_{\iota,w}(\Gamma(\boldsymbol{\theta}_w))$ – given within the fifth iteration of the algorithm – is 56.2 seconds. In extrapolating, this implies approximately 2.9 hours would be required to attain the exact value of $p^{\circ}_{\iota,w}$ over Γ. Moreover, there is a negligible difference in the estimate of $p^{\circ}_{\iota,w}$ between the final two iterations of the algorithm (column 4), indicating superior precision in the estimate of $p^{\circ}_{\iota,w}$ for the chosen value of ϵ. Therefore, this example illustrates the efficiency of utilizing a truncated unconditional reference set over that of Γ, when one is willing to incur a minute amount of error in estimating the elements of $\mathbf{P}$. ■

Table 3.4: Summary Measures for the Implementation of Algorithm 3.4 Applied to Example 3.1.

Iteration	$e(\Gamma(\boldsymbol{\theta}_w))$	$n(\Gamma_0(\boldsymbol{\theta}_w))$	$p_{\iota,w}(\Gamma(\boldsymbol{\theta}_w))^{\dagger}$	δ_s	Marginal Time (s)
1	5.55×10^{-11}	26728	2.26021	2.5×10^{-14}	28.3
2	5.53×10^{-12}	29233	2.26305	2.5×10^{-15}	6.1
3	5.52×10^{-13}	31731	2.26363	2.5×10^{-16}	6.7
4	5.42×10^{-14}	34241	2.26373	2.5×10^{-17}	7.3
5	5.53×10^{-15}	36705	2.26375	2.5×10^{-18}	7.8

†Depicted values are divided by 10^{-10}.

Example 3.2.

Consider a balanced GWAS of $n = 10$K study subjects, and $\boldsymbol{\theta}_w = \left(\pi^{aa}, \pi^{Aa}\right) = (0.25, 0.50)$. We assign ϵ to be the value, such that $100\epsilon\% = 99.999\%$, where $\tau_\iota = 40$ for $\mathbf{P}$. We utilize Algorithm 3.4 to define a truncated unconditional reference set $\Gamma(\boldsymbol{\theta}_w)$ for the CATT statistic, assuming the additive genetic model of inheritance under $H_a^{(j)}$, some $j = 1, \ldots, m$.

Table 3.5 summarizes the application of the algorithm for this example. Overall, to attain the desired precision in the estimate of $p^{\circ}_{u,w}$ for the given value of ϵ, all $u = 1, \ldots, \mathcal{T}$, we find that each of the collections $\Gamma_y(\boldsymbol{\theta}_w)$ comprises only about 1.6% (202111/12507501) of the total elements for the collection Γ_y, $y \in \mathcal{Y}$. Moreover, the number of elements comprising the truncated unconditional

reference set is about 0.03% $(202111/12507501)^2$ of that for Γ, and the integrity of the PPT remains intact. The total time needed to generate the estimate $p_{\iota,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$ – given within the sixth iteration of the algorithm – is 29.4 minutes. In extrapolating, this implies approximately 78 days would be required to attain the exact value of $p^{\circ}_{\iota,w}$ over Γ. Moreover, there is a negligible difference in the estimate of $p^{\circ}_{\iota,w}$ between the final two iterations of the algorithm (column 4), indicating superior precision in the estimate of $p^{\circ}_{\iota,w}$ for the chosen value of ϵ. Therefore, as with Example 3.1, this example illustrates the efficiency of utilizing a truncated unconditional reference set over that of Γ, when one is willing to incur a minute amount of error in estimating the elements of $\mathbf{P}$. ■

Table 3.5: Summary Measures for the Implementation of Algorithm 3.4 Applied to Example 3.2.

Iteration	$e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$	$n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right)$	$p_{\iota,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)^{\dagger}$	δ_s	Marginal Time (m)
1	2.80×10^{-11}	138508	2.478877	2.5×10^{-15}	12.9
2	2.80×10^{-12}	151247	2.480751	2.5×10^{-16}	2.8
3	2.80×10^{-13}	163974	2.481116	2.5×10^{-17}	3.0
4	2.81×10^{-14}	176666	2.481180	2.5×10^{-18}	3.3
5	2.79×10^{-15}	189412	2.481191	2.5×10^{-19}	3.6
6	2.80×10^{-16}	202111	2.481193	2.5×10^{-20}	3.8

†Depicted values are divided by 10^{-10}.

Example 3.3.

Here, we desire to compare the performance of Algorithm 3.4 at $\tau_\iota = 40$, across: (1) balanced GWAS samples of sizes $n \in \{1\text{K}, 2\text{K}, 5\text{K}, 10\text{K}\}$; (2) $\boldsymbol{\theta}_w \in \{(0.25, 0.50), (0.063, 0.38), (0.01, 0.18)\}$ (equivalent to genotypes at three SNP loci adhering to HWE with respective MAFs 0.50, 0.25, and 0.10); and (3) $100\epsilon\% \in \{99.99\%, 99.999\%\}$. We utilize said algorithm to define the truncated unconditional reference set $\Gamma\left(\boldsymbol{\theta}_w\right)$ for the CATT statistic, assuming the additive genetic model of inheritance under $H_a^{(j)}$, some $j = 1, \ldots, m$.

Table 3.6 summarizes the application of the algorithm for this example. For a fixed choice in ϵ, these data indicate that Algorithm 3.4 becomes increasingly more efficient (here, efficiency is relative to generating $p^{\circ}_{\iota,w}$ from the entire collection Γ): as the minor allele frequency decreases, across the sample sizes chosen for this example. For example, consider $\epsilon = 0.9999$ and $n = 1\text{K}$. Relative to a locus with population MAF equal to 0.50, the size of Γ_0 for loci with respective population MAFs equal to 0.25 and 0.01 are 60% (11133/18477) and 19%; and, for a fixed minor allele frequency, increasing sample size. For example, consider $\epsilon = 0.9999$ and MAF equal to 0.50. Relative to the sample size $n = 1\text{K}$, the magnitude in the ratio $n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right) n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right) / n\left(\Gamma\right)$ for the respective sample sizes $n = 2\text{K}$, $n = 5\text{K}$, and $n = 10\text{K}$ are 25% $(\{36705 \times 125751/(501501 \times 18477)\}^2)$, 4%, and

Table 3.6: Summary Measures for the Implementation of Algorithm 3.4 Applied to Example 3.3.

n	$\boldsymbol{\theta}_w$	$100\epsilon\%$	Iterations	$e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)^{\dagger}$	$n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right)$	$p_{\iota,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)^{\ddagger}$	Time (s)
1K	$(0.25, 0.50)$	99.99%	5	2.6	18477	2.036553	14.7
		99.999%	5	0.26	19675	2.036555	16.4
	$(0.063, 0.38)$	99.99%	4	8.9	11133	1.885308	5.2
		99.999%	4	0.88	11856	1.885314	6.4
	$(0.01, 0.18)$	99.99%	4	5.4	3476	1.255795	0.7
		99.999%	4	0.56	3709	1.255797	0.8
2K	$(0.25, 0.50)$	99.99%	5	5.5	36705	2.263751	56.2
		99.999%	5	0.55	39195	2.263754	67.0
	$(0.063, 0.38)$	99.99%	5	3.5	23947	2.163368	24.5
		99.999%	5	0.35	25514	2.163370	28.0
	$(0.01, 0.18)$	99.99%	4	10.0	6966	1.795663	2.8
		99.999%	4	1.0	7426	1.795667	3.1
5K	$(0.25, 0.50)$	99.99%	5	14.0	90017	2.424471	353.0
		99.999%	5	1.4	96357	2.424477	408.5
	$(0.063, 0.38)$	99.99%	5	9.0	59060	2.384602	155.1
		99.999%	5	0.89	63149	2.384606	187.7
	$(0.01, 0.18)$	99.99%	5	2.6	19390	2.215479	21.7
		99.999%	5	0.26	20585	2.215481	24.1
10K	$(0.25, 0.50)$	99.99%	6	2.8	189412	2.481191	1594.4
		99.999%	6	0.28	202111	2.481193	1761.9
	$(0.063, 0.38)$	99.99%	5	18.0	116141	2.460379	600.7
		99.999%	5	1.8	124373	2.460387	668.9
	$(0.01, 0.18)$	99.99%	5	5.8	39377	2.373231	88.8
		99.999%	5	0.58	42025	2.373234	96.7

†Depicted values are divided by 10^{-15}.
‡Depicted values are divided by 10^{-10}.

1%. Finally, these data indicate very subtle differences in the estimates of $p^{\circ}_{\iota,w}$ between the chosen values of ϵ across both the chosen values of MAF and the chosen sample sizes. This suggests that the choice of ϵ equal to 0.9999 will suffice for high precision in the estimates of $p^{\circ}_{u,w}$ over $\mathbf{P}$.

Figure 3.11 displays the ratio (negative natural logarithm thereof) in the number of elements comprising the truncated unconditional reference set (from applying Algorithm 3.4) to that of the unconditional reference set, against population MAF for several balanced GWAS samples sizes, assuming HWE amongst population genotype frequencies and $\epsilon = 0.9999$. This plot is in direct agreement with our observations regarding Table 3.6. Namely, for any fixed value in the population MAF, these data indicate the efficiency of Algorithm 3.4 increases for increasing sample size. This notion is seen by the greater values in $-\log\left(n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right) n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right)/n\left(\Gamma\right)\right)$ for greater values in n. Moreover, this plot also suggests the efficiency for the algorithm increases for decreasing values in MAF, irrespective of sample size. This is seen by way of the decreasing trend in any of the colored

curves for increasing values in MAF. Overall, in assigning $\epsilon = 0.9999$ within Algorithm 3.4, this example suggests the ascertainment of high precision estimates of $p^{o}_{u,w}$ over $\mathbf{P}$. ■

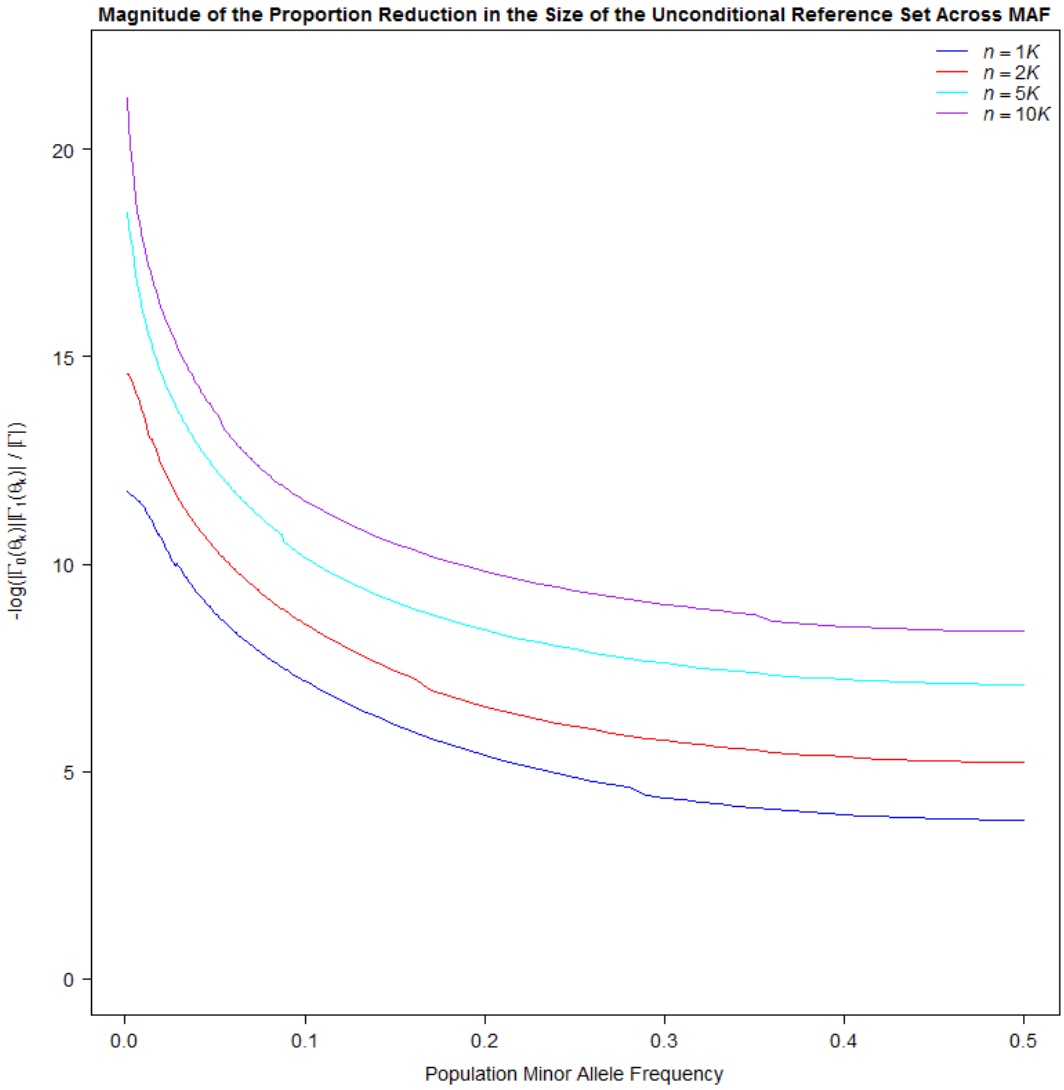

Fig. 3.11: Plot of Negative $\log\left(n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right) n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right) / n\left(\Gamma\right)\right)$ – the Ratio (Natural Log Scale) of the Number of Elements Comprising the Truncated Unconditional Reference Set to That of the Unconditional Reference Set – Against Population Minor Allele Frequency for Balanced GWAS Samples of Varying Sizes, Assuming Hardy-Weinberg Equilibrium among Population Genotype Frequencies and $\epsilon = 0.9999$.

3.5.3.6 An Algorithm to Rapidly Generate the Estimated *P*-value Lookup Table

Having motivated the utility of a truncated unconditional reference set in generating the values upon any column of the estimated PPT $\mathbf{P}$, and having developed an algorithm for the con-

struction of a truncated unconditional reference set, we now develop a parallel processing approach to efficiently generate all values comprising $\mathbf{P}^{\epsilon}$, where $\mathbf{P}^{\epsilon}$ denotes the estimate of $\mathbf{P}$ given the user specified value of ϵ. To motivate a parallel approach over that of a strictly serial approach, note that the latter approach to constructing $\mathbf{P}^{\epsilon}$ could entail the following pseudocode:

1: **for** each $\boldsymbol{\theta}_w \in \mathcal{O}$ **do**
2: Construct $\Gamma(\boldsymbol{\theta}_w)$ in accordance with Algorithm 3.4.
3: **for** each $\tau_u \in \mathcal{T}$ **do**
4: $p_{u,w}(\Gamma(\boldsymbol{\theta}_w)) \leftarrow 0$. {Initialize the p-value estimate to zero}.
5: **for** each table $(\mathbf{z}_0, \mathbf{z}_1) \in \Gamma(\boldsymbol{\theta}_w)$ **do**
6: **if** $(\mathbf{z}_0, \mathbf{z}_1) \in (\Gamma(\boldsymbol{\theta}_w) \cap \Gamma_A(\tau_u))$ **then**
7: $p_{u,w}(\Gamma(\boldsymbol{\theta}_w)) \leftarrow p_{u,w}(\Gamma(\boldsymbol{\theta}_w)) + g(\mathbf{z}_0, \mathbf{z}_1|\boldsymbol{\theta}_w)$. {Table contributes to the p-value. Increment the p-value by the probability of the table being realized under $\mathcal{H}_0$}.
8: **end if**
9: **end for**
10: **end for**
11: **end for**

There are several problems with this approach. First, this strictly serial approach in generating $\mathbf{P}^{\epsilon}$ is implausible whenever the number of elements comprising a particular truncated unconditional reference set is large, because excessive computational time would be required to generate the corresponding column upon $\mathbf{P}^{\epsilon}$. This is due to the third loop within the above pseudocode (lines 5-9 therein) being comprised of a large number of elements in such circumstances.
Based upon the results obtained within the examples (§3.5.3.5) – particularly, the observed values of $n(\Gamma_0(\boldsymbol{\theta}_w))$ depicted within Tables 3.4-3.6 – we see that the value of $n(\Gamma(\boldsymbol{\theta}_w))$ is likely to be considerably large, because said value is the product of the [quite large] factors $n(\Gamma_0(\boldsymbol{\theta}_w))$ and $n(\Gamma_1(\boldsymbol{\theta}_w))$. Moreover, based upon these examples, we see that this notion is increasingly exacerbated for increasing values in MAF and appears to hold true even for small (i.e., 0.10) values in MAF.

Second, while the programming code of §D.1 is well suited to construct the collections $\Gamma(\boldsymbol{\theta}_w)$, all $w = 1 \ldots, n(\mathcal{O})$, in the circumstance for which the column dimension of $\mathbf{P}$ is small (say, not larger than the value of ten (10)), it is computationally intractable for our purposes, since here we assume the column dimension of $\mathbf{P}$ to be in the thousands. We assume such a column dimension over $\mathbf{P}$,

because we desire high precision in our estimates of $p_{j,r}$ within the minP MTP, all $j = 1, \ldots, m$ and $r = 1, \ldots, R$ (see Algorithm 3.3). For example, consider $n = 2\text{K}$ upon a balanced GWAS, $\epsilon = 0.9999$, and $n(\mathcal{O}) \geq 1000$. Hypothetically speaking, suppose the average computational time to construct $\Gamma(\boldsymbol{\theta}_w)$ – applying the programming code of §D.1 – all $w = 1, \ldots, n(\mathcal{O})$, is 29 seconds (the arithmetic average of the computational times presented upon rows 7 and 11 of Table 3.6). Under this presumption, the appropriate truncated unconditional reference sets, $\{\Gamma(\boldsymbol{\theta}_w)\}_{w=1,\ldots,n(\mathcal{O})}$, would be constructed in roughly 8 hours time for $n(\mathcal{O}) = 1000$. This is too much time to allot in generating the truncated unconditional reference sets over $\mathbf{P}$.

As an alternative to the above serial computing approach in generating the elements upon $\mathbf{P}^{\epsilon}$, we propose a parallel computing approach, based upon the CUDA C programming model, comprised of five CUDA kernels. The first three kernels, denoted TURK1 (TURK is shorthand for truncated unconditional reference set kernel), TURK2, and TURK3, respectively, work in collaboration to generate the collection $\{\delta_w^*\}_{w=1,\ldots,n(\mathcal{O})}$. The final two kernels, denoted PPTK1 (PPTK is shorthand for pointwise p-value table kernel) and PPTK2, respectively, work collaboratively in deriving the values $p_{u,w}(\Gamma(\boldsymbol{\theta}_w))$ from the defined collection $\{\delta_w^*\}_{w=1,\ldots,n(\mathcal{O})}$, for all $u = 1, \ldots, n(\mathcal{T})$ and $w = 1, \ldots, n(\mathcal{O})$. For the underlying details encompassing these kernels, see Algorithm B.7 within §B.2 of Appendix B. Algorithm 3.5 outlines our approach to generating $\mathbf{P}^{\epsilon}$.

Algorithm 3.5 Generating the Estimated Pointwise P-value Lookup Table

---

1. Let $\mathcal{W} = \{1, \ldots, n(\mathcal{O})\}$ warehouse the column indices upon $\mathbf{P}$; let $\epsilon > 0$ be given; and let the value of $\delta = \delta_1$, where δ_1 is as specified within step 1 of Algorithm 3.4. We assume the elements within $\mathcal{T}$ are ordered such that

$$\tau_1 < \tau_2 < \cdots < \tau_{\iota},$$

where ι is as defined within §3.5.3.4. Within steps 2–4 to follow, we are essentially invoking Algorithm 3.4, simultaneously, upon all $\boldsymbol{\theta}_w \in \mathcal{O}$.

2. For each $y \in \mathcal{Y}$, we invoke the TURK1 kernel, where the return thereof is the collection

$$\left\{ \sum_{\mathbf{z}_y \in \Gamma_y'(\boldsymbol{\theta}_w)} h(\mathbf{z}_y | \boldsymbol{\theta}_w) \right\}_{w \in \mathcal{W}}.$$

3. For each $w \in \mathcal{W}$:

   a. Compute the value of $e\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$, by evaluating the two sums within (3.14).

   b. For each $y \in \mathcal{Y}$, we invoke the TURK2 kernel, where the return thereof is the collection $\Gamma_y\left(\boldsymbol{\theta}_w\right)$.

   c. We invoke the TURK3 kernel, where the return is the value $p_{\iota,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$.

   d. Evaluate the condition (3.16), where $u = \iota$ therein. If the condition is true, then let $\delta_w^* = \delta$ and update the collection $\mathcal{W}$ to $\mathcal{W}\backslash\{w\}$ (i.e., extract the element w from said collection).

4. If $\mathcal{W} = \emptyset$, then the collection $\{\delta_w^*\}_{w=1,\ldots,n(\mathcal{O})}$ has been determined. Let $\mathcal{W} = \{1,\ldots,n\left(\mathcal{O}\right)\}$ and proceed to step 5; otherwise, update the value of δ to one-tenth its value and proceed to step 2 (essentially, we are iterating through Algorithm 3.4 upon the remaining elements contained within $\mathcal{W}$).

5. For each $w \in \mathcal{W}$:

   a. For each $y \in \mathcal{Y}$, invoke the TURK2 kernel where the return thereof is the collection $\Gamma_y\left(\boldsymbol{\theta}_w\right)$.

   b. Invoke the PPTK1 kernel as follows:

      A. For each $\mathbf{z}_0 \in \Gamma_0\left(\boldsymbol{\theta}_w\right)$ and $\mathbf{z}_1 \in \Gamma_1\left(\boldsymbol{\theta}_w\right)$ under $H_0^{(j)}$: compute the realization in the CATT statistic – denote it by $T\left(\mathbf{z}_0, \mathbf{z}_1\right)$; and compute the table probability corresponding to the element $\left(\mathbf{z}_0, \mathbf{z}_1\right) \in \Gamma$, $g\left(\mathbf{z}_0, \mathbf{z}_1 | \boldsymbol{\theta}_w\right)$.

      B. For each $\mathbf{z}_0 \in \Gamma_0\left(\boldsymbol{\theta}_w\right)$ and $\mathbf{z}_1 \in \Gamma_1\left(\boldsymbol{\theta}_w\right)$, determine the value of $\lambda\left(\mathbf{z}_0, \mathbf{z}_1\right)$, where

      $$\lambda\left(\mathbf{z}_0, \mathbf{z}_1\right) = \min\left\{\min\left\{u, u = 1,\ldots,n\left(\mathcal{T}\right) : T\left(\mathbf{z}_0, \mathbf{z}_1\right) \leq \tau_u,\ \ \tau_u \in \mathcal{T}\right\}, \iota\right\}. \tag{3.21}$$

      C. Return the collections $\left\{g\left(\mathbf{z}_0, \mathbf{z}_1 | \boldsymbol{\theta}_w\right)\right\}_{\left(\mathbf{z}_0,\mathbf{z}_1\right)\in\Gamma\left(\boldsymbol{\theta}_w\right)}$ and $\left\{\lambda\left(\mathbf{z}_0, \mathbf{z}_1\right)\right\}_{\left(\mathbf{z}_0,\mathbf{z}_1\right)\in\Gamma\left(\boldsymbol{\theta}_w\right)}$.

   c. Invoke the PPTK2 kernel as follows:

      A. For each $\tau_u \in \mathcal{T}$, we compute the value $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$. It is,

      $$p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) = \sum_{\left(\mathbf{z}_0,\mathbf{z}_1\right)\in\Gamma\left(\boldsymbol{\theta}_w\right)} I\left(\lambda\left(\mathbf{z}_0, \mathbf{z}_1\right) \geq u\right) g\left(\mathbf{z}_0, \mathbf{z}_1 | \boldsymbol{\theta}_w\right). \tag{3.22}$$

3.6 Proof of Concept

To illustrate the integrity of the proposed synergistic methodologies – outlined within §3.4 and §3.5 – in providing balanced and accurate control of the FWER at the 5% level for the CATT statistic, we conducted a large-scaled simulation investigation (§3.6.1); and to illustrate the application of the methodologies in practice, we applied them against a large GWAS data set (§3.6.2).

3.6.1 Simulation

Under assumed HWE for genotype frequencies across loci, by way of simulation we will suggest: (1) the exact unconditional distribution $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ for the CATT statistic under $\mathcal{H}_0$ within the minP MTP, provides balanced and accurate control of the FWER at the 5% level;[8] and (2) the exact unconditional distribution $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$, correctly identifies the null distribution for the CATT statistic under $H_0^{(j)}$.

3.6.1.1 Methods

Indeed, to illustrate the aforementioned two notions, we analyzed – simultaneously, alongside the simulations conducted within §3.2.4.1 – the identical $D = 100$K simulated balanced (recall, denoted simulation 1 (S1)) and unbalanced (recall, a two to one ratio of controls to cases and denoted simulation 2 (S2)) GWAS data sets, each data set comprised of $n = 1200$ subjects and $m = 10$K SNP loci. The CATT statistic was used to test each of the null hypotheses, $H_0^{(j)}$, assuming the additive genetic model of inheritance under the corresponding two-sided alternative hypothesis. The minP and Šidák MTPs were utilized to control the FWER within each data set, where pointwise p-values were calculated under $\left\{Q_{0j}^{(*H)}\right\}_{j=1,\ldots,m}$, taking the unknown population minor allele frequency (π_j) at its corresponding MLE value $\hat{\pi}_j$, all $j = 1, \ldots, m$. For each data set, $R = 2048$ (i.e., two blocks of 2^{10} permutations) random shuffles of the column indices upon the genotype matrix $\mathbf{G}^{(*\rho)}$ were applied within the GPER algorithm (see §2.4) for the minP MTP. Recall, each of the D data sets was simulated – each simulation, S1 and S2 – under assumed HWE among population genotype frequencies. Thus, $Q_{0j}^{(*H)}(\pi_j, n_0, n_1)$ correctly identifies the true underlying distribution for the

[8] Here, we assume HWE for both simplicity in illustration and to serve as an analogue to the HWE assumption made upon the simulation conducted within §3.2.4.1.

CATT statistic under $H_0^{(j)}$, across the D data sets. Hence, we expect the results here to confirm the two notions of §3.6.1.

To resolve the computational problem – in utilizing $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ to compute pointwise p-values under $H_0^{(j)}$ – we adopted a modified version of the computational tools outlined within Algorithm 3.3 (i.e., steps 1-5 therein). The method is as follows: step 1 – as mentioned within the aforementioned paragraph, within each simulated data set we estimated the parameter π_j with its MLE $\hat{\pi}_j$. To define our subspace about the parameter space of π_j, we first took note of the fact that the parameter space for π_j is the interval of the reals, $[0, 0.5]$, although our simulation was restricted to $\pi_j \in \{0.01, 0.02, \ldots, 0.1\}$. Hence, we recognized that our subspace was to be a subset of the interval $[0, 0.5]$, of which we denoted by $[s_l, s_u]$. At the onset of the simulation, we did not know off-hand what range of values the MLEs $\{\hat{\pi}_j\}_{j=1,\ldots,m}$ would comprise across the D simulated data sets (true of each simulation, S1 and S2), which made formulation of this subspace – through the values of s_l and s_u – a bit non-trivial. However, under the HWE model for genotype frequencies, it can be shown (see Proposition A.7 within Appendix A) that $E(\hat{\pi}_j) = \pi_j$ and $Var(\hat{\pi}_j) = \pi_j(1-\pi_j)/(2n)$. By the Central Limit Theorem, it follows – since $\hat{\pi}_j$ is a function involving a sum of random variables – that $\hat{\pi}_j \dot{\sim} N\left(\pi_j, \sqrt{\pi_j(1-\pi_j)/(2n)}\right)$. We arbitrarily considered assigning $s_l = 0.001$ and $s_u = 0.12$. Hence,

$$\max\left\{\Pr\left(\hat{\pi}_j \leq s_l | \pi_j = 0.01\right), \Pr\left(\hat{\pi}_j \geq s_u | \pi_j = 0.1\right)\right\} \doteq 5 \times 10^4, \tag{3.23}$$

for which we determined that our defined subspace, $[s_l, s_u] = [0.001, 0.12]$, should possess excellent coverage in capturing the majority of elements over the collection $\{\hat{\pi}_j\}_{j=1,\ldots,m}$ across the D simulated data sets. We partitioned this subspace into equal length subintervals, each of length 0.001, which yielded the collection of unconditional distributions for the CATT statistic, $\left\{Q_0^{(*H)}(w/1000, 600, 600)\right\}_{w=1,\ldots,120}$ (for S1) and $\left\{Q_0^{(*H)}(w/1000, 800, 400)\right\}_{w=1,\ldots,120}$ (for S2), to be utilized within step 2 of Algorithm 3.3 (to be elaborated upon below), where $Q_0^{(*H)}(\pi, n_0, n_1)$ denotes the exact unconditional distribution for the CATT statistic under the null hypothesis of no genotype-phenotype association at a locus with population minor allele frequency π, such that genotype frequencies at the locus adhere to HWE.

Next, to define our partition of the domain for the CATT statistic, we first took note that said domain is all positive real numbers. Rather than partitioning the entire positive real line into disjoint intervals, we decided to partition a subset of the positive reals, say $[0, \tau_\iota] \subset \mathbb{R}$, where the

value of τ_ι (recall, equal to the max over $\mathcal{T}$) was to be assigned such that the likelihood of the CATT statistic attaining a value at least as extreme as τ_ι, just by chance under $\mathcal{H}_0$, is essentially zero. We then recognized that under $\mathcal{H}_0$, the likelihood of observing extremely large realizations of this statistic over the interval $\pi_j \in [s_l, s_u]$, say $t_j \geq 40$, is very small, where this notion holds even upon the aggregated data sets comprising 500K mutually independent SNP loci (recall, we randomly aggregated ten/fifty data sets – each comprised of $m = 10$K loci – together to obtain simulated data sets comprised of 100K/500K mutually independent SNP loci under $\mathcal{H}_0$). To illustrate, under $\mathcal{H}_0$, the expected number of [mutually independent] SNP loci to attain a realization in the CATT statistic at least equal to the value forty, just by chance upon S1, is given by

$$\begin{aligned}(m)\Pr\left(T_j \geq 40 | Q_{0j}^{(*H)}(\pi_j, 600, 600)\right) &\leq (m) \sup_{\pi \in [s_l, s_u]} \Pr\left(T_j \geq 40 | Q_{0j}^{(*H)}(\pi, 600, 600)\right) \\ &= (m)\Pr\left(T_j \geq 40 | Q_{0j}^{(*H)}(s_u, 600, 600)\right) \\ &= (m)(1.6 \times 10^{-10}) \\ &\leq 7.8 \times 10^{-5} \qquad \text{(taking } m = 500\text{K)},\end{aligned} \tag{3.24}$$

where $T_j \sim Q_{0j}^{(*H)}(\pi_j, 600, 600)$ denotes a random value of the CATT statistic for locus j under $\mathcal{H}_0$. We extended the interpretation of expression (3.24) amongst our entire 2K mutually independent aggregated data sets (each comprised of 500K mutually independent SNP loci), so that prior to conducting the simulation we expected to see less than one realization of the CATT statistic within the observed (non-permuted) data taking a value at least as large as 40.[9] Hence, we defined the upper bound for our interval of the domain for the CATT statistic, $[0, \tau_\iota]$, by $\tau_\iota = 40$. We chose to partition the interval, $[0, \tau_\iota]$, into disjoint subintervals, each of length 0.05. Our reasoning was that the pointwise p-values should be very similar for CATT statistic realizations t and t', such that $|t - t'| = 0.05$. This yielded the collection of CATT statistic realizations, $\{5u/100\}_{u=1,\ldots,800}$, by which to compute pointwise p-values within step 2 of Algorithm 3.3.

Steps 2 and 3 – taking $\epsilon = 0.99999$ within Algorithm 3.5, we computed each of the pointwise p-values over the elements comprising: the joint collection over $\left\{Q_0^{(*H)}(w/1000, 600, 600)\right\}_{w=1,\ldots,120}$ and $\{5i/100\}_{i=1,\ldots,800}$ (for S1); and, the joint collection over $\left\{Q_0^{(*H)}(w/1000, 800, 400)\right\}_{w=1,\ldots,120}$ and $\{5u/100\}_{u=1,\ldots,800}$ (for S2); step 4 – in other words, our estimated PPT upon each simulation

[9] Note that in performing $R = 2048$ random shuffles of the phenotype labels in the construction of the minP MTP permutation null distributions across the D data sets, just by chance, one would expect to find several realizations $t_{j,r}$ at least equal to forty, where $t_{j,r}$ is as defined within step 3 of the minP pseudocode (Algorithm 3.2). However, this has essentially zero probability of affecting the results of the simulation.

scenario, $\mathbf{P}^{\epsilon}$, was comprised of $800 \times 120 = 96\text{K}$ elements. Table 3.7 summarizes the computational time needed to generate the lookup table $\mathbf{P}^{\epsilon}$ across the two simulation scenarios for two selections in the value of ϵ. These data indicate application of Algorithm 3.5 to be exceptionally efficient, since just slightly more than a single minute of GPU computing time (worst case) was needed to generate the 96K elements upon $\mathbf{P}^{\epsilon}$; and, finally step 5 – within the r^{th} permutation of the minP MTP, we estimated the pointwise p-value $p_{j,r}$, for the realization of the CATT statistic $t_{j,r}$ under $\mathcal{H}_0$, by the $(u,w)^{\text{th}}$ element of our table, $[\mathbf{P}^{\epsilon}]_{u,w}$, where

$$u = \min\left\{\max\left\{\lfloor 20 t_{j,r} \rfloor, 1\right\}, 800\right\}; \text{ and}$$
$$w = \begin{cases} \min\left\{\lceil 1000\hat{\pi}_j \rceil, 120\right\}, & \text{if S1} \\ \min\left\{\max\{\lfloor 1000\hat{\pi}_j \rfloor, 1\}, 120\right\}, & \text{if S2,} \end{cases} \tag{3.25}$$

Table 3.7: Computational Time (Seconds) Needed to Generate $\mathbf{P}^{\epsilon}$ for the Simulations of §3.6.1, Applying Algorithm 3.5.

			Time (s)		
n_1	n_0	ϵ	To Process Steps 1-4	To Process Step 5	Total
400	800	0.99990	4.5	36.0	40.5
400	800	0.99999†	5.4	41.5	46.9
600	600	0.99990	14.6	44.4	59.0
600	600	0.99999†	15.7	49.2	64.9

†Value of ϵ utilized for generation of $\mathbf{P}^{\epsilon}$.

where, recall $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the respective floor and ceiling functions for the argument $(\cdot)$. In an analogous manner, we estimated the pointwise p-value p_j, for the realization of the CATT statistic t_j under $\mathcal{H}_0$ for the observed (non-permuted) data, by the element $[\mathbf{P}^{\epsilon}]_{u,w}$, where the value of u was obtained by substituting the value t_j within (3.25) in lieu of $t_{j,b}$ therein – note: the value of w only needed to be computed once per SNP locus, since said value is permutation invariant. In this regard, pursuant to visual interpretation (e.g., Figure 3.1 (balanced GWAS) and Figure 3.2 (unbalanced GWAS)) and empirical evidence over the generated estimated PPT table, $\mathbf{P}^{\epsilon}$ (upon each of S1 and S2), each of our estimated pointwise p-values – permutation derived or otherwise – was a slightly conservative estimate of their corresponding true underlying bootstrap pointwise p-value counterparts.

3.6.1.2 Results

Table 3.8 summarizes the number of observed Type I errors for the simulated data at the true

underlying 5% level in the FWER, cross-classified by: MTP (minP or Šidák) under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ and $\mathcal{H}_0$; the marker density (m); and the assumed population minor allele frequency (π_j). The data depicted within this table should be compared to those of Table 3.1, on a row-to-row basis. As expected, in ignoring π_j (i.e., collapsing over π_j) these data support the notion that each of the minP and Šidák MTPs under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ control the FWER at the 5% level, since the three 95% confidence intervals cover said level in the FWER across marker densities. Note the remarkable distinction in the observations between the two tables, regarding control of the 5% FWER for the Šidák MTP – this MTP is suggestive of being overly conservative in controlling the 5% FWER when assuming $\tilde{Q}_0$ as the underlying null distribution for the CATT statistic under $\mathcal{H}_0$, where the conservatism appears to be increasing in m; this MTP is suggestive of properly controlling the 5% FWER when assuming $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ as the underlying null distribution for the CATT statistic under $\mathcal{H}_0$, where this notion seems to hold across marker densities. Moreover, these data indicate that the minP and Šidák MTPs under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ control the 5% FWER in a *balanced* manner across π_j, since the number of observed Type I errors for each of these MTPs is roughly uniform across π_j, where this notion holds across the marker densities depicted within the table. Again, we note a remarkable distinction in the observations between the two tables, namely, in assuming $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)/\tilde{Q}_0$ as the underlying null distribution for the CATT statistic under $\mathcal{H}_0$, the minP/maxT MTPs seem to provide balanced/unbalanced control of the 5% FWER across π_j. Finally, the number of Type I errors for these MTPs (minP and Šidák under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$) are roughly identical across π_j and marker densities. This suggests that the minP and Šidák adjusted p-values are essentially identical. In turn, this suggests that the distribution $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ essentially identifies the true underlying null distribution for the CATT statistic under $\mathcal{H}_0$ – at least in the circumstance for which the loci are independent. Overall, these data indicate: (1) that the minP MTP under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ is nearly balanced in its control of the 5% FWER under $\mathcal{H}_0$; and (2) that the distribution $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ essentially identifies the true underlying null distribution for the CATT statistic under $H_0^{(j)}$. These two notions are vital elements when conducting multiplicity adjustment within a GWAS and extensions thereof – particularly, studies entailing loci possessing a rare variant allele – as they ensure the proper reporting of false-positives. Note that under our simulation conditions (i.e., independent SNP loci within a GWAS data set), we have suggested that $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ correctly identifies Q_0 for the CATT statistic under $H_0^{(j)}$.

Figures 3.3 and 3.4 display simultaneous exact Clopper-Pearson 95% confidence intervals for

Table 3.8: Number of Data Sets Exhibiting Some Type I Error Cross-Classified by Multiple Testing Procedure (MTP), the Marker Density (m), and Assumed Minor Allele Frequency (π_j; MAF), Within a Population Whose Genotype Frequencies at Each SNP Locus Adhere to Hardy-Weinberg Equilibrium, Assuming the Cochran-Armitage Trend Test Statistic Is Distributed as $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ under $\mathcal{H}_0$. The True Underlying Family-wise Type I Error Rate (FWER) Is 5%. Assuming Type I Errors Are Independent of MAF, the Expected Number of Type I Errors by MAF Are 500 (m = 10K), Fifty (m = 100K), and Ten (m = 500K). 95% Exact Clopper-Pearson Confidence Intervals (CI) Are for Control in the Overall True Underlying FWER[†].

MTP (m)	Minor Allele Frequency										Totals	Observed FWER (95% CI)
	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1		
minP												
(10K)	562	503	477	536	502	474	467	475	496	452	4944	4.94% (4.8%, 5.1%)
(100K)	53	44	41	54	56	43	45	43	52	51	482	4.82% (4.4%, 5.3%)
(500K)	10	11	9	10	10	10	9	9	7	8	93	4.65% (3.8%, 5.7%)
Šidák												
(10K)	533	540	483	541	510	487	480	491	511	468	5044	5.04% (4.9%, 5.2%)
(100K)	52	43	45	54	54	43	46	41	52	54	484	4.84% (4.4%, 5.3%)
(500K)	10	12	9	9	10	9	9	9	8	9	94	4.70% (3.8%, 5.8%)
minP												
(10K)	502	504	488	490	486	500	498	500	506	498	4972	4.97% (4.8%, 5.1%)
(100K)	51	50	47	45	53	51	50	46	48	56	497	4.97% (4.6%, 5.4%)
(500K)	10	10	12	9	13	10	9	10	11	11	105	5.25% (4.3%, 6.3%)
Šidák												
(10K)	510	506	502	503	487	502	510	506	510	510	5046	5.05% (4.9%, 5.2%)
(100K)	55	49	48	49	55	54	52	48	48	55	513	5.13% (4.7%, 5.6%)
(500K)	10	11	12	9	11	10	10	10	11	11	105	5.25% (4.3%, 6.3%)

[†]Initial [final] six rows correspond to simulation 1 (S1) [simulation 2 (S2)].

control in the FWER across π_j, upon the maxT and Šidák MTPs assuming the CATT statistic is distributed as $\tilde{Q}_0$ under $\mathcal{H}_0$ (the circles and squares depicted within the figure denote the respective observed FWER for these MTPs) and upon the minP and Šidák MTPs assuming the CATT statistic is distributed as $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ under $\mathcal{H}_0$ (the respective triangles and crosses depicted within the figure denote the respective observed FWER for these MTPs) for S1; Figures 3.5 and 3.6 display the analogous information, but for S2 as opposed to S1. These figures illustrate analogous observations to those made for the initial six rows of empirical data of Table 3.8. Namely, we see that the minP and Šidák MTPs under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ each indicate: (1) control of the true underlying 5% FWER across marker densities, since their respective observed FWER is nearly that expected 5%; (2) balanced control of the 5% FWER across π_j and marker densities, since their respective observed FWER seems to randomly cycle above and below the 5% FWER reference line across π_j and marker densities; and (3) almost identical observed FWER across π_j and marker densities, an indication that the minP and Šidák adjusted p-values are likely identical. Overall, combining the three observations, these data suggest that $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ correctly identifies the true underlying

null distribution for the CATT statistic under $\mathcal{H}_0$.

3.6.2 Application

3.6.2.1 Methods

To illustrate the utility of our proposed statistical method in practice, we applied it against a GWAS data set investigating Bipolar disorder (this is the same Bipolar GWAS data set, first introduced within §2.6) [105,106]. Prior to subject and SNP filtering, there were $n_1 = 1001$ cases of Bipolar disorder, $n_0 = 1034$ control subjects, and 875 578 SNP markers within this data set. Pursuant to consulting advise,[10] we excluded subjects from the final GWAS sample who were missing at least 5% genotype data amongst the 875 578 markers and we excluded SNP markers from the final GWAS sample which were either: missing amongst at least 5% of the 2035 subjects; and/or failed (i.e., we rejected the null hypothesis) – at significance level 10^{-6} – the exact test of the null hypothesis of HWE among control subjects. By convention, SNPs possessing a rare variant allele (less than 0.01 MAF, among those study subjects included within the final GWAS data set) are excluded from a final GWAS sample, recommended based upon statistical power to detect associations [104] – we speculate the author's to imply 'low power' to detect associations for minute values in π_j. However, there are several problems with this MAF filtering approach: (1) in excluding these rare variant allele loci, one is throwing away precious data, the loci of which could potentially be part of the genetic component within the disease etiology; (2) the notion of low statistical power assumes that proper statistical tools are utilized for inference purposes. The assumption of $\tilde{Q}_0$ for the CATT statistic for these rare variant allele loci is highly suspect, insofar as the 2×3 contingency tables summarizing the data thereof will be quite sparse. In other words, this notion of low statistical power for rare variant loci should not be of such concern when adopting the proper null distribution for the CATT statistic under $H_0^{(j)}$; and (3) by excluding these rare variant allele loci, one is inadvertently increasing the power (i.e., introducing a bias) to detect associations at common variant allele loci. Nonetheless, since our method correctly identifies the test statistics null distribution for the CATT statistic, we employed a less stringent MAF filter and included – within our final GWAS sample – those SNP loci possessing a sample estimated minor allele frequency at least equal to 0.001. PLINK (version 1.07; October 2009) [63] was utilized to carry out this subject/marker

[10] Provided by Peter Zandi, PhD, Director, Psychiatric Epidemiology Training Program of John Hopkins Bloomberg School of Public Health.

filtering recipe. After employing this procedure, we obtained our final GWAS sample, comprised of $n_1 = 1001$ case subjects, $n_0 = 1034$ control subjects, and $m = 769672$ SNP loci. In coherence with the reasoning provided within §2.6, for data compression we chose the value of ρ to be four (4), so that our genotype matrix $\mathbf{G}^{(*4)}$ was of dimension 192418×2035.

Unconditional Distribution of the CATT Statistic: HWE

We tested our proposed methodology in two ways: (1) assuming HWE among population genotype frequencies across loci; and (2) making no assumption whatsoever regarding the distributional characteristics governing population genotype frequencies at locus j, all $j = 1, \ldots, m$. In both circumstances, the CATT statistic was used to test the null hypothesis $H_0^{(j)}$, assuming the additive genetic model of inheritance under the two-sided alternative hypothesis. We first consider the former of the two approaches.

The minP and Bonferroni MTPs were utilized to control the FWER, where pointwise p-values were calculated under $\left\{Q_{0j}^{(*H)}\right\}_{j=1,\ldots,m}$, taking the unknown population minor allele frequency (π_j) at its corresponding MLE value under $H_0^{(j)}$, $\hat{\pi}_j$, for all $j = 1, \ldots, m$. Since we desired to compare the performance of our method against that of conventional (i.e., assuming $\tilde{Q}_0$ for the CATT statistic under $\mathcal{H}_0$) GWAS MHT practice, we also employed the maxT and Bonferroni MTPs under $\tilde{Q}_0$. We applied $R = 102400$ random shuffles of the column labels upon $\mathbf{G}^{(*\rho)}$ within the GPER algorithm (§2.6) for each of the maxT and minP MTPs.

As with the simulation analysis conducted within §3.6.1, to resolve the computational problem here – in utilizing $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ to compute pointwise p-values under $H_0^{(j)}$ – we adopted a modified version of the computational tools outlined within Algorithm 3.3. The methodology follows: step 1 – we defined our subspace for π_j to be equivalent to the parameter space thereof. That is, our subspace was defined by the interval $[0, 0.5]$. Our reasoning was due to the fact that we expected – based upon inspection of Figure 3.9 – the MLEs, $\{\hat{\pi}_j\}_{j=1,\ldots,m}$, to "fill-in" much of this compact interval of the reals (this notion was confirmed empirically through the data). Next, we observed that our GWAS sample is nearly balanced ($n_0 \approx n_1; n \approx 2\text{K}$). Hence, we assumed the dependency of the exact unconditional probability of Type I error for the CATT statistic to resemble that of the blue curves depicted within Figure 3.1. Insofar as the slopes for these curves appear to be exacerbated – relative to the 5% FWER reference line – for minute values in π_j, say $\pi_j \leq 0.1$, to maintain the veracity of $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ in computing pointwise p-values under $H_0^{(j)}$ for the CATT

statistic we decided to partition our subspace (i.e., the interval $[0, 0.5]$) into two subintervals, $[0, 0.1]$ and $(0.1, 0.5]$, respectively. We partitioned the interval $[0, 0.1]$, the region within the parameter space of π_j for which we assumed – based upon our interpretation of the slopes for the curves depicted within Figure 3.1 over this region of the parameter space – the integrity of the computed UPTE under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ is particularly sensitive to misclassification over $\hat{\pi}_j$, into equal length subintervals, each of length 0.001; and we partitioned the interval $(0.1, 0.5]$, the region within the parameter space of π_j for which we assumed – based upon our visual interpretation of the slopes for the curves within Figure 3.1 tending to "flatten out" within this region of the parameter space – the integrity of the computed UPTE under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ is likely not as sensitive to misclassification over $\hat{\pi}_j$, into equal length subintervals, each of length 0.005. This yielded the collection of unconditional distributions for the CATT statistic, $\left\{Q_0^{(*H)}(d_w/1000, n_0, n_1)\right\}_{w=1,\ldots,180}$, to be utilized within step 2 of Algorithm 3.3, where $Q_0^{(*H)}(\pi, n_0, n_1)$ is as previously defined within §3.6.1 and d_w is defined by $d_w = w + 4(w - 100)I(w > 100)$, for all $w = 1, \ldots, 180$. Finally, we partitioned the domain of the CATT statistic in an identical manner as that conducted within §3.6.1, which yielded the collection of CATT statistic realizations, $\{5u/100\}_{u=1,\ldots,800}$, by which to compute pointwise p-values within step 2 of Algorithm 3.3, of which we now state; steps 2 and 3 – the methodology is identical to that outlined within §3.6.1, replacing the collection $\left\{Q_0^{(*H)}(w/1000, 600, 600)\right\}_{w=1,\ldots,120}$ (simulation 1), say, with $\left\{Q_0^{(*H)}(d_w/1000, n_0, n_1)\right\}_{w=1,\ldots,180}$ therein; step 4 – prior to conducting statistical inference for this Bipolar GWAS data set, taking $\epsilon = 0.9999$ we formulated – in coherence with Algorithm 3.5 – an 800×180 two-dimensional lookup table, $\mathbf{P}^{\epsilon}$, comprised of 144K computed elements $p_{u,w}(\Gamma(\boldsymbol{\theta}_w))$, $u = 1, \ldots, 800$ and $w = 1, \ldots, 180$, for utility within the Bonferroni and minP MTPs. The first row of Table 3.9 summarizes the required computation time to generate $\mathbf{P}^{\epsilon}$, applying Algorithm 3.5; and step 5 – identical to that within §3.6.1, with the exception that

$$w = \min\left\{I(\hat{\pi}_j > 0.1)\left(\lceil 200(\hat{\pi}_j - 0.1)\rceil + 100 - \lceil 1000\hat{\pi}_j\rceil\right) + \lceil 1000\hat{\pi}_j\rceil, 180\right\}$$

should be substituted in lieu of $w = \min\{\lceil 1000\hat{\pi}_j\rceil, 120\}$ within expression (3.25). In this regard, each of our estimated pointwise p-values – permutation derived or otherwise – was a slightly conservative estimate of their corresponding true underlying bootstrap pointwise p-value counterpart.

Table 3.9: Computational Time (Hours) Needed to Generate $\mathbf{P}^{\epsilon}$ for Application to the Bipolar GWAS Data Set, Applying Algorithm 3.5.

Distribution Under $\mathcal{H}_0$	ϵ	Time (h)		
		To Process Steps 1-4	To Process Step 5	Total
$Q_0^{(*H)}$	0.9999	0.02	0.9	0.9
Q_0^{*}	0.9999	0.52	14.1†	14.6†

†Utilizing a cluster of two GPUs.

Unconditional Distribution of the CATT Statistic

The minP and Bonferroni MTPs were utilized to control the FWER, where pointwise p-values were calculated under $\left\{Q_{0j}^{*}\right\}_{j=1,\ldots,m}$, taking the unknown population genotype frequencies, π_j^{aa} and π_j^{Aa}, at their respective MLEs under $H_0^{(j)}$, $\hat{\pi}_j^{aa}$ and $\hat{\pi}_j^{Aa}$, all $j = 1, \ldots, m$. To remain consistent with the minP MTP approach above, we applied $R = 102\,480$ random shuffles of the column labels upon $\mathbf{G}^{(*\rho)}$ within GPER. To resolve the computational problem – in utilizing the distribution $Q_{0j}^{*}\left(\hat{\pi}_j^{aa}, \hat{\pi}_j^{Aa}, n_0, n_1\right)$ to compute pointwise p-values under $H_0^{(j)}$ – we employed the computational tools outlined within Algorithm 3.3 as follows. First, we partitioned the domain of the CATT statistic in an identical manner to that previously conducted (e.g., §3.6.1), which yielded the collection of CATT statistic realizations, $\{5u/100\}_{u=1,\ldots,800}$, by which to compute p-values within step 2 of said algorithm.

Next, we formulated our subspace of the parameter space for $\boldsymbol{\theta}$. We estimated the parameter vector $\boldsymbol{\theta}_j$ at its corresponding MLE under $\mathcal{H}_0$, $\hat{\boldsymbol{\theta}}_j$, for all $j = 1, \ldots, m$. We then sorted the MLE vectors in increasing order, by first ordering over the estimated values $\left\{\hat{\pi}_j^{aa}\right\}_{j=1,\ldots,m}$ and then ordering over the estimated values $\left\{\hat{\pi}_j^{Aa}\right\}_{j=1,\ldots,m}$. For clarity, let $\hat{\boldsymbol{\theta}}_{(j)}$ denote the j^{th} element upon the ordered MLE vectors (i.e., if $\hat{\boldsymbol{\theta}}_{(j)} = \left(\hat{\pi}_{(j)}^{aa}, \hat{\pi}_{(j)}^{Aa}\right)$, then $\hat{\pi}_{(j)}^{aa} \leq \hat{\pi}_{(j+1)}^{aa}$; and, if $\hat{\pi}_{(j)}^{aa} = \hat{\pi}_{(j+1)}^{aa}$, then $\hat{\pi}_{(j)}^{Aa} \leq \hat{\pi}_{(j+1)}^{Aa}$, for all $j = 1, \ldots, m-1$). So as to obtain high precision in our estimates over the collection $\left\{\left\{p_{j,r}\right\}_{j=1,\ldots,m}\right\}_{r=1,\ldots,R}$ within the minP MTP, we partitioned the ordered MLE vectors into 20 disjoint groups and considered a pointwise p-value lookup table (PPT) for each group. Here, the initial 19 of these groups were each comprised of 40 000 MLE vectors, such that the k^{th} group was comprised of the vectors $\left\{\hat{\boldsymbol{\theta}}_{(j)}\right\}_{j=40000(k-1)+1,\ldots,40000k}$, $k = 1, \ldots, 19$; and the final group was comprised of the largest (relative to our sorting methodology) 9672 MLE vectors over the collection $\left\{\hat{\boldsymbol{\theta}}_{(j)}\right\}_{j=1,\ldots,m}$. For clarity in discussion, we denote these collections of ordered MLE vectors by $\hat{\Theta}_k$, $k = 1, \ldots, 20$. Within the [triangular] parameter space of $\boldsymbol{\theta}$ we constructed 20 rectangles, each formed to encapsulate the elements for a unique collection $\hat{\Theta}_k$, some $k = 1, \ldots, 20$. Geometrically,

if $k = 1, \ldots, 20$ indexes the groups, and if the vector (A_k, B_k, C_k, D_k) denotes the vertices of the constructed rectangle for group k, then we defined the vertices for rectangle $A_k B_k C_k D_k$ by

$$A_k = \left(\pi_1^{aa}, \pi_1^{Aa}\right), \quad B_k = \left(\pi_1^{aa}, \pi_2^{Aa}\right), \quad C_k = \left(\pi_2^{aa}, \pi_2^{Aa}\right), \quad \text{and} \quad D_k = \left(\pi_2^{aa}, \pi_1^{Aa}\right),$$

such that

$$\begin{aligned} \pi_1^{aa} &= \min\left\{\hat{\pi}_j^{aa} : \left(\hat{\pi}_j^{aa}, \hat{\pi}_j^{Aa}\right) \in \hat{\Theta}_k\right\}, \quad \pi_2^{aa} = \max\left\{\hat{\pi}_j^{aa} : \left(\hat{\pi}_j^{aa}, \hat{\pi}_j^{Aa}\right) \in \hat{\Theta}_k\right\}, \\ \pi_1^{Aa} &= \min\left\{\hat{\pi}_j^{Aa} : \left(\hat{\pi}_j^{aa}, \hat{\pi}_j^{Aa}\right) \in \hat{\Theta}_k\right\}, \quad \pi_2^{Aa} = \max\left\{\hat{\pi}_j^{Aa} : \left(\hat{\pi}_j^{aa}, \hat{\pi}_j^{Aa}\right) \in \hat{\Theta}_k\right\}. \end{aligned} \tag{3.26}$$

Our subspace – of the parameter space over $\boldsymbol{\theta}$ – was defined by the aggregation of those $\boldsymbol{\theta}$ encapsulated within some rectangle. Figure 3.12 depicts the MLE vectors $\hat{\boldsymbol{\theta}}_j$ (red dots), for all $j = 1, \ldots, m$, and the rectangles (the four-sided polygons with edges depicted in blue), $A_k B_k C_k D_k$, for all $k = 1, \ldots, 20$.

Each of the formulated rectangles was partitioned by way of equally spaced horizontal and vertical line segments, such that the length of the spacing between sequential horizontal/vertical line segments was proportional to the width/height of the corresponding rectangle. Ordered pairs – in π^{aa} and π^{Aa} within the parameter space for $\boldsymbol{\theta}$ – were formulated within each of the rectangles, such that each ordered pair was formed from an intersection of: a horizontal line segment and a vertical line segment; a horizontal or vertical line segment and some edge of the accompanying rectangle; or, two edges of the rectangle (i.e., a vertex upon the rectangle), where we limited the number of ordered pairs within each rectangle to 200. We limited this number of ordered pairs by rectangle for two reasons: (1) if (x, y) denotes some formed ordered pair of the parameter vector $\boldsymbol{\theta}$ within rectangle $A_k B_k C_k D_k$, $k = 1, \ldots, 20$, this upper limit in the number of ordered pairs gave us good coverage for estimating the distribution $Q_{0j}^*\left(\hat{\pi}_j^{aa}, \hat{\pi}_j^{Aa}, n_0, n_1\right)$ with that of $Q_0^*(x, y, n_0, n_1)$, where $\left(\hat{\pi}_j^{aa}, \hat{\pi}_j^{Aa}\right) \in \hat{\Theta}_k$; and (2) to reduce the computational burden in generating the PPT upon each of the rectangles. This partitioning of the rectangles, yielded our partition of the subspace over $\boldsymbol{\theta}$. Table 3.10 summarizes our partitioning methodology over the 20 constructed rectangles comprising our defined subspace of the parameter space for $\boldsymbol{\theta}$.

For each $k = 1, \ldots, 20$, let $\mathcal{O}_k$ denote the collection of ordered pairs (x, y) corresponding with the aforementioned partitioning of rectangle $A_k B_k C_k D_k$. Taking $\epsilon = 0.9999$, for each $k = 1, \ldots, 20$ we formulated – in coherence with Algorithm 3.5 – a pointwise p-value lookup table (PPT), denoted $\mathbf{P}_k^{\epsilon}$, whose rows corresponded to the realizations in the CATT statistic $\{5u/100\}_{u=1,\ldots,800}$ and whose

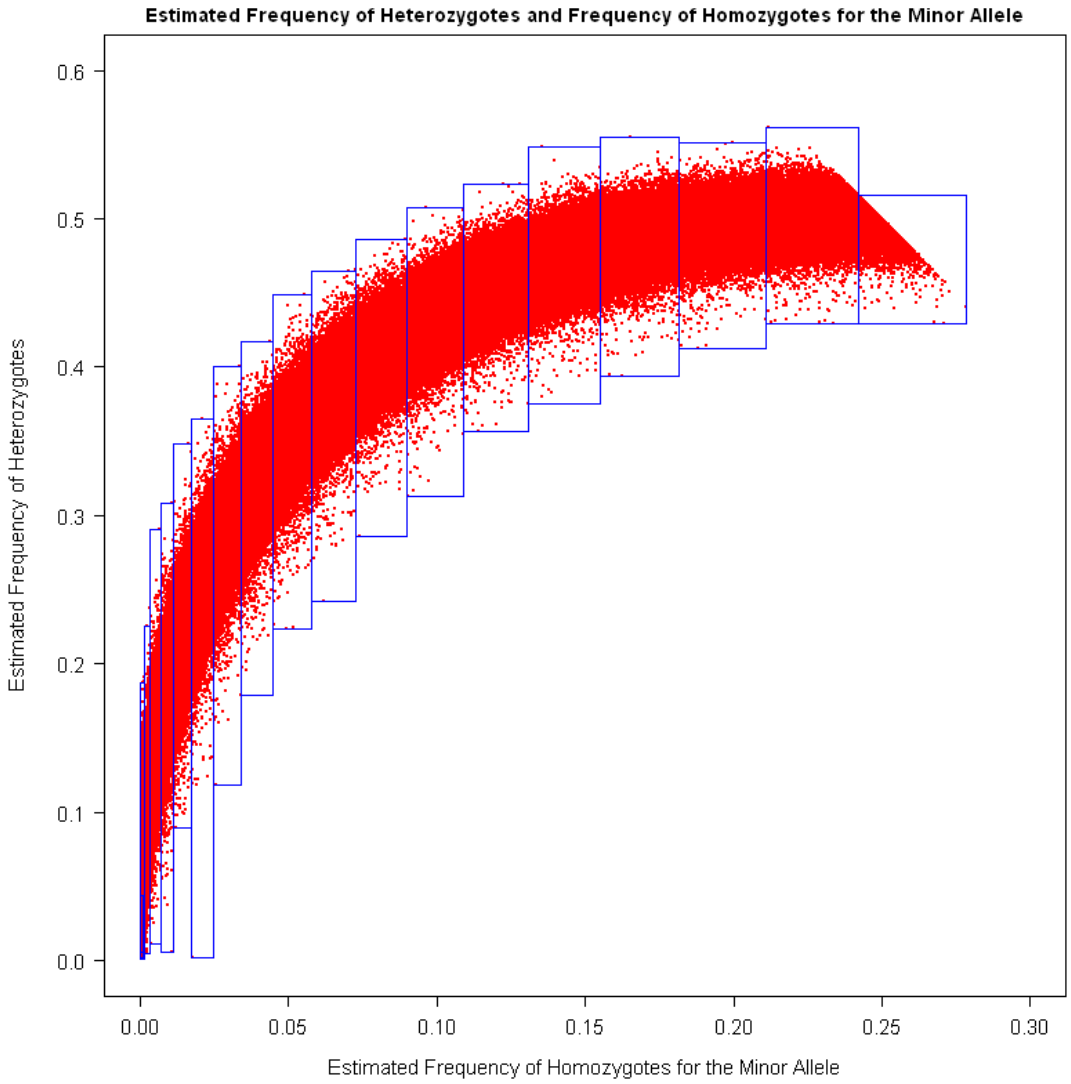

Fig. 3.12: Estimated $\boldsymbol{\theta}_j$ (Red Dots) for the $m = 769672$ SNP Loci of a Case-Control GWAS Investigating Bipolar Disorder, and the Defined Subspace of the Parameter Space over $\boldsymbol{\theta}$ (Blue Rectangles).

columns corresponded to the elements contained within the collection $\mathcal{O}_k$. The value depicted upon row k of the final column of Table 3.10, corresponds to the number of elements contained within $\mathcal{O}_k$, $k = 1, \ldots, 20$. The second row of Table 3.9 summarizes the required computational time to generate all 20 of the PPTs. Using a cluster of two NVIDIA GeForce GTX 470 GPUs, the roughly 3 million elements (800×3821) comprising the aggregate of the 20 PPTs were generated in about 14.6 hours.

For each $k = 1, \ldots, 20$, the values comprising table $\mathbf{P}_k^{\epsilon}$ were utilized to estimate the pointwise

Table 3.10: Summary of the Subspace Partitioning over the Parameter Space for $\boldsymbol{\theta}$, as Applied Within Algorithm 3.3 for the Bipolar GWAS Data Set[†].

	Rectangle $A_kB_kC_kD_k$				Resolution[(a)] in		Number of Ordered Pairs (x, y)[(b)]
k	π_1^{aa}	π_2^{aa}	π_1^{Aa}	π_2^{Aa}	π^{aa}	π^{Aa}	Within Rectangle $A_kB_kC_kD_k$
1	0.000	0.001	0.002	0.006	0.003	0.02	198
2	0.000	0.001	0.006	0.044	0.020	0.19	198
3	0.000	0.002	0.002	0.187	0.368	3.67	153
4	0.002	0.003	0.005	0.225	0.495	4.94	180
5	0.003	0.007	0.011	0.291	0.701	7.00	200
6	0.007	0.011	0.006	0.309	0.886	8.85	170
7	0.011	0.017	0.090	0.348	0.923	9.22	196
8	0.017	0.025	0.003	0.366	1.238	12.4	174
9	0.025	0.034	0.118	0.401	1.159	11.6	200
10	0.034	0.045	0.179	0.417	1.191	11.9	200
11	0.045	0.058	0.224	0.449	1.253	12.5	198
12	0.058	0.073	0.242	0.464	1.369	13.7	187
13	0.073	0.090	0.286	0.486	1.333	13.3	195
14	0.090	0.109	0.313	0.507	1.390	13.9	196
15	0.109	0.131	0.357	0.523	1.386	13.9	192
16	0.131	0.155	0.376	0.548	1.513	15.1	192
17	0.155	0.182	0.394	0.555	1.484	14.8	198
18	0.182	0.211	0.412	0.551	1.450	14.5	200
19	0.211	0.242	0.429	0.561	1.467	14.7	198
20	0.242	0.278	0.429	0.516	1.304	13.0	196
Total:							3821

[†] π_s^{aa} and π_s^{Aa} as defined within (3.26), $s \in \{1, 2\}$.
[(a)] Increment between line segments within the rectangle $A_kB_kC_kD_k$; Depicted values are divided by 10^{-3}.
[(b)] (x, y) denotes an ordered pair of intersecting line segments within the rectangle partitioning.

p-values computed under the minP MTP, for those SNP loci whose MLE vector $\hat{\boldsymbol{\theta}}$ resided within $\hat{\Theta}_k$. Specifically, if $\|\mathbf{x}\|$ denotes the Euclidian norm for some vector $\mathbf{x}$, then within the r^{th} permutation of the minP MTP, we estimated pointwise p-value $p_{j,r}$ – corresponding to the realization of the CATT statistic $t_{j,r}$ under $\mathcal{H}_0$ – by the $(u, w)^{\text{th}}$ element of $\mathbf{P}_k^{\epsilon}$, such that the value of u was obtained by way of (3.25) and the value of w satisfied

$$w \in \left\{ \omega, 1 \leq \omega \leq n\left(\mathcal{O}_k\right) : \boldsymbol{\theta}_{(\omega)} \in \mathcal{O}_k \text{ satisfies } \boldsymbol{\theta}_{(\omega)} = \arg\min_{\boldsymbol{\theta} \in \mathcal{O}_k} \|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}\| \right\}, \tag{3.27}$$

where $\hat{\boldsymbol{\theta}}_j \in \hat{\Theta}_k$, some $k = 1, \ldots, 20$, and $\boldsymbol{\theta}_{(s)}$ is the s^{th} ordered element within the collection $\mathcal{O}_k$. In an analogous manner, we estimated the pointwise p-value p_j, for the realization of the CATT statistic t_j under $\mathcal{H}_0$ for the observed data, by the element $[\mathbf{P}_k^{\epsilon}]_{u,w}$, where the value of u was obtained by substituting the value t_j within (3.25) in lieu of $t_{j,r}$ therein, and the value of w satisfied (3.27).

3.6.2.2 Results

Figure 3.13 displays a Manhattan plot of the pointwise p-values for the CATT statistic, computed under $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$, with reference line (black; $(-1)\log_{10}\left(p_{(\lfloor \alpha R \rfloor)}\right)$) added ($p_{(\lfloor \alpha R \rfloor)}$ as defined within step 4 of the minP pseudocode (§3.4)), taking $\alpha = 0.05$ and $R = 102\,400$. By inspection of this plot, we find that the genotypes for 14 SNP loci – represented by the dots lying above the reference line – to be statistically significantly associated with Bipolar disorder at the 5% FWER, after MHT correction within the minP MTP, assuming the CATT statistic to be distributed by $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$ under $\mathcal{H}_0$. Indeed, Table 3.11 summarizes select summary measures for these 14 SNP loci. The first striking observation we make, is that these data indicate each of the markers to possess a rare variant allele within the population, assuming $\mathcal{H}_0$ to be true, as the range of MLEs in π_j lie within the interval $[0.007, 0.02]$. Next, these data indicate that the genotype frequencies among the controls at these loci adhere very close to the Hardy-Weinberg equilibrium assumption, insofar as the MLEs in the inbreeding coefficient of our alternative genotype frequency model (f_j, see §3.3) attain values essentially equal to zero. Thus, these data suggest that the distribution $Q_{0j}^{(*H)}(\pi_j, n_0, n_1)$ to almost certainly identify the CATT statistic under $H_0^{(j)}$ at these 14 loci, for which we expect the corresponding pointwise p-values derived under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ at these loci to be very accurate (i.e., assume values approximately equal to those under Q_{0j}^{*}). Indeed, comparing the Bonferroni adjusted p-values computed under $Q_{0j}^{(*H)}(\hat{\pi}_j, n_0, n_1)$ (column 7) on a row-to-row basis to those computed under $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$ (column 8), we find the values depicted upon the two columns to be nearly one in the same. Next, as expected these data indicate that the pointwise p-values computed under $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$ are remarkably different (in fact, considerably smaller) than their counterparts computed under $\tilde{Q}_0$, since the ratio of latter to the former pointwise p-values take extremely large values as shown by the final column of the table. This is not a surprising result, per our discussions (e.g., §3.2.2 and §3.2.3) regarding the conservative behavior in assuming $\tilde{Q}_0$ under $\mathcal{H}_0$ for the CATT statistic over minute values of π_j (see also Figure 3.1) upon a [almost] balanced GWAS. However, drawing attention to the [exceptionally large] values within the final column of the table upon SNP-IDs 2, 3, and 5, we can speculate – based upon the respective small and large realizations in $\hat{\pi}_j$ and t_j for these loci – that the conservatism in the assumption of $\tilde{Q}_0$ for the CATT statistic must surely be considerable as the marker density increases towards that of genome-wide coverage (i.e., a genetic marker sample of approximately 3 billion base pairs), due to the extremely large p-value ratios depicted within the table for these loci. Finally, we note that the genotypes for

five (5) of these markers are statistically significantly associated with Bipolar disorder at the 5% FWER after Bonferroni correction under $\tilde{Q}_0$, whereas the genotypes for thirteen (13) of the markers are statistically significantly associated with Bipolar disorder at the 5% FWER after Bonferroni correction under $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$. These data illustrate the potential increase in statistical power which can be achieved by applying our method (i.e., applying $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$ in lieu of $\tilde{Q}_0$) upon a [nearly] balanced GWAS data set in practice, even when applying a conservative MTP such as the Bonferroni.

Table 3.12 displays risk estimates and permutation $\big($maxT MTP under $\tilde{Q}_0$, and minP MTP under each of $Q_{0j}^{(*H)}\left(\hat{\pi}_j, n_0, n_1\right)$ and $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)\big)$ adjusted p-values for the fourteen (14) SNP loci whose genotypes were determined to be statistically significantly associated with Bipolar disorder at the 5% FWER, after adjustment for MHT applying the minP MTP, assuming the CATT statistic to be distributed by $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$ under $\mathcal{H}_0$. These data illustrate two key notions: (1) as expected – pursuant to the argument presented within §3.2.3 regarding the *unbalanced* nature of the maxT MHT adjustment – these data indicate an inflated Type II error rate for the maxT MTP under $\tilde{Q}_0$ upon loci suggestive to possess a rare variant allele, since this MTP reports only seven of these 14 loci to be statistically significantly associated with Bipolar disorder at the 5% FWER; and (2) pursuant to the argument presented within the second paragraph of §1.5, that rare variant allele loci could makeup part of the genetic component within the etiology of complex diseases. It is worth noting here that at these loci, both cases and controls possess at least one copy of the major allele (the only exception being the locus whose SNP-ID is 4). This could indicate that possessing a single copy of the minor allele at these loci leads to a gain (or, loss)-of-function within the biological mechanism responsible for the incidence of Bipolar disorder. In other words, a mutation at one (or, more) of these loci during ones' lifetime, could significantly alter the biological mechanism responsible for the incidence of Bipolar disease. Finally, as with the observation made regarding the Bonferroni adjusted p-values computed under each of the unconditional distributions of the CATT statistic discussed within this Dissertation, here we note the minP adjusted p-values computed under $Q_{0j}^{(*H)}\left(\hat{\pi}_j, n_0, n_1\right)$ (assuming HWE among population genotype frequencies across SNP loci) are remarkably similar to those computed under $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$. When combined, these observations regarding similarity in each of the respective Bonferroni and minP adjusted p-values across the null distributions $Q_{0j}^{(*H)}\left(\hat{\pi}_j, n_0, n_1\right)$ and $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$, suggests that the unconditional distribution $Q_{0j}^{(*H)}\left(\hat{\pi}_j, n_0, n_1\right)$ is robust to deviations from HWE among population genotype frequencies.

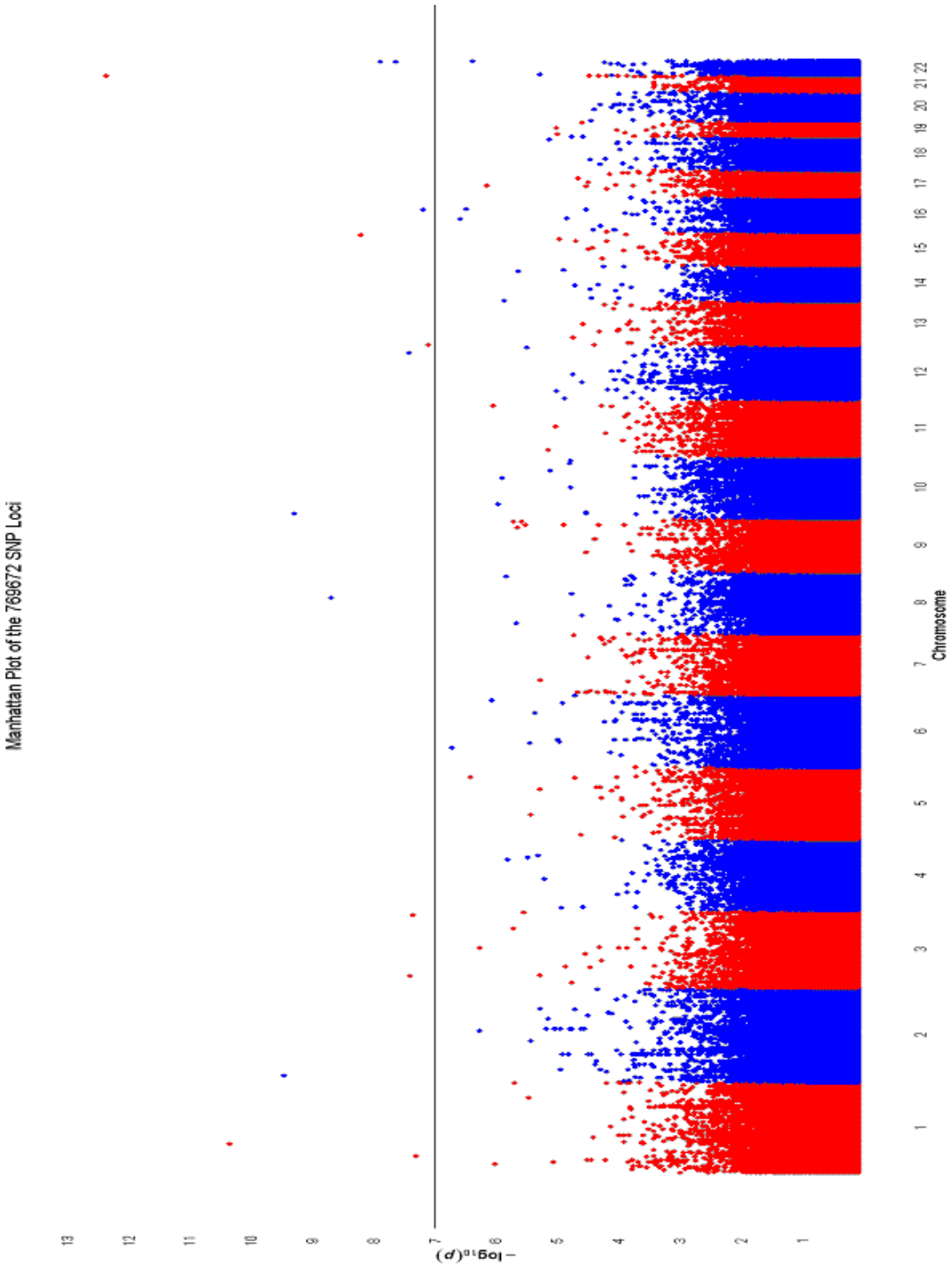

Fig. 3.13: Manhattan Plot of the Pointwise P-values for the 769 672 SNP Loci of the Bipolar GWAS Data Set, Computed under $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$. The Black Reference Line Denotes the Value $-\log_{10}\left(p_{(\lfloor \alpha R \rfloor)}\right)$ for the MinP MTP, Taking $R = 102\,400$ and $\alpha = 0.05$.

Table 3.11: Summary Statistics for Markers Within the $m = 769\,672$ Bipolar SNP Panel Resulting in a Statistically Significant Association with Bipolar Disorder at the 5% FWER, after Multiplicity Adjustment by Way of the MinP MTP under $Q_{0j}^* \left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)^{\dagger}$.

SNP ID	Chr	$\hat{\pi}_j$	$\hat{f}_j$	t_j	Bonferroni Corrected P-value			P-value Ratio
					$\tilde{Q}_0$	$Q_{0j}^{(*H)}\left(\hat{\pi}_j, n_0, n_1\right)$	$Q_{0j}^* \left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$	
1	21	0.012	0.00	43.5	< 0.001	< 0.001	< 0.001	8
2	1	0.011	0.00	36.6	0.001	< 0.001	< 0.001	36
3	2	0.010	0.00	33.3	0.006	< 0.001	< 0.001	23
4	10	0.020	0.00	35.7	0.002	< 0.001	< 0.001	16
5	8	0.007	0.00	28.6	0.069	0.002	0.001	51
6	15	0.012	0.00	30.3	0.029	0.005	0.005	6
7	22	0.009	-0.02	27.9	0.101	0.013	0.010	10
8	22	0.009	0.00	27.0	0.160	0.019	0.015	11
9	12	0.007	0.00	25.2	0.390	0.031	0.032	12
10	3	0.009	0.00	26.2	0.242	0.032	0.024	10
11	3	0.008	0.00	25.5	0.347	0.036	0.032	11
12	1	0.011	0.00	26.6	0.193	0.040	0.039	5
13	16	0.008	-0.01	25.1	0.425	0.054	0.042	10
14	13	0.011	0.00	25.8	0.288	0.065	0.062	5

$^{\dagger}\hat{\pi}_j$ is estimated among all study subjects under $H_0^{(j)}$; $\hat{f}_j$ is estimated among control subjects; and the p-value ratio is the Bonferroni corrected p-value under $\tilde{Q}_0$ divided by that under $Q_{0j}^* \left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$.

Table 3.12: Risk Estimates and Permutation Based Adjusted P-values for Markers Within the $m = 769\,672$ Bipolar SNP Panel Resulting in a Statistically Significant Association with Bipolar Disorder at the 5% FWER, after Multiplicity Adjustment by Way of the MinP MTP under $Q_{0j}^* \left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$.

SNP ID	OR†	95% CI for OR‡	Permutation Corrected P-value		
			maxT	minP	
			$\tilde{Q}_0$	$Q_{0j}^{(*H)}\left(\hat{\pi}_j, n_0, n_1\right)$	$Q_{0j}^* \left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$
1	47.6	(6.5, 346.0)	< 0.001	< 0.001	< 0.001
2	21.6	(5.2, 89.9)	0.001	< 0.001	< 0.001
3	19.9	(4.8, 82.9)	0.003	< 0.001	< 0.001
4	13.6	(4.2, 44.2)	0.001	< 0.001	< 0.001
5	58.0	(3.5, 952.7)$^{(a)}$	0.035	0.001	0.001
6	8.9	(3.5, 22.6)	0.016	0.003	0.003
7	0.06	(0.01, 0.25)	0.051	0.006	0.007
8	16.6	(4.0, 69.6)	0.079	0.012	0.010
9	28.6	(3.9, 211.0)	0.182	0.020	0.020
10	16.1	(3.8, 67.7)	0.116	0.021	0.015
11	15.8	(3.8, 66.3)	0.165	0.025	0.021
12	8.1	(3.2, 20.7)	0.094	0.028	0.026
13	0.04	(0.01, 0.26)	0.198	0.034	0.027
14	7.9	(3.1, 20.2)	0.138	0.040	0.040

†Odds ratio of Bipolar disorder, comparing carriers of the minor allele to non-carriers of the minor allele.

‡Asymptotic Wald-based confidence interval, uncorrected for MHT.

$^{(a)}$Odds ratio and confidence interval were estimated by adding 0.5 to the cells of the applicable 2×2 table.

3.7 Conclusions and Future Directions

Correct identification of the test statistics null distribution under $\mathcal{H}_0$ (Q_0) is arguably the most vital step in implementing an MTP, as improper identification of Q_0 could lead to control in the FWER at a level other than that intended. Our results have illustrated – both probabilistically over the PMF of the exact unconditional distribution for the CATT statistic, and empirically by way of simulation – that the widely accepted asymptotic chi-square assumption for the CATT statistic under $\mathcal{H}_0$ is not realistic in a GWAS. This is due to the exceptionally small pointwise significance level on a per-marker basis, ensuing from the scale in the number of tested null hypotheses. Under these conditions (i.e., a large value in m), our analysis suggests that the unconditional probability of Type I error (UPTE) for the CATT statistic is dependent upon a number of factors for the GWAS sample and its underlying population, including the values for $\boldsymbol{\theta}_j$, n, n_0, n_1, and m. When these factors are ignored, improper control in the FWER at level α for the Bonferroni MTP can result, insofar as pointwise p-values are computed under an incorrectly identified null distribution for the CATT statistic. Based upon our analysis for a balanced/unbalanced GWAS and assumed 5% FWER, under $\tilde{Q}_0$ and $\mathcal{H}_0$ for the CATT statistic the potential for under/over reporting of false-positives is prevalent. Moreover, because the UPTE for the CATT statistic is suggestive to be dependent upon several factors of a GWAS sample and its underlying population, the chi-square assumption for this statistic under $\mathcal{H}_0$ can lead to the GWAS gold standard maxT and minP MTPs to produce unbalanced MHT adjustment. This notion is problematic for several reasons, including: the potential for lack of replication among several GWAS investigating a common binary phenotypic trait; and future genetic association studies where rare variant alleles may be of particular interest – for a balanced/unbalanced GWAS (and, extensions thereof), our results indicate that the maxT (or, minP) MTP, applied against the CATT statistic under $\tilde{Q}_0$, would under/over-state Type I errors for those SNP loci possessing a rare variant allele.

On the other hand, the exact unconditional distribution for the CATT statistic under $H_0^{(j)}$, Q_{0j}^*, correctly identifies Q_0 for SNP locus j and removes the underlying asymptotic chi-square assumption thereof. The joint implementation of $\left\{Q_{0j}^*\left(\boldsymbol{\theta}_j, n_0, n_1\right)\right\}_{j=1,\ldots,m}$ within a GWAS, by way of say the minP MTP, accounts for the correlation among the hypotheses encompassing $\mathcal{H}_0$, thereby resulting in: high statistical power; and accurate, balanced, strong control of the FWER at level α. Moreover, our abolishing of the underlying asymptotic chi-square assumption for the CATT statistic, implies that one does not need to be concerned about incorrectly interpreting the statistical inference of said

statistic for sparse tables, as the results thereof are surely accurate. As a result, we now possess the proper tools for testing $\mathcal{H}_0$ in GWAS and extensions (e.g., integrating inference of rare variant allele loci with common variant loci within the same study, as demonstrated through the Bipolar GWAS example of §3.6.2) thereof. This implies that one can: revisit [each-and-every] historically reported GWAS, apply our method, and where applicable provide the necessary corrections to the reported results of these studies; and correctly report the statistical results of future GWAS and extensions thereof. This latter notion is particularly critical, because future large-scaled genetic association studies are likely – due to improvements in genotyping technology – to possess: ever increasing sizes in their respective sampled SNP panels; and increasing presence of SNP loci possessing a rare variant allele. Our results indicate that the UPTE for the CATT statistic depends upon each of these parameters (i.e., m and $\boldsymbol{\theta}_j$) of the GWAS sample, motivating the necessity for the application of our developed tools. We have illustrated the utility of Q_{0j}^* in practice, against a GWAS data set investigating genetic associations with Bipolar disorder. Overall, when compared to the maxT MTP under $\tilde{Q}_0$, at the 5% FWER we found an additional seven markers – double that of the maxT MTP – statistically significantly associated with Bipolar disorder applying the minP MTP under $Q_{0j}^*\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$. In turn, we have demonstrated the realized potential for increased statistical power upon a [nearly] balanced GWAS data set, when utilizing Q_{0j}^* as the assumed test statistics null distribution for the CATT statistic under $H_0^{(j)}$, in lieu of $\tilde{Q}_0$ thereof.

Implementing Q_{0j}^* in practice within the minP MTP, introduces a rather profuse computational problem. In fact, this computational problem is orders of magnitude higher than that of the maxT computational problem resolved within Chapter 2. As a simple illustration, within §3.5 we argued that the statistical analysis under Q_{0j}^*, applied against a balanced GWAS comprised of $n = 2\text{K}$ subjects and $m = 500\text{K}$ SNP loci, would take more than 1580 computational years, assuming $R = 100\text{K}$ minP permutations and one second of computational time being required to compute a single pointwise p-value. Upon clustering $R = 100\text{K}$ maxT permutations (to two GPUs) within GPER to 500K [randomly selected] SNP loci (of the $m = 769\,672$ total loci) for the aforementioned Bipolar GWAS, we found that GPER was able to complete this task in roughly 73 minutes.[11] Hence, the computational problem for the implementation of the GPER algorithm within the minP MTP under Q_{0j}^* is upwards of 11.5 million times (the ratio of: 1580 years, converted to minutes; and 73 minutes) that of the GPER algorithm application within the maxT MTP under $\tilde{Q}_0$ – note:

[11] Note that the sample size for this Bipolar GWAS is $n = 2035 \approx 2000$, and is nearly balanced in the numbers of its sampled cases and controls. Thus, the results of GPER applied to this data set should be very representative of those for a balanced GWAS of $n = 2\text{K}$ subjects.

this estimate is likely exceptionally conservative (i.e., understated), insofar as it assumes a mere single second of computational time is required to compute the pointwise p-value under Q_{0j}^{*} for each realization of the CATT statistic.

To address the computational problem, we developed the computational tools of §3.5. These tools embrace the central theme of Algorithm 3.3, namely that one can utilize a p-value lookup table (PPT) to ascertain accurate estimates of the true underlying pointwise p-values computed within the minP MTP. Albeit, generating a p-value lookup table in-and-of-itself presents a steep computational problem (§3.5.1.1), primarily due to the size of the support encompassing the unconditional reference set for Q_{0j}^{*}, Γ. To lessen this computational problem, we proposed the notion of generating an estimate of the PPT by way of a truncated unconditional reference set (§3.5.1.1). To efficiently generate a truncated unconditional reference set, we developed an iterative algorithm (Algorithm 3.4); to generate the estimates of the PPT, we developed a parallel processing algorithm (Algorithm 3.5). The estimates of the PPT were then utilized within the minP MTP. Overall, based upon the results obtained within the Bipolar GWAS application, our proposed approach to estimating pointwise p-values – by way of the estimates over the PPT – within the minP MTP is suggestive to be: exceptionally efficient, as the roughly 3 million estimated pointwise p-values over the PPTs (20 p-value lookup tables total, each generated over a specific rectangle within the parameter space of $\boldsymbol{\theta}$ – see §3.6.2.1) were generated in 14.6 hours using our parallel processing approach of Algorithm 3.5; and exceptionally accurate, as demonstrated through the examples presented within §3.5.3.5.

An unfortunate consequence of implementing Q_{0j}^{*} in practice is due to the parameter values, encompassing the parameter vector $\boldsymbol{\theta}_j$, likely being unknown and a nuisance. Hence, prior to implementing the PMF for Q_{0j}^{*} in computing pointwise p-values for the CATT statistic, we must handle these nuisance parameters in some manner. One approach is to condition on a sufficient statistic for the parameter vector $\boldsymbol{\theta}_j$ under $H_0^{(j)}$. For example, we could condition on the observed genotype margin of the 2×3 contingency table – cross classifying genotype-phenotype – at locus j, since the values of this margin are sufficient for $\boldsymbol{\theta}_j$. In doing this, it can be readily shown that under $H_0^{(j)}$, the resulting *conditional* PMF is free of $\boldsymbol{\theta}_j$. However, this approach seems untenable, as it would remove the random component in the exposure (i.e., genotype) status at the locus. Alternatively, we could estimate each element of the nuisance parameter vector $\boldsymbol{\theta}_j$. Indeed, we have proposed a methodology for the implementation of Q_{0j}^{*} within a GWAS, such that the nuisance parameter vector $\boldsymbol{\theta}_j$ is to be estimated at its MLE, $\hat{\boldsymbol{\theta}}_j$. In this regard, the pointwise p-values

computed under $Q_{0j}^{*}\left(\hat{\boldsymbol{\theta}}_j, n_0, n_1\right)$ are called bootstrap p-values and are approximate. However, this approach seems tenable, insofar as the calculation of the Cochran-Armitage trend test statistic at locus j (2.5) itself involves estimating a nuisance parameter at its MLE under $\mathcal{H}_0$ (for details, see page 150 of [62]). Although future research – beyond the body of research comprising this dissertation – is needed for developing a methodology to transform these approximate p-values to their exact counterparts, based upon the simulation results presented within §3.6.1 we are optimistic that the approximations are sufficient for accurate strong control of the FWER in a GWAS. This notion holds particularly true to the minP MTP, as any discrepancies between the approximate and exact p-values will result in: an unbalanced multiplicity adjustment, as a worst case scenario; and should not compromise the overall control in the FWER for this MTP. Moreover, recent research investigating the distributional properties of bootstrap p-values, particularly within the realm of discrete data, suggests the accuracy of these p-values to be quite remarkable (see e.g., [140]), which is in direct agreement with the results we obtained within our simulation (§3.6.1).

According to a recent article published within the American Journal of Psychiatry, the cost of assaying the genotype data for a GWAS is approximately $500 per study subject [144]. Thus, collecting the genotype data upon a GWAS of $n = 2$K subjects, amounts to a cost of approximately one million dollars. Upon tendering this quantity of funds for data, one most certainly desires the application of the very best quality statistical analysis tools available. Apparently, GWAS practice does not conform to this notion, for two reasons: (1) many GWAS rely on application of an overly conservative MTP to control the FWER, such as the Bonferroni MTP; and (2) these studies rely upon näive distributional assumptions for the test statistics (e.g., the asymptotic chi-square assumption for the CATT statistic under $\mathcal{H}_0$), utilized as the inference tools in testing the null hypothesis of no genotype-phenotype association on a per-marker basis. Chapters 2 and 3 of this dissertation address and correct upon each of these two notions. Within Chapter 2, we proposed GPER, an algorithm for the rapid implementation of the maxT and minP MTPs applied against a GWAS data set. In this regard, said chapter addresses and corrects the former of these two notions, insofar as the maxT and minP MTPs are amongst the most powerful MTPs controlling the FWER – recall, the goal of the MHT problem is to control some Type I error rate in the strong sense, while simultaneously maximizing statistical power to reject false null hypotheses (§2.1.1). Within Chapter 3, we proposed the test statistics null distribution $Q_{0j}^{*}\left(\boldsymbol{\theta}_j, n_0, n_1\right)$ for the CATT statistic under $\mathcal{H}_0$. Its utility within the minP MTP, provides high statistical power, and accurate, balanced, strong

control of the FWER within a GWAS and extensions thereof. In this regard, said chapter addresses and corrects upon the latter notion. When combined, the inference tools developed within said chapters of this dissertation surely provide the very best quality statistical analysis tools available for a GWAS.

CHAPTER 4

A PERMUTATION APPROACH TO DETECT GENE-ENVIRONMENT INTERACTION IN GENETIC ASSOCIATION STUDIES

4.1 Introduction

As mentioned within §1.2, many common, complex diseases are believed to be a result of the collective effects of genetic and environmental factors, and their interactions. For example, [145] showed a significant interaction between smoking status and the apurinic/apyrimidinic endonuclease 1 protein coding gene (*APE1*) for lung cancer. The article of [146] demonstrated smoking status to be an effect modifier for the association between the XPD codon 751 polymorphism and risk of bladder cancer. Understanding the relationship between genetic polymorphisms and environmental exposures can aid in identifying high-risk subgroups of a population and can provide better perception into the causative pathway mechanisms for complex diseases.

Within this chapter, we adapt the gene-gene interaction testing framework of [109] to include tests for gene-environment interaction. We enhance the framework by: relinquishing the asymptotic approximation upon the appropriate test statistics null distribution; and, for control of the family-wise Type I error rate (FWER), we implement the resampling-based maxT multiple testing procedure (MTP) of [62]. For control of the FWER at the 5% level, within the context of a case-control study we demonstrate by way of simulation that our proposed method offers greater statistical power to detect gene-environment interaction, when compared to several competing approaches to assess this type of interaction.

4.1.1 Existing Methods to Detect Gene-Environment Interaction

Analyses of interactions between genetic variants (here, we consider genetic variation upon SNP loci) and a single binary exposure (environmental factor) in case-control studies, entails comparisons of the estimated genetic relative risks (RRs) for the exposed and unexposed subjects; or, exposure RRs for genetically susceptible and non-susceptible subjects [147]. The conventional approach to estimating these RRs lies with modeling the genetic and environmental effects, along with their interaction effect, through an appropriate multiple logistic regression model. In this context, a standard approach to test for gene-environment interaction (henceforth, GxE interaction) would be

to perform a likelihood ratio test (LRT), for example, assigning the appropriate parameter(s) of the logistic regression model to nullity under the null hypothesis. While this is the conventional approach for assessing the significance of GxE interaction for the single genetic variant paired with a single environmental exposure, assessing GxE interaction across multiple genetic markers – where the LRT is conducted upon each pairing of the environmental factor with that of a genetic factor – introduces a multiple testing problem. A correction for multiple testing is required, to ensure the proper reporting of false-positive significant GxE interaction findings. Here, we assume the case-control study design, but our arguments extend to other population-based genetic study designs.

As elucidated within the preceding chapters, a popular approach to correcting for the multiple testing problem in genetic association studies (e.g., GWAS) is control of the FWER by way of the Bonferroni MTP. While this approach is simple to implement, within the context of multiple hypothesis testing (MHT) of GxE interaction it ignores the correlation among the LRT statistics – resulting from testing for GxE interaction across multiple genetic markers paired with a single environmental factor. As a consequence of ignoring correlation amongst the test statistics, this MTP can be overly conservative in its control of the FWER [20,21,22] – *assumes correct selection of the test statistics null distribution, as elucidated within Chapter 3.* Correlation most certainly exists amongst these LRT statistics, insofar as the environmental factor is a common ingredient amongst the tests conducted across the SNP loci. To circumvent the conservatism of the Bonferroni approach, one might consider a more powerful approach to correcting the multiple testing problem, such as the resampling-based maxT MTP proposed within [62]. By accounting for the correlation within the data, the maxT MTP can result in greater statistical power – with respect to the Bonferroni MTP and other MTPs which control the FWER – to detect true GxE interactions. However, this resampling approach is generally not tenable when testing hypotheses involving coefficient(s) of a multiple regression model (e.g., testing the null hypothesis of no GxE interaction upon an appropriately constructed multiple logistic regression model), insofar as strong control of the FWER is not guaranteed [148, 149]. In fact, the article [150] argues that exact permutation tests of gene-environment (or, gene-gene) interaction are typically not possible to construct in genetic association studies. Furthermore, we note that the above assumes one has implemented the correct genetic model of inheritance (GMI; e.g., additive, recessive, dominant) into the logistic regression model across the loci. Should this notion fail to hold true (e.g., upon a novel complex disease in which etiology is unknown), two apparent resolutions present themselves: (a) abide with the results ascertained from

testing for GxE interaction upon the chosen GMI across the loci. The problem with this approach is that one likely forfeits detecting GxE interaction upon loci for which the GMI is incorrectly specified; or, (b) at each locus, test for GxE interaction against several GMIs. However, testing for GxE interaction upon several GMIs at each locus, exacerbates the multiple testing problem for the conventional case-control LRT approach.

Alternatively, one might consider taking a case-only analysis as the approach to identifying interactions within genetic association studies. In fact, the case-only test has been previously shown to be more powerful than a case-control analysis for detecting interactions (see e.g., [151, 152]). However, the case-only approach depends upon the assumption of G-E independence within the population, the assumption in which – because of the rather profuse number of genetic markers under study – seems unsupported within the context of GWAS and candidate gene association studies. In the circumstance for which a population association exists between genetic and environmental factors under study, a case-only analysis will result in an inflated observed FWER [151, 153]. Further, like the conventional case-control LRT approach, the case-only approach assumes one has implemented the correct GMI into the appropriate model.

As an alternative approach to the conventional case-control and case-only LRT testing, [3] proposed a 2-step logistic regression approach for detecting significant GxE interactions within the context of a case-control GWAS. Within step 1 of their method (screening step), a modification to the case-only test of GxE interaction is employed, such that the entire case-control sample is included within the hypothesis test. Those SNP loci determined to be statistically significantly associated with the environmental factor are then carried forward to the second step of their method, where the standard 1-df LRT test for GxE interaction is carried out and a Bonferroni correction applied for multiple testing. When compared to the conventional 1-step LRT, this 2-step method can attain more power to detect GxE interactions. More recently, [27] proposed a *hybrid* approach to detect GxE interaction within the context of a case-control GWAS, combining the screening effects of their aforementioned 2-step method along with that of the [154] screening method – the screening approach of [154] was proposed to detect gene-gene (GxG) interactions within the context of GWAS. While these approaches seem to work well for a GWAS, there are potential problems with their respective approaches for assessing GxE interactions upon genetic association studies in general, such as: the user having to specify the typically unknown GMI upon the genetic markers under investigation; control of the FWER is based upon the conservative Bonferroni MTP; p-values are computed in

reference to an asymptotic test statistics null distribution from the corresponding LRT statistics; and the significance threshold, within the screening step for each of these procedures, is subjective. In this regard, the utility of these approaches within the context of a candidate gene study – involving say [at most] a few hundred sampled genetic markers – is likely limited because too stringent a value for the significance threshold within the screening step could result in few (if, any) markers filtering through step 1 and being assessed for GxE interaction within step 2 of the approach.

An alternative approach – of which has been given considerable attention recently within GxG interaction studies and will be considered here – to logistic regression modeling in genetic association studies, considers *logical patterns* in the genetic and environmental factors. If G_j denotes the random variable corresponding to the number of copies for the minor allele (i.e., $G_j \in \{0,1,2\} = \mathcal{G}$) at SNP locus j, $j = 1, \ldots, m$, and $E \in \{0,1,\ldots,\varepsilon-1\} = \mathcal{E}_\varepsilon$, $\varepsilon \geq 2$, is the random variable pertaining to the level of exposure upon some environmental factor, we consider logical patterns in G_j and E of the form, for example,

$$L_A = (G_j \in \{0,1\}) \wedge (E = 1)\,, \tag{4.1}$$

of which may bring about higher or lower risk of developing a particular complex disease when compared to some alternative logical pattern (denoted, L_B; in most circumstances this will be the complement logical pattern to L_A). A logical rule, such as that given within (4.1), is denoted as a *pattern*. A pattern in G_j and E which is to be utilized in assessment for a main effect in G_j, a main effect in E, or a GxE interaction, is henceforth denoted a *candidate pattern*. Identifying significant GxE interactions within high dimensional data (e.g., candidate gene studies and GWAS) is a non-trivial endeavor. Several search algorithms have been proposed in recent years, within the context of SNP and gene expression data for assessing GxG interactions. Some are based upon logic regression and use a simulated Monte Carlo approach to search the space of all possible interactions (see e.g., [155,156,157,158]). Tree-based methods could also be employed to search for interactions. For example, [159] apply a random forest approach to assess interactions within SNP data.

Unlike traditional logistic regression modeling, these data mining approaches are not based upon parametric additive models. They can typically identify main effects; but, in the presence of two main effects fail to detect the additional interaction effect [109]. Also, within the context of GxE interaction, properties related to Type I error rates and statistical power have not been thoroughly compared with the more conventional approach of tests based upon parametric additive

models (see [160] for comparative analysis of these data mining approaches within the context of assessing epistasis (GxG interaction)). Nonetheless, these approaches tend to be easier to interpret than regression models involving interaction terms. Because of the high-dimensionality search, assessing statistical significance among patterns thought to be relevant to the disease causative pathway, is a critical issue. Cross-validation is commonly employed. In their testing for SNP-SNP interactions, [109] examines this assessment from a multiple testing viewpoint, which is the approach we take here in our assessment of GxE interactions.

Finally, one might argue that even if a disease locus only affects disease risk among those exposed to an environmental factor, the locus will likely still have a detectable main effect with disease [161], so analysis can be carried out in the absence of (or, ignoring) data on environmental exposure. Albeit, the relative statistical power of these various approaches to assessing GxE interaction will depend upon the true penetrance model of the disease, for which we have little information *a priori* for complex diseases [162]. If a locus truly only affects disease risk among exposed (or, unexposed) individuals upon the environmental factor within the population, the locus may or may not have a detectible main genetic effect, dependent upon the prevalence of exposure within the population and the magnitude of the genetic effect [162]. For example, under the dominant genetic model, if the risk of disease among non-carriers of the risk allele is unaffected by exposure status to the environmental factor, the statistical power in detecting the main genetic effect at the locus will be dependent upon: the prevalence of exposure within the population, among individuals carrying at least one copy of the risk allele at the locus; and, the magnitude of the genetic effect at the locus (see Proposition A.8 for illustration). On the other hand, if the locus affects risk among both exposed and unexposed individuals, then tests to detect the main genetic effect may be more powerful than tests to detect GxE interaction, even when genotype odds ratios differ between the populations of exposed and unexposed individuals [116, 163]. The performance of data-mining procedures – including statistical power to detect disease susceptible loci (DSLs), Type I error rate, and mean prediction error – is not generally clear. Complex disease DSLs likely have incomplete penetrance, modest effects, and high phenocopy[1] rates, making them poor risk indicators of disease [162]. Even with an odds ratio upwards of three (3) – considerably larger than anticipated for complex disease DSLs – a common allele can be a poor risk indicator of disease [164]. Apparently, in the presence of genetic and environmental heterogeneity, data-mining procedures designed to identify DSLs may

[1] A phenocopy is an individual whose phenotype is determined under a particular environmental condition, and is identical to that of another individual whose phenotype is determined by the genotype of some genetic locus.

have low statistical power to detect these DSLs for many complex disease [116, 162], unless each DSL acts in a simple Mendelian manner upon a subset of individuals defined by measured genetic or environmental factors (i.e., genetic and environmental homogeneity exists upon some subset of individuals within the population) [162].

4.1.2 Approach

Here, we adapt the logical pattern SNP-SNP interaction framework of [109] (denoted LPCV – shorthand for Logical Patterns within Categorical Variables) in assessing SNP-Environment interaction (assumed synonymous with GxE interaction) within genetic association studies, where the environmental factor is assumed categorical in nature. Given a set of q-fold candidate pattern ordered pairs $\{(L_{A_l}, L_{B_l})\}_{l=1,\ldots,q}$ involving the categorical random variables G_j and E, the LPCV approach to assessing interaction between these two random variables works as follows:

(a) for each $l = 1, \ldots, q$, we collect the subset of data pertaining to subjects satisfying either of the patterns, L_{A_l} or L_{B_l};

(b) this subset of data is dichotomized according to the candidate pattern L_{A_l} – i.e., subjects satisfying this candidate pattern form one group, while subjects satisfying the candidate pattern L_{B_l} form the alternative group;

(c) for each candidate pattern ordered pair, the test of the null hypothesis of no association between disease status (denoted by the binary random variable Y) and the dichotomized indicator random variable – pertaining to whether or not a subject is a member of the group satisfying candidate pattern L_{A_l}, say – is carried out, yielding a chi-square test statistic;

(d) the maximum of these test statistics is selected; and

(e) for control of the FWER – over the multiple testing problem invoked upon carrying out step (c) of this procedure – [109] exploit statistical inference over the asymptotic distribution of the maximum chi-square test statistic.

Here, we abstain from making any asymptotic assumptions encompassing the maximum chi-square test statistic altogether, and propose multiple testing correction by way of the permutation-based maxT MTP [62]. Our resampling approach offers a vital advantage over that of an asymptotic assumption governing the maximum chi-square test statistic – namely, conditional on the observed

data, this approach correctly estimates the true underlying null distribution for the maximum chi-square test statistic at [or, across] the SNP locus [loci], resulting in accurate control of the FWER. Moreover, for the circumstance in which $m = 1$ and E is binary (i.e., assessment of GxE interaction for a single SNP locus when $\varepsilon = 2$), we propose correction for multiple testing by way of the exact conditional (i.e., permutation) null distribution for the maximum chi-square test statistic. We perform the applicable computations for the multiple hypothesis testing correction without resorting to simulation, by modifying the network algorithm of [142]. This approach is equivalent to implementing the maxT MTP upon said test statistic, but does not require resampling. Without the uncertainty associated with simulating a null distribution, this approach provides the highest accuracy for control of the FWER over the permutation null distribution of the maximum test statistic. Moreover, because this approach is based upon a resampling MTP, it can result in greater statistical power over other MTPs controlling the FWER, such as the Bonferroni.

In addition to assessing GxE interaction – corresponding with each of the dominant and recessive GMIs – at a particular SNP locus, our approach [simultaneously] assesses the main effect upon the environmental factor and the main genetic effect corresponding to each of the dominant and recessive GMIs at the locus. Because the maxT MTP corrects for the multiple testing problem across genetic markers, our approach: is applicable to a wide range of genetic association studies, including single locus association studies, candidate gene studies, and GWAS; and, fully accounts for correlation in the tests across SNP loci, resulting in high statistical power for control over the FWER. We develop our methodological framework within the context of a case-control study (i.e., retrospective design), but the approach is also applicable to other genetic association study designs (e.g, a prospective cohort). We denote our method by GEM (shorthand for detecting GxE interaction by way of maxT). In the circumstance for which a binary environmental factor has been sampled, by way of simulation we demonstrate that GEM properly controls the FWER at the 5% level under a variety of conditions and achieves greater statistical power over a number of competing approaches used to assess GxE interaction. An `R` [103] package, tentatively denoted *GEM*, is under development for our method and its application is outlined within Appendix C.

4.2 Formulation of Candidate Patterns

Suppose that among n_1 sampled cases ($Y = 1$) and n_0 sampled controls ($Y = 0$), a fixed number of c_{jk} subjects are observed at level k upon the random variable X_j, where this variable is defined such that each element within its support represents a specific level to the combination of

G_j and E, all $j = 1, \ldots, m$ and $k \in \{1, \ldots, 3\varepsilon\} = \mathcal{X}_\varepsilon$. Precisely, for each $G_j \in \{0, 1, 2\} = \mathcal{G}$ and $E \in \{0, 1, \ldots, \varepsilon - 1\} = \mathcal{E}_\varepsilon$, we have

$$X_j = 1 + G_j + 3E. \tag{4.2}$$

Note that any pattern of G_j and E can be expressed through combination(s) of element(s) from the support of X_j, $\mathcal{X}_\varepsilon$. These element(s) form a subset of the collection $\mathcal{X}_\varepsilon$, which we denote by the subscript of the pattern L_C, $C \in \{A, B\}$. For example, consider the pattern given by (4.1). It is,

$$L_A = (G_j \in \{0, 1\}) \wedge (E = 1) \qquad \Longleftrightarrow \qquad X_j \in \{4, 5\},$$

for which $A = \{4, 5\}$.

When studying associations between disease status and explanatory variables by way of a parametric additive model (e.g., logistic regression), our interest centers on determining which variables belong in the model and estimating their corresponding effect size (measured by way of the appropriate coefficient of the model). On the other hand, when studying these associations within the context of logic patterns – as is the circumstance here – we seek logic expressions which are associated with disease status. If W_l represents the indicator random variable with success/failure defined as those elements within the support of X_j lying within the collection A_l/B_l, some $l = 1, \ldots, q$, we consider the test of the null hypothesis of no association between disease status (Y) and W_l, versus the alternative hypothesis for the existence of some association between these variables. Rejection of the null hypothesis, suggests that the odds of disease statistically significantly differs between the two levels in W_l. In turn, this indicates that the logical pattern ordered pair (L_{A_l}, L_{B_l}) is associated with disease status.

In the context of the aforementioned test of hypotheses, we essentially seek the logical pattern ordered pair (L_{A_l}, L_{B_l}) which yields the smallest p-value (i.e., strongest association signal) – computed under the null hypothesis – amongst all possible logical pattern ordered pairs thereof. Without loss of generality, assume that L_{A_l} and L_{B_l} are chosen such that the sets A_l and B_l form a binary partition of the collection $\mathcal{X}_\varepsilon$ – that is, we assume $A_l \cup B_l = \mathcal{X}_\varepsilon$ such that $A_l \cap B_l = \emptyset$; equivalently, L_{A_l} and L_{B_l} are assumed complementary patterns. There are in fact $2^{(3\varepsilon-1)} - 1$ distinct binary partitions of the collection $\mathcal{X}_\varepsilon$, each relating to a unique logical pattern ordered pair (L_{A_l}, L_{B_l}). Thus, the consideration of the logical pattern ordered pairs pertaining to all partitions

of $\mathcal{X}_\varepsilon$, leads one to conducting a total of $2^{(3\varepsilon-1)}-1$ inordinate tests of the null hypothesis of no association between W_l and disease status at each locus.

However, from a biological perspective, it makes sense to restrict attention upon logical patterns in G_j and E connected by the $\wedge$ operator (e.g., (4.1)), particularly when searching for GxE interaction. This said, we construct our q-fold set of candidate pattern ordered pairs $\{(L_{A_l}, L_{B_l})\}_{l=1,\ldots,q}$ by formulating patterns upon G_j and E in a systematic manner, with the aims of assessing GxE interaction as well as assessing the main effects upon the genetic and environmental factors. Since heterozygous individuals most often have an intermediate phenotype, or the identical phenotype to that of the homozygous variant individuals (dominant GMI)[2] or the homozygous wild-type individuals (recessive GMI) [109], here we consider the heterozygous genotype at locus j (i.e., $G_j = 1$) as an intermediate to the two homozygous genotypes. Hence, we consider the following four combinations of G_j: $G_j \in \{0,1\}$, $G_j \in \{1,2\}$, $G_j = 0$, and $G_j = 2$. Coalescing these combinations in G_j with those for the environmental factor (e.g., $G_j = 0$ combined with $E = 0$), we obtain a total of 4ε candidate patterns, denoted $L_{A_1}, \ldots, L_{A_{4\varepsilon}}$ (for clarity, we index candidate patterns by l). For each $l = 1, \ldots, 4\varepsilon$, the candidate pattern L_{B_l} is defined to be the complement of L_{A_l}. Each of these candidate pattern ordered pairs formulate a distinct random variable W_l, by which to test the null hypothesis of no association between W_l and disease status. Taken collectively, the hypothesis tests involving the random variables $\{W_l\}_{l=1,\ldots,4\varepsilon}$ assess the effect of GxE interaction.

Following the line of regression modeling – which incorporates both main effects and interaction effects – we can also incorporate candidate patterns to assess genetic and environment main effects. The candidate patterns for assessing the genetic main effect are those encompassing the dominant and recessive genetic models, given by

$$L_{A_{4\varepsilon+1}} = (G_j \in \{1,2\}) \wedge (E \in \mathcal{E}_\varepsilon)$$

and

$$L_{A_{4\varepsilon+2}} = (G_j = 2) \wedge (E \in \mathcal{E}_\varepsilon)\,,$$

respectively, where the candidate pattern L_{B_l} is defined to be the complement of L_{A_l}, each $l \in \{4\varepsilon+1, 4\varepsilon+2\}$. Each of these candidate pattern ordered pairs formulate a distinct random variable W_l, by which to test the null hypothesis of no association between W_l and disease status. Taken collectively,

[2] Unless otherwise specified, henceforth when we speak of a GMI upon a SNP locus it is assumed in terms of the minor allele for the locus, as we have so defined here for the dominant and recessive GMIs.

the hypothesis tests involving the random variables $W_{4\varepsilon+1}$ and $W_{4\varepsilon+2}$ assess the main effect for G_j. On the other hand, consider now the candidate patterns essential for assessing the main effect in E. Insofar as we make no assumption regarding intermediate effects for the environmental factor, we consider this factor as a nominal categorical variable.[3] We model our candidate pattern ordered pairs in an analogous manner to that of dummy coding a qualitative predictor in regression modeling, where level zero is our baseline group in E. In this regard, whenever $\varepsilon > 2$ the candidate patterns L_{A_l} and L_{B_l} will not be complements of one another. More precisely, for each $l = 4\varepsilon+3, \ldots, 5\varepsilon+1$, the candidate patterns for assessing the environment main effect are defined by

$$L_{A_l} = (G_j \in \mathcal{G}) \wedge (E = l - (4\varepsilon + 2))$$

and

$$L_{B_l} = (G_j \in \mathcal{G}) \wedge (E = 0)\,.$$

Each of these candidate pattern ordered pairs formulate a distinct random variable W_l, by which to test the null hypothesis of no association between W_l and disease status. Taken collectively, the hypothesis tests involving the random variables $\{W_l\}_{l=4\varepsilon+3,\ldots,5\varepsilon+1}$ assess the main effect for the environmental factor. Table 4.1 summarizes our proposed candidate patterns for assessing – upon SNP locus j – each of the genetic and environmental main effects, and GxE interaction. Overall, a total of $q = 5\varepsilon + 1$ candidate pattern ordered pairs (L_{A_l}, L_{B_l}) are considered within our GEM approach presented here.

4.3 Chi-Square Tests

Having defined the collection of q-fold candidate patterns, $\{(L_{A_l}, L_{B_l})\}_{l=1,\ldots,q}$, we now define the notation which will be used to assess main effects over each of the random variables G_j and E, and the effect for GxE interaction. We consider testing the null hypothesis of no association between disease status and the random variable W_l (denoted, $H_0^{(j,l)}$), against the alternative hypothesis that some association exists between these variables (denoted $H_a^{(j,l)}$), for all $j = 1, \ldots, m$ and $l = 1, \ldots, q$. Rejection of null hypothesis $H_0^{(j,l)}$ indicates: a significant GxE interaction between G_j and E, whenever $l \leq 4\varepsilon$; a significant genetic main effect in G_j, whenever $l = 4\varepsilon + 1, 4\varepsilon + 2$; or, a significant main environmental effect, whenever $l = 4\varepsilon + 3, \ldots, q$.

[3]Note that GEM offers flexibility in assumptions governing formulation of candidate patterns. In this regard, formulation of candidate patterns over an ordinal environment factor is a general extension of the approach presented here.

Table 4.1: Summary of the Candidate Patterns for Assessing the Main Effect in Each of the Genetic and Environmental Factors, and GxE Interaction.

l	Effect	Candidate Pattern (L_{A_l})	A_l
1	GxE	$(G_j = 0) \wedge (E = 0)$	$\{1\}$
2		$(G_j = 0) \wedge (E = 1)$	$\{4\}$
$\cdots$		$\cdots$	$\cdots$
ε		$(G_j = 0) \wedge (E = \varepsilon - 1)$	$\{3\varepsilon - 2\}$
$\varepsilon + 1$		$(G_j \in \{0, 1\}) \wedge (E = 0)$	$\{1, 2\}$
$\varepsilon + 2$		$(G_j \in \{0, 1\}) \wedge (E = 1)$	$\{4, 5\}$
$\cdots$		$\cdots$	$\cdots$
2ε		$(G_j \in \{0, 1\}) \wedge (E = \varepsilon - 1)$	$\{3\varepsilon - 2, 3\varepsilon - 1\}$
$2\varepsilon + 1$		$(G_j \in \{1, 2\}) \wedge (E = 0)$	$\{2, 3\}$
$2\varepsilon + 2$		$(G_j \in \{1, 2\}) \wedge (E = 1)$	$\{5, 6\}$
$\cdots$		$\cdots$	$\cdots$
3ε		$(G_j \in \{1, 2\}) \wedge (E = \varepsilon - 1)$	$\{3\varepsilon - 1, 3\varepsilon\}$
$3\varepsilon + 1$		$(G_j = 2) \wedge (E = 0)$	$\{3\}$
$3\varepsilon + 2$		$(G_j = 2) \wedge (E = 1)$	$\{6\}$
$\cdots$		$\cdots$	$\cdots$
4ε		$(G_j = 2) \wedge (E = \varepsilon - 1)$	$\{3\varepsilon\}$
$4\varepsilon + 1$	G	$(G_j = 0) \wedge (E \in \mathcal{E}_\varepsilon)$	$\{1, 4, \ldots, 3\varepsilon - 2\}$
$4\varepsilon + 2$		$(G_j \in \{0, 1\}) \wedge (E \in \mathcal{E}_\varepsilon)$	$\{1, 2, 4, 5, \ldots, 3\varepsilon - 2, 3\varepsilon - 1\}$
$4\varepsilon + 3$	E	$(G_j \in \mathcal{G}) \wedge (E = 1)$	$\{4, 5, 6\}$
$\cdots$		$\cdots$	$\cdots$
$q = 5\varepsilon + 1$		$(G_j \in \mathcal{G}) \wedge (E = \varepsilon - 1)$	$\{3\varepsilon - 2, 3\varepsilon - 1, 3\varepsilon\}$

Amongst the population of individuals with $X_j \in (A_l \cup B_l)$, let π_{jA_l1} and π_{jA_l0} denote the respective conditional probabilities of observing $X_j \in A_l$ for given cases and controls. That is, for each $y \in \{0, 1\} = \mathcal{Y}$, $\pi_{jA_ly} = \Pr(X_j \in A_l | Y = y, X_j \in (A_l \cup B_l))$. The test of hypotheses $H_0^{(j,l)}$ versus $H_a^{(j,l)}$ can therefore be written as

$$\begin{aligned} H_0^{(j,l)} &: \pi_{jA_l1} = \pi_{jA_l0} \\ H_a^{(j,l)} &: \pi_{jA_l1} \neq \pi_{jA_l0} \end{aligned} . \tag{4.3}$$

Let X_{j1k} and X_{j0k} denote the respective random numbers of cases and controls observed at level k of X_j, and let $\mathbf{X}_j$ be the $2 \times 3\varepsilon$ table comprised of the random vectors, $\mathbf{X}_{jy} = \left(X_{jy1}, \ldots, X_{jy\{3\varepsilon\}}\right)$, $y \in \mathcal{Y}$. Table 4.2 depicts table $\mathbf{X}_j$ at SNP locus j.

Table 4.2: Cross-Classification of Disease Status and Level in X_j

	Level in X_j			Total
	1	$\cdots$	3ε	
Cases	X_{j11}	$\cdots$	$X_{j1\{3\varepsilon\}}$	n_1
Controls	X_{j01}	$\cdots$	$X_{j0\{3\varepsilon\}}$	n_0
Total	c_{j1}	$\cdots$	$c_{j\{3\varepsilon\}}$	n

For each $y \in \mathcal{Y}$ we have

$$\left(X_{jy1}, \ldots, X_{jy\{3\varepsilon\}}\right) \sim \text{Multinomial}\left(n_y, \boldsymbol{\pi}_{jy} = \left(\pi_{jy1}, \ldots, \pi_{jy\{3\varepsilon\}}\right)\right), \tag{4.4}$$

where $\pi_{jyk} = \Pr\left(X_j = k | Y = y\right)$ for all $k \in \mathcal{X}_\varepsilon$. For each $j = 1, \ldots, m$, $y \in \mathcal{Y}$, and $l = 1, \ldots, q$, we define the random variable D_{jyl} by

$$D_{jyl} = \sum_{k \in (A_l \cup B_l)'} X_{jyk}.$$

Note that the random variables D_{j1l} and D_{j0l} represent the respective random numbers of cases and controls whose value in X_j lies outside of both collections A_l and B_l. Also, each of these random variables (for a given value in l) will always assume the value of zero whenever L_{A_l} and L_{B_l} are complementary candidate patterns. For the sake of clarity in discussion, we consider D_{jyl} as a fixed known constant whenever the collection $\left(A_l \cup B_l\right)'$ is not empty, for each $j = 1, \ldots, m$, $y \in \mathcal{Y}$, and every $l = 1, \ldots, q$ (the consideration of these variables in their random state of nature is an open research question). Thus, for each $y \in \mathcal{Y}$ it follows by [165] (see pg. 167 therein) that the conditional distribution

$$\sum_{k \in A_l} X_{jyk} | D_{jyl} = d_{jyl} \sim \text{Binomial}\left(n_y - d_{jyl}, \pi_{jA_ly}\right). \tag{4.5}$$

Here, given $D_{jyl} = d_{jyl}$ for each $y \in \mathcal{Y}$, we consider testing the hypotheses (4.3) by applying the Wald-based statistic

$$T_{jA_l} = \hat{\pi}_{jA_l1} - \hat{\pi}_{jA_l0}, \tag{4.6}$$

against the data depicted within the 2×2 contingency table (Table 4.3), where

$$\hat{\pi}_{jA_ly} = \left(\frac{1}{n_y - d_{jyl}}\right) \sum_{k \in A_l} X_{jyk}$$

is the maximum likelihood estimator (MLE) of π_{jA_ly}.

Table 4.3: Collapsed $2 \times 3\varepsilon$ Table for Testing the Hypotheses (4.3).

	W_l		Total
	A_l	B_l	
Cases	$\sum_{k \in A_l} X_{j1k}$	$\sum_{k \in B_l} X_{j1k}$	$n_1 - d_{j1l}$
Controls	$\sum_{k \in A_l} X_{j0k}$	$\sum_{k \in B_l} X_{j0k}$	$n_0 - d_{j0l}$
Total	$\sum_{k \in A_l} c_{jk}$	$\sum_{k \in B_l} c_{jk}$	$n - d_{j0l} - d_{j1l}$

Under the null hypothesis (4.3), the standard error of (4.6),

$$se\left(T_{jA_l}\right) = \sqrt{\pi_{jA_l}\left(1 - \pi_{jA_l}\right)\left(\frac{1}{n_0 - d_{j0l}} + \frac{1}{n_1 - d_{j1l}}\right)}, \tag{4.7}$$

is estimated by

$$\hat{se}\left(T_{jA_l}\right) = \sqrt{\hat{\pi}_{jA_l}\left(1 - \hat{\pi}_{jA_l}\right)\left(\frac{1}{n_0 - d_{j0l}} + \frac{1}{n_1 - d_{j1l}}\right)}, \tag{4.8}$$

where $\pi_{jA_l} = \Pr\left(X_j \in A_l | X_j \in (A_l \cup B_l)\right)$ such that

$$\hat{\pi}_{jA_l} = \left(\frac{1}{n - d_{j0l} - d_{j1l}}\right) \sum_{k \in A_l} c_{jk}. \tag{4.9}$$

It is well known that under the null hypothesis of (4.3), the standardized Wald-based test statistic

$$Z_{jA_l} = \frac{T_{jA_l} - E\left(T_{jA_l}\right)}{\hat{se}\left(T_{jA_l}\right)}, \tag{4.10}$$

converges to the standard normal distribution, where $E\left(T_{jA_l} | H_0^{(j,l)}\right) = 0$. Under said null hypothesis, $Z_{jA_l}^2$ is asymptotically distributed as chi-square with one degree of freedom. Thus, if p_{jl} denotes the p-value for the test of hypotheses (4.3), it is

$$p_{jl} \doteq \Pr\left(\chi_1^2 \geq z_{jA_l}^2 | H_0^{(j,l)}\right),$$

where $z_{jA_l}^2$ denotes a realization of $Z_{jA_l}^2$ computed under $H_0^{(j,l)}$.

4.4 Multiple Hypothesis Testing Correction

If we conduct the q-fold test of hypotheses (4.3) across the index of l at locus j, a multiple testing problem is induced. Multiple testing correction to the p-values $p_{j1}, \ldots, p_{jq}$ could be performed by way of, say the Bonferroni MTP. However, because the test statistics $Z^2_{jA_1}, \ldots, Z^2_{jA_q}$ are likely correlated, implementation of the maxT MTP may result in a much more powerful adjustment [60]. Thus, if

$$Z^2_{j\max} = \max\left\{Z^2_{jA_1}, \ldots, Z^2_{jA_q}\right\}, \tag{4.11}$$

denotes the maximum chi-square test statistic under $\mathcal{H}_0^{(j)} = \cap_{l=1}^{q} H_0^{(j,l)}$ (here, called the complete null hypothesis at locus j), the maxT adjusted p-value for p_{jl}, denoted $\tilde{p}_{jl\sigma}$ (the 'σ' within the subscript is shorthand for single locus adjustment), is given by

$$\tilde{p}_{jl\sigma} = \Pr\left(Z^2_{j\max} \geq z^2_{jA_l} | \mathcal{H}_0^{(j)}\right) = \Pr\left(V \geq 1 | \mathcal{H}_0^{(j)}\right) = \text{FWER}, \tag{4.12}$$

where the final equality holds assuming $\mathcal{H}_0^{(j)}$ to in fact be the underlying truth regarding the null hypotheses $H_0^{(j,1)}, \ldots, H_0^{(j,q)}$, such that the random variable V corresponds to the number of Type I errors committed in testing $\mathcal{H}_0^{(j)}$. Hence, at the α level in the FWER, null hypothesis $H_0^{(j,l)}$ is rejected whenever $\tilde{p}_{jl\sigma} \leq \alpha$.

Furthermore, if we conduct the complete null hypothesis $\mathcal{H}_0^{(j)}$ across the m SNP loci and if

$$Z^2_{\max} = \max\left\{Z^2_{1\max}, \ldots, Z^2_{m\max}\right\}, \tag{4.13}$$

denotes the maximum chi-square test statistic under $\mathcal{H}_0 = \cap_{j=1}^{m} \mathcal{H}_0^{(j)}$ (called the complete null hypothesis), the maxT adjusted p-value for p_{jl}, denoted $\tilde{p}_{jl\mu}$ (the 'μ' within the subscript is shorthand for multiple loci adjustment), is given by

$$\tilde{p}_{jl\mu} = \Pr\left(Z^2_{\max} \geq z^2_{jA_l} | \mathcal{H}_0\right) = \Pr\left(V \geq 1 | \mathcal{H}_0\right) = \text{FWER}, \tag{4.14}$$

where the final equality holds assuming $\mathcal{H}_0$ to in fact be the underlying truth regarding the null hypotheses $\{H_0^{(j,1)}, \ldots, H_0^{(j,q)}\}_{j=1,\ldots,m}$, such that the random variable V corresponds to the number of Type I errors committed in testing $\mathcal{H}_0$. Therefore, in testing all $m \times q$ null hypotheses represented by $\mathcal{H}_0$, at the α level in the FWER null hypothesis $H_0^{(j,l)}$ is rejected whenever $\tilde{p}_{jl\mu} \leq \alpha$.

4.5 A Permutation Approach to the Multiple Testing Problem

Note that, as presented, each of the expressions (4.12) and (4.14) assume *weak control* over the FWER (see §1.3 for definitions of weak and strong control of Type I error rates). Strong control of the FWER will be assured to hold upon each of these expressions, provided that we can establish GEMs adherence to the property of *subset pivotality* (see §2.2.4 for definition of this property). If $\mathbf{P}_j = \left(P^*_{j1}, \ldots, P^*_{jq}\right)$ denotes the vector of unadjusted p-values corresponding with the test statistics $Z^*_{jA_1}, \ldots, Z^*_{jA_q}$ – where, for all $l = 1, \ldots, q$, $Z^*_{jA_l}$ is given by (4.10) with (4.7) substituted in lieu of (4.8) therein – we show within Proposition A.9 that the distribution of $\mathbf{P}_j$ fails adherence to the property of subset pivotality for the circumstance in which E is binary. An unfortunate consequence of this result is that the maxT MTP, as implemented within GEM upon a sampled binary environmental factor, can be assumed to control the FWER only in the weak sense [62]. However, while the subset pivotality condition is sufficient for strong control of the FWER upon the maxT MTP, it is not clear whether the condition is necessary for strong control of the FWER. Within §4.9 we conduct a simulation study to examine this notion closer for GEM.

4.5.1 Single Genetic Marker

Consider the maximum test statistic over the q-fold test of hypotheses (4.3) conducted at locus j, $Z^2_{j\max}$ (4.11). In order to compute the adjusted p-value $\tilde{p}_{jl\sigma}$ (4.12), $l = 1, \ldots, q$ – with the intention of making a decision of whether to reject or fail to reject $H_0^{(j,l)}$ – we could approximate the distribution of $Z^2_{j\max}$ under $\mathcal{H}_0^{(j)}$ by way of an appropriately parameterized multivariate normal distribution (MVN). This could be accomplished by simply modifying the MVN framework of [109] (utilized for testing SNP-SNP interactions) to suit the candidate patterns corresponding with our GEM setup. However, there are several problems with this approach (for clarity in presentation, see §4.11 for details). Here, we abstain from approximating the distribution of $Z^2_{j\max}$ under $\mathcal{H}_0^{(j)}$ with an asymptotic MVN distribution.

Alternatively, we consider computing the adjusted p-value $\tilde{p}_{jl\sigma}$ (4.12) by way of the permutation null distribution of $Z^2_{j\max}$ under $\mathcal{H}_0^{(j)}$. Conditional on the [assumed fixed] values of the margins for the $2 \times 3\varepsilon$ table (Table 4.2), the permutation null distribution of $Z^2_{j\max}$ under $\mathcal{H}_0^{(j)}$ can be determined by enumerating every possible arrangement of the cell values in the table. We call this set of tables

the conditional reference set

$$\Gamma_{\mathbf{c}_j} = \left\{ \mathbf{x}_j : \sum_{y \in \mathcal{Y}} x_{jyk} = c_{jk} \ \forall k \in \mathcal{X}_\varepsilon, \ \sum_{k \in \mathcal{X}_\varepsilon} x_{jyk} = n_y \ \forall y \in \mathcal{Y} \right\}, \tag{4.15}$$

where $\mathbf{x}_j$ is a realization of the random table $\mathbf{X}_j$ and $\mathbf{c}_j = \left(c_{j1}, \ldots, c_{j\{3\varepsilon\}}\right)$ is the vector of values pertaining to the column margin of the $2 \times 3\varepsilon$ table. Under $\mathcal{H}_0^{(j)}$, each table within this set has an affiliated probability of being realized and a corresponding realization of $Z^2_{j\max}$. Computing the *exact conditional* maxT adjusted p-value under $\mathcal{H}_0^{(j)}$ at realization $z^2_{jA_l}$, namely $\tilde{p}_{jl\sigma}$, involves finding the exact tail area for the distribution of $Z^2_{j\max}$ over the conditional reference set.

Enumerating the set $\Gamma_{\mathbf{c}_j}$ presents a difficult computational problem, irrespective of the number of levels (i.e., ε) to the environmental factor under study. In the circumstance for which a binary environmental factor ($\varepsilon = 2$) is under study, within §4.6 we present a network algorithm approach to tackling the computational problem, whereby we are able to compute the exact conditional maxT adjusted p-value $\tilde{p}_{jl\sigma}$ under $\mathcal{H}_0^{(j)}$, for all $l = 1, \ldots, q$. For all other circumstances encompassing the environmental factor (i.e., $\varepsilon > 2$), as a compromise to enumerating $\Gamma_{\mathbf{c}_j}$ in its entirety we propose estimating $\tilde{p}_{jl\sigma}$ by way of sampling from the permutation null distribution of $Z^2_{j\max}$ under $\mathcal{H}_0^{(j)}$. We provide the underlying details for doing this within §4.5.3.

4.5.2 Multiple Genetic Markers

Consider the maximum test statistic over the q-fold tests of hypotheses (4.3) conducted across the m loci, $Z^2_{\max}$ (4.13). In order to compute the adjusted p-value $\tilde{p}_{jl\mu}$ (4.14), each $j = 1, \ldots, m$ and $l = 1, \ldots, q$, we consider application of the permutation null distribution of $Z^2_{\max}$ under $\mathcal{H}_0$. In an analogous manner to §4.5.1, we would like to draw inference for this distribution by conditioning on the fixed values of the margins upon an appropriate two-way table. However, this is a bit complicated here, because it requires evaluation of the appropriate conditional reference set over the joint distribution of the vector of random variables $(X_1, \ldots, X_m, Y)$. It is not immediately clear how to represent this random vector by a two-way table. We proceed by noting that for any categorical variable X, the joint distribution of (X, Y) can be conveniently depicted by a two-way contingency table. Thus, if we can collectively summarize the X_j by some random variable X, this would allow for closed-form formulation of the conditional reference (constructed about a two-way table) set for the permutation null distribution of $Z^2_{\max}$ under $\mathcal{H}_0$. Indeed, with the spirit of (4.2) in mind, we

consider

$$X = 1 + 3^m E + \sum_{j=1}^{m} 3^{j-1} G_j, \tag{4.16}$$

where $X \in \{1, \ldots, 3^m\varepsilon\} = \mathcal{X}_{\varepsilon m}$. Note: each $X \in \mathcal{X}_{\varepsilon m}$ corresponds to a unique specification of the random vector $(G_1, \ldots, G_m, E)$ – the proof here is a slight modification to that given within Proposition A.2; and X indirectly summarizes the random variables X_j, by way of the random variables G_j and E.

Let X_{1k} and X_{0k} denote the respective random numbers of cases and controls observed at level k of the random variable X, and let $\mathbf{X}$ denote the $2 \times 3^m\varepsilon$ table comprised of the random vectors $\left(X_{y1}, \ldots, X_{y\{3^m\varepsilon\}}\right)$, $y \in \mathcal{Y}$. Table 4.4 depicts table $\mathbf{X}$, where it is assumed no missing data values are prevalent. Conditional on the values of the margins for the $2 \times 3^m\varepsilon$ table (Table 4.4), the permutation null distribution of $Z^2_{\max}$ under $\mathcal{H}_0$ can be determined by enumerating the conditional reference set

$$\Gamma_{\mathbf{c}} = \left\{ \mathbf{x} : \sum_{y \in \mathcal{Y}} x_{yk} = c_k \ \forall k \in \mathcal{X}_{\varepsilon m}, \ \sum_{k \in \mathcal{X}_{\varepsilon m}} x_{yk} = n_y \ \forall y \in \mathcal{Y} \right\}, \tag{4.17}$$

where $\mathbf{x}$ is a realization of the random table $\mathbf{X}$ and $\mathbf{c} = (c_1, \ldots, c_{3^m\varepsilon})$ is the vector of values pertaining to the column margin of the $2 \times 3^m\varepsilon$ table. Under $\mathcal{H}_0$, each table within this set has an affiliated probability of being realized and a corresponding realization of $Z^2_{\max}$. Computing the *exact conditional* maxT adjusted p-value under $\mathcal{H}_0$ at realization $z^2_{jA_l}$, namely $\tilde{p}_{jl\mu}$, involves finding the exact tail area for the distribution of $Z^2_{\max}$ over the conditional reference set.

Table 4.4: Cross-Classification of Disease Status and X

	Level in X			Total
	1	$\cdots$	$3^m\varepsilon$	
Cases	X_{11}	$\cdots$	$X_{1\{3^m\varepsilon\}}$	n_1
Controls	X_{01}	$\cdots$	$X_{0\{3^m\varepsilon\}}$	n_0
Total	c_1	$\cdots$	$c_{3^m\varepsilon}$	n

Enumerating the set $\Gamma_{\mathbf{c}}$ presents an exceptionally difficult computational problem, where the magnitude of the computational burden is positively associated with each of the values in ε and m. This is due to the number of columns for the $2 \times 3^m\varepsilon$ table increasing whenever either of the values ε or m increase. In fact, each table within the conditional reference set $\Gamma_{\mathbf{c}_j}$ (4.15) relates to at least

one table within $\Gamma_{\mathbf{c}}$, for all $j = 1, \ldots, m$. As a compromise to enumerating $\Gamma_{\mathbf{c}}$ in its entirety, we propose estimating $\tilde{p}_{jl\mu}$ by way of sampling from the permutation null distribution of $Z^2_{\max}$ under $\mathcal{H}_0$. We provide the underlying details for doing this within §4.5.3.

4.5.3 Sampling from the Permutation Null Distribution

For each $i = 1, \ldots, n$ and $j = 1, \ldots, m$, let ge_{ji} (here, ge is used to signify the level in GxE) denote the realization in X_j (4.2) for study subject i, and let $\mathbf{ge}_i = (ge_{1i}, \ldots, ge_{mi})'$ denote the realized profile of the vector of random variables $(X_1, \ldots, X_m)$ for said subject, where we assign missing data in the ge_{ji} to the value of zero. In accordance with §2.3.1, let $\mathbf{y}^*$ correspond to the specified ordering of the case-control responses (2.11), and let $\mathbf{GE}^*$ denote the matrix of ordered profiles in the $\mathbf{ge}_i$ corresponding with the chosen $\mathbf{y}^*$. The maxT adjusted p-values (4.12) and (4.14) can be estimated by utility of Algorithm 4.1.

Algorithm 4.1 A Permutation Approach for GEM

1. Initialize the $q \times 3\varepsilon$ matrix $\mathbf{I}$, whose $(l,k)^{\text{th}}$ entry (denoted $[\mathbf{I}]_{(l,k)}$) is an indicator for membership of realization $X_j = k$ to A_l, B_l, or neither of these two collections. Specifically, for all $l = 1, \ldots, q$ and $k \in \mathcal{X}_\varepsilon$, we define $\mathbf{I}$ by

$$[\mathbf{I}]_{(l,k)} = I\left(k \in A_l\right) - I\left(k \in (A_l \cup B_l)'\right).$$

   Initialize the $m \times (3\varepsilon + 3)$ matrix $\mathbf{M}$, whose $(j,k)^{\text{th}}$ entry (denoted $[\mathbf{M}]_{(j,k)}$) warehouses pertinent [permutation invariant] information for construction of the 2×2 table (depicted by Table 4.3) at locus j. Namely, upon row j of $\mathbf{M}$: the initial 3ε elements warehouse the values of the column margin upon Table 4.2; elements $3\varepsilon + 1$ and $3\varepsilon + 2$ warehouse the values of the row margin upon Table 4.2 (i.e., the non-missing case and control data upon row j of $\mathbf{GE}^*$); and element $3\varepsilon + 3$ warehouses the number of non-missing data values upon row j of $\mathbf{GE}^*$. Specifically, for all $j = 1, \ldots, m$ and $k = 1, \ldots, 3\varepsilon + 3$, it is

$$[\mathbf{M}]_{(j,k)} = \begin{cases} \sum_{i=1}^{n} I\left(ge^*_{ji} = k\right), & \text{if } k \leq 3\varepsilon \\ \sum_{i=1}^{n_0} I\left(ge^*_{ji} > 0\right), & \text{if } k = 3\varepsilon + 1 \\ \sum_{i=n_0+1}^{n} I\left(ge^*_{ji} > 0\right), & \text{if } k = 3\varepsilon + 2 \\ \sum_{i=1}^{n} I\left(ge^*_{ji} > 0\right), & \text{if } k = 3\varepsilon + 3 \end{cases},$$

where ge^*_{ji} is the $(j,i)^{\text{th}}$ element of $\mathbf{GE}^*$.

2. Compute the realization in the test statistic (4.10) for the observed (i.e., non-permuted) data, as follows. For $j = 1, \ldots, m$:

   (a) Formulate the values of the case row, say, upon the $2 \times 3\varepsilon$ table (Table 4.2), by appropriate evaluation of the latter n_1 elements upon row j of $\mathbf{GE}^*$. Specifically, for each $k \in \mathcal{X}_\varepsilon$, if x_{j1k} denotes the realization in X_{j1k}, we evaluate the following expression:

$$x_{j1k} = \sum_{i=n_0+1}^{n} I\left(ge^*_{ji} = k\right).$$

   (b) For $l = 1, \ldots, q$:

      A. Formulate the values of the column margin upon the 2×2 table (Table 4.3), by way of the following formulas

$$\sum_{k \in A_l} c_{jk} = \sum_{k \in \mathcal{X}_\varepsilon} \left(I\left([\mathbf{I}]_{(l,k)} = 1 \right) [\mathbf{M}]_{(j,k)} \right)$$
$$\sum_{k \in B_l} c_{jk} = \begin{cases} [\mathbf{M}]_{(j,3\varepsilon+3)} - \sum_{k \in A_l} c_{jk}, & \text{if } \varepsilon = 2 \\ \sum_{k \in \mathcal{X}_\varepsilon} \left(I\left([\mathbf{I}]_{(l,k)} = 0 \right) [\mathbf{M}]_{(j,k)} \right), & \text{if } \varepsilon > 2 \end{cases}.$$

      B. Formulate the values of the case row upon the 2×2 table (Table 4.3), by way of the following formulas

$$\sum_{k \in A_l} x_{j1k} = \sum_{k \in \mathcal{X}_\varepsilon} \left(x_{j1k} I\left([\mathbf{I}]_{(l,k)} = 1 \right) \right)$$
$$\sum_{k \in B_l} x_{j1k} = [\mathbf{M}]_{(j,3\varepsilon+2)} - \sum_{k \in A_l} x_{j1k}.$$

      C. Formulate the values of the control row upon said 2×2 table by subtracting the values of the case row (computed in step 2(b)B above) from those of the column margin (computed in step 2(b)A above). Utilizing these computed values, along within those computed within steps 2(b)A and 2(b)B, and those over $\left\{ [\mathbf{M}]_{(j,3\varepsilon+k)} \right\}_{k=1,2,3}$, compute the realization of Z_{jA_l} and denote it by z_{jA_l}.

3. We consider permuting the columns upon $\mathbf{GE}^*$. Permuting in this manner (i.e., retaining the observed configuration in the response vector $\mathbf{y}^*$), ensures that the phenotype data is

independent of the genetic and environment data. As a result, we are: simulating each of the complete null hypotheses over $\left\{\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(m)}, \mathcal{H}_0\right\}$; and maintaining the correlation structure in $\mathbf{GE}^*$. Let R denote the desired number of permutations over the columns upon $\mathbf{GE}^*$.

4. For $r = 1, \ldots, R$:

   (a) Shuffle the columns upon $\mathbf{GE}^*$. For each $j = 1, \ldots, m$ and $l = 1, \ldots, q$, compute the realization in the test statistic (4.10) for the permuted data, by repeating step 2 above upon the permuted matrix $\mathbf{GE}^*$, with the following modifications: change the formulas presented within step 2(a)B to

$$\sum_{k \in A_l} x_{j1k} = \sum_{k \in \mathcal{X}_\varepsilon} \left(x_{j1k} I\left([\mathbf{I}]_{(l,k)} = 1 \right) \right)$$
$$\sum_{k \in B_l} x_{j1k} = \sum_{k \in \mathcal{X}_\varepsilon} \left(x_{j1k} I\left([\mathbf{I}]_{(l,k)} = 0 \right) \right);$$

   and, change the second sentence within step 2(b)C to: utilizing these computed values, along with those computed within steps 2(b)A and 2(b)B, and – due to potential missing values in the ge^*_{ji} – possibly those over $\left\{ [\mathbf{M}]_{(j,3\varepsilon+k)} \right\}_{k=1,2,3}$, compute the realization of Z_{jA_l} and denote it by $z^{(r)}_{jA_l}$. Let $z^{(r)}_{\max}$, and for each $j = 1, \ldots, m$, $z^{(r)}_{j\max}$, be defined by

$$z^{(r)}_{j\max} = \max\left\{ |z^{(r)}_{jA_1}|, \ldots, |z^{(r)}_{jA_q}| \right\}, \qquad \text{and} \qquad z^{(r)}_{\max} = \max\left\{ z^{(r)}_{1\max}, \ldots, z^{(r)}_{m\max} \right\}.$$

5. For each $j = 1, \ldots, m$ and $l = 1, \ldots, q$, let $\tilde{p}^*_{jl\sigma}$ and $\tilde{p}^*_{jl\mu}$ denote the respective estimates of the maxT permutation adjusted p-values for (4.12) and (4.14), for our having to sample from the permutation null distributions for the respective test statistics $Z^2_{j\max}$ and $Z^2_{\max}$. These values are given by

$$\tilde{p}^*_{jl\sigma} = \frac{\sum_{r=1}^{R} I\left(z^{(r)}_{j\max} \geq |z_{jA_l}| \right)}{R}, \tag{4.18}$$

   and

$$\tilde{p}^*_{jl\mu} = \frac{\sum_{r=1}^{R} I\left(z^{(r)}_{\max} \geq |z_{jA_l}| \right)}{R}. \tag{4.19}$$

4.6 An Exact Approach to Assessing GxE Interaction upon a Single Genetic Marker and a Binary Environment Factor

Consider the simplest circumstance in which one could assess GxE interaction in a case-control study using GEM. Namely, it is the assessment of GxE interaction upon a single SNP marker ($m = 1$) and a binary environment factor ($\varepsilon = 2$). Under this condition, cross-classification of disease status and the random variable X (4.16) (equivalent to X_1 (4.2)) can be conveniently depicted by a 2×6 contingency table, as shown by Table 4.5. Conditional on the assumed fixed values of the margins for this table, the permutation null distributions under $\mathcal{H}_0$ (equivalent to $\mathcal{H}_0^{(1)}$) for the [equivalent] test statistics, $Z^2_{1\max}$ (4.11) and $Z^2_{\max}$ (4.13), can be determined by enumerating the conditional reference set (4.17), where $\mathbf{c} = \mathbf{c}_1 = (c_{11}, \ldots, c_{16})$ and $\mathcal{X}_{2\{1\}} = \mathcal{X}_2 = \{1, \ldots, 6\}$. Under $\mathcal{H}_0$, the exact conditional maxT adjusted p-value at realization $z^2_{1A_l}$, $\tilde{p}_{jl\mu}$ (equivalent to $\tilde{p}_{jl\sigma}$) , can be computed from the permutation null distribution of $Z^2_{\max}$, provided: that we have access to an explicit formula for the conditional probability mass function (PMF) of $\mathbf{X}_1$, given $\mathbf{x} \in \Gamma_{\mathbf{c}}$, so that we can compute exact tail area from said null distribution; and, a formula which relates this conditional PMF to computing the exact conditional probability of Type I error for the test statistic $Z^2_{\max}$ under $\mathcal{H}_0$.

Table 4.5: Cross-Classification of Disease Status and Level in X_1 for a Binary Environmental Factor.

	Level in X_1			Total
	1	$\cdots$	6	
Cases	X_{111}	$\cdots$	X_{116}	n_1
Controls	X_{101}	$\cdots$	X_{106}	n_0
Total	c_{11}	$\cdots$	c_{16}	n

We first derive the conditional probability mass function for the table $\mathbf{X}_1$ under $\mathcal{H}_0$, given $\mathbf{x} \in \Gamma_{\mathbf{c}}$, which will be used to compute the exact conditional maxT adjusted p-value, $\tilde{p}_{1l\mu}$, for all $l = 1, \ldots, q$. It can be shown (see Proposition A.10 of Appendix A) that $\mathcal{H}_0$ is equivalent to the null hypothesis of no association between X_1 and Y (disease status), which can be written as

$$H_0 : \pi_{10k} = \pi_{11k} \ \forall k \in \mathcal{X}_2 \qquad \Longleftrightarrow \qquad H_0 : \boldsymbol{\pi}_{10} = \boldsymbol{\pi}_{11}, \tag{4.20}$$

where for each $y \in \mathcal{Y}$ and $k \in \mathcal{X}_2$, π_{1yk} and $\boldsymbol{\pi}_{1y}$ are defined by (4.4). Because the data arise from two independent multinomial populations (see (4.4)), under $\mathcal{H}_0$ (equivalent to (4.20)) the probability

mass function of $\mathbf{X}_1$, conditional on $\mathbf{x} \in \Gamma_{\mathbf{c}}$, is given by

$$
\begin{aligned}
h\left(\mathbf{x}|\mathbf{c}, \mathcal{H}_0\right) &= \Pr\left(\mathbf{X}_1 = \mathbf{x}|\mathbf{X}_{11} + \mathbf{X}_{10} = \mathbf{c}, \mathcal{H}_0\right) \\
&= \frac{\Pr\left(\mathbf{X}_{11} = \mathbf{x}_{11}\right) \Pr\left(\mathbf{X}_{10} = \mathbf{c} - \mathbf{x}_{11}\right)}{\Pr\left(\mathbf{X}_{11} + \mathbf{X}_{10} = \mathbf{c}\right)} \\
&= \frac{\prod_{k \in \mathcal{X}_2} \binom{c_{1k}}{x_{11k}}}{\binom{n}{n_1}}.
\end{aligned} \tag{4.21}
$$

Next, we relate this conditional PMF to computing the exact conditional probability of Type I error for the test statistic $Z^2_{\max}$ under $\mathcal{H}_0$. Let $\mathbf{x} \in \Gamma_{\mathbf{c}}$ be an arbitrarily chosen table from the conditional reference set, and let $T\left(\mathbf{x}\right)$ denote the realization in the test statistic $Z^2_{\max}$ corresponding to the table, computed under $\mathcal{H}_0$. Thus, the critical region of the test, corresponding with our observed test statistic $z^2_{1A_l}$, denoted $\Gamma_{\mathbf{c}}\left(z^2_{1A_l}\right)$, is given by

$$
\Gamma_{\mathbf{c}}\left(z^2_{1A_l}\right) = \left\{\mathbf{x} \in \Gamma_{\mathbf{c}} : T\left(\mathbf{x}\right) \geq z^2_{1A_l}\right\}. \tag{4.22}
$$

Therefore, in terms of (4.21) and (4.22), the exact conditional maxT adjusted p-value, $\tilde{p}_{1l\mu}$, is given by

$$
\tilde{p}_{1l\mu} = \sum_{\mathbf{x} \in \Gamma_{\mathbf{c}}\left(z^2_{1A_l}\right)} h\left(\mathbf{x}|\mathbf{c}, \mathcal{H}_0\right) \;\; \forall l = 1, \ldots, q. \tag{4.23}
$$

For a relatively small random sample of cases and controls the exact conditional maxT adjusted p-value (4.23) can be computed, by explicitly enumerating all possible tables within $\Gamma_{\mathbf{c}}$. For example, the data depicted within Table 4.6, represent a hypothetical random sample of $n_1 = n_0 = 4$ cases and controls – there are 32 tables within $\Gamma_{\mathbf{c}}$ for these data. However, upon larger case-control samples, explicit enumeration of $\Gamma_{\mathbf{c}}$ becomes computationally prohibitive. For example, even for a reasonably small random sample of $n_1 = n_0 = 100$ cases and controls, the number of tables comprising $\Gamma_{\mathbf{c}}$ lies in the millions (assumes the distribution of the elements upon $\mathbf{c}$ here is the same as that for the example of $n_1 = n_0 = 4$); the computational problem is greatly exacerbated for case-control samples comprising thousands of study subjects. Given these considerations, it is necessary that we possess a tool which provides computational efficiency in practice, and in which can also accommodate the restrictions imposed by the conditioning over the elements upon $\Gamma_{\mathbf{c}}$. A network algorithm is such a

tool.

Table 4.6: Cross-Classification of Disease Status and X_1 for a Small Case-Control Sample.

	Level in X_1						Total
	1	2	3	4	5	6	
Cases	1	1	1	1	0	0	4
Controls	0	1	0	0	2	1	4
Total	1	2	1	1	2	1	8

4.6.1 A Network Algorithm

The pioneer work of network algorithms can be traced back to two articles published in the early 1980's: the article of [166] examined the circumstance of exploiting a network algorithm for conducting exact inference upon $2 \times k$ contingency tables; and, the article of [167] generalized this network approach to include exact inference upon $r \times k$ contingency tables. The network algorithms of these papers have been applied to a number of other computationally challenging problems, such as exact inference over $2 \times r \times k$ contingency tables [168], and exact inference for ordered $r \times k$ contingency tables [169]. More recent applications of network algorithms can be found, for example, within the articles of [141, 142, 143, 170].

A network algorithm is an efficient means by which to develop and process a conditional reference set. Specifically, the network representation of a conditional reference set for a $2 \times k$ table is a directed acyclic network of nodes and arcs [143], originating upon a single node (called the initial node) and ending upon a single node (called the terminal node). A path across the network: is defined as a series of k connected arcs which emanate from the initial node and reach the terminal node, passing through $k - 1$ intermediate nodes; represents a unique table within the conditional reference set; and, has an affiliated probability under $\mathcal{H}_0$. Here, the problem of defining the critical region of the test (4.22), reduces to finding all paths across the network which adhere to a prespecified condition. When compared to explicit enumeration of the conditional reference set, there are two advantages to the network approach in computing the p-value (4.23). First, there are typically far fewer nodes in the network than tables in the conditional reference set. This provides a condensed means by which to portray said set. Second, the calculation of the p-value involves summing up the probabilities of all of the paths satisfying the prespecified condition, with no need to explicitly enumerate each path. The decision as to whether or not a particular path contributes to the p-value is made upon the nodes

of the network as it (the network) is being processed. In this regard, considerable computational savings can be realized by not having to explicitly enumerate the paths of the network.

The conditional reference set for a $2 \times k$ contingency table with fixed column margin $\mathbf{c} = (c_1, \ldots, c_k)$ and fixed row margin (n_1, n_0), $\Gamma_{\mathbf{c}}$ (here, within (4.17) we replace $\mathcal{X}_{\varepsilon m}$ with the collection $\{1, \ldots, k\}$), can be represented by a directed acyclic network of nodes and arcs. The nodes are structured over $k+1$ stages, the stages in which are labelled $0, \ldots, k$. In any stage s, there is a nonempty set of nodes and each of them is labelled by a pair of integer values (s, m_s), such that

$$\max\left\{0, n_1 - n + c_{(s)}\right\} \leq m_s \leq \min\left\{c_{(s)}, n_1\right\},$$

where $n = n_0 + n_1$, $c_{(s)} = c_1 + \cdots + c_s$, and $s = 0, \ldots, k$. The value of m_s is the sum over the initial s columns of the first row upon some table(s) in $\Gamma_{\mathbf{c}}$. In particular, within stage 0 there is a single node $(0, m_0)$ with $m_0 = 0$ (initial node) and in stage k there is also a single node (k, m_k) with $m_k = n_1$ (terminal node). Arcs emanate upon each node of stage s, such that each arc (upon a given node) is directed towards a particular node of stage $s+1$, for all $s = 0, \ldots, k-1$. A node of stage $s+1$, say $(s+1, m_{s+1})$, which is joined by an arc with a node (s, m_s) of stage s is called a direct successor of node (s, m_s), and we write $(s, m_s) \to (s+1, m_{s+1})$ to signify that these nodes are joined by an arc. The collection of direct successors of node (s, m_s) are the elements of the set $\Psi(s, m_s)$, where

$$\Psi(s, m_s) = \left\{(s+1, w) : \max\left\{m_s, n_1 - c_{(k)} + c_{(s+1)}\right\} \leq w \leq \min\left\{n_1, m_s + c_{s+1}\right\}\right\}. \tag{4.24}$$

Here, for some $s = 0, \ldots, k-1$ and $b = 1, \ldots, k-s$, we say that node $(s+b, m_{s+b})$ is a successor node of (s, m_s) if and only if there exists some collection of nodes $\{(s, m_s), \ldots, (s+b, m_{s+b})\}$ satisfying

$$(s+a, m_{s+a}) \to (s+a+1, m_{s+a+1}) \;\; \forall a = 0, \ldots, b-1,$$

in which case a subpath between nodes (s, m_s) and $(s+b, m_{s+b})$ is defined by the series of arcs connecting said collection of nodes. Thus, the collection of nodes $\{(s, m_s)\}_{s=0,\ldots,k}$ forms a path through the network if and only if for every pair of nodes over this collection, say (a, m_a) and (b, m_b) depicting some pair of nodes, it holds that the latter node is a successor node of the former node whenever $0 \leq a < b \leq k$. Figure 4.1 illustrates the 32 paths for the network representation of

the conditional reference set for Table 4.6, where the line segment connecting nodes (s, m_s) and $(s+1, m_{s+1})$ depicts the arc for $(s+1, m_{s+1}) \in \Psi(s, m_s)$. The dashed path corresponds to the observed table.

Now, for implementation of the network algorithm approach to GEM upon a sampled binary environment factor, we consider $k = n(\mathcal{X}_2)$ (here, $n(\cdot)$ denotes the cardinality of the set $(\cdot)$). Let $\mathbf{I}$ be the $q \times k$ matrix, whose $(l,w)^{\text{th}}$ element (denoted $[\mathbf{I}]_{(l,w)}$) is an indicator for realization $X_j = w$, $w = 1, \ldots, k$, to the collection A_l ($[\mathbf{I}]_{(l,w)} = 1$) or B_l ($[\mathbf{I}]_{(l,w)} = 0$), where it is noted that L_{A_l} and L_{B_l} are complementary candidate patterns for all $l = 1, \ldots, q$. Under $\mathcal{H}_0$, the arc joining nodes (s, m_s) and $(s+1, m_{s+1}) \in \Psi(s, m_s)$, $s = 0, \ldots, k-1$, has a q-vector of rank lengths defined by

$$\mathbf{r}_{s+1} = \left([\mathbf{I}]_{(1,s+1)} (m_{s+1} - m_s), \ldots, [\mathbf{I}]_{(q,s+1)} (m_{s+1} - m_s) \right), \tag{4.25}$$

and – in accordance with (4.21) – associated probability length

$$p_{s+1} = \binom{c_{s+1}}{m_{s+1} - m_s}. \tag{4.26}$$

The probability of table $\mathbf{x} \in \Gamma_{\mathbf{c}}$, $h(\mathbf{x}|\mathbf{c}, \mathcal{H}_0)$ (4.21), is therefore recovered by taking the product of the probability lengths of the arcs which comprise the corresponding path through the network and subsequently dividing this resultant by the binomial coefficient n choose n_1; and the q-vector of rank lengths for this table, denoted $\mathbf{r}(\mathbf{x})$, is obtained by summing over the k q-vectors of rank lengths of the arcs which comprise the corresponding path through the network.

Having constructed the network, we could traverse through it to identify the tail area of the permutation null distribution of $Z^2_{\max}$ under $\mathcal{H}_0$. However, there are two problems with this approach here, each leading to the potential for increased computations. First, note that for any $q' < q$ and $\Lambda_{q'}$ some q'-size proper subset of $\{1, \ldots, q\}$, (4.14) can be written as

$$\tilde{p}_{1l\mu} = \Pr\left(Z^2_{\max} \geq z^2_{1A_l} | \mathcal{H}_0\right) = \Pr\left(\bigcup_{v=1}^{q} Z^2_{1A_v} \geq z^2_{1A_l} | \mathcal{H}_0 \right) \geq \Pr\left(\bigcup_{v \in \Lambda_{q'}} Z^2_{1A_v} \geq z^2_{1A_l} | \mathcal{H}_0 \right), \tag{4.27}$$

with strict inequality whenever the $Z^2_{1A_l}$, $l = 1, \ldots, q$ are not perfectly correlated. Assuming these test statistics are not perfectly correlated, this implies that the collection of tables within $\Gamma_{\mathbf{c}}$ contributing to the critical region for the permutation null distribution of $Z^2_{\max}$, is a richer set than that for the permutation null distribution of the maximum test statistic over $\{Z_{1A_v}\}_{v \in \Lambda_{q'}}$ for all

$q' < q$ and any $\Lambda_{q'} \subset \{1, \ldots, q\}$. But, the computational burden in computing (4.14) is positively associated with the number of tables within $\Gamma_{\mathbf{c}}$ contributing to the critical region of the permutation null distribution of $Z^2_{\max}$, and – in light of (4.27) – so also positively associated with the value of q. This could lead to the potential for an increase in computations here, since $q = 11$ for a binary environment factor (see Table 4.1 for the relationship between ε and q). Second, our goal in processing the network is to abridge paths which do not contribute to the calculation of (4.14). Having to do this amidst the union over q (4.27), requires scrupulous pruning of paths over the joint distribution of the test statistics $\left\{Z^2_{1A_l}\right\}_{l=1,\ldots,q}$ under $\mathcal{H}_0$. This could lead to an increase in the computational demand when processing the network.

Alternatively, here we consider processing the network to determine the value of the complement of (4.14). It is,

$$\tilde{q}_{1l\mu} = 1 - \tilde{p}_{1l\mu} = \Pr\left(\bigcap_{v=1}^{q} Z^2_{1A_v} < z^2_{1A_l} | \mathcal{H}_0\right) = \Pr\left(\bigcap_{v=1}^{q} |Z_{1A_v}| < |z_{1A_l}| | \mathcal{H}_0\right), \tag{4.28}$$

for all $l = 1, \ldots, q$. The intersections over q within this expression are exceptionally attractive, because – as illustrated below within step 4(b) of the forward induction pass of Algorithm 4.2 – we can prune paths over the marginal permutation null distribution of $Z^2_{1A_l}$ for some $l = 1, \ldots, q$. Working upon marginal distributions in this regard is aesthetically appealing when compared to working upon the joint distribution of $\left\{Z^2_{1A_l}\right\}_{l=1,\ldots,q}$. Moreover, calculation of $\tilde{p}_{1l\mu}$ (4.14) from the computed value (4.28) is a trivial exercise in arithmetic.

We now outline the network algorithm approach for GEM. For each $l = 1 \ldots, q$, in terms of Table 4.5, let $X_{1A_l} = \sum_{v \in A_l} X_{11v}$ be the random number of cases – for an arbitrary table within the conditional reference set – whose values of X_1 lie within A_l and let $c_{1A_l} = \sum_{v \in A_l} c_{1v}$ be the number of sampled subjects whose values of X_1 lie within A_l. We make the observation that for any $t > 0$, under $\mathcal{H}_0$ it holds

$$|Z_{1A_l}| < t \qquad \Longleftrightarrow \qquad X_{1A_l} \in \left(\frac{n_1 c_{1A_l} - tK_l}{n}, \frac{n_1 c_{1A_l} + tK_l}{n}\right), \tag{4.29}$$

where

$$K_l = \sqrt{\frac{n_0 n_1 c_{1A_l} (n - c_{1A_l})}{n}}.$$

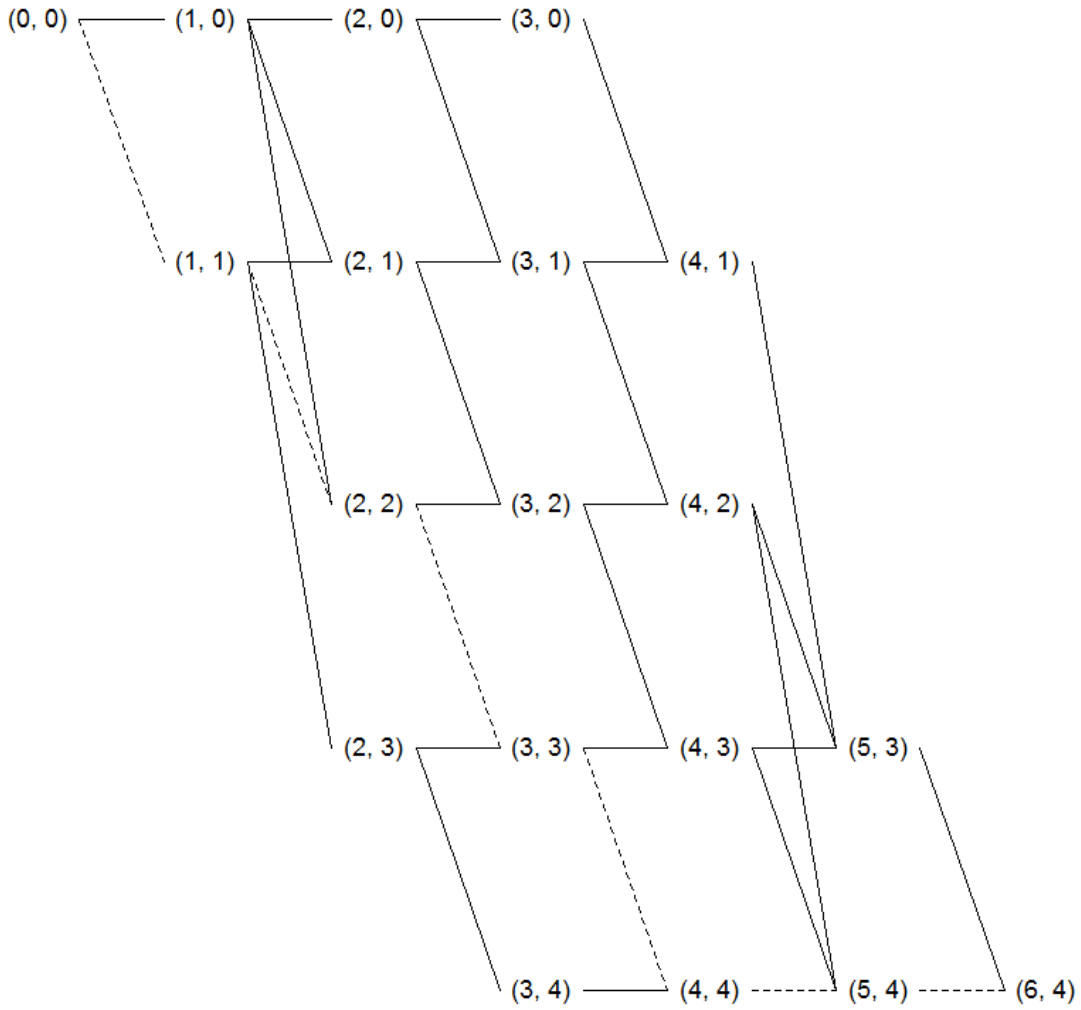

Fig. 4.1: The Network Representation of the Conditional Reference Set for the 2×6 Contingency Table Depicted by Table 4.6. The Observed Table Is Represented by the Dashed Path.

For some $t > 0$, let $\psi_1(l,t)$ and $\psi_2(l,t)$ be the functions defined by

$$\begin{aligned} \psi_1(l,t) &= \min\left\{w = 0,\ldots,n_1 : w > \frac{n_1 c_{1A_l} - tK_l}{n}\right\} \\ \psi_2(l,t) &= \max\left\{w = 0,\ldots,n_1 : w < \frac{n_1 c_{1A_l} + tK_l}{n}\right\} \end{aligned} . \tag{4.30}$$

Thus, for any $t > 0$, it holds that

$$\begin{aligned} & \mathbf{x} \in \Gamma_{\mathbf{c}}(t) \\ (4.31) \qquad \Longleftrightarrow \quad & \exists\ l = 1, \ldots, q \text{ such that } X_{1A_l} \in [0, \psi_1(l,t)) \cup (\psi_2(l,t), n_1] \\ \Longleftrightarrow \quad & \exists\ r_l \in \mathbf{r}(\mathbf{x}),\ \text{some } l = 1, \ldots, q \text{ such that } r_l \in [0, \psi_1(l,t)) \cup (\psi_2(l,t), n_1], \end{aligned}$$

where $\Gamma_{\mathbf{c}}(\cdot)$ is defined by (4.22) and $r_l \in \mathbf{r}(\mathbf{x})$ denotes the l^{th} element upon the q-vector of rank lengths corresponding to the path of $\mathbf{x}$, $\mathbf{r}(\mathbf{x})$. Given $t > 0$, the network algorithm for GEM computes (4.28)[4] by identifying and summing up the probability lengths of all paths in the network *failing adherence* to the conditions imposed by (4.31), but with no need to explicitly enumerate each path. The decision as to whether or not the paths of the network contribute to the value of (4.28) occurs upon the nodes of the network. Specifically, given $t > 0$, for each $(s+1, m_{s+1}) \in \Psi(s, m_s)$ [direct successor] of node (s, m_s), $s = 0, \ldots, k-1$, we must check if one of the following conditions holds:

$$(4.32) \qquad \underbrace{\sum_{v=1}^{s} [\mathbf{r}_v]_l}_{(A)} + \underbrace{[\mathbf{r}_{s+1}]_l}_{(B)} + [\mathbf{SP}(s+1, m_{s+1})]_l > \psi_2(l,t)$$

and

$$(4.33) \qquad \sum_{v=1}^{s} [\mathbf{r}_v]_l + [\mathbf{r}_{s+1}]_l + [\mathbf{LP}(s+1, m_{s+1})]_l < \psi_1(l,t),$$

for some $l = 1, \ldots, q$, where: $\mathbf{SP}(s+1, m_{s+1})$ and $\mathbf{LP}(s+1, m_{s+1})$ are q-vectors, such that $[\mathbf{SP}(s+1, m_{s+1})]_l$ and $[\mathbf{LP}(s+1, m_{s+1})]_l$ are the rank lengths of the shortest and longest subpath, respectively, from node $(s+1, m_{s+1})$ to the terminal node, corresponding with the l^{th} element upon each of the vectors of rank length, $\{\mathbf{r}_{s+2}, \ldots, \mathbf{r}_k\}$, $l = 1, \ldots, q$; (A) corresponds to the sum of the rank lengths upon element l across the vectors $\{\mathbf{r}_1, \ldots, \mathbf{r}_s\}$; and (B) corresponds to the rank length for the arc joining node (s, m_s) to the direct successor node $(s+1, m_{s+1})$. If Q is the set of all paths which pass through the node $(s+1, m_{s+1}) \in \Psi(s, m_s)$ and which have a common subpath rank length equal to the value of (A) (4.32) upon reaching node (s, m_s) from the initial node, then no path of Q will contribute to (4.28) if either of the conditions (4.32) or (4.33) holds for some

[4]The network algorithm we develop here assumes realizations in (4.10), such that $|z_{1A_1}| = \cdots = |z_{1A_q}| = t$. The complement of the adjusted p-value, $\tilde{q}_{1l\mu}$, can thus be computed by substituting the true underlying realization $|z_{1A_l}|$ in lieu of t within the network algorithm, for all $l = 1, \ldots, q$ (i.e., upwards of q implementations of the network algorithm could be required to obtain the values for all q of the complements to the adjusted p-values).

$l = 1, \ldots, q$. In this circumstance, the paths of Q are not considered again in the network algorithm, and this forms the basis for the pruning of paths within the algorithm.

We process the network in two steps. First, we conduct a backward pass (called, the *backward induction pass*) through the network, beginning at the terminal node and ending at the initial node. During this pass through the network, we construct the vectors $\mathbf{SP}(s, m_s)$ and $\mathbf{LP}(s, m_s)$ upon the nodes of the network, so that the conditions (4.32) and (4.33) can be evaluated. Upon completion of the backward pass, the network is processed in the forward direction (called, the *forward induction pass*), beginning at the initial node and ending at the terminal node. During this second pass through the network, we essentially prune paths in accordance to the aforementioned conditions (4.32) and (4.33) and compute (4.28) for some $t > 0$. Without further delay, we now state the network algorithm for GEM.

Algorithm 4.2 Network Algorithm for GEM

Backward Induction Pass

1. Insofar as the alternative hypothesis of (4.3) is two-sided, the statistical inference encompassing the test statistic (4.10) under $\mathcal{H}_0$ is invariant to the labeling of the random variable W_l, for all $l = 1, \ldots, q = 11$. Indeed, in accordance with Table 4.1, for $l \in \{6, 8, 11\}$ we swap the collections A_l and B_l. Specifically, let

$$A_l = \begin{cases} A_l, \text{ if } l \in \{1, 2, 3, 4, 5, 7, 9, 10\} \\ B_l, \text{ if } l \in \{6, 8, 11\} \end{cases},$$

   and since L_{A_l} and L_{B_l} are complementary candidate patterns, let $B_l = \{1, \ldots, q\} \backslash A_l$, for all $l = 1, \ldots, q$. We consider $\mathbf{I}$, the $q \times k$ (here, $k = 6$) matrix as previously defined within this section. See step 4(b) of the forward induction pass to follow, for details motivating the current step of this algorithm.

2. For stage $s = 0, \ldots, k - 1 = 5$:

   (a) For each node (s, m_s), let $\mathcal{T}_{\mathbf{c}}(s, m_s)$ be the set of all subpaths from node (s, m_s) to the terminal node.

   (b) Let $\mathbf{SP}(s, m_s)$ and $\mathbf{LP}(s, m_s)$ be as defined above, the q-vectors where $[\mathbf{SP}(s, m_s)]_l$ and $[\mathbf{LP}(s, m_s)]_l$ are the rank lengths of the shortest and longest subpath, respectively, over

$\mathcal{T}_{\mathbf{c}}(s, m_s)$, corresponding with the l^{th} element upon each of the vectors of rank length, $\{\mathbf{r}_{s+1}, \ldots, \mathbf{r}_k\}$, $l = 1, \ldots, q$, where $\mathbf{r}_{s+1}$ is as defined within (4.25).

3. Let $\mathbf{SP}(k, m_k)$ and $\mathbf{LP}(k, m_k)$ each be q-vectors, whose l^{th} elements are defined by $[\mathbf{SP}(k, m_k)]_l = [\mathbf{LP}(k, m_k)]_l = 0$, for all $l = 1, \ldots, q$.

4. For stage $s = k - 1, \ldots, 0$:

   For each node (s, m_s):

   For $l = 1, \ldots, q$:

   (a) $[\mathbf{SP}(s, m_s)]_l = \min_{\Psi(s,m_s)} \{[\mathbf{r}_{s+1}]_l + [\mathbf{SP}(s+1, m_{s+1})]_l\}$.

   (b) $[\mathbf{LP}(s, m_s)]_l = \max_{\Psi(s,m_s)} \{[\mathbf{r}_{s+1}]_l + [\mathbf{LP}(s+1, m_{s+1})]_l\}$.

Forward Induction Pass

1. Let $t > 0$ be some observed value of the statistic $|Z_{\max}| = \sqrt{Z^2_{\max}}$, where $Z^2_{\max}$ is defined by (4.13).

2. For stage $s = 1, \ldots, k$:

   (a) For each node (s, m_s), let $\mathcal{I}_{\mathbf{c}}(s, m_s)$ denote the set of all subpaths originating at the initial node and ending at node (s, m_s).

   (b) Consider $\eta \in \mathcal{I}_{\mathbf{c}}(s, m_s)$. Let $\mathbf{r}(\eta)$ be the q-vector of rank lengths over $\{\mathbf{r}_1, \ldots, \mathbf{r}_s\}$, corresponding to subpath η, whose l^{th} element is defined by

$$[\mathbf{r}(\eta)]_l = \sum_{v=1}^{s} [\mathbf{r}_v]_l \quad \forall l = 1, \ldots, q,$$

   where $\mathbf{r}_s$ is defined by (4.25). Also, let $p(\eta)$ denote the probability length for subpath η under $\mathcal{H}_0$. Specifically, it is

$$p(\eta) = \prod_{v=1}^{s} p_v,$$

   where p_s is defined by (4.26).

(c) For each node (s, m_s), let $\mathcal{I}^*_{\mathbf{c}}(s, m_s) \subseteq \mathcal{I}_{\mathbf{c}}(s, m_s)$ denote the refined set of subpaths of $\mathcal{I}_{\mathbf{c}}(s, m_s)$, such that $\eta \in \mathcal{I}^*_{\mathbf{c}}(s, m_s)$ if and only if for every $l = 1, \ldots, q$ and for each $\phi_l \in \{[\mathbf{SP}(s, m_s)]_l, [\mathbf{LP}(s, m_s)]_l\}$, it holds

$$\psi_1(l, t) \leq [\mathbf{r}(\eta)]_l + \phi_l \leq \psi_2(l, t),$$

where $\psi_1(l, t)$ and $\psi_2(l, t)$ are given by (4.30).

(d) For each node (s, m_s), if $\mathcal{L}$ denotes the set of all unique vectors $\mathbf{r}(\eta)$ such that $\eta \in \mathcal{I}^*_{\mathbf{c}}(s, m_s)$, then we define the set of records $\mathcal{R}(s, m_s) = \{(\boldsymbol{\nu}, \pi(\boldsymbol{\nu})) : \boldsymbol{\nu} \in \mathcal{L}\}$, where

$$\pi(\boldsymbol{\nu}) = \sum_{\substack{\eta \in \mathcal{I}^*_{\mathbf{c}}(s, m_s) : \\ \mathbf{r}(\eta) = \boldsymbol{\nu}}} p(\eta).$$

3. Let $\mathcal{R}(0, m_0) = \{(\mathbf{0}, \pi(\mathbf{0}) = 1)\}$, where $\mathbf{0}$ is the q-vector comprised of all entries equal to zero.

4. For stage $s = 0, \ldots, k-1$:

   For each node (s, m_s):

   For each $(s+1, m_{s+1}) \in \Psi(s, m_s)$:

   For each record $(\boldsymbol{\nu}, \pi(\boldsymbol{\nu})) \in \mathcal{R}(s, m_s)$:

   (a) Evaluate the conditions imposed by (4.32) and (4.33), for each $l = 1, \ldots, q$, replacing the summands within each of these expressions with $[\boldsymbol{\nu}]_l$. If either one of these conditions holds for some l, then continue to next record within $\mathcal{R}(s, m_s)$; otherwise, continue to step 4(b).

   (b) For $l = 1, \ldots, q$: If $\sum_{w=s+2}^{k} [\mathbf{I}]_{(l,w)} = 0$ and the following condition holds

   $$[\boldsymbol{\nu}]_l + [\mathbf{r}_{s+1}]_l \in [0, \psi_1(l, t)) \cup (\psi_2(l, t), n_1],$$

   then continue to next record within $\mathcal{R}(s, m_s)$; otherwise, continue to step 4(c). Note that by our defining the A_l in the manner in which we did (within step 1 of the backward induction pass), the first condition of the premise here (i.e, the sum evaluating to zero) holds true for $l \in \{1, 3, 6, 8, 11\}$ upon the respective values of $s \in \{0, 1, 2, 3, 4\}$ – this

allows for the pruning of paths over the marginal permutation null distributions of $Z^2_{1A_l}$ for said values in l and can enhance computational performance of the network algorithm.

(c) Pass the record $(\boldsymbol{\nu}, \pi(\boldsymbol{\nu}))$ to the direct successor $(s+1, m_{s+1})$. We consider the modified record $(\boldsymbol{\nu}^*, \pi^*)$, where $\boldsymbol{\nu}^* = \boldsymbol{\nu} + \mathbf{r}_{s+1}$ and $\pi^* = \pi(\boldsymbol{\nu}) p_{s+1}$.

(d) If there exists a record $(\boldsymbol{\mu}, \pi(\boldsymbol{\mu})) \in \mathcal{R}(s+1, m_{s+1})$, such that $\boldsymbol{\mu} = \boldsymbol{\nu}^*$, then update the record $(\boldsymbol{\mu}, \pi(\boldsymbol{\mu}))$ with $(\boldsymbol{\mu}, \pi(\boldsymbol{\mu}) + \pi^*)$ and continue to next record within $\mathcal{R}(s, m_s)$; otherwise, continue to step 4(e).

(e) Insert $(\boldsymbol{\nu}^*, \pi^*)$ into $\mathcal{R}(s+1, m_{s+1})$ as a new record.

5. It follows that

$$\tilde{q} = \Pr(|Z_{\max}| < t | \mathcal{H}_0) = \left(\frac{n_1! n_0!}{n!}\right) \sum_{(\boldsymbol{\nu}, \pi(\boldsymbol{\nu})) \in \mathcal{R}(k, m_k)} \pi(\boldsymbol{\nu}),$$

for which taking $|z_{1A_l}| = t$, some $l = 1, \ldots, q$, (4.23) evaluates to

$$\tilde{p}_{1l\mu} = 1 - \tilde{q}.$$

4.7 Simulation Study: Statistical Power to Detect GxE Interaction in General

We performed a simulation analysis with the aims of: (1) demonstrating that our proposed GEM method controls the FWER at the 5% level under the complete null hypothesis $\mathcal{H}_0$, where we compare control of the FWER at this level across a number of competing methods to assess GxE interaction; and (2) under various conditions for which the complete null hypothesis is not true (i.e., $H_0^{(j,l)}$ is false for some $l = 1, \ldots, q$ and $j = 1, \ldots, m$), compare the statistical power of our proposed GEM method to those for a number of competing methods. Unless otherwise specified, the investigation of statistical power is assumed at the 5% level of the FWER.

4.7.1 Data Setup

We assumed a sample size of $n = 1$K per data set throughout the simulation investigation, where each data set was comprised of: a binary response, a single binary environmental factor, and m biallelic SNP markers, such that $m \in \{1, 2, 5, 10\}$. Let π_{G_j} denote the population minor

allele frequency (MAF) for the j^{th} SNP of the data set, $j = 1, \ldots, m$, and let the population prevalence of exposure be denoted by $\pi_E = \Pr(E = 1)$. The random variables $E, G_1, \ldots, G_m$ were simulated mutually independent of one another, where we assumed $E \sim \text{Binomial}\,(n, \pi_E)$ and $G_j \sim \text{Multinomial}\left(n, 3, \left((1-\pi_{G_j})^2, 2\pi_{G_j}(1-\pi_{G_j}), \pi_{G_j}^2\right)\right)$. That is, the random variable G_j was assumed to adhere to Hardy-Weinberg equilibrium within the population. For simplicity, we assumed the π_{G_j} satisfy the condition $\pi_{G_1} = \cdots = \pi_{G_m}$, where the parameters $\left(\pi_{G_j}, \pi_E\right)$ were assumed to reside within the selected collection $\{(0.05, 0.4), (0.2, 0.2), (0.2, 0.4), (0.4, 0.2), (0.5, 0.5)\}$. In order to investigate the behavior in the FWER under the complete null and the statistical power under some partial null hypothesis, we varied the distributional assumptions governing the random variable Y in six different ways (for clarity, we denote the accompanying simulations as A, B, ..., F, respectively):

FWER: The random variable Y was simulated independently of the random variables $E, G_1, \ldots, G_m$, such that the conditional probabilities $\Pr(Y = 1 | X_j = k) = 0.5$, for all $k \in \mathcal{X}_2$ and $j = 1, \ldots, m$. Assigning this common value in the conditional probabilities, corresponds to simulating data sets in coherence with a balanced case-control study; by simulating Y independently of the genetic and environmental factors, we modelled the complete null hypothesis $\mathcal{H}_0$.

Power (Main Effect G_j Recessive Model): To simulate a main effect for G_j, we assumed the recessive genetic model of inheritance such that the conditional probabilities

$$\Pr(Y = 1 | X_j \in \{3, 6\}) = 0.60 \text{ (corresponds to } G_j = 2),$$

and – to preserve a balanced case-control sampling design –

$$\Pr(Y = 1 | X_j \in \{1, 2, 4, 5\}) = \left(0.5 - 0.6\pi_{G_j}^2\right) / \left(1 - \pi_{G_j}^2\right).$$

Power (Main Effect G_j Dominant Model): To simulate a main effect for G_j, we assumed the dominant genetic model of inheritance such that the conditional probabilities

$$\Pr(Y = 1 | X_j \in \{2, 3, 5, 6\}) = 0.55 \text{ (corresponds to } G_j \in \{1, 2\}),$$

and – to preserve a balanced case-control sampling design –

$$\Pr(Y = 1 | X_j \in \{1, 4\}) = \left(0.5 - 0.55\left(1 - \left(1 - \pi_{G_j}\right)^2\right)\right) / \left(1 - \pi_{G_j}\right)^2.$$

Power (Main Effect E): To simulate a main effect for E, we assumed the conditional probabilities

$$\Pr\left(Y = 1 | X_j \in \{4, 5, 6\}\right) = 0.55 \text{ (corresponds to } E = 1),$$

and – to preserve a balanced case-control sampling design –

$$\Pr\left(Y = 1 | X_j \in \{1, 2, 3\}\right) = \left(0.5 - 0.55\pi_E\right) / \left(1 - \pi_E\right).$$

Power (GxE Recessive Model): To simulate GxE where the SNP adheres to the recessive genetic model, we assumed the conditional probabilities

$$\Pr\left(Y = 1 | X_j = 6\right) = 0.70 \text{ (corresponds to logical pattern } (G_j = 2) \wedge (E = 1)),$$

and – to preserve a balanced case-control sampling design –

$$\Pr\left(Y = 1 | X_j \in \{1, 2, 3, 4, 5\}\right) = \left(0.5 - 0.70\pi_E\pi_{G_j}^2\right) / \left(1 - \pi_E\pi_{G_j}^2\right).$$

Power (GxE Dominant Model): To simulate GxE where the SNP adheres to the dominant genetic model, we assumed the conditional probabilities

$$\Pr\left(Y = 1 | X_j \in \{5, 6\}\right) = 0.65 \text{ (corresponds to logical pattern } (G_j \in \{1, 2\}) \wedge (E = 1)),$$

and – to preserve a balanced case-control sampling design –

$$\Pr\left(Y = 1 | X_j \in \{1, 2, 3, 4\}\right) = \left(0.5 - 0.65\pi_E\left(1 - (1 - \pi_G)^2\right)\right) / \left(1 - \pi_E\left(1 - (1 - \pi_G)^2\right)\right).$$

Within each of the five power conditions, the random variables $G_2, \ldots, G_m$ were assumed independent of the random variables Y and E, whenever $m > 1$. Our intention was to investigate the power to detect either the main genetic effect or the GxE effect – in the case where one of these effects is present upon exactly one SNP marker – adjusting for the multiplicity problem across multiple SNP markers simultaneously being assessed for GxE. Interestingly, this approach is analogous to that taken by [3] within their simulation investigating statistical power to detect GxE interaction. For each of these six variations in the distributional properties of Y across the support

of the random variable X_j, for each $m \in \{1, 2, 5, 10\}$, and for each $\left(\pi_{G_j}, \pi_E\right)$ within the collection $\{(0.05, 0.4), (0.2, 0.2), (0.2, 0.4), (0.4, 0.4), (0.5, 0.5)\}$, we simulated $D = 10\text{K}$ mutually independent data sets. To obtain adequate estimates for $\tilde{p}_{jl\sigma}$ (4.12) and $\tilde{p}_{jl\mu}$ (4.14), within each simulated data set we carried out the permutation procedure of Algorithm 4.1, assigning the value of R (i.e., the number of random shuffles upon the columns of $\mathbf{GE}^*$) therein to 10K.

4.7.2 Competing Methods for Detecting GxE Interaction

To remain consistent with the simulation methodology undertaken within [109], we compared our proposed GEM method to six competing approaches for detecting GxE – note: because GEM is able to perform multiplicity adjustment for $m \geq 1$ (the simulations conducted within [109] were carried out assuming $m \equiv 1$), some of these approaches have been modified (i.e., to accommodate the circumstance for which $m > 1$) for benchmarking our GEM method. The seven methods, including GEM, are described as follows (here, we reference the seven methods as competing methods):

Raw GEM (RGEM): The maximum value of the proposed test statistic Z_{jA_l} (4.10) over all hypothesis tests for the data set $\left(\text{absolute value thereof; i.e., } \max\left\{|Z_{jA_1}|, \ldots, |Z_{jA_q}|\right\}_{j=1,\ldots,m}\right)$ is determined. The computed value is then assumed – under the complete null hypothesis ($\mathcal{H}_0$) – asymptotically distributed by the standard normal distribution, and the corresponding two-sided p-value, denoted p_{RGEM}, is computed from the CDF of said distribution.

Bonferroni Raw GEM (BRGEM): The p-value computed under RGEM is adjusted for multiple hypothesis testing, by applying the Bonferroni correction for having tested the $m \times q$ null hypotheses represented by $\mathcal{H}_0$. That is, the BRGEM p-value is given by $\min\{(mq)p_{\text{RGEM}}, 1\}$.

Pearson Chi-Square Test (PCT): Extending (4.20) to the circumstance in which $m \geq 1$, it follows that the null hypothesis of no association between the random variables X_j and Y is equivalent to $\mathcal{H}_0^{(j)}$, for any $j = 1, \ldots, m$. We conduct the Pearson chi-square test of no association between X_j and Y, against the corresponding two-sided alternative hypothesis, for each $j = 1, \ldots, m$; the corresponding p-value is computed under $\mathcal{H}_0^{(j)}$, by referring to the chi-square distribution with five degrees-of-freedom. The smallest of these p-values is then selected and a Bonferroni correction is applied for having tested the m null hypotheses represented by $\mathcal{H}_0$.

GEM: The maxT adjusted p-value $\tilde{p}_{jl\sigma}$ $(m = 1)$ or $\tilde{p}_{jl\mu}$ $(m > 1)$ is estimated by its respective permutation counterpart, (4.18) or (4.19), in accordance with Algorithm 4.1, for all $j = 1, \ldots, m$ and $l = 1, \ldots, q$. The smallest of these maxT permutation adjusted p-values is selected.

Nominal Likelihood Ratio Test (NLRT): The conventional approach to assess whether an association exists between genetic/environmental factors and a binary response in genetic association studies consists of constructing various logistic regression models. These models can involve one or both of the genetic or environmental factors, and may possibly involve term(s) for testing GxE interaction [147, 171, 172]. Indicator random variables are typically employed, to distinguish the levels for each of the genetic and environmental factors. Here, for $j = 1, \ldots, m$, we define the two indicator random variables for the j^{th} SNP locus, G_{j1} and G_{j2}, where

$$G_{jg} = I\left(G_j \geq g\right),$$

each $g = 1, 2$. That is, G_{j1} and G_{j2} are the respective indicator random variables corresponding to the dominant and recessive genetic models for locus j. Since our GEM method assesses patterns for both main effects and GxE interaction, as a benchmarking tool for GEM we construct the following seven *nested* logistic regression models: the three simple logistic regression models, each distinguishable from the remaining two models and comprised of a single predictor variable from the collection $\{G_{j1}, G_{j2}, E\}$; the two [main effect] multiple logistic regression models, each comprised of a unique indicator random variable from G_j modeled along with the environmental factor indicator random variable E; and the two GxE multiple logistic regression models, each comprised of a unique indicator random variable from G_j modeled along with the environmental factor indicator random variable E and the appropriate GxE interaction term. The likelihood ratio test (LRT) is carried out for each of these seven models – against the null model consisting of solely an intercept regression parameter, for fair comparison – at each SNP locus. The p-value for each LRT is determined by referring to the appropriate chi-square distribution.[5] The minimum of the p-values, denoted p_{NLRT}, is selected from amongst all of those computed across the m SNP loci, insofar as the underlying genetic model is generally unknown for a novel complex disease.

[5] The degrees-of-freedom for the chi-square distribution is equal to the difference between the number of regression parameters for the model under the alternative hypothesis and that for the null model.

Bonferroni Likelihood Ratio Test (BLRT): The p-value computed under NLRT, p_{NLRT}, is adjusted by applying the Bonferroni correction for having tested $7 \times m$ null hypotheses. That is, the BLRT p-value is given by $\min\{(7m)p_{\text{NLRT}}, 1\}$.

Global Likelihood Ratio Test (GLRT): The random variable G_j is modeled as a qualitative predictor, categorized by the two indicator random variables G_{j1} and G_{j2}, respectively, where

$$G_{jg} = I\left(G_j = g\right),$$

each $g = 1, 2$. For each $j = 1, \ldots, m$, we consider the multiple logistic regression model

$$\text{logit}\left(\Pr\left(Y = 1 | G_{j1}, G_{j2}, E\right)\right) = \beta_0 + \beta_e E + \sum_{g \in \{1,2\}} \left(\beta_{jg} G_{jg} + \gamma_{jg} G_{jg} E\right), \tag{4.34}$$

where logit$(\cdot)$ is the [natural] log odds of $(\cdot)$ and each of the regression parameters of this model is assumed unknown. The LRT statistic – computed under the null hypothesis that all predictor coefficients equal zero upon this multiple logistic regression model, against the two-sided alternative hypothesis that some coefficient is different from zero – is computed, each $j = 1, \ldots, m$. The corresponding p-value for each of these LRT statistics is computed by referring to the chi-square distribution with five degrees-of-freedom. The minimum of these p-values is selected and a Bonferroni correction is applied for having tested a total of m null hypotheses.

4.7.3 Type I Error Rate and Power

For each of the aforementioned competing methods to detect GxE interaction, let V_d be the indicator random variable with success defined as some Type I error being observed within simulation A upon data set d, where a Type I error is assumed to occur whenever the p-value (as defined by the competing method) falls below the value 0.05; and, upon each of the simulations B thru F, let S_d be the indicator random variable with success occurring whenever the p-value (as defined by the competing method) falls below the value of 0.05, for all $d = 1, \ldots, D = 10\text{K}$. Since the data sets are simulated independently of one another, it follows that

$$\sum_d V_d \sim \text{Binomial}\left(D, \alpha_{\text{F}}\right),$$

and

$$\sum_d S_d \dot{\sim} \text{Binomial}\left(D, 1-\beta\right),$$

where α_{F} denotes the true underlying FWER and – assuming control of the FWER is at the 5% level – β denotes the true underlying Type II error rate.[6] Taking the parameters α_{F} and β at their respective MLEs, our estimates of the FWER and the power of the test are given by

$$\hat{\alpha}_{\text{F}} = \frac{\sum_d V_d}{D}, \tag{4.35}$$

and

$$1-\hat{\beta} \doteq \frac{\sum_d S_d}{D}, \tag{4.36}$$

respectively.

4.7.4 Results

Table 4.7 depicts the proportion of the [$D = 10$K total] p-values which fall below the value of 0.05, across the levels in m by competing method to detect GxE interaction (rows) and parameterization of the ordered pair $\left(\pi_{G_j}, \pi_E\right)$ (columns) for simulation A. That is, these values are the estimated FWER (4.35) by competing method under the complete null hypothesis ($\mathcal{H}_0$). Each of the RGEM and NLRT methods have inflated observed FWER rates when compared to the expected 5% value, irrespective of the value of m and parameterization of the ordered pair $\left(\pi_{G_j}, \pi_E\right)$, where the discrepancy of the observed FWER rate from expected is exacerbated as m increases. This result is not unexpected, since the corresponding p-values arising from these methods are not adjusted for multiple hypothesis testing. Application of these methods in practice would therefore lead one to extreme misrepresentation of results, as the reporting of some Type I errors is very likely. Because these two methods clearly fail to control the FWER at the 5% level, for clarity in discussion these two methods are henceforth no longer considered competing methods.

On the other hand, these data indicate that the GLRT method controls the FWER at the 5% level, in general, whereas each of the methods BRGEM, PCT (for the circumstances in which

[6] The estimator $\sum_d S_d/D$ does not [technically] represent the estimated power of the test, since $S_d = 1$, for $d = 1, \ldots, D$, could indicate either a Type I error or a correct rejection of a false null hypothesis. However, said estimator should provide a reasonable estimate for the power of the test, and – provided the true underlying level of control in the FWER is the same across competitive methods which assess GxE interaction – provides an adequate means by which to compare the power of the test across said methods.

$m > 1$), and BLRT possess observed FWERs falling [well] below the 5% expected level, in general, suggesting that these latter three methods are conservative in their respective control of the FWER at the 5% level. This notion is particularly true of the BLRT method, suggesting that this method is overly conservative in its control of the FWER at the 5% level. In fact, of the five methods which control the FWER at a level not exceeding the true underlying 5% level, namely BRGEM, PCT, GEM, BLRT, and GLRT, only GEM is unbiased[7] in its control of the FWER at the 5% level for every combination of chosen ordered pair $\left(\pi_{G_j}, \pi_E\right)$ and chosen number of SNP markers (m) within the simulated data sets. This is not an unexpected result, since the multiple testing correction for GEM is based upon the permutation null distribution of $Z^2_{\max}$ (4.14), as opposed to the multiple testing correction upon the four alternative methods being based upon an asymptotic assumption governing the test statistics null distribution.

The conservatism in the control of the FWER at the 5% level is particularly interesting upon the ordered pair $\left(\pi_{G_j}, \pi_E\right) = (0.05, 0.4)$. With the exception of BRGEM, the conservatism for each of the four competing methods based upon the Bonferroni MTP seems to be positively associated with the magnitude in m. The observed FWERs upon the PCT method, for example, are $\{0.029, 0.014, 0.012, 0.008\}$ for the respective values of $m \in \{1, 2, 5, 10\}$. In light of this, we expect the statistical power for the PCT, BLRT, and GLRT competing methods to be lower than that of both GEM and BRGEM. Conversely, these data indicate an apparent increasing trend in the observed FWER for the BRGEM method as the value of m increases. The observed FWERs upon this competing method are $\{0.028, 0.030, 0.035, 0.039\}$, for the respective values of $m \in \{1, 2, 5, 10\}$. This suggests that the veracity in the asymptotic normal assumption governing the distribution of the test statistic Z_{jA_l} (4.10) under $\mathcal{H}_0$, may be dependent upon the value of m. Further research investigating this notion is needed, but we do point out here that this observation is similar to that made within Chapter 3 regarding the distribution of the Cochran-Armitage trend test statistic (under the complete null hypothesis therein) being dependent upon sample characteristics, including the magnitude in m.

Figures 4.2 and 4.3 display the estimated statistical power (4.36) to detect a main genetic effect (blue symbols, simulation B; red symbols, simulation C), a main environmental effect (orange symbols, simulation D), and GxE interaction (purple symbols, simulation E; black symbols, simulation F) at the 5% level in the FWER by competing method and selected values in the ordered pairs $\left(\pi_{G_j}, \pi_E\right)$, upon a single SNP locus ($m = 1$; top panel of the former figure), two loci ($m = 2$; lower

[7] 95% Exact Clopper-Pearson confidence interval for the true underlying FWER covers the value of 0.05.

Table 4.7: Observed FWER (4.35) by Competing Method to Assess GxE Interaction and Selected Values in the Ordered Pair $\left(\pi_{G_j}, \pi_E\right)$ for Simulation A (Complete Null Hypothesis). The True Underlying FWER Is 5%.

		$\left(\pi_{G_j}, \pi_E\right)$				
Method	m	(0.05, 0.4)	(0.2, 0.2)	(0.2, 0.4)	(0.4, 0.2)	(0.5, 0.5)
RGEM	1	0.206	0.289	0.298	0.312	0.317
BRGEM	1	0.028	0.044	0.040	0.048	0.046
PCT	1	0.029	0.048	0.047	0.049	0.052
GEM	1	0.049	0.052	0.047	0.052	0.053
NLRT	1	0.245	0.244	0.230	0.226	0.232
BLRT	1	0.025	0.042	0.037	0.035	0.039
GLRT	1	0.026	0.059	0.051	0.051	0.053
RGEM	2	0.324	0.442	0.473	0.487	0.480
BRGEM	2	0.030	0.045	0.042	0.049	0.042
PCT	2	0.014	0.043	0.045	0.047	0.046
GEM	2	0.050	0.051	0.050	0.052	0.049
NLRT	2	0.404	0.389	0.383	0.380	0.367
BLRT	2	0.021	0.038	0.034	0.031	0.032
GLRT	2	0.024	0.053	0.052	0.049	0.047
RGEM	5	0.497	0.640	0.671	0.696	0.687
BRGEM	5	0.035	0.046	0.043	0.052	0.044
PCT	5	0.012	0.038	0.041	0.047	0.045
GEM	5	0.049	0.048	0.046	0.052	0.050
NLRT	5	0.621	0.603	0.585	0.581	0.588
BLRT	5	0.023	0.033	0.032	0.032	0.030
GLRT	5	0.022	0.053	0.049	0.050	0.046
RGEM	10	0.795	0.900	0.920	0.922	0.914
BRGEM	10	0.039	0.056	0.054	0.056	0.043
PCT	10	0.008	0.036	0.042	0.040	0.044
GEM	10	0.046	0.049	0.050	0.050	0.049
NLRT	10	0.907	0.898	0.883	0.875	0.868
BLRT	10	0.025	0.034	0.035	0.031	0.031
GLRT	10	0.016	0.053	0.054	0.046	0.047

panel of the former figure), five loci ($m = 5$; top panel of the latter figure), and ten loci ($m = 10$; lower panel of the latter figure) paired with a binary environmental factor. These data indicate that the statistical power for GEM, in general, exceeds that of all competing methods chosen for this investigation. The only two apparent exceptions to this notion occur upon simulation E (power to detect GxE for the recessive genetic model of inheritance; purple symbols), where $\left(\pi_{G_j}, \pi_E\right)$ equals $(0.2, 0.2)$ (second pane) and $(0.2, 0.4)$ (third pane). However, the statistical power is low for all competing methods upon simulation E within these two panes upon each of the two panel plots within each figure. In light of the conservatism in its control of the FWER at the 5% level (Table 4.7), as expected these data indicate that the BLRT method possesses the weakest statistical power amongst the competing methods, in general, within each of the simulation conditions B–F across the

five panes of each figure. Furthermore, when compared to GEM, the BLRT method has particularly low power in detecting GxE interaction upon the dominant genetic model and in detecting a main environmental effect.

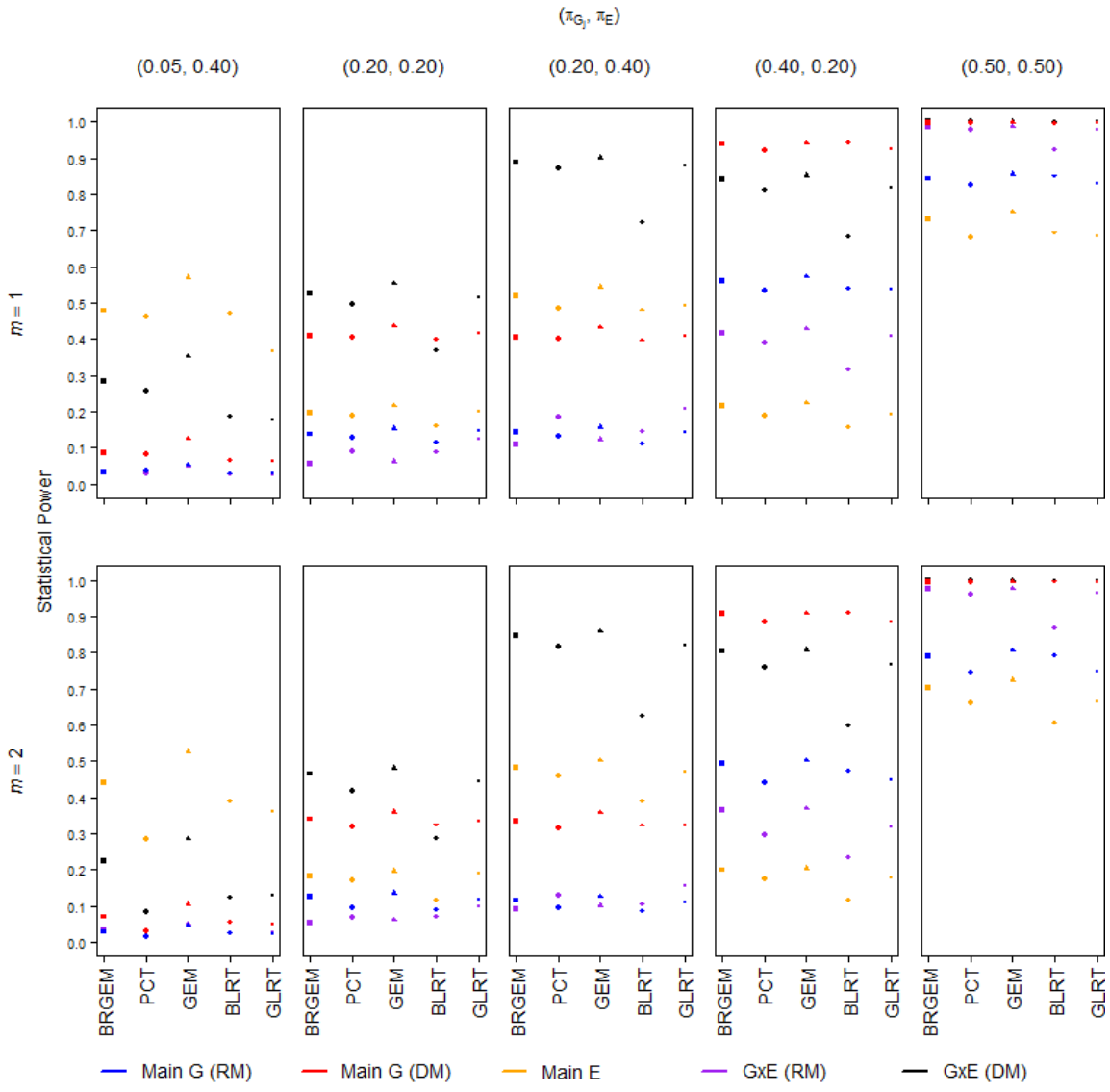

Fig. 4.2: Estimated Statistical Power (4.36) to Detect a Main Genetic Effect (Blue Symbols, Simulation B; Red Symbols, Simulation C), a Main Environment Effect (Orange Symbols, Simulation D), and GxE Interaction (Purple Symbols, Simulation E; Black Symbols, Simulation F) at the 5% Level in the FWER for $m = 1$ (Upper Panel) and $m = 2$ (Lower Panel), by Competing Method (Various Symbol Types) to Assess GxE Interaction and Selected Values for the Ordered Pair (π_{G_j}, π_E) (Panes). G = Genetic Effect; RM = Recessive Genetic Model; DM = Dominant Genetic Model; and E = Environment Effect.

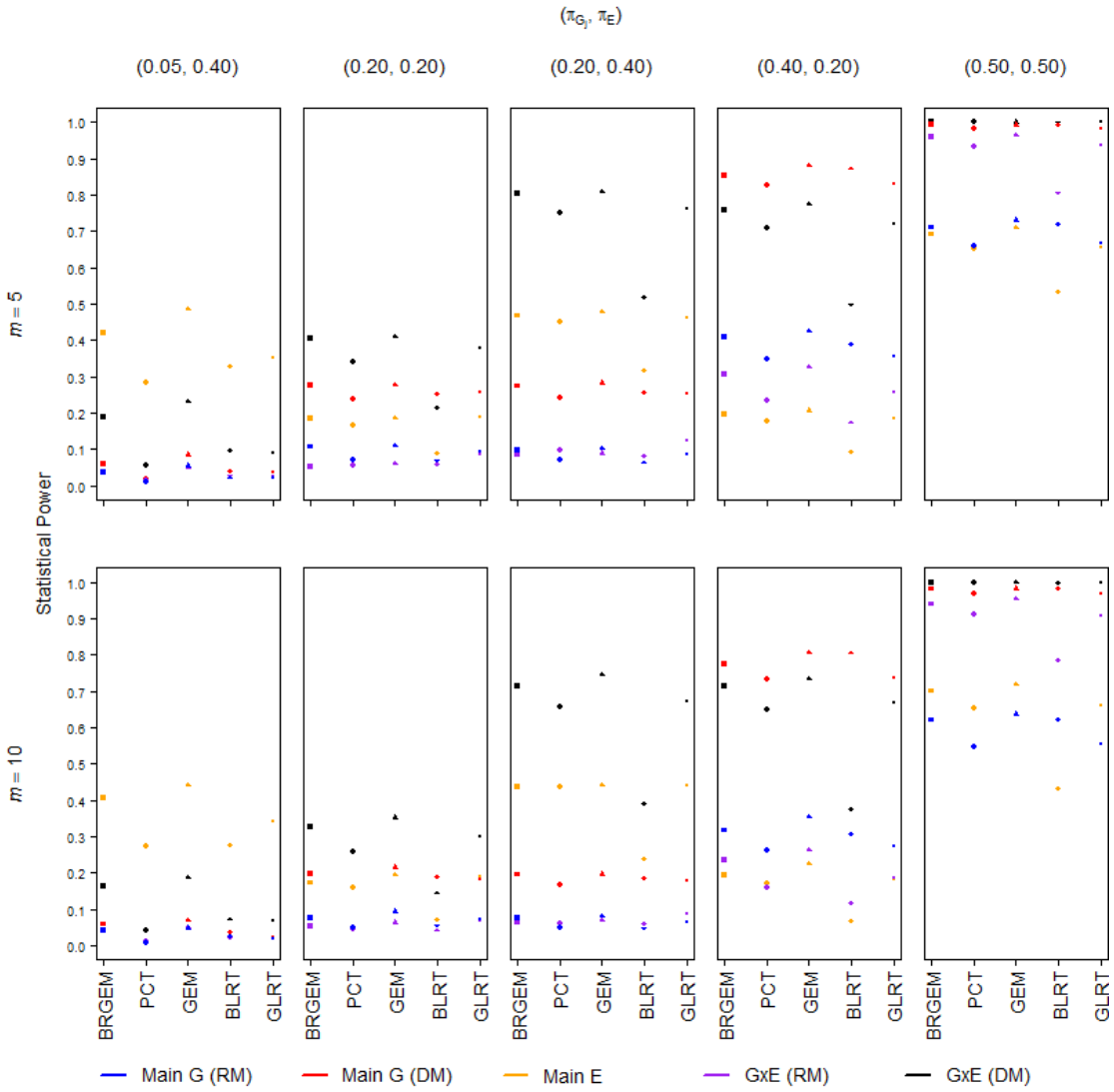

Fig. 4.3: Estimated Statistical Power (4.36) to Detect a Main Genetic Effect (Blue Symbols, Simulation B; Red Symbols, Simulation C), a Main Environment Effect (Orange Symbols, Simulation D), and GxE Interaction (Purple Symbols, Simulation E; Black Symbols, Simulation F) at the 5% Level in the FWER for $m = 5$ (Upper Panel) and $m = 10$ (Lower Panel), by Competing Method (Various Symbol Types) to Assess GxE Interaction and Selected Values for the Ordered Pair (π_{G_j}, π_E) (Panes). G = Genetic Effect; RM = Recessive Genetic Model; DM = Dominant Genetic Model; and E = Environment Effect. Note: Only Those Competing Methods Which Control the FWER at the 5% Level Are Presented.

This observation of low statistical power for the BLRT method is particularly interesting, because the NLRT method – for which the BLRT is based upon – is essentially the conventional approach one would undertake in exploratory data analysis, for determining an equation which describes the relationship between the binary response (Y) and the three predictor variables G_{j1}, G_{j2}, and E upon a novel complex disease. These data suggest that alternative competing methods (to BLRT),

such as GEM, can yield greater statistical power in such circumstances. While each of the PCT and GLRT methods seem to possess statistical power slightly lower than that of GEM within the second thru fifth panes upon each panel plot within each figure, the statistical power for each of these two methods appears to suffer considerably within the first pane of each figure. In fact, within the first pane of the upper panel plot of the former figure (i.e., taking $\left(\pi_{G_j}, \pi_E\right) = (0.05, 0.4)$ for $m = 1$), the statistical power of GEM is at least 20% greater than that of all competing methods, where the minimum relative statistical power of GEM to all competing methods occurs within simulation D (assessing a main effect in the environmental factor), comparing the statistical power of GEM (57.5%) to that of BRGEM (47.5%). This suggests that GEM likely possesses greater ability to detect true associations upon SNP loci with rare population MAF, when compared to the competing methods of this investigation.

Indeed, for a rare (a value not exceeding 0.05) minor allele frequency in π_{G_j}, we examined [to a closer extent] the statistical power to detect the main effect of the environmental factor (simulation D) and GxE interaction (dominant GMI; simulation F) for the competing methods, over the domain in the population prevalence of exposure, $\pi_E \in (0, 1)$, taking $m = 1$. We examined these two effects specifically, since the first pane of the plot within the upper panel of Figure 4.2 suggests these effects: to yield adequate power to carry out this task; and, to be of particular interest, due to the apparent vertical separation in power amongst the competing methods. We assumed the identical distributional assumptions for Y and data set characteristics, as specified within §4.7.1. To carry out this power investigation, we first examined each of the competing methods for adherence to control of the FWER at the 5% level under the complete null hypothesis (i.e., simulation A), taking $\pi_{G_j} \in \{0.01, 0.02, 0.04, 0.05\}$ to serve as a proxy for SNP loci possessing rare MAFs and $\pi_E \in \{0.01, 0.02, \ldots, 0.99\}$. Figure 4.4 illustrates the observed FWER for the competing methods ($D = 10$K data sets simulated upon each ordered pair $\left(\pi_{G_j}, \pi_E\right)$ and each plot depicted within a panel of the figure), taking $\pi_{G_j} = 0.01$ (upper left panel), $\pi_{G_j} = 0.02$ (upper right panel), $\pi_{G_j} = 0.04$ (lower left panel), and $\pi_{G_j} = 0.05$ (lower right panel). These data indicate that control of the FWER at the 5% level to be conservative for the PCT, BRGEM, and GLRT, irrespective of the value of π_{G_j} chosen here, where for a given value in π_E the conservatism upon the latter two methods seems to increase for decreasing values in π_{G_j}. On the other hand, for $\pi_E \leq 0.60$, interestingly the BLRT method is suggestive to be conservative in its control of the FWER at the 5% level for $\pi_{G_j} \in [0.04, 0.05]$, to be somewhat liberal in its control of this error rate for $\pi_{G_j} \leq 0.02$,

the anti-conservatism for this method to be negatively associated with the value of π_{G_j} over this latter range in the parameter π_{G_j}, and is suggestive to possess liberal control of the FWER at the 5% level for $\pi_E > 0.60$ and $\pi_{G_j} < 0.05$.

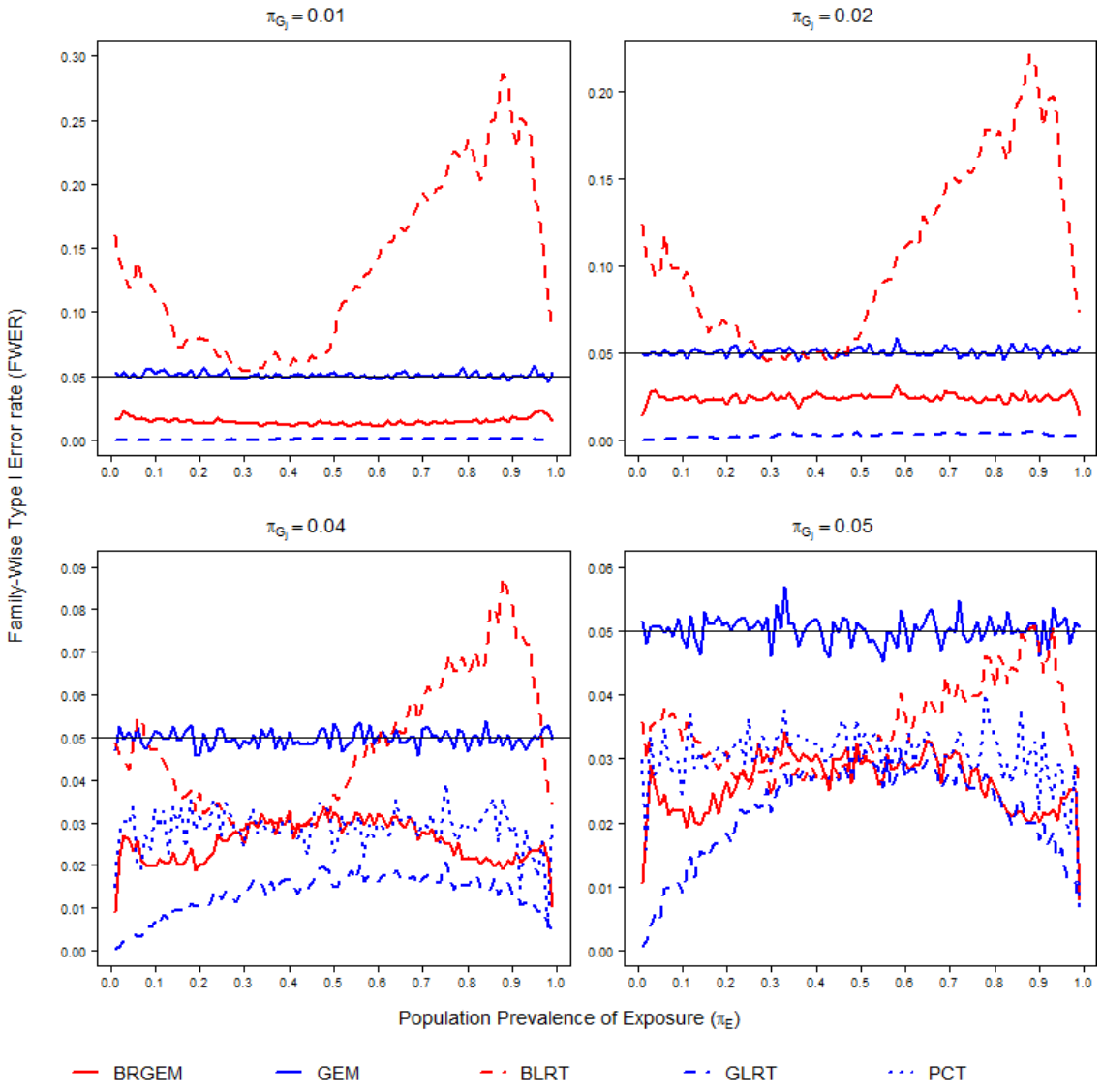

Fig. 4.4: Observed FWER (4.35) by Competing Method to Assess GxE Interaction under Simulation A for $\pi_E \in (0, 1)$, Taking $\pi_{G_j} = 0.01$ (Upper Left Panel), $\pi_{G_j} = 0.02$ (Upper Right Panel), $\pi_{G_j} = 0.04$ (Lower Left Panel), and $\pi_{G_j} = 0.05$ (Lower Right Panel), Where the True Underlying FWER Is 5%. Wald-based Methods Are Depicted by Solid Curves; LRT-based Methods by Heavy Dashed Curves; and, the PCT Method by the Light Dashed Blue Curve. Due to Data Sparsity, the PCT Method Comprised a Large Proportion of Non-Calculable Test Statistics and as Such Is Only Shown Within the Lower Two Panels of the Figure.

Finally, as expected, these data indicate that GEM properly controls the FWER at the 5% level, as demonstrated by the solid blue curve within these plots adhering very close to the expected 5% thin black reference line, where it is noted that the 95% Clopper-Pearson confidence intervals for the true FWER (not shown) cover the value of 0.05 across the domain of π_E within each of the panel plots of the figure.

Since each competing method appears to control the FWER at the 5% level across the domain of π_E for $\pi_{G_j} = 0.05$, to remain consistent with the chosen ordered pairs $\left(\pi_{G_j}, \pi_E\right)$ of our power study over simulations A–F, we decided to examine the statistical power of the competing methods upon simulations D and F taking $\pi_{G_j} = 0.05$. At the 5% level in the FWER, Figure 4.5 portrays the estimated statistical power to detect GxE interaction (dominant GMI; upper panel) and a main effect in the environmental factor (lower panel), where $D = 10$K data sets were simulated upon the conditions depicted within each panel of the figure at each ordered pair $\left(\pi_{G_j}, \pi_E\right)$ for $\pi_E \in \{0.01, 0.02, \ldots, 0.99\}$. These data indicate that the statistical power to detect GxE interaction for GEM is greater than that of all competing methods, whenever $\pi_E \geq 0.20$. Moreover, over this range of π_E, with the exception of BRGEM, the rate at which power increases for GEM (for increasing π_E) appears to be accelerated when compared to the competing methods, as seen by the increasing vertical separation between the appropriate curves within the upper panel plot. Also, even though power for GEM appears lower than some competing methods for $\pi_E < 0.20$, the power to detect GxE interaction is low for all competing methods over this range of population prevalence of exposure. These data also indicate that GEM possesses the highest statistical power – over the competing methods chosen for this investigation – to detect a main effect upon the environmental factor (lower panel), irrespective of the chosen value in π_E.[8] Finally, it is worth mentioning here that in applying the Bonferroni MTP (in lieu of the maxT MTP), for multiple testing correction upon the GEM methodology (i.e., BRGEM), these data indicate a considerable loss in statistical power can be incurred when failing to account for the correlation amongst the test statistics, $Z_{jA_1}, \ldots, Z_{jA_q}$, as shown by the vertical separation between the solid curves within each panel plot of this figure.

An interesting feature – of which amongst all competing methods (BRGEM set aside) is unique to GEM – is the ability of said approach to identify the correct candidate pattern upon the appropriate simulation condition (i.e., simulations B–F), independently of the obtained p-value. To illustrate this notion, upon each of the simulation conditions B–F we simulated $D = 2$K data sets

[8] Note that pursuant to the setup for simulation D – assessing statistical power for a main effect in the environmental factor (see §4.7.1) – the value of π_E is restricted to a maximum of $0.5/0.55 \approx 0.9$.

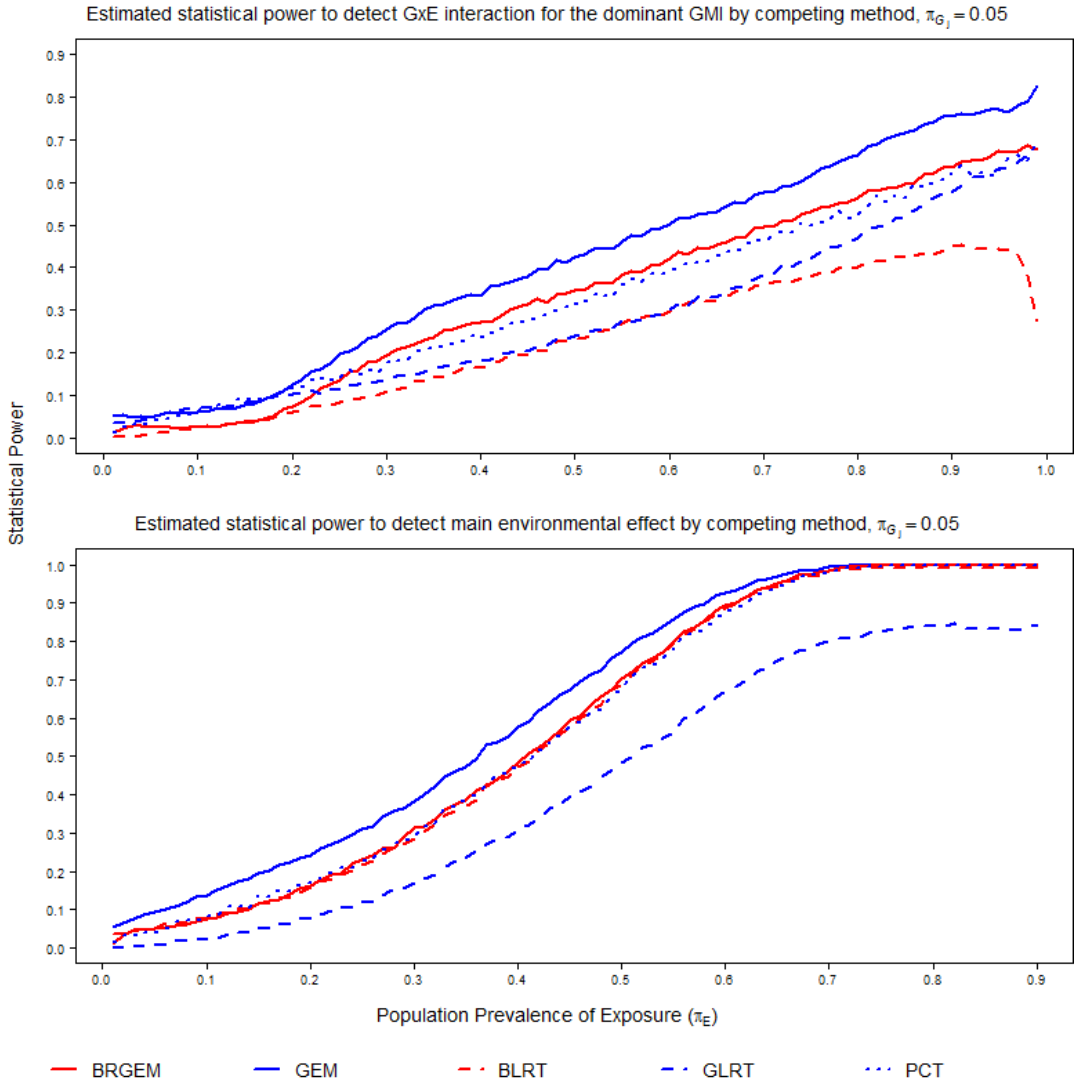

Fig. 4.5: Estimated Statistical Power (4.36) to Assess GxE Interaction (Upper Panel) Assuming the Dominant GMI for $\pi_E \in (0, 1)$, and the Main Effect in the Environmental Factor (Lower Panel) for $\pi_E \in (0, 0.9)$, Taking $\pi_{G_j} = 0.05$, Where the True Underlying FWER Is 5%. Wald-based Methods Are Depicted by Solid Curves; LRT-based Methods by Heavy Dashed Curves; and, the PCT Method by the Light Dashed Blue Curve.

at each combination of $\pi_E \in \{0.01k\}_{k=1,\ldots,99}$ (the index k was limited to the value of 90 upon simulation D – see footnote 8 above) and $\pi_{G_j} \in \{0.05, 0.10, 0.20, 0.40\}$, assigning the value of R within Algorithm 4.1 to 5K. The panel plots shown within Figure 4.6, depict the proportion of the simulated data sets – amongst those exhibiting some rejected null hypothesis upon the collection $\{H_0^{(j,1)}, \ldots, H_0^{(j,q)}\}_{j=m=1}$ at the 5% FWER level – for which GEM correctly identified the appropriate candidate pattern (y-axis) versus the population prevalence of exposure for the environmental

factor (x-axis), across simulations B–F (shown by the various colored curves) for select choices in π_{G_j} (panels). Here, for clarity we denote said proportion as the "success rate for GEM" (SRG). These data indicate that SRG is an increasing function of π_E/π_{G_j} upon simulations D–F (these simulations exclude the genetic main effect)/B–C (simulations entailing a main genetic effect), irrespective of the value of π_{G_j}/π_E. These results are not unexpected, because as π_E/π_{G_j} increases, the appropriate cells within the 2×6 table (i.e., Table 4.2 with $\varepsilon = 2$ therein) should exhibit a value in the statistic T_{jA_l} increasingly deviating from that expected under $H_0^{(j,l)}$ $\left(E\left(T_{jA_l}|H_0^{(j,l)}\right) = 0\right)$, thereby increasing the evidence in favor of $H_a^{(j,l)}$. Finally, upon simulations D (main effect in the environmental factor)/B-C (main effect in the genetic factor), for a fixed value in π_E/π_{G_j} the data suggest SRG to be constant across the values of π_{G_j}/π_E. These results make sense, in light of the fact that the genetic and environmental factors are simulated independently of one another.

4.8 Simulation Study: Statistical Power to Detect Cross-Interaction

Gene-environment interactions can portray several different patterns of association, as previously described within §1.2.1. Of particular interest is the cross-interaction pattern. A cross-interaction between a binary environmental factor and a genetic factor, for example, will exhibit opposite effects within the two exposure groups of the former factor. This pattern of GxE interaction is prevalent in the literature. For example, in a study of delinquency (phenotype of interest) among a sample of 1825 high school students, self-reported maltreatment in gender (binary environmental factor) cross-interacted with the alleles upon a functional 30-base pair repeat polymorphism in the promoter region ($MAOA$-$VNTR$) of the human $MAOA$ gene (genetic factor) [58]. In a case-control study involving 735 cases of coronary artery disease (phenotype) and 519 healthy controls, gender and presence of hypertension (environmental factors) cross-interacted with a haplotype of six SNPs within the AGT gene (genetic factor) [173]. Upon this three-way gene-environment-environment cross-interaction, risk of coronary artery disease: increased in women with hypertension; decreased in men with hypertension; and, the haplotype effect was not significant in men nor women without the presence of hypertension.

It is loosely stated within the article [3] that genetic markers which cross-interact with an environmental factor will not show a main genetic effect. This statement is not correct for two reasons. First, given a significance level, one could find a significant main genetic effect by chance. The statement does not allow for a chance finding and is thus incorrect. Second, although the statement is too strong to be valid upon all possible ways of modeling cross-interaction (see Table 4.8 for sampled

data which contradict said statement), it allegorizes that scanning for solely main genetic effects (e.g., the GWAS approach) may fail to detect genetic-phenotype associations upon those genetic markers exhibiting cross-interaction with some environmental factor.

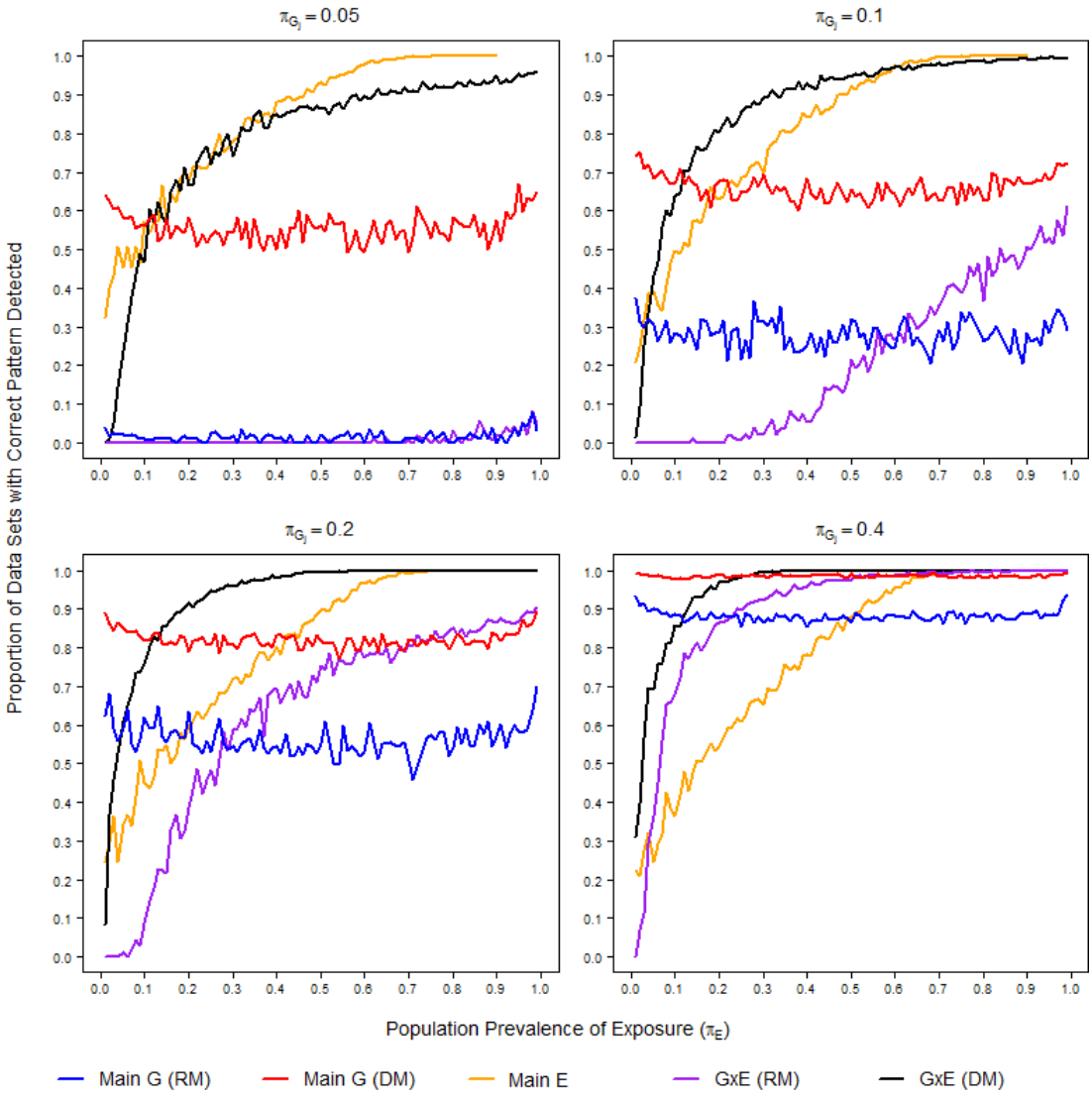

Fig. 4.6: Proportion of the Simulated Data Sets – among Those Exhibiting Some Rejected Null Hypothesis at the 5% Level in the FWER – for Which GEM Correctly Detected the Logical Pattern (y-axis) Versus the Prevalence of the Environmental Exposure (x-axis), upon Selected Values in π_{G_j} (Panel Plots) for $m = 1$. Curves for the Main Genetic Effect Are Shown in Blue (Simulation B) and Red (Simulation C); for the Main Environmental Effect Are Shown in Orange (Simulation D); and for GxE Interaction Are Shown in Purple (Simulation E) and Black (Simulation F). G = Genetic Effect; RM = Recessive Genetic Model; DM = Dominant Genetic Model; E = Environmental Effect.

Table 4.8: Hypothetical Case-Control Sample Showing a Cross-Interaction Pattern of GxE Interaction Between Binary Genetic (G) and Environmental (E) Factors, as Seen by the Opposite Effects in the Estimates of the Odds Ratio ψ Across the Two Levels of Exposure in E. These Data Exhibit a Main Genetic Effect, as Seen by the Estimate of ψ Within the Pooled Data Deviating from the Null Value of One.[†]

	$E=1$			$E=0$			Pooled		
	$Y=1$	$Y=0$	Total	$Y=1$	$Y=0$	Total	$Y=1$	$Y=0$	Total
$G=1$	50	100	150	200	100	300	250	200	450
$G=0$	100	50	150	100	200	300	200	250	450
Total	150	150	300	300	300	600	450	450	900
	$\hat{\psi}=0.25$			$\hat{\psi}=4$			$\hat{\psi}=1.56$		

[†] ψ is the odds ratio of disease ($Y=1$), comparing subjects with $G=1$ to subjects with $G=0$; $\hat{\psi}$ is the MLE of ψ.

In fact, failure to account for cross-interaction can result in overall poor replicability of genetic-phenotype associations. For example, when environmental exposures are not considered, associations are not seen between the alleles of SNP rs2569190 within the *CD14* gene and risk of asthma (see §1.2.1 for cross-interaction between this gene and an environmental factor in risk of asthma) [26, 134, 135, 136]. In light of the importance of the cross-interaction pattern of gene-environment interaction, here we extend upon the above simulation (§4.7) with the specific aim of investigating the statistical power of GEM in its ability to detect the cross-interaction pattern of GxE interaction. We conduct this investigation under conditions for which the genetic factor is always independent of the phenotype (i.e., a main genetic effect is hereby assumed to not exist), with the intention of illustrating the aforementioned statement of [3].

4.8.1 Methods

Letting Y be as previously defined – an indicator of disease – we considered a true penetrance model of the form

$$\text{logit}\left(\Pr\left(Y=1|G,E\right)\right)=\beta_0+\beta_g G+\beta_e E+\gamma_{ge}GE, \tag{4.37}$$

where G is some genotype coding upon a SNP locus and $E=1$ or 0 for respective exposed or unexposed subjects. For clarity in exposition of concept we considered a dominant genotype coding at the locus, where $G=1$ for carriers of the risk allele and $G=0$ for non-carriers. Here, among unexposed subjects, $\exp\left(\beta_g\right)=\text{OR}_g$ is the odds ratio (OR) of disease, comparing carriers of at least one risk allele with non-carriers; among non-carriers of the risk allele, $\exp\left(\beta_e\right)=\text{OR}_e$ is the odds

ratio of disease, comparing exposed with unexposed subjects; and, $\exp(\gamma_{ge}) = \mathrm{OR}_{ge}$ is the ratio of the genetic odds ratios, comparing exposed with unexposed subjects (i.e., $\mathrm{OR}_{g|E=1}/\mathrm{OR}_{g|E=0}$) – equivalently, OR_{ge} is the ratio of the exposure odds ratios, comparing carriers of the risk allele with non-carriers of the risk allele (i.e., $\mathrm{OR}_{e|G=1}/\mathrm{OR}_{e|G=0}$). If this ratio is equal to 1 (equivalently, $\gamma_{ge} = 0$), we say that there is no interaction between genotype at this locus and the environmental factor in their synergistic effect towards risk of disease. Figure 4.7 portrays the model (4.37) in the circumstance of cross-interaction between E and G.

To remain consistent with the notation of §4.7.1, let π_G denote the population MAF at the locus. We assumed that genotypes at the locus are in Hardy-Weinberg equilibrium within the population (i.e., under the dominant GMI, $\Pr(G = 1) = 1 - (1 - \pi_G)^2$), and we assumed that the genetic and environmental factors are independent within the population. We noted that the model (4.37) could be written as

$$\begin{aligned}
&\mathrm{logit}\left(\Pr\left(Y = 1|G, E = 0\right)\right) = \beta_0 + \beta_g G, \text{ and} \\
&\mathrm{logit}\left(\Pr\left(Y = 1|G, E = 1\right)\right) = \left(\beta_0 + \beta_e\right) + \left(\beta_g + \gamma_{ge}\right) G,
\end{aligned}$$

which, for $x \in \{0, 1\}$, in turn can be expressed by

$$\mathrm{logit}\left(\Pr\left(Y = 1|G, E = x\right)\right) = \beta_{x0} + \beta_{xg} G, \tag{4.38}$$

where

$$\beta_{00} = \beta_0, \qquad \beta_{0g} = \beta_g, \qquad \beta_{10} = \beta_0 + \beta_e, \qquad \text{and} \qquad \beta_{1g} = \beta_g + \gamma_{ge}.$$

To invoke a cross-interaction effect between G and E, we assumed the slope parameters of the model (4.38) adhered to the relationship $0 \leq \beta_{1g} = -\beta_{0g}$. Now, since we assumed G and E are independent, for $g \in \{0, 1\}$, it holds

$$\begin{aligned}
\Pr\left(Y = 1, G = g\right) &= \sum_{x \in \{0,1\}} \Pr\left(Y = 1, G = g, E = x\right) \\
&= \sum_{x \in \{0,1\}} \Pr\left(Y = 1|G = g, E = x\right) \Pr\left(G = g\right) \Pr\left(E = x\right).
\end{aligned}$$

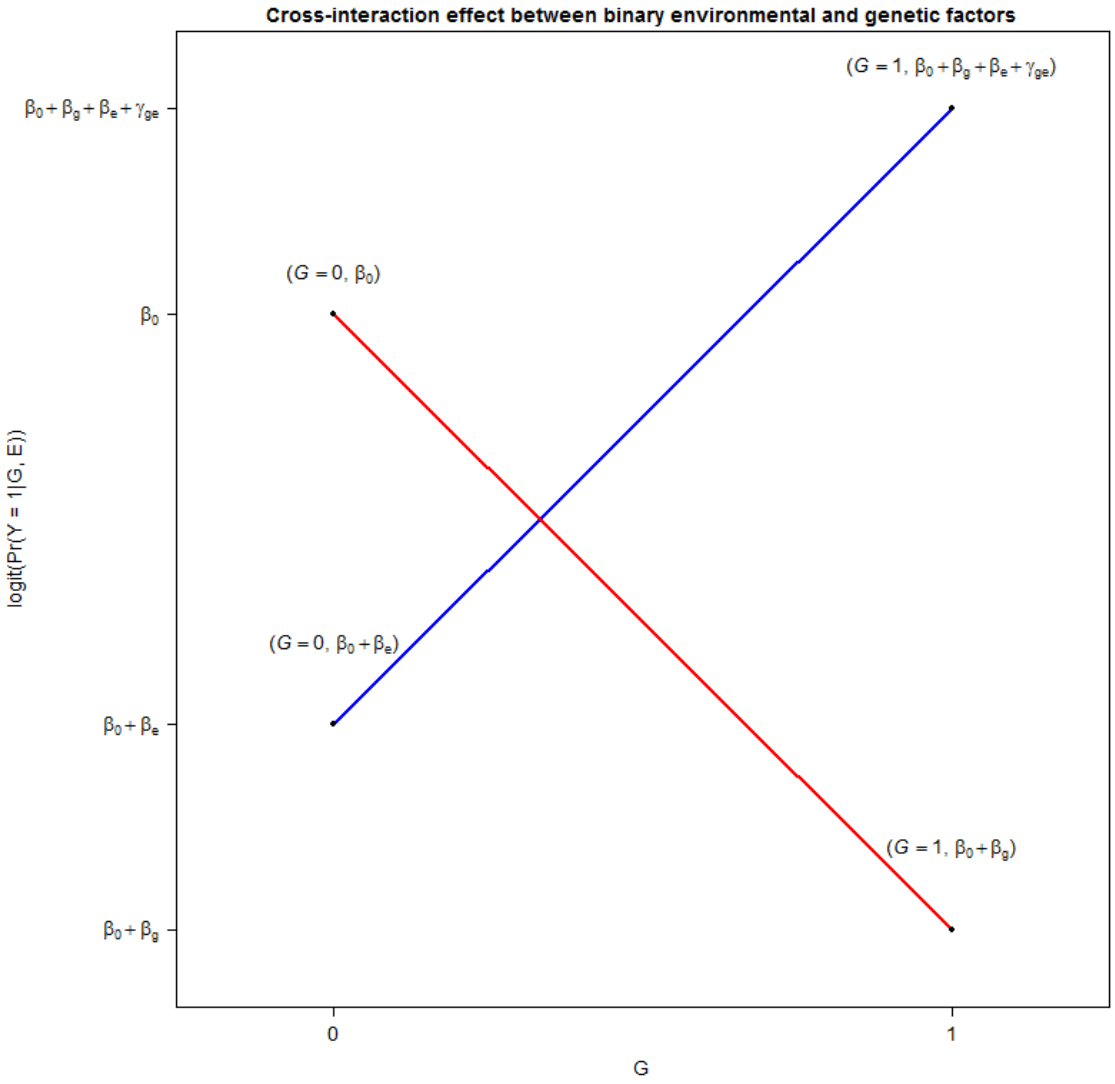

Fig. 4.7: Cross-Interaction Between Binary Environmental and Genetic Factors with Respect to the Assumed Penetrance Model of Disease Risk (4.37). The Blue and Red Lines Represent Risk among Respective Exposed and Unexposed Subjects Within the Population.

Conditioning the left-hand-side of this expression on G, we get

$$\begin{aligned}\Pr\left(Y=1|G=g\right) &= \sum_{x\in\{0,1\}} \Pr\left(Y=1|G=g,E=x\right)\Pr\left(E=x\right) \\ &= \sum_{x\in\{0,1\}} \left(\frac{\exp\left(\beta_{x0}+\beta_{xg}G\right)}{1+\exp\left(\beta_{x0}+\beta_{xg}G\right)}\right)\Pr\left(E=x\right). \end{aligned} \tag{4.39}$$

Since we desired to illustrate cross-interaction between G and E, upon a genetic factor which is assumed independent of Y, we evaluated sufficient conditions upon (4.39) for which $\Pr(Y=1|G=0) = \Pr(Y=1|G=1)$ (i.e., Y and G are independent if this equality holds). It can be readily shown – by way of (4.39) – that Y and G are independent if one of the following two conditions holds: (1) $\beta_{01} = \beta_{11} = 0$; or, (2)

$$\Pr(E=1) = \frac{\operatorname{expit}(\beta_0) - \operatorname{expit}(\beta_0+\beta_g)}{\operatorname{expit}(\beta_0) + \operatorname{expit}(\beta_0+\beta_g+\beta_e+\gamma_{ge}) - \operatorname{expit}(\beta_0+\beta_g) - \operatorname{expit}(\beta_0+\beta_e)}, \tag{4.40}$$

where $\operatorname{expit}(\cdot) = \exp(\cdot)(1+\exp(\cdot))^{-1}$. The former condition implies that $\beta_g = \gamma_{ge} = 0$, for which GxE interaction would not be prevalent. Since we desired a cross-interaction effect between G and E, we assumed the latter condition held.

To evaluate the power to detect cross-interaction using GEM, we conducted a simulation analysis. Throughout the simulation, we assumed a sample size of $n = 1\text{K}$ and – for sake of simplicity – held the parameter β_{10} constant at the value $\operatorname{logit}(0.4)$. We simulated data in coherence with the model (4.38), where we considered a range of values for MAF, $\pi_G \in \{0.05, 0.10, 0.20\}$, effect sizes in E, $\beta_e \in \{0.0, -0.1, -0.25\}$, and GxE interaction effects, $\gamma_{ge} \in \{-2\beta_e + 0.02k\}_{k=0,\ldots,50}$. The parameters β_0 and β_g for the model (4.38) were recovered by referencing the respective relations $\beta_0 = \operatorname{logit}(0.4) - \beta_e$ and $\beta_g = -\gamma_{ge}/2$; the population exposure prevalence, $\Pr(E=1)$, was calculated in coherence with (4.40). For each parameterization in π_G, β_e, and γ_{ge}, a total of $D = 10\text{K}$ data sets were simulated mutually independent of one another. Upon each simulated data set, the adjusted p-value $\tilde{p}_{jl\sigma}$ (4.12) was estimated by (4.18), taking $R = 10\text{K}$ permutations within Algorithm 4.1, for all $l = 1, \ldots, 11$ and $j = m = 1$. We compared the statistical power of GEM to that of the competing methods PCT, BLRT, and GLRT, where these methods are as outlined within §4.7.1. The assessment of the main effect in G was carried out using the Wald-based GEM test statistic upon candidate pattern nine (Z_{jA_9}) – since the dominant genotype coding was assumed for G – and referencing the standard normal distribution for appropriate p-value computation. Statistical significance was set at the 5% level in the FWER, and statistical power was estimated by way of expression (4.36).

4.8.2 Results

Figure 4.8 shows the estimated statistical power to detect: GxE interaction (various blue and red curves) for the competing methods, at the 5% level in the FWER; and, the main effect in the

genetic factor (black curves) at the nominal 5% significance level. As expected, these data indicate that the statistical power to detect a genetic main effect is at the nominal 5% level in the Type I error rate, irrespective of the chosen values upon the elements within the triplicate $(\pi_G, \beta_e, \gamma_{ge})$, as shown by the black curves (essentially horizontal lines) lying at the 5% power level within each of the panel plots of the figure. Thus, these data confirm the notion that a GxE interaction may in fact be prevalent between the genotypes of some SNP locus and exposure status of some binary environmental factor – in the particular form of cross-interaction – where genotype is independent of the phenotype (i.e., a phenotype-genotype association is not necessary for GxE interaction). Under these conditions, by testing for solely main genetic effects (e.g., the GWAS approach), one will most surely fail in detecting GxE interaction. That is, only by a chance finding will genotype at such loci be deemed statistically significantly associated with phenotype, whereupon further study upon these loci – which could include testing for GxE interaction – would be considered. These data also suggest that the statistical power of GEM, for the most part, to be: at least as high as that of the competing methods when no marginal effect[9] in E is present (i.e., $\beta_e = 0$) upon the multiple logistic regression model (4.37), as depicted by the first column panel plots within the figure; to be on par with the competing methods when a small marginal effect in E is present (i.e., $\beta_e = -0.1$) upon said regression model, as shown by the second column panel plots within the figure; and, to be slightly lower that of the competing methods when a moderate marginal effect in E is present (i.e., $\beta_e = -0.25$) upon the regression model, as illustrated by the third column panel plots within the figure. Interestingly, when compared to the other competing methods, the statistical power for the BLRT and GLRT methods seem to suffer upon minute values in β_e and π_G, respectively; the statistical power for each of these methods improves for increasing values in these respective parameters. These observations make sense in light of how these competing methods were defined within §4.7.2. For example, under the assumed dominant genotype coding, the regression coefficients of the GLRT model (4.34) reduce to the assumed penetrance model (4.37). By construction of our cross-interaction model, the power to detect a main effect in the environmental factor is expected to be considerably higher amongst carriers of the risk allele when compared to non-carriers of the allele whenever $\gamma_{ge} > 0$ – note: this notion can be seen visually by way of Figure 4.7. In turn, all else being equal, whenever $\gamma_{ge} > 0$ we expect the power to detect a main effect in the environmental factor to depend upon the population minor allele frequency at the locus – the larger the value in π_G, the

[9] To be clear in discussion, a marginal effect in a covariate is assumed to be the effect (i.e., magnitude of association with the phenotype) due to the covariate, after accounting for the effects of all other covariates of the model.

larger we expect the power to detect the main effect in the environmental factor. In other words, not only is genotype considered an effect modifier for the phenotype-environment relationship, but the degree to which it is an effect modifier for the relationship depends upon the distribution of genotype frequencies at the locus.

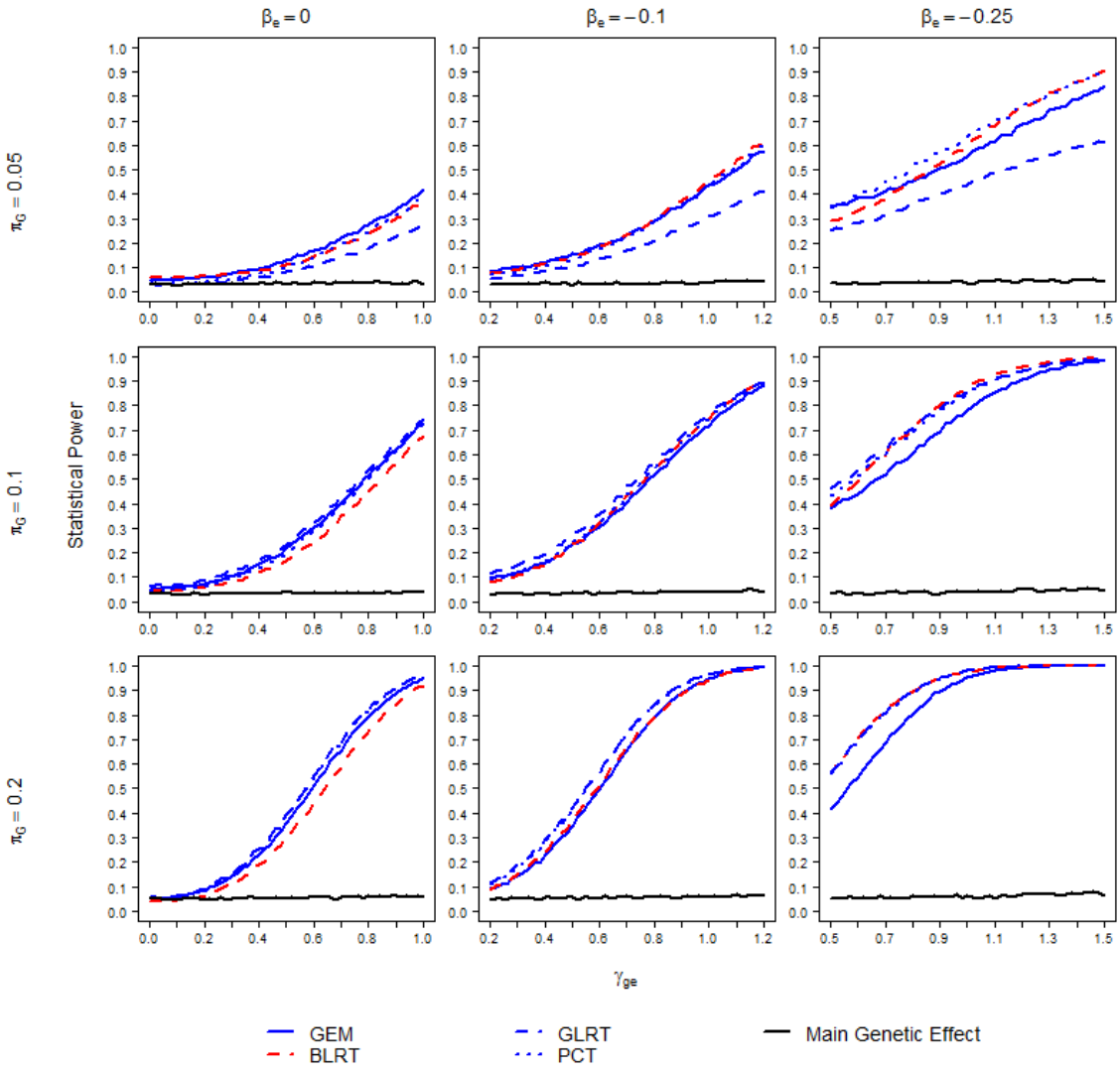

Fig. 4.8: Estimated Statistical Power (y-axis) to Detect GxE Interaction or the Main Effect in the Genetic Factor Versus the Interaction Parameter (γ_{ge}) of the Assumed Penetrance Model (4.37) (x-axis) for a Range of Environmental Marginal Effects (β_e; Columns upon the Panel Plots) and Locus Minor Allele Frequencies (π_G; Rows upon the Panel Plots). Assumed FWER Is 5%; GEM Depicted by Solid Blue Curves; LRT-based Methods by Heavy Dashed Curves; PCT Method by the Light Dashed Blue Curves; and the Power to Detect the Main Genetic Effect Is Depicted by the Black Curves.

From a mathematical vantage point, if δ represents the difference between $\Pr(Y = 1|E = x)$ for x taking the respective values zero and one, then in accordance with the model (4.37), it is

$$\begin{aligned}\delta &= p_G\left(\text{expit}\left(\beta_0 + \beta_g\right) - \text{expit}\left(\beta_0 + \beta_g + \beta_e + \gamma_{ge}\right)\right) + \left(1 - p_G\right)\left(\text{expit}\left(\beta_0\right) - \text{expit}\left(\beta_0 + \beta_e\right)\right) \\ &= p_G\left(\text{expit}\left(\beta_0 - \gamma_{ge}/2\right) - \text{expit}\left(\beta_0 + \beta_e + \gamma_{ge}/2\right)\right) + \left(1 - p_G\right)\left(\text{expit}\left(\beta_0\right) - \text{expit}\left(\beta_0 + \beta_e\right)\right),\end{aligned}$$

where $p_G = \Pr(G = 1) = 1 - (1 - \pi_G)^2$. Differentiating this expression with respect to p_G, we have

$$\frac{d\delta}{dp_G} = \text{expit}\left(\beta_0 - \gamma_{ge}/2\right) + \text{expit}\left(\beta_0 + \beta_e\right) - \text{expit}\left(\beta_0 + \beta_e + \gamma_{ge}/2\right) - \text{expit}\left(\beta_0\right) \leq 0,$$

for any fixed triplicate $(\beta_0, \beta_e, \gamma_{ge})$ of our simulation study. Since $\delta = 0$ represents the null hypothesis of no association between the phenotype and the environmental factor, this result implies that the statistical power to detect the main effect for the environmental factor is an increasing function in p_G (so also an increasing function in π_G by a Chain Rule result). Therefore, the simulation results for the GLRT method are confirmed – namely, increasing statistical power for this method for increasing values in π_G – both from a model (4.37) perspective and a mathematical perspective.

Figure 4.9 shows the estimated statistical power for GEM to detect GxE interaction or the main effect in the environmental factor by candidate pattern at the 5% level in the FWER, where these candidate patterns are as defined within Table 4.1, taking $\varepsilon = 2$ therein for a binary environmental factor. Unsurprisingly, these data indicate that the statistical power to detect GxE interaction is highest for GEM upon candidate patterns $L_{A_5} = (G \in \{1, 2\}) \wedge (E = 0)$ and $L_{A_6} = (G \in \{1, 2\}) \wedge (E = 1)$, since these candidate patterns pertain to GxE interaction for the dominant genetic model. Interestingly, for fixed marginal effects in β_e and γ_{ge}, these data suggest an increasing trend in the statistical power to detect the main effect in E for increasing π_G. This phenomenon can be explained by the arguments presented within the preceding paragraph for the GLRT competing method, and also empirically by comparing the magnitude in the conditional probability

$$\Pr(E = 1|Y = y) = \frac{\sum_{g\in\{0,1\}} \Pr(Y = y|E = 1, G = g) \Pr(E = 1) \Pr(G = g)}{\sum_{x,g\in\{0,1\}} \Pr(Y = y|G = g, E = x) \Pr(E = x) \Pr(G = g)},$$

for each $y \in \{0, 1\}$, whereupon one would substitute the appropriate parameter values – from those given upon the simulation conditions, as defined within the final paragraph of §4.8.1 – within this expression. Taking $\beta_e = 0$ and $\gamma_{ge} = 1.0$, for example, we find that the absolute value in the

difference $\Pr(E = 1|Y = 0) - \Pr(E = 1|Y = 1)$ to equal 0.024, 0.047, and 0.088, for the respective values of π_G equal to 0.05, 0.10, and 0.20. This suggests that the magnitude in the association between Y and E to be increasing for increasing π_G, which is precisely the trend seen within the first column panel plots of the figure for this association.

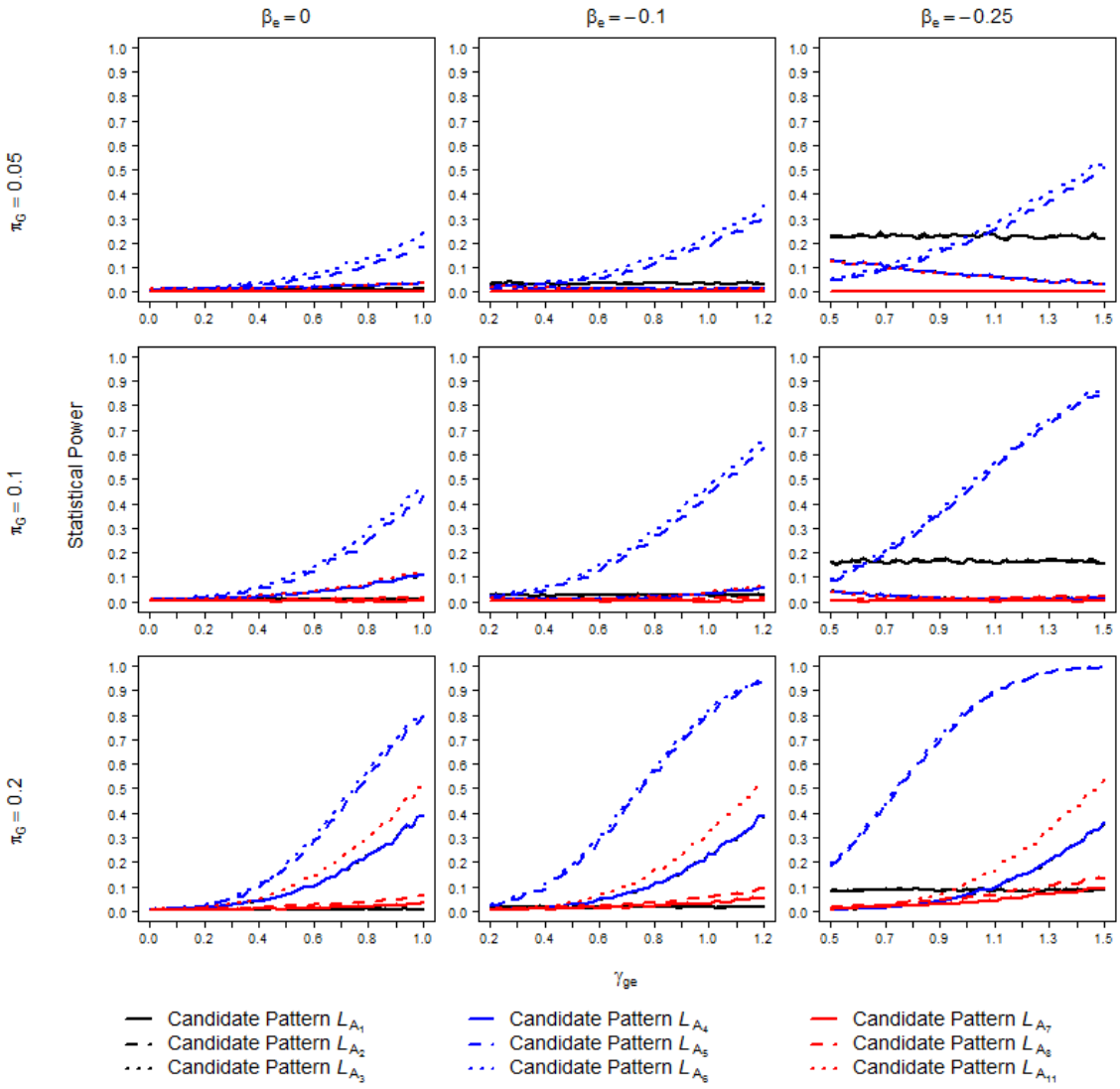

Fig. 4.9: Estimated Statistical Power (y-axis) for GEM to Detect GxE Interaction or the Main Effect in the Environmental Factor by Candidate Pattern Versus the Interaction Parameter (γ_{ge}) of the Assumed Penetrance Model (4.37) (x-axis) for a Range of Environmental Marginal Effects (β_e; Columns upon the Panel Plots) and Locus Minor Allele Frequencies (π_G; Rows upon the Panel Plots), at the 5% Level in the FWER. The Candidate Pattern(s) Corresponding to: GxE Interaction Are Depicted by $L_{A_l}, l = 1, \ldots, 8$; the Main Effect in the Environmental Factor Is Depicted by $L_{A_{11}}$, Where the L_{A_l} Are Specified Within Table 4.1.

4.9 Simulation Study: Control of the FWER under Partial Null Hypotheses

Insofar as GEM can be assumed to control the FWER only in the weak sense (see §4.5 and Proposition A.9), it is important for one to keep in mind that the veracity upon the power comparisons of §4.7 and §4.8, assumes that our proposed method properly controls the FWER at the 5% level under the appropriate partial null hypothesis thereto. For example, since our data setup for each of the aforementioned simulation studies assumes the genetic and environmental factors to be mutually independent, when assessing for a main effect in the environmental factor upon simulation D of the former simulation study, the null hypotheses involving the candidate patterns upon the main effect in the genetic factor, namely $H_0^{(j,9)}$ and $H_0^{(j,10)}$, are in fact true, whereas $H_0^{(j,l)}$, for all $l \in \{1,\ldots,8,11\}$ are in fact false. In this circumstance, unless GEM properly controls the FWER at the 5% level upon these two true null hypotheses (a partial null hypothesis), the credibility in the statistical power for GEM to detect the main effect in the environmental factor is essentially compromised, and the statistical power comparisons could be biased in favor of GEM over some competing methods. Here, we demonstrate [empirically] that GEM appears to properly control the FWER at the 5% level amongst the simulation studies conducted within §4.7 and §4.8, for which $j = m = 1$. Furthermore, we extend upon these simulation studies and examine the ability of GEM to properly control the FWER under several different scenarios governing the elements over the multinomial probability vectors $\boldsymbol{\pi}_{j0}$ and $\boldsymbol{\pi}_{j1}$ (see (4.4)), each scenario in which entails some partial null hypothesis(es) over the collection $\{H_0^{(j,1)},\ldots,H_0^{(j,11)}\}_{j=m=1}$.

4.9.1 Methods

Since we assumed $j = m = 1$, here for clarity in discussion we drop the superscript j from the notation upon the null hypothesis $H_0^{(j,l)}$, and the subscript from: the vector $\boldsymbol{\pi}_{jy}$; each element within this vector; the unadjusted p-value p_{jl}; and, the adjusted p-value $\tilde{p}_{jl\sigma}$ (4.12) and its permutation sampling estimate $\tilde{p}^*_{jl\sigma}$ (4.18). Specifically, for all $l = 1,\ldots,q = 11$, we define $H_0^{(l)}$, p_l, $\tilde{p}_{l\sigma}$, and $\tilde{p}^*_{l\sigma}$ to be equivalent to $H_0^{(j,l)}$, p_{jl}, $\tilde{p}_{jl\sigma}$, and $\tilde{p}^*_{jl\sigma}$, respectively, and for each $y \in \{0,1\} = \mathcal{Y}$ and $k \in \mathcal{X}_2$, we define $\boldsymbol{\pi}_y = (\pi_{1y},\ldots,\pi_{6y}) = \boldsymbol{\pi}_{jy}$, where $\pi_{ky} = \Pr(X = k|Y = y)$, such that the random variable X is as given by (4.16). Based upon the results of Proposition A.9, we considered partial null hypotheses which encompass the true null hypotheses $H_0^{(3)}$ and $H_0^{(9)}$. That is, if $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ denotes the true partial null hypothesis over the collection of null hypotheses $\{H_0^{(1)},\ldots,H_0^{(11)}\}$, it is

$$\{H_0^{(3)}, H_0^{(9)}\} \subseteq \tilde{\mathcal{H}}_0^{\mathrm{p}}.$$

We considered a biallelic SNP locus with population minor allele frequency (MAF) π_G, and a binary environmental factor with a population prevalence of exposure given by $\pi_E = \Pr(E = 1)$. We assumed the genotype frequencies at the locus adhere to Hardy-Weinberg equilibrium (HWE) within the population, and that the genotypes at the locus are independent of the environmental factor.

To generate partial null hypotheses, we considered two scenarios governing the probability vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$: (a) we assumed that genotype frequencies at the locus adhere to HWE among population controls, and that the elements within the vector $\boldsymbol{\pi}_1$ could be expressed by

$$\pi_{k1} = \begin{cases} \pi_{k0} + \delta, & \text{if } k \in \{1, 5\} \\ \pi_{k0} - \delta, & \text{if } k \in \{2, 4\} \\ \pi_{k0}, & \text{if } k \in \{3, 6\} \end{cases}, \tag{4.41}$$

for some real value of δ, such that $0 \leq \pi_{k1} \leq 1$ for all $k \in \mathcal{X}_2$. This setup for the elements over $\boldsymbol{\pi}_1$ ensured that $\{H_0^{(3)}, H_0^{(9)}\} \subseteq \tilde{\mathcal{H}}_0^{\mathrm{p}}$, and assumed – among other things – that risk of disease is the same, comparing subjects carrying two copies of the risk allele at the locus to those carrying not more than one copy of the risk allele at the locus. In fact, one can trivially show that

$$\tilde{\mathcal{H}}_0^{\mathrm{p}} = \{H_0^{(3)}, H_0^{(4)}, H_0^{(7)}, H_0^{(8)}, H_0^{(9)}, H_0^{(10)}, H_0^{(11)}\}; \tag{4.42}$$

and, (b) we assumed that genotype frequencies at the locus adhere to HWE among population controls, and that the elements within the vector $\boldsymbol{\pi}_1$ could be expressed by

$$\pi_{k1} = \begin{cases} \pi_{k0} + \delta_1, & \text{if } k \in \{1\} \\ \pi_{k0} - \delta_1, & \text{if } k \in \{2, 4\} \\ \pi_{k0} + \delta_2, & \text{if } k \in \{3\} \\ \pi_{k0} + \delta_3, & \text{if } k \in \{5\} \\ \pi_{k0} + \delta_1 - \delta_2 - \delta_3, & \text{if } k \in \{6\} \end{cases}, \tag{4.43}$$

for real numbers δ_1, δ_2, and δ_3, such that $0 \leq \pi_{k1} \leq 1$ for all $k \in \mathcal{X}_2$. This setup for the elements over $\boldsymbol{\pi}_1$ ensured that $\{H_0^{(3)}, H_0^{(9)}\} \subseteq \tilde{\mathcal{H}}_0^{\mathrm{p}}$, and – when compared to the first setup of $\boldsymbol{\pi}_1$ (4.41) – allowed for greater generalizability in the behavior of the control over the FWER for GEM.

Given the above assumptions governing the population characteristics over the random variables Y and – by way of G and E – X, to evaluate the integrity of GEM in its ability to control the FWER

under partial null hypotheses, we conducted a simulation study. As with the previous simulation studies (§4.7 and §4.8), here we assumed a sample size of $n = 1\text{K}$ per simulated data set. We also assumed a balanced case-control study design, so that $n_0 = n_1 = 500$. We considered a range of values for MAF, $\pi_G \in \{0.05, 0.10, 0.20, 0.40\}$, and a range of values for the exposure prevalence, $\pi_E \in \{0.10, 0.20, 0.30, 0.40, 0.50\}$. Upon the first scenario governing the probability vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ (4.41), for each combination of π_G and π_E we considered two values for δ – one negative and one positive – at the plausible extremities over the real interval $[-0.1, 0.1]$. For example, consider $\pi_G = 0.05$ and $\pi_E = 0.10$. Under HWE and G-E independence, it follows that

$$\begin{aligned}
\pi_{11} &= \pi_{10} + \delta = \Pr(G = 0)\Pr(E = 0) + \delta = (1 - \pi_G)^2 (1 - \pi_E) + \delta = 0.8123 + \delta \\
\pi_{21} &= \pi_{20} - \delta = \Pr(G = 1)\Pr(E = 0) - \delta = 2\pi_G (1 - \pi_G)(1 - \pi_E) - \delta = 0.0855 - \delta \\
\pi_{41} &= \pi_{40} - \delta = \Pr(G = 0)\Pr(E = 1) - \delta = (1 - \pi_G)^2 \pi_E - \delta = 0.09025 - \delta \\
\pi_{51} &= \pi_{50} + \delta = \Pr(G = 1)\Pr(E = 1) + \delta = 2\pi_G (1 - \pi_G)\pi_E + \delta = 0.0095 + \delta
\end{aligned},$$

so that δ is confined to lie within the interval $[-0.0095, 0.0855]$. Thus, for this example, we considered $\delta \in \{-0.0095, 0.0855\}$. Overall, we considered a total of 40 ordered triples in (π_G, π_E, δ) for the first scenario governing the probability vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$. Upon the latter scenario governing these probability vectors (4.43), we carried out the following procedure:

1. For each combination of π_G and π_E, we formulated the vector $\boldsymbol{\pi}_0$ under the assumptions of HWE and G-E independence, and considered assigning plausible ordered triples in $(\delta_1, \delta_2, \delta_3)$ over the collection

$$\Delta = \left\{(a_1, a_2, a_3) \in \mathbb{R}^3 : a_i = -0.10 + 0.01 s_i, \text{ where } s_i = 0, \ldots, 20, \text{ for all } i = 1, 2, 3\right\}.$$

2. We substituted each element of this collection within (4.43) and determined whether the following condition held: $0 \leq \pi_{1k} \leq 1$ for all $k \in \mathcal{X}_2$. Let $\Delta(\pi_G, \pi_E) \subseteq \Delta$ denote those elements of Δ which satisfied said condition.

3. For each element within the collection $\Delta(\pi_G, \pi_E)$, we formulated the vector $\boldsymbol{\pi}_1$ and determined its accompanying partial null hypothesis, $\tilde{\mathcal{H}}_0^{\text{P}}$. Let $\tilde{\mathcal{H}}_0^{\text{P}}(\pi_G, \pi_E)$ denote the collection of unique partial null hypotheses upon those constructed over the collection $\Delta(\pi_G, \pi_E)$.

4. We considered $\tilde{\mathcal{H}}_0^{\text{P}} \in \tilde{\mathcal{H}}_0^{\text{P}}(\pi_G, \pi_E)$. Now, since $\left\{H_0^{(3)}, H_0^{(9)}\right\} \subseteq \tilde{\mathcal{H}}_0^{\text{P}}$, we were particularly

interested in the $\boldsymbol{\pi}_1$ – formulated from some element within $\Delta(\pi_G, \pi_E)$, and yielding the partial null hypothesis $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ – for which the magnitude of the difference

$$Cov\left(Z^*_{A_3}, Z^*_{A_9} | \tilde{\mathcal{H}}_0^{\mathrm{p}}\right) - Cov\left(Z^*_{A_3}, Z^*_{A_9} | \mathcal{H}_0\right) \tag{4.44}$$

is most extreme, where $Z^*_{A_l}$ denotes the test statistic (4.10) under $H_0^{(l)}$, with (4.7) substituted in lieu of (4.8) therein. We were interested in these $\boldsymbol{\pi}_1$, because extreme values of (4.44) signify that the joint distribution of the test statistics $Z^*_{A_3}$ and $Z^*_{A_9}$ under $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ is very different from that under $\mathcal{H}_0$. That is, if one could assign a magnitude upon the extent to which the subset pivotality condition is violated, in this circumstance – with regard to solely the joint distribution of the test statistics $Z^*_{A_3}$ and $Z^*_{A_9}$ – said condition is 'violated to an extreme extent,' and we conjectured that this could have an adverse consequence towards GEMs ability to properly control the FWER under $\tilde{\mathcal{H}}_0^{\mathrm{p}}$. So, for each $\tilde{\mathcal{H}}_0^{\mathrm{p}} \in \tilde{\mathcal{H}}_0^{\mathrm{p}}(\pi_G, \pi_E)$, we selected a pair of ordered triples from $\Delta(\pi_G, \pi_E)$, such that: each ordered triple yielded $\tilde{\mathcal{H}}_0^{\mathrm{p}}$; and, collectively the pair of ordered triples yielded the most extreme negative and positive values of (4.44), respectively, amongst all ordered triples of $\Delta(\pi_G, \pi_E)$ yielding $\tilde{\mathcal{H}}_0^{\mathrm{p}}$.

Overall, we considered a total of 408 ordered quintuples in $(\pi_G, \pi_E, \delta_1, \delta_2, \delta_3)$ for the second scenario governing the probability vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$. Table 4.9 summarizes the distributions of the collections $\Delta(\pi_G, \pi_E)$ and $\tilde{\mathcal{H}}_0^{\mathrm{p}}(\pi_G, \pi_E)$ for each combination of π_G and π_E. For each ordered pair (π_G, π_E), this provided us at least five unique partial null hypotheses to consider (fourth column) and provided us with a wide range to the size of the elements (i.e., the number of hypotheses included within $\tilde{\mathcal{H}}_0^{\mathrm{p}}$) of the collection $\tilde{\mathcal{H}}_0^{\mathrm{p}}(\pi_G, \pi_E)$ (final column). For example, consider $\pi_G = 0.05$ and $\pi_E = 0.10$. Here,

(4.45)

$$\begin{aligned}\tilde{\mathcal{H}}_0^{\mathrm{p}}(\pi_G, \pi_E) = \; & \Big\{\{H_0^{(3)}, H_0^{(9)}\}, \{H_0^{(3)}, H_0^{(8)}, H_0^{(9)}\}, \{H_0^{(3)}, H_0^{(7)}, H_0^{(9)}, H_0^{(11)}\}, \\ & \{H_0^{(3)}, H_0^{(5)}, H_0^{(6)}, H_0^{(8)}, H_0^{(9)}\}, \{H_0^{(3)}, H_0^{(4)}, H_0^{(7)}, H_0^{(8)}, H_0^{(9)}, H_0^{(10)}, H_0^{(11)}\}\Big\}\end{aligned},$$

the collection in which includes five unique and varying size partial null hypotheses to consider. In contrast, the first scenario governing the probability vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ (4.41), provided us with only a single common partial null hypothesis (4.42) to consider for each-and-every ordered pair (π_G, π_E).

For each of the 40 ordered triples (π_G, π_E, δ) and each of the 408 ordered quintuples $(\pi_G, \pi_E, \delta_1, \delta_2, \delta_3)$, we simulated a total of $D = 2\mathrm{K}$ mutually independent data sets. For each data

set, we estimated the adjusted p-value $\tilde{p}_{l\sigma}$ (4.12) with (4.18), for all $l = 1, \ldots, 11$, taking $R = 5\text{K}$ permutations within Algorithm 4.1. We considered rejection of null hypothesis $H_0^{(l)}$ – under the assumption that the complete null hypothesis held (see §4.4) – if and only if the estimate of the adjusted p-value $\tilde{p}_{l\sigma}$ assumed a value not larger than $\tilde{\alpha}_{\text{F}}$, for all $l = 1, \ldots, 11$, some $\tilde{\alpha}_{\text{F}} \in (0, 1)$ – here, $\tilde{\alpha}_{\text{F}}$ is the user *assumed* level in the FWER being controlled under GEM. If α_{F} and β_{F} denote the respective *true* underlying FWER and Type II error rates for GEM, at the assumed $\tilde{\alpha}_{\text{F}}$ FWER level for a given ordered triple (π_G, π_E, δ) or ordered quintuple $(\pi_G, \pi_E, \delta_1, \delta_2, \delta_3)$, we estimated these true parameters by their respective MLEs

$$\hat{\alpha}_{\text{F}} = \frac{\sum_d V_d}{D}, \tag{4.46}$$

and

$$\hat{\beta}_{\text{F}} = \frac{\sum_d U_d}{D}, \tag{4.47}$$

where for data set d, $d = 1, \ldots, D$,

$$\begin{aligned} V_d &= I\left(\tilde{p}^*_{l\sigma} \leq \tilde{\alpha}_{\text{F}},\ \text{for some } H_0^{(l)} \in \tilde{\mathcal{H}}_0^{\text{p}}\right) \sim \text{Bernoulli}\left(\alpha_{\text{F}}\right), \text{ and} \\ U_d &= I\left(\tilde{p}^*_{l\sigma} > \tilde{\alpha}_{\text{F}},\ \text{for all } H_0^{(l)} \in \mathcal{H}_0 \backslash \tilde{\mathcal{H}}_0^{\text{p}}\right) \sim \text{Bernoulli}\left(\beta_{\text{F}}\right). \end{aligned}$$

Statistical significance was set at the 5% level in the FWER (i.e., $\tilde{\alpha}_{\text{F}} = 0.05$).

4.9.2 Results

Table 4.10 provides summary measures for the 40 ordered triples (π_G, π_E, δ), upon the simulation conducted over the first scenario governing the probability vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ (4.41). These data indicate that GEM controls the FWER at the 5% level under the partial null hypothesis $\tilde{\mathcal{H}}_0^{\text{p}}$ (4.42), since for each ordered triple (π_G, π_E, δ): the observed FWER ($\hat{\alpha}_{\text{F}}$; fifth column) lies below the value 0.05; and, the corresponding 95% exact Clopper-Pearson confidence interval for α_{F} either covers the value of 0.05, or encapsulates values below that of 0.05. Moreover, the data suggest GEM controls the FWER at the 5% level, even for extreme differences in the joint distribution between the test statistics $Z^*_{A_3}$ and $Z^*_{A_9}$ under the assumed partial null hypothesis (4.42) and the complete null hypothesis (fourth column). For example, consider $(\pi_G, \pi_E, \delta) = (0.05, 0.10, 0.08)$. Under $\tilde{\mathcal{H}}_0^{\text{p}}$ (4.42) we find $Cov\left(Z^*_{A_3}, Z^*_{A_9}\right) = 0.50$, while under $\mathcal{H}_0$ we find the value of this covariance to be

0.02. The joint distribution between the test statistics $Z^*_{A_3}$ and $Z^*_{A_9}$ is considerably different under $\tilde{\mathcal{H}}_0^{\mathrm{P}}$ and $\mathcal{H}_0$, yet the data indicate that GEM controls the FWER at the 5% level ($\hat{\alpha}_{\mathrm{F}} = 0.024$; 95% CI for α_{F} (0.017, 0.031)).

Table 4.9: Summary of the Partial Null Hypotheses Considered for the Second Scenario Governing the Probability Vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ (4.43).†

π_G	π_E	$n\left(\Delta\left(\pi_G, \pi_E\right)\right)$	$n\left(\tilde{\mathcal{H}}_0^{\mathrm{P}}\left(\pi_G, \pi_E\right)\right)$	$n\left(\tilde{\mathcal{H}}_0^{\mathrm{P}}\right) : \tilde{\mathcal{H}}_0^{\mathrm{P}} \in \tilde{\mathcal{H}}_0^{\mathrm{P}}\left(\pi_G, \pi_E\right)$
0.05	0.10	165	5	$\{2, 3, 4, 5, 7\}$
	0.20	165	8	$\{2, 3, 4, 5, 7, 8\}$
	0.30	165	9	$\{2, 3, 4, 5, 7, 8\}$
	0.40	165	9	$\{2, 3, 4, 5, 7, 8\}$
	0.50	165	9	$\{2, 3, 4, 5, 7, 8\}$
0.10	0.10	220	8	$\{2, 3, 4, 5, 7, 8\}$
	0.20	550	9	$\{2, 3, 4, 5, 7, 8\}$
	0.30	781	9	$\{2, 3, 4, 5, 7, 8\}$
	0.40	1056	9	$\{2, 3, 4, 5, 7, 8\}$
	0.50	1210	9	$\{2, 3, 4, 5, 7, 8\}$
0.20	0.10	455	12	$\{2, 3, 4, 5, 6, 7, 8\}$
	0.20	1474	12	$\{2, 3, 4, 5, 6, 7, 8\}$
	0.30	2066	12	$\{2, 3, 4, 5, 6, 7, 8\}$
	0.40	2294	12	$\{2, 3, 4, 5, 6, 7, 8\}$
	0.50	2499	12	$\{2, 3, 4, 5, 6, 7, 8\}$
0.40	0.10	1275	12	$\{2, 3, 4, 5, 6, 7, 8\}$
	0.20	4211	12	$\{2, 3, 4, 5, 6, 7, 8\}$
	0.30	6090	12	$\{2, 3, 4, 5, 6, 7, 8\}$
	0.40	6260	12	$\{2, 3, 4, 5, 6, 7, 8\}$
	0.50	6367	12	$\{2, 3, 4, 5, 6, 7, 8\}$

† $n(\cdot)$ represents the cardinality of the set $(\cdot)$.

In light of the fact that many estimates of α_{F} within this table assume values considerably smaller than expected (namely, 0.05), one might be inclined to conjecture that GEM is overly conservative in its control of the FWER at the 5% level. However, this is likely not the circumstance, and can be attributed to the underlying characteristics in which the maxT MTP controls the FWER under the partial null hypothesis $\tilde{\mathcal{H}}_0^{\mathrm{P}}$ (4.42). To see this, assuming the unadjusted p-value P_l – for test statistic (4.10) – is distributed as $U(0, 1)$ under $H_0^{(l)}$, for all $H_0^{(l)} \in \mathcal{H}_0^{\mathrm{P}}$, some $\mathcal{H}_0^{\mathrm{P}} \subseteq \mathcal{H}_0$, here we note that the FWER for the Bonferroni MTP is given by

$$\alpha_{\mathrm{F}} = \Pr\left(V \geq 1 | \mathcal{H}_0^{\mathrm{P}}\right) = \Pr\left(\bigcup_{H_0^{(l)} \in \mathcal{H}_0^{\mathrm{P}}} \{P_l^{\mathrm{B}} \leq \tilde{\alpha}_{\mathrm{F}}\}\right) \leq \sum_{H_0^{(l)} \in \mathcal{H}_0^{\mathrm{P}}} \Pr\left(P_l \leq \frac{\tilde{\alpha}_{\mathrm{F}}}{11}\right) \leq \frac{n\left(\mathcal{H}_0^{\mathrm{P}}\right) \tilde{\alpha}_{\mathrm{F}}}{11}, \tag{4.48}$$

where V denotes the number of Type I errors committed in testing the partial null hypothesis $\mathcal{H}_0^{\mathrm{P}}$,

P_l^{B} is the Bonferroni adjusted p-value for null hypothesis $H_0^{(l)}$, and $n(\cdot)$ denotes the cardinality of the collection $(\cdot)$.

Table 4.10: Estimated Family-wise Type I Error Rate and Statistical Power for GEM under Partial Null Hypotheses over Various Parametrizations of the Ordered Triple $(\pi_G, \pi_E, \delta)^{\dagger}$.

π_G	π_E	δ	$Cov\left(Z^*_{A_3}, Z^*_{A_9}\right)$	Observed FWER (95% CI for α_{F})	Power $(1-\hat{\beta}_{\mathrm{F}})$
0.05	0.10	−0.01	−0.03/0.02	0.044 (0.035, 0.053)	0.30
	0.20	−0.02	−0.06/0.02	0.034 (0.026, 0.042)	0.88
	0.30	−0.03	−0.09/0.01	0.034 (0.026, 0.043)	0.97
	0.40	−0.04	−0.12/0.01	0.021 (0.015, 0.028)	1.00
	0.50	−0.05	−0.15/0.01	0.027 (0.020, 0.035)	1.00
0.10	0.10	−0.02	−0.01/0.06	0.044 (0.035, 0.054)	0.77
	0.20	−0.04	−0.07/0.04	0.033 (0.026, 0.042)	0.99
	0.30	−0.05	−0.12/0.03	0.040 (0.032, 0.050)	1.00
	0.40	−0.07	−0.16/0.03	0.042 (0.033, 0.051)	1.00
	0.50	−0.09	−0.21/0.02	0.030 (0.023, 0.038)	0.93
0.20	0.10	−0.03	0.04/0.14	0.043 (0.035, 0.053)	1.00
	0.20	−0.06	−0.06/0.10	0.030 (0.023, 0.038)	1.00
	0.30	−0.10	−0.13/0.08	0.033 (0.026, 0.042)	1.00
	0.40	−0.10	−0.15/0.06	0.032 (0.024, 0.040)	1.00
	0.50	−0.10	−0.16/0.05	0.035 (0.027, 0.043)	0.98
0.40	0.10	−0.05	0.14/0.25	0.042 (0.034, 0.052)	1.00
	0.20	−0.10	−0.01/0.20	0.033 (0.026, 0.042)	1.00
	0.30	−0.10	−0.04/0.17	0.037 (0.029, 0.046)	1.00
	0.40	−0.10	−0.06/0.14	0.024 (0.017, 0.031)	1.00
	0.50	−0.10	−0.09/0.12	0.032 (0.024, 0.040)	1.00
0.05	0.10	0.08	0.50/0.02	0.024 (0.017, 0.031)	1.00
	0.20	0.08	0.33/0.02	0.023 (0.016, 0.030)	1.00
	0.30	0.07	0.25/0.01	0.027 (0.020, 0.035)	1.00
	0.40	0.06	0.21/0.01	0.029 (0.022, 0.037)	1.00
	0.50	0.05	0.17/0.01	0.028 (0.021, 0.036)	1.00
0.10	0.10	0.08	0.39/0.06	0.035 (0.027, 0.044)	1.00
	0.20	0.10	0.36/0.04	0.034 (0.026, 0.042)	1.00
	0.30	0.10	0.31/0.03	0.032 (0.024, 0.040)	1.00
	0.40	0.10	0.28/0.03	0.032 (0.025, 0.041)	1.00
	0.50	0.09	0.25/0.02	0.041 (0.032, 0.050)	1.00
0.20	0.10	0.06	0.33/0.14	0.031 (0.024, 0.040)	1.00
	0.20	0.10	0.35/0.10	0.034 (0.026, 0.042)	1.00
	0.30	0.10	0.30/0.08	0.040 (0.031, 0.049)	1.00
	0.40	0.10	0.28/0.06	0.034 (0.026, 0.043)	1.00
	0.50	0.10	0.26/0.05	0.033 (0.025, 0.041)	1.00
0.40	0.10	0.04	0.34/0.25	0.045 (0.036, 0.055)	1.00
	0.20	0.07	0.36/0.20	0.038 (0.030, 0.047)	1.00
	0.30	0.10	0.38/0.17	0.034 (0.026, 0.043)	1.00
	0.40	0.10	0.35/0.14	0.038 (0.030, 0.047)	1.00
	0.50	0.10	0.33/0.12	0.043 (0.034, 0.052)	1.00

†The null hypotheses $H_0^{(l)}$, for $l \in \{3,4,7,8,9,10,11\}$, are in fact true for the given parameterizations in (π_G, π_E, δ) within this table (see (4.42)); Depicted value of $Cov\left(Z^*_{A_3}, Z^*_{A_9}\right)$ is calculated under $\tilde{\mathcal{H}}_0^{\mathrm{p}}/\mathcal{H}_0$, where $Z^*_{A_l}$ is defined by (4.44); 95% confidence intervals (CI) for α_{F} are exact Clopper-Pearson; Assumed FWER is 5% ($\tilde{\alpha}_{\mathrm{F}} = 0.05$).

Substituting $\tilde{\alpha}_{\mathrm{F}} = 0.05$ and $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ (4.42) within (4.48), we find that the Bonferroni MTP controls the FWER at a level not exceeding 0.032. This apparent conservative control of the FWER for the Bonferroni MTP under $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ can be attributed to its underlying characteristics for achieving strong control of the FWER, as seen through (4.48). A similar argument for the maxT MTP can explain the apparent conservative estimates of α_{F} given within the table (Table 4.10). Finally, these data suggest that GEM controls the FWER at the 5% level, irrespective of the statistical power to detect true associations, as seen by the variety of estimates within the final column of the table.

Table 4.11 provides summary measures for the 408 ordered quintuples $(\pi_G, \pi_E, \delta_1, \delta_2, \delta_3)$, for the simulation conducted over the second scenario governing the probability vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ (4.43). Although some estimates of α_{F} suggest GEMs control of the FWER at a level exceeding the 5% level (e.g., $\max\{\hat{\alpha}_{\mathrm{F}}\} = 0.056$ for $(\pi_G, \pi_E) = (0.05, 0.30)$), after accounting for sampling variation, these data indicate that GEM controls the FWER at the 5% level, insofar as all 95% confidence intervals for α_{F}, either: cover the value 0.05; or, cover values strictly falling below that of 0.05. The fourth column of the table provides the average of $\hat{\alpha}_{\mathrm{F}}$ for the selected pair(s) of ordered triples $(\delta_1, \delta_2, \delta_3)$ from $\Delta(\pi_G, \pi_E)$ yielding a partial null hypothesis $\tilde{\mathcal{H}}_0^{\mathrm{p}} \in \tilde{\mathcal{H}}_0^{\mathrm{p}}(\pi_G, \pi_E)$ by value of $n\left(\tilde{\mathcal{H}}_0^{\mathrm{p}}\right)$, such that the first and second elements of the accompanying superscript indicate the respective values of $n\left(\tilde{\mathcal{H}}_0^{\mathrm{p}}\right)$ and the number of selected pair(s) of ordered triples $(\delta_1, \delta_2, \delta_3)$ from $\Delta(\pi_G, \pi_E)$ yielding a partial null hypothesis comprised of $n\left(\tilde{\mathcal{H}}_0^{\mathrm{p}}\right)$ true null hypotheses.[10] For example, consider $\pi_G = 0.05$ and $\pi_E = 0.10$, for which $\tilde{\mathcal{H}}_0^{\mathrm{p}}(\pi_G, \pi_E)$ is given by (4.45). So, the value $0.024^{(2,1)}$ given in the table, implies that the arithmetic average of the two values in $\hat{\alpha}_{\mathrm{F}}$ – corresponds to the single (second superscript element) selected pair of ordered triples $(\delta_1, \delta_2, \delta_3)$ from $\Delta(\pi_G, \pi_E)$ with $n\left(\tilde{\mathcal{H}}_0^{\mathrm{p}}\right) = 2$ (first superscript element; i.e., $\tilde{\mathcal{H}}_0^{\mathrm{p}} = \{H_0^{(3)}, H_0^{(9)}\}$) – is equal to 0.024. For the most part, the data indicate that [average] $\hat{\alpha}_{\mathrm{F}}$ is an increasing function in $n\left(\tilde{\mathcal{H}}_0^{\mathrm{p}}\right)$ (the initial element upon the superscripts) for any given ordered pair (π_G, π_E), particularly for values in $\pi_G \geq 0.10$ and $\pi_E \geq 0.20$. This result is not unexpected and can be explained by an analogous argument to that presented above encompassing (4.48). Specifically, for a fixed value of $\tilde{\alpha}_{\mathrm{F}}$, the upper bound of said expression is increasing in $n\left(\tilde{\mathcal{H}}_0^{\mathrm{p}}\right)$. In this circumstance, we expect estimates of α_{F} to be increasing in $n\left(\tilde{\mathcal{H}}_0^{\mathrm{p}}\right)$ for the Bonferroni MTP – so also, the maxT MTP – and the data precisely support this notion (through the arithmetic mean of the $\hat{\alpha}_{\mathrm{F}}$). Finally, these data support the notion that GEM

[10] For a given ordered pair (π_G, π_E), note that: taking the union of the initial element over the superscripts, yields the corresponding set depicted upon the final column of Table 4.9; and, summing over the second element of the superscripts yields the corresponding cell value depicted within the fourth column of said table.

controls the FWER at the 5% level, irrespective of the statistical power to detect true associations, as seen by the assortment of estimates within the final column of the table.

Table 4.12 summarizes GEMs control over the FWER at the assumed 5% level ($\tilde{\alpha}_{\mathrm{F}} = 0.05$), upon: the final simulation study conducted within §4.7 (i.e., the investigation of the "success rate for GEM" (SRG) – see final paragraph within §4.7.4), depicted by the first half of the table; and, the simulation conducted within §4.8 (i.e., the investigation of GEMs ability to detect the cross-interaction pattern of GxE interaction), depicted by the latter half of the table. These data indicate that GEM controls the FWER at the 5% level, upon the simulations summarized within the table, since: nearly all of the estimates of α_{F} lie below the value of 0.05 (the only exceptions being $\hat{\alpha}_{\mathrm{F}} \geq 0.05$ upon the simulations conducted over §4.8, taking $\gamma_{ge} = \beta_e = 0$ therein); and, after considering sampling variation, all 95% exact Clopper-Pearson confidence intervals for α_{F} either cover the value 0.05, or cover values strictly falling below that of 0.05. These observations imply that, when comparing the statistical power of GEM to that of competing methods within these simulation studies, the veracity in these comparisons are essentially sound. Furthermore, the fact that the estimates of α_{F} assume values approximately equal to/slightly exceeding that of $\tilde{\alpha}_F = 0.05$ upon the simulations conducted over §4.8, for which $\gamma_{ge} = \beta_e = 0$, can be attributed to the fact that the complete null hypothesis is in fact true in these circumstances. Finally, the apparent conservative control of the FWER at the 5% level upon the maxT MTP (e.g., the average $\hat{\alpha}_{\mathrm{F}} = 0.007$, taking $\pi_G = 0.40$ for the main genetic effect of the dominant genetic model) can be explained by an analogous argument to that presented above regarding (4.48).

4.10 Application

4.10.1 Methods

To illustrate application of our proposed GEM method in practice, we applied it against two population-based case-control study samples. The first, a study of colon cancer, included $n_1 = 1555$ cases of cancer and $n_0 = 1956$ healthy controls; the second, a study of rectal cancer, included $n_1 = 754$ cases of cancer and $n_0 = 959$ healthy controls. Details encompassing the sampling characteristics for these studies can be found within the respective articles [174, 175]. A candidate pathway, consisting of genes involved in modulating reactive oxygen species (ROS; chemically reactive molecules carrying oxygen), was constructed over the four genes: eosinophil peroxidase (*EPX*); myeloperoxidase (*MPO*); hypoxia-inducible factor-1A (*HIF1A*); and nitric oxide synthase (*NOS2A*)

Table 4.11: Estimated Family-wise Type I Error Rate and Statistical Power for GEM under Partial Null Hypotheses over Various Parametrizations of the Ordered Quintuple $(\pi_G, \pi_E, \delta_1, \delta_2, \delta_3)^{\dagger}$.

		$\max\{\hat{\alpha}_{\mathrm{F}}\}$	Average	
π_G	π_E	(95% CI for α_{F})	$\hat{\alpha}_{\mathrm{F}}^{(\mathrm{a})}$	Power $(1-\hat{\beta}_{\mathrm{F}})$
0.05	0.10	0.039 (0.031, 0.049)	$0.024^{(2,1)}$, $0.023^{(3,1)}$, $0.024^{(4,1)}$ $0.039^{(5,1)}$, $0.027^{(7,1)}$	0.70
	0.20	0.047 (0.038, 0.057)	$0.018^{(2,1)}$, $0.021^{(3,1)}$, $0.023^{(4,2)}$ $0.038^{(5,2)}$, $0.029^{(7,1)}$, $0.046^{(8,1)}$	0.71
	0.30	0.056 (0.046, 0.067)	$0.018^{(2,1)}$, $0.022^{(3,1)}$, $0.029^{(4,3)}$ $0.036^{(5,2)}$, $0.022^{(7,1)}$, $0.046^{(8,1)}$	0.81
	0.40	0.047 (0.038, 0.057)	$0.020^{(2,1)}$, $0.020^{(3,1)}$, $0.026^{(4,3)}$ $0.035^{(5,2)}$, $0.029^{(7,1)}$, $0.044^{(8,1)}$	0.82
	0.50	0.046 (0.037, 0.056)	$0.023^{(2,1)}$, $0.021^{(3,1)}$, $0.025^{(4,3)}$ $0.034^{(5,2)}$, $0.027^{(7,1)}$, $0.040^{(8,1)}$	0.86
0.10	0.10	0.050 (0.041, 0.061)	$0.019^{(2,1)}$, $0.017^{(3,1)}$, $0.027^{(4,2)}$ $0.030^{(5,2)}$, $0.034^{(7,1)}$, $0.048^{(8,1)}$	0.66
	0.20	0.042 (0.034, 0.052)	$0.013^{(2,1)}$, $0.018^{(3,1)}$, $0.029^{(4,3)}$ $0.030^{(5,2)}$, $0.032^{(7,1)}$, $0.040^{(8,1)}$	0.76
	0.30	0.046 (0.038, 0.057)	$0.015^{(2,1)}$, $0.018^{(3,1)}$, $0.029^{(4,3)}$ $0.029^{(5,2)}$, $0.032^{(7,1)}$, $0.045^{(8,1)}$	0.85
	0.40	0.054 (0.044, 0.064)	$0.015^{(2,1)}$, $0.016^{(3,1)}$, $0.023^{(4,3)}$ $0.030^{(5,2)}$, $0.036^{(7,1)}$, $0.049^{(8,1)}$	0.88
	0.50	0.044 (0.035, 0.054)	$0.013^{(2,1)}$, $0.018^{(3,1)}$, $0.025^{(4,3)}$ $0.031^{(5,2)}$, $0.037^{(7,1)}$, $0.042^{(8,1)}$	0.88
0.20	0.10	0.044 (0.036, 0.054)	$0.020^{(2,1)}$, $0.017^{(3,1)}$, $0.023^{(4,4)}$, $0.031^{(5,2)}$ $0.033^{(6,2)}$, $0.033^{(7,1)}$, $0.044^{(8,1)}$	0.82
	0.20	0.043 (0.035, 0.053)	$0.016^{(2,1)}$, $0.020^{(3,1)}$, $0.024^{(4,4)}$, $0.027^{(5,2)}$ $0.034^{(6,2)}$, $0.035^{(7,1)}$, $0.040^{(8,1)}$	0.86
	0.30	0.046 (0.037, 0.056)	$0.014^{(2,1)}$, $0.023^{(3,1)}$, $0.025^{(4,4)}$, $0.029^{(5,2)}$ $0.035^{(6,2)}$, $0.038^{(7,1)}$, $0.044^{(8,1)}$	0.79
	0.40	0.042 (0.034, 0.052)	$0.016^{(2,1)}$, $0.020^{(3,1)}$, $0.025^{(4,4)}$, $0.033^{(5,2)}$ $0.035^{(6,2)}$, $0.031^{(7,1)}$, $0.041^{(8,1)}$	0.79
	0.50	0.051 (0.042, 0.062)	$0.013^{(2,1)}$, $0.018^{(3,1)}$, $0.024^{(4,4)}$, $0.028^{(5,2)}$ $0.038^{(6,2)}$, $0.034^{(7,1)}$, $0.044^{(8,1)}$	0.86
0.40	0.10	0.039 (0.031, 0.048)	$0.013^{(2,1)}$, $0.021^{(3,1)}$, $0.025^{(4,4)}$, $0.032^{(5,2)}$ $0.029^{(6,2)}$, $0.035^{(7,1)}$, $0.039^{(8,1)}$	0.81
	0.20	0.049 (0.040, 0.059)	$0.013^{(2,1)}$, $0.020^{(3,1)}$, $0.024^{(4,4)}$, $0.033^{(5,2)}$ $0.034^{(6,2)}$, $0.040^{(7,1)}$, $0.047^{(8,1)}$	0.97
	0.30	0.042 (0.033, 0.051)	$0.012^{(2,1)}$, $0.022^{(3,1)}$, $0.022^{(4,4)}$, $0.032^{(5,2)}$ $0.034^{(6,2)}$, $0.033^{(7,1)}$, $0.041^{(8,1)}$	0.98
	0.40	0.041 (0.033, 0.051)	$0.012^{(2,1)}$, $0.022^{(3,1)}$, $0.024^{(4,4)}$, $0.028^{(5,2)}$ $0.031^{(6,2)}$, $0.037^{(7,1)}$, $0.034^{(8,1)}$	1.00
	0.50	0.041 (0.033, 0.051)	$0.014^{(2,1)}$, $0.017^{(3,1)}$, $0.024^{(4,4)}$, $0.029^{(5,2)}$ $0.033^{(6,2)}$, $0.035^{(7,1)}$, $0.039^{(8,1)}$	1.00

†Maximum estimated FWER and average power are computed over all selected pairs of ordered triples $(\delta_1, \delta_2, \delta_3)$ considered over all $\tilde{\mathcal{H}}_0^{\mathrm{P}} \in \tilde{\mathcal{H}}_0^{\mathrm{P}}(\pi_G, \pi_E)$; Average = Arithmetic mean; 95% confidence intervals (CI) for α_{F} are exact Clopper-Pearson; Assumed FWER is 5% ($\tilde{\alpha}_{\mathrm{F}} = 0.05$).

$^{(\mathrm{a})}$Average of $\hat{\alpha}_{\mathrm{F}}$ is the arithmetic mean of the $\hat{\alpha}_{\mathrm{F}}$ over all selected pairs of ordered triples $(\delta_1, \delta_2, \delta_3)$ from $\Delta(\pi_G, \pi_E)$ yielding a partial null hypothesis $\tilde{\mathcal{H}}_0^{\mathrm{P}}$, such that $\tilde{\mathcal{H}}_0^{\mathrm{P}}$ is some partial null hypothesis for which the depicted first superscript element denotes the value of $n\left(\tilde{\mathcal{H}}_0^{\mathrm{P}}\right)$ – $n(\cdot)$ denotes the cardinality of the set $(\cdot)$. The second superscript element denotes the total number of selected pairs of ordered triples $(\delta_1, \delta_2, \delta_3)$ from $\Delta(\pi_G, \pi_E)$ yielding a partial null hypothesis $\tilde{\mathcal{H}}_0^{\mathrm{P}}$, comprised of a total of $n\left(\tilde{\mathcal{H}}_0^{\mathrm{P}}\right)$ true null hypotheses.

Table 4.12: Estimated Family-wise Type I Error Rate for GEM under Partial Null Hypotheses over the Simulation Studies of §4.7 and §4.8.

Simulation Study of §4.7†					
		π_G			
Effect of Interest	$l : H_0^{(l)} \in \tilde{\mathcal{H}}_0^{\mathrm{p}}$	0.05	0.10	0.20	0.40
Main Genetic (DOM Genetic Model)	$\{11\}$	0.013	0.010	0.008	0.007
		(0.021)	(0.016)	(0.013)	(0.014)
Main Genetic (REC Genetic Model)	$\{11\}$	0.013	0.010	0.008	0.007
		(0.021)	(0.016)	(0.017)	(0.016)
Main Environment	$\{9, 10\}$	0.013	0.018	0.015	0.014
		(0.026)	(0.031)	(0.031)	(0.033)
Simulation Study of §4.8‡					
		π_G			
β_e	$l : H_0^{(l)} \in \tilde{\mathcal{H}}_0^{\mathrm{p}}$	0.05	0.10	0.20	0.40
0	$\{1, 2, 9, 10\}$	0.032	0.033	0.027	–
		(0.053)	(0.050)	(0.050)	–
-0.10	$\{9, 10\}$	0.014	0.017	0.014	–
		(0.028)	(0.029)	(0.024)	–
-0.25	$\{9, 10\}$	0.012	0.017	0.014	–
		(0.018)	(0.034)	(0.036)	–

†For the SRG simulation outlined within the final paragraph of §4.7.4; Depicted values represent the arithmetic mean of $\hat{\alpha}_{\mathrm{F}}$ (parenthetic values are $\max\{\hat{\alpha}_{\mathrm{F}}\}$) over the $D = 2\mathrm{K}$ simulated data sets across the 99 (main genetic effect)/90 (main environment effect) selected values of π_E over the collection $\{0.01k\}_{k=0,\ldots,99}$; DOM = Dominant; REC = Recessive; Assumed FWER is 5% ($\tilde{\alpha}_{\mathrm{F}} = 0.05$).

‡Depicted values represent the arithmetic mean of $\hat{\alpha}_{\mathrm{F}}$ (parenthetic values are $\max\{\hat{\alpha}_{\mathrm{F}}\}$) over the $D = 10\mathrm{K}$ simulated data sets across the 51 selected values of γ_{ge} over the collection $\{-2\beta_e + 0.02k\}_{k=0,\ldots,50}$; note that $\tilde{\mathcal{H}}_0^{\mathrm{p}} = \mathcal{H}_0$ for the simulation in which $\gamma_{ge} = \beta_e = 0$; Assumed FWER is 5% ($\tilde{\alpha}_{\mathrm{F}} = 0.05$).

– for details, see the article [176]. Here, we examined the association between genetic variation within these candidate genes and risk of incidence for colon and rectal cancer. Furthermore, given their respective associations with ROS [176], we evaluated the associations upon each of the respective two binary environmental factors, recent use of aspirin (NSAIDs) (yes/no) and recent consumption of cigarettes (yes/no), with risk of incidence for colon and rectal cancer. Our interest lied not solely upon the main effects of these candidate genes and lifestyle (environmental) factors, but more importantly on their synergistic effect towards the risk of cancer.

Twenty-nine tSNPs (see footnote 3 within Chapter 1 for definition) were selected and genotyped upon the four candidate genes as follows: eight markers for EPX; two markers for MPO; four markers for *HIF1A*; and fifteen markers for *NOS2A*. Subjects missing genotype data at a particular SNP locus were excluded from the analysis for that marker. Analysis for GxE interaction was based upon these selected tSNP markers and the aforementioned binary environmental factors, and assessed using two competing approaches: by way of our novel GEM method; and – as a benchmarking tool

for GEM – by way of the methodology outlined within the following paragraph.

If G_M denotes the genetic coding for each genotype at a SNP locus, such that the label M identifies the genetic model of inheritance (GMI), we considered a model of the form

$$\text{logit}\left(\Pr\left(Y=1|G_j,E,M\right)\right)=\beta_0+\beta_{GM}G_M+\beta_{EM}E+\gamma_M G_M E, \tag{4.49}$$

where the GMI was assumed to follow either the dominant (DOM) or recessive (REC) genetic models, and

$$G_M = I\left(G_j=2\right)+I\left(G_j=1, M=\text{DOM}\right).$$

Given M, a standard approach to test for GxE interaction would be to perform a 1-df test of the null hypothesis $H_0:\gamma_M=0$ – versus the two-sided alternative hypothesis – based upon the model (4.49), for each of the 29 tSNP markers comprising the candidate pathway. We assumed M was not known, and thus we conducted an LRT under each of the dominant and recessive GMIs upon each of the 29 SNP loci. The largest of the two LRT statistics for each SNP locus was used to assess GxE interaction at the locus. For clarity, we denote this competing approach in assessing GxE interaction as the LRT of interaction (LRTI).

Correction for multiple hypothesis testing was carried out at the gene level (i.e., taking into account all hypothesis tests conducted upon the tSNPs within a given gene), by permutation for GEM and by the pACT approach [20] for LRTI. We performed the MHT correction at the gene level, because we conjecture that SNPs exhibiting GxE interaction within a gene lends to increasing likelihood of the gene itself exhibiting GxE interaction. For the permutation approach of GEM we assigned the value $R=100\text{K}$, and estimated the respective maxT adjusted p-values (4.12) and (4.14) by (4.18) and (4.19) using Algorithm 4.1. For the pACT approach, we assigned each value within the genotype vectors to their appropriate predictor in G_ME of model (4.49), and we assigned each value within the covariate vectors to their appropriate predictor in E or G_M of model (4.49) (see pg. 1160 of [20] for definitions of the terms genotype vector and covariate vector). Note that this procedure for pACT was deliberately carried out, because here we were particularly interested in testing for GxE interaction of the model (4.49), accounting for the marginal effects in E and G_M, but were not directly interested in testing for an association between Y and G_M of said model – accounting for the effects of E and G_ME – for which the pACT method is based upon. Statistical inference was conducted within the `R` (version 2.13.1; July 2011) statistical software environment [103]. For GEM,

we compiled the C code corresponding with our proposed *GEM* `R` package (see Appendix C) to a dynamic link library (DLL) and interfaced – by way of the code given within §D.2 of Appendix D – the resulting DLL with `R`, and for pACT we sourced the `p_ACT_seq.R`[11] file within `R`. Statistical significance was set at the 5% level in the FWER prior to conducting analysis.

4.10.2 Results

A summary of the 29 tSNPs for the genes *EPX*, *MPO*, *HIF1A*, and *NOS2A* is provided within Table 4.13. After multiple testing correction, by way of the false discovery rate (FDR) method of [61] using the `multtest` package of `R` [103], none of the allele frequencies upon these markers statistically significantly deviated from Hardy-Weinberg equilibrium (FDR adjusted $p \geq 0.9$ for colon cancer; FDR adjusted $p \geq 0.16$ for rectal cancer). Here, unless otherwise specified, the value of j indexes the SNP ID column of this table. After multiplicity correction by way of GEM (at the SNP locus level), no marginal genetic effects were found to be statistically significantly associated with either cancer ($\tilde{p}^*_{j9\sigma}, \tilde{p}^*_{j\{10\}\sigma} \geq 0.21$ for colon, $\tilde{p}^*_{j9\sigma}, \tilde{p}^*_{j\{10\}\sigma} \geq 0.12$ for rectal). Recent NSAID use was found to exhibit a statistically significant protective effect for each cancer (colon: OR 0.65; 95% CI for OR (0.57,0.75); $\tilde{p}^*_{j\{11\}\sigma} < 0.0001$ for all $j = 1, \ldots, 29$, rectal: OR 0.69; 95% CI for OR (0.57,0.84); $\tilde{p}^*_{j\{11\}\sigma} < 0.0001$ for all $j = 1, \ldots, 29$)[12], where 31.5% of the cases and 41.4% of the controls within the colon cancer study were recent NSAID users, and these respective proportions were 36.2% and 45.1% within the rectal cancer study. On the other hand, after multiplicity correction by way of GEM, recent cigarette consumption was not statistically significantly associated with either cancer (colon: OR 1.20; 95% CI for OR (1.01,1.42); $\tilde{p}^*_{j\{11\}\sigma} \geq 0.14$, rectal: OR 1.33; 95% CI for OR (1.03,1.70); $\tilde{p}^*_{j\{11\}\sigma} \geq 0.10$), where 20.5% of the cases and 17.7% of the controls within the colon cancer study were recent cigarette consumers, and these respective proportions were 19.7% and 15.6% within the rectal cancer study.

Table 4.14 summarizes the statistically significant GxE interactions between recent NSAID use and each of the 29 genetic markers in their respective effect towards risk of colon cancer, after multiplicity correction at the gene level by way of GEM ($\tilde{p}^*_{jl\mu} < 0.001$ for some $l \leq 8$, all $j = 1, \ldots, 29$). While GEM was able to detect statistically significant GxE interaction upon all 29 SNP markers, prior to application of the pACT MHT correction the genotypes upon each of two markers exhibited a statistically significant interaction with recent NSAID use in their respective effect with

[11] pACT version 1.2, retrieved November 11, 2011 from `http://csg.sph.umich.edu/boehnke/p_act/p_ACT_1.2/p_act.html`.

[12] Unless otherwise specified, reported confidence intervals are unadjusted for multiple hypothesis testing.

colon cancer risk ($p = 0.04$ SNP ID 6 and $p = 0.05$ SNP ID 3) using the LRTI approach. However, the results for each of these markers under LRTI can be attributed to a chance finding amongst the 58 tests (two tests for each marker) of the null hypothesis of no GxE interaction conducted upon this approach for these 29 markers. In fact, the pACT MHT corrected LRTI approach failed to identify any statistically significant GxE interactions amongst these markers (pACT adjusted $p > 0.2$).

Table 4.13: Profiles of the 29 TagSNPs Studied upon the Genes *EPX*, *MPO*, *HIF1A*, and *NOS2A*[†].

Gene	Chromosome Location	Genome SNP ID	SNP Index	Major/Minor Allele	MAF Colon/Rectal Cancer
EPX	17q23.1	rs12602891	1	T/C	0.45/0.42
		rs11079339	2	A/G	0.15/0.14
		rs10853004	3	A/G	0.32/0.29
		rs2240815	4	A/G	0.47/0.46
		rs12602498	5	A/G	0.33/0.32
		rs9892223	6	A/G	0.46/0.44
		rs8077426	7	G/A	0.22/0.20
		rs2302313	8	G/A	0.10/0.11
HIF1A	14q21-q24	rs1951795	9	C/A	0.20/0.21
		rs2301113	10	A/C	0.24/0.25
		rs11549465	11	C/T	0.10/0.10
		rs6573399	12	G/T	0.15/0.15
MPO	17q23.1	rs2759	13	A/G	0.03/0.03
		rs2243828	14	A/G	0.20/0.23
NOS2A	17q11.2-q12	rs7406657	15	G/C	0.24/0.26
		rs9906835	16	A/G	0.39/0.41
		rs2297518	17	G/A	0.20/0.19
		rs2274894	18	G/T	0.39/0.38
		rs2314810	19	G/C	0.05/0.06
		rs4795067	20	A/G	0.34/0.34
		rs3729508	21	G/A	0.41/0.39
		rs3730017	22	C/T	0.03/0.03
		rs944725	23	C/T	0.39/0.40
		rs3794763	24	G/A	0.22/0.23
		rs8072199	25	C/T	0.45/0.43
		rs16949	26	T/C	0.24/0.23
		rs3730013	27	C/T	0.32/0.33
		rs10459953	28	G/C	0.36/0.36
		rs2779248	29	T/C	0.38/0.37

[†]Chromosome location, genome SNP identifiers, and nucleotide base-pair coding for major/minor alleles (A, C, G, T) taken from [176]; Minor allele frequencies (MAF) calculated – at their maximum likelihood estimates, in coherence with Proposition A.7 – from sample controls upon the colon/rectal cancer data sets, assuming population Hardy-Weinberg equilibrium.

After applying GEM, nine markers demonstrated their respective strongest association with risk of colon cancer upon candidate pattern $L_{A_3} = (G_j \in \{0, 1\}) \wedge (E = 0)$ (see Table 4.1 for a summary of candidate pattern formulation), while the remaining 20 markers demonstrated strongest association

with risk of colon cancer upon candidate pattern $L_{A_4} = (G_j \in \{0,1\}) \wedge (E = 1)$. To obtain a better understanding of these GxE interactions, we created numerous graphs. Figures 4.10 and 4.11 portray the relationships between genotypes and risk of colon cancer, stratified by the status of recent NSAID use, for the corresponding SNPs having demonstrated strongest association with risk of colon cancer upon the respective candidate patterns L_{A_3} and L_{A_4}.

Table 4.14: Statistically Significant Interactions Between Recent Use of NSAIDs and the Genes *EPX*, *MPO*, *HIF1A*, and *NOS2A* in Their Effect Towards Risk of Colon Cancer, at the 5% FWER Level as Determined by GEM[†].

Gene	SNP ID	*P*-value LRTI Raw (GMI)	*P*-value LRTI pACT	*P*-value GEM $\min_{1\le l\le 8}\{\tilde{p}^*_{jl\sigma}\}$ (l)	*P*-value GEM $\tilde{p}^*_{jl\mu}$	Odds Ratio (95% CI)
EPX	1	0.18 (REC)	0.84	< 0.001 (4)	< 0.001	0.66 (0.56, 0.76)
	2	0.63 (DOM)	1.00	< 0.001 (3)	< 0.001	1.54 (1.33, 1.77)
	3	0.05 (DOM)	0.45	< 0.001 (3)	< 0.001	1.54 (1.34, 1.76)
	4	0.18 (DOM)	0.83	< 0.001 (4)	< 0.001	0.66 (0.56, 0.77)
	5	0.57 (REC)	1.00	< 0.001 (4)	< 0.001	0.66 (0.57, 0.77)
	6	0.04 (REC)	0.35	< 0.001 (4)	< 0.001	0.63 (0.54, 0.73)
	7	0.07 (DOM)	0.50	< 0.001 (3)	< 0.001	1.53 (1.33, 1.76)
	8	0.49 (REC)	0.99	< 0.001 (3)	< 0.001	1.53 (1.33, 1.77)
HIF1A	9	0.40 (REC)	0.92	< 0.001 (4)	< 0.001	0.64 (0.55, 0.74)
	10	0.60 (REC)	1.00	< 0.001 (4)	< 0.001	0.64 (0.55, 0.74)
	11	0.71 (DOM)	1.00	< 0.001 (4)	< 0.001	0.65 (0.57, 0.75)
	12	0.58 (REC)	1.00	< 0.001 (3)	< 0.001	1.51 (1.31, 1.74)
MPO	13	0.07 (REC)	0.24	< 0.001 (4)	< 0.001	0.65 (0.56, 0.75)
	14	0.25 (REC)	0.53	< 0.001 (4)	< 0.001	0.66 (0.57, 0.77)
NOS2A	15	0.06 (DOM)	0.75	< 0.001 (4)	< 0.001	0.66 (0.57, 0.76)
	16	0.09 (DOM)	0.86	< 0.001 (4)	< 0.001	0.67 (0.58, 0.78)
	17	0.59 (REC)	1.00	< 0.001 (3)	< 0.001	1.51 (1.31, 1.74)
	18	0.19 (REC)	0.97	< 0.001 (4)	< 0.001	0.63 (0.54, 0.74)
	19	0.27 (REC)	0.99	< 0.001 (4)	< 0.001	0.65 (0.56, 0.75)
	20	0.12 (DOM)	0.90	< 0.001 (4)	< 0.001	0.67 (0.58, 0.78)
	21	0.71 (DOM)	1.00	< 0.001 (4)	< 0.001	0.67 (0.57, 0.77)
	22	0.19 (REC)	0.97	< 0.001 (4)	< 0.001	0.65 (0.57, 0.75)
	23	0.62 (REC)	1.00	< 0.001 (4)	< 0.001	0.69 (0.59, 0.80)
	24	0.15 (REC)	0.94	< 0.001 (3)	< 0.001	1.48 (1.29, 1.71)
	25	0.11 (REC)	0.89	< 0.001 (4)	< 0.001	0.65 (0.56, 0.76)
	26	0.12 (DOM)	0.89	< 0.001 (3)	< 0.001	1.52 (1.32, 1.75)
	27	0.49 (DOM)	1.00	< 0.001 (4)	< 0.001	0.67 (0.58, 0.77)
	28	0.68 (DOM)	1.00	< 0.001 (4)	< 0.001	0.66 (0.57, 0.77)
	29	0.12 (DOM)	0.89	< 0.001 (3)	< 0.001	1.48 (1.29, 1.70)

[†]Odds ratio = Odds of colon cancer for individuals over candidate pattern L_{A_l}, compared to that for individuals over candidate pattern L_{B_l}; Raw p-value for LRTI is unadjusted for MHT, where GMI is the genetic model yielding the largest LRT statistic for this approach; 95% Fisher's exact-based confidence intervals are uncorrected for multiple comparisons.

These plots indicate that recent NSAID use may be an effect modifier for the relationship between risk of colon cancer and genotype for many of the tSNP loci, as seen by the deviation in parallelism

of the lines connecting adjacent genotype groups (e.g., the line connecting genotypes AA and Aa) across the two strata in recent NSAID use. For example, consider the plot within the lower left panel of the former figure (corresponds with SNP rs3794763). Now, recent NSAID use will not be an effect modifier for the relationship between risk of colon cancer and genotype at this locus, provided that the odds ratio of colon cancer – based upon the comparison of any two genotype groups – is the same across the strata of NSAID use. Here, among non-recent NSAID users, these data indicate that the odds of colon cancer among subjects with genotype Aa (heterozygote) is 1.1 times that of subjects with genotype AA (homozygote wildtype), and this odds ratio is 0.9 comparing subjects with genotype aa (homozygote variant) to subjects with genotype Aa – here, A/a is used to denote the respective major/minor allele at a particular SNP locus. Conversely, among recent NSAID users, these odds ratios are 1.0 and 0.6, respectively. Since these genotype odds ratios appear to depend upon the status of recent NSAID use, said environmental factor may be an effect modifier for the relationship between risk of colon cancer and genotype at this SNP locus.

Moreover, there is evidence of cross-interaction at play upon some loci, as seen by opposite genotype-phenotype effects between the two NSAID strata across genotype levels. For example, SNPs rs10853004 (the plot within the panel of the first row and second column upon the former figure), rs2759 (the plot within the panel of the second row and third column upon the latter figure), and rs12602891 (the plot within the panel of the first row and column upon the latter figure). To illustrate, we consider the first of these markers. Among non-recent NSAID users, the odds of colon cancer among subjects with genotype Aa is 0.89 times that of subjects with genotype AA, this odds ratio is 0.81 comparing subjects with genotype aa to subjects with genotype Aa, and this odds ratio is 0.72 comparing subjects with genotype aa to subjects with genotype AA. Conversely, among recent NSAID users, these respective odds ratios are 1.12, 1.04, and 1.16. When comparing the two strata of recent NSAID use, it seems that the within-NSAID strata genotype effects are in opposite directions across the genotype levels at this locus, which suggests cross-interaction is at play. Table 4.15 summarizes the observed cross-interaction pattern of GxE interaction, among the SNP loci determined to exhibit statistically significant GxE interaction at the 5% FWER level by GEM (SNP locus level multiple testing adjustment). In particular, the initial six rows of this table, depict the observed cross-interactions upon assessing GxE interaction between recent NSAID use and genotypes upon six SNP loci in their synergistic effect towards risk of colon cancer. Here, we note that amongst the SNP markers depicted within this table, none exhibit a statistically significant

genetic main effect, as assessed by testing the null hypothesis, $H_0 : \gamma_g = 0$ upon the model

$$\text{logit}\left(\Pr\left(Y = 1|G_j\right)\right) = \gamma_0 + \gamma_g G_j, \tag{4.50}$$

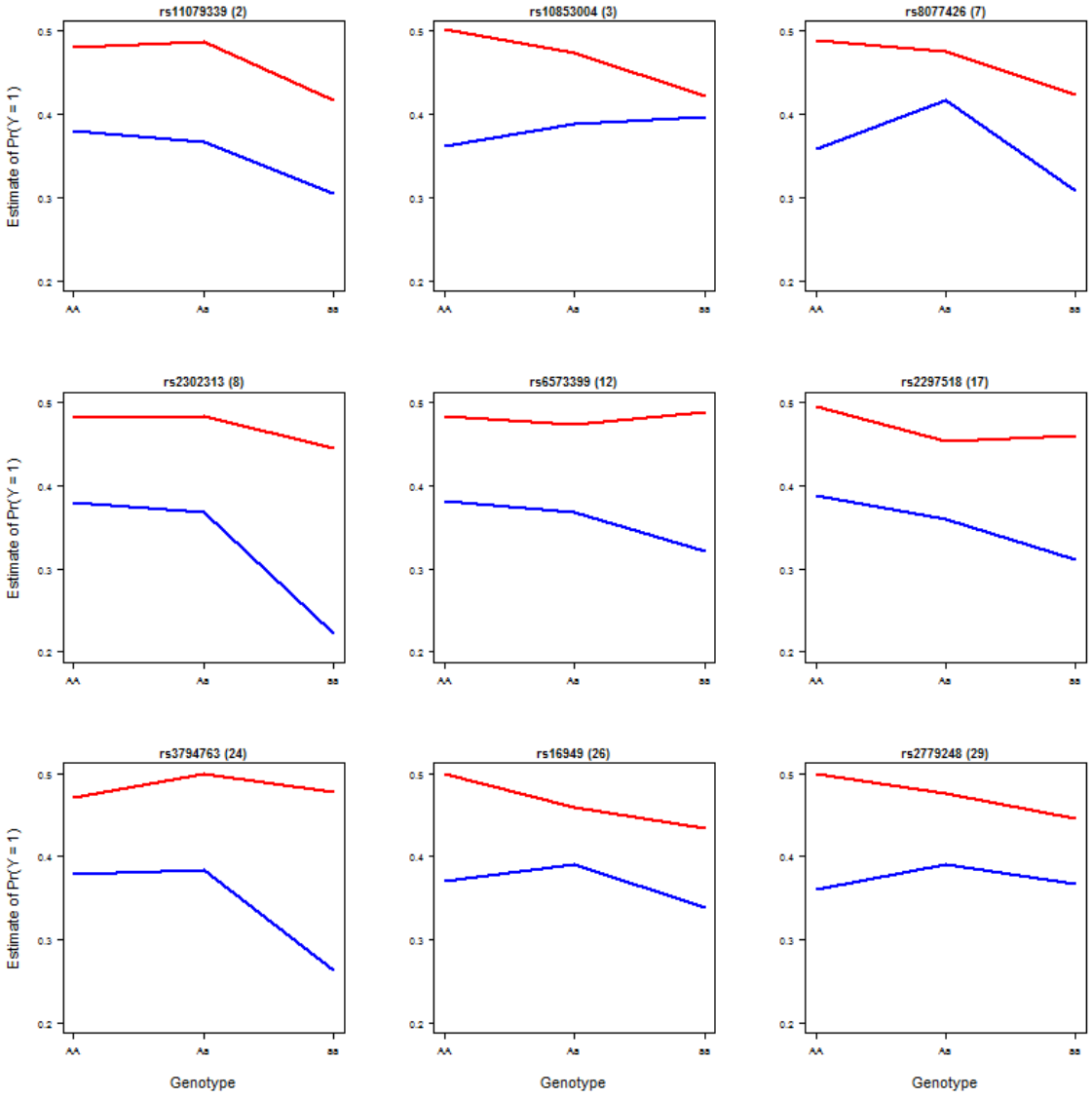

Fig. 4.10: Relationships Between Genotype and Risk of Colon Cancer, Stratified by the Levels of Exposure to Recent NSAID Use, Amongst the 9 SNPs Determined to Possess the Strongest Association Signal Within GEM upon Candidate Pattern $L_{A_3} = (G_j \in \{0, 1\}) \wedge (E = 0)$. Blue Curves Correspond to Recent NSAID Users and Red Curves to Non-Recent NSAID Users. The Genome SNP ID and SNP Index (in Parentheses) Are Shown above Each Plot, for a SNP with Respective Major and Minor Alleles A and a.

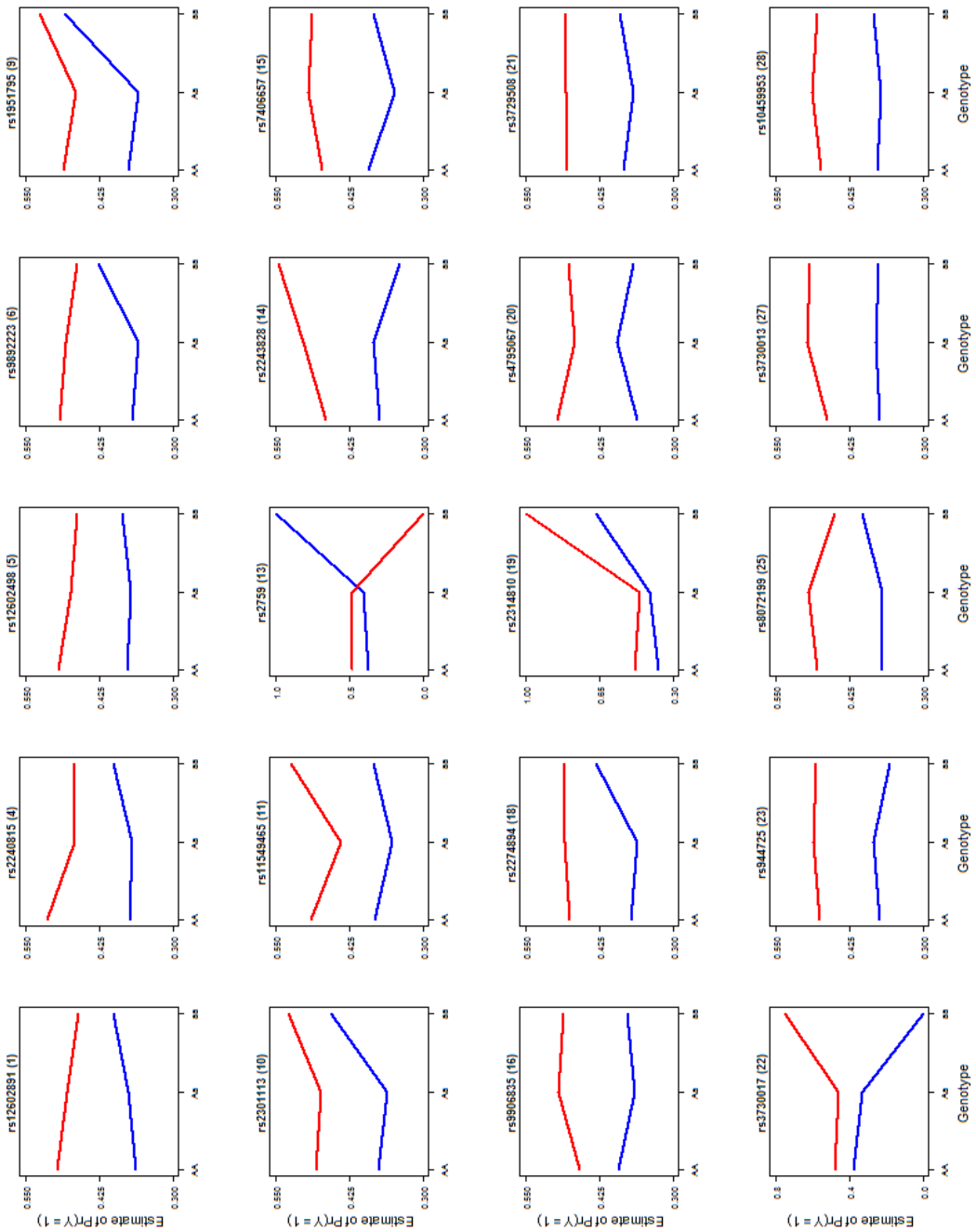

Fig. 4.11: Relationships Between Genotype and Risk of Colon Cancer, Stratified by the Levels of Exposure to Recent NSAID Use, Amongst the 20 SNPs Determined to Possess the Strongest Association Signal Within GEM upon Candidate Pattern $L_{A_4} = (G_j \in \{0,1\}) \wedge (E = 1)$. Blue Curves Correspond to Recent NSAID Users and Red Curves to Non-Recent NSAID Users. The Genome SNP ID and SNP Index (in Parentheses) Are Shown above Each Plot, for a SNP with Respective Major and Minor Alleles A and a.

under each of the dominant and recessive GMIs.[13] Thus, these data support the notion introduced within §4.8. Namely, scanning for solely main genetic effects may fail to detect a genetic-phenotype association upon genetic markers exhibiting cross-interaction; as a consequence in doing this, important risk factors – both genetic and environmental – for disease may be overlooked. Moreover, we point out here that a rather large proportion of the GxE interactions upon these data show the cross-interaction pattern, as 20.7% (6/29) of the GxE interactions between recent NSAID use and the genotypes of these 29 markers show cross-interaction in their effect towards risk of colon cancer.

Figures 4.12 and 4.13 portray the relationships between recent NSAID use status and risk of colon cancer, stratified by genotype, for the corresponding SNPs having demonstrated strongest association with risk of colon cancer upon the respective candidate patterns L_{A_3} and L_{A_4}. These plots indicate that genotype may be an effect modifier for the relationship between recent NSAID status and risk of colon cancer, as seen by the deviation in parallelism among the three lines depicted within each plot of these figures.[14] To illustrate, we again consider SNP locus rs3794763 (the plot within the lower left panel of the former figure). Upon subjects with genotype AA (homozygote wildtype) at this locus, these data indicate that the odds of colon cancer among non-recent NSAID users is 1.45 times that of recent NSAID users; upon subjects with genotype Aa (heterozygote), the odds of colon cancer among non-recent NSAID users is 1.60 times that of recent NSAID users; and, upon subjects with genotype aa (homozygote mutant), the odds of colon cancer among non-recent NSAID users is 2.55 times that of recent NSAID users. Since these odds ratios differ across genotype levels, this suggests that the relationship between recent NSAID use and risk of colon cancer may be modified by genotype at this SNP locus.

Furthermore, through examination of these two figures, we obtain a sense of clarification as to the rationale for GEMs reporting the significant associations upon candidate patterns L_{A_3} and L_{A_4}, in assessing GxE interaction upon these 29 tSNPs with status of recent NSAID use. Namely, comparing any pair of genotypes, the consensus between the risk for colon cancer seems to be maximum upon the two genotypes AA and Aa, for non-recent NSAID users within the former figure and recent NSAID users within the latter figure. Consider SNP rs11079339 (the plot within the

[13] For each SNP locus, the 1-df LRT was conducted upon the model (4.50) under each of the dominant and recessive GMIs. The larger of the two test statistics was selected, and the p-value – under the null hypothesis $H_0 : \gamma_g = 0$ – was computed by referring to the chi-square distribution with 1-df. Correction for multiple testing was performed upon each locus by way of the Bonferroni MTP.

[14] Note that this effect modification can be equivalently seen within Figures 4.10 and 4.11, by comparing the differences in the vertical spread between the two colored line plots across genotype levels within the appropriate plot. If these differences differ across genotype, this suggests genotype to be an effect modifier for the relationship between recent NSAID use and risk of colon cancer.

upper left panel of the former figure) for example. By inspection of this plot, given status in recent NSAID use, it is clear that the consensus in colon cancer risk is maximized comparing genotypes AA and Aa. Indeed, upon non-recent NSAID users, the odds of colon cancer among subjects with genotype Aa is 1.03 times that of subjects with genotype AA; and, among recent NSAID users, this odds ratio is 0.94. Since the former of these two odds ratios is closer in magnitude to the value of

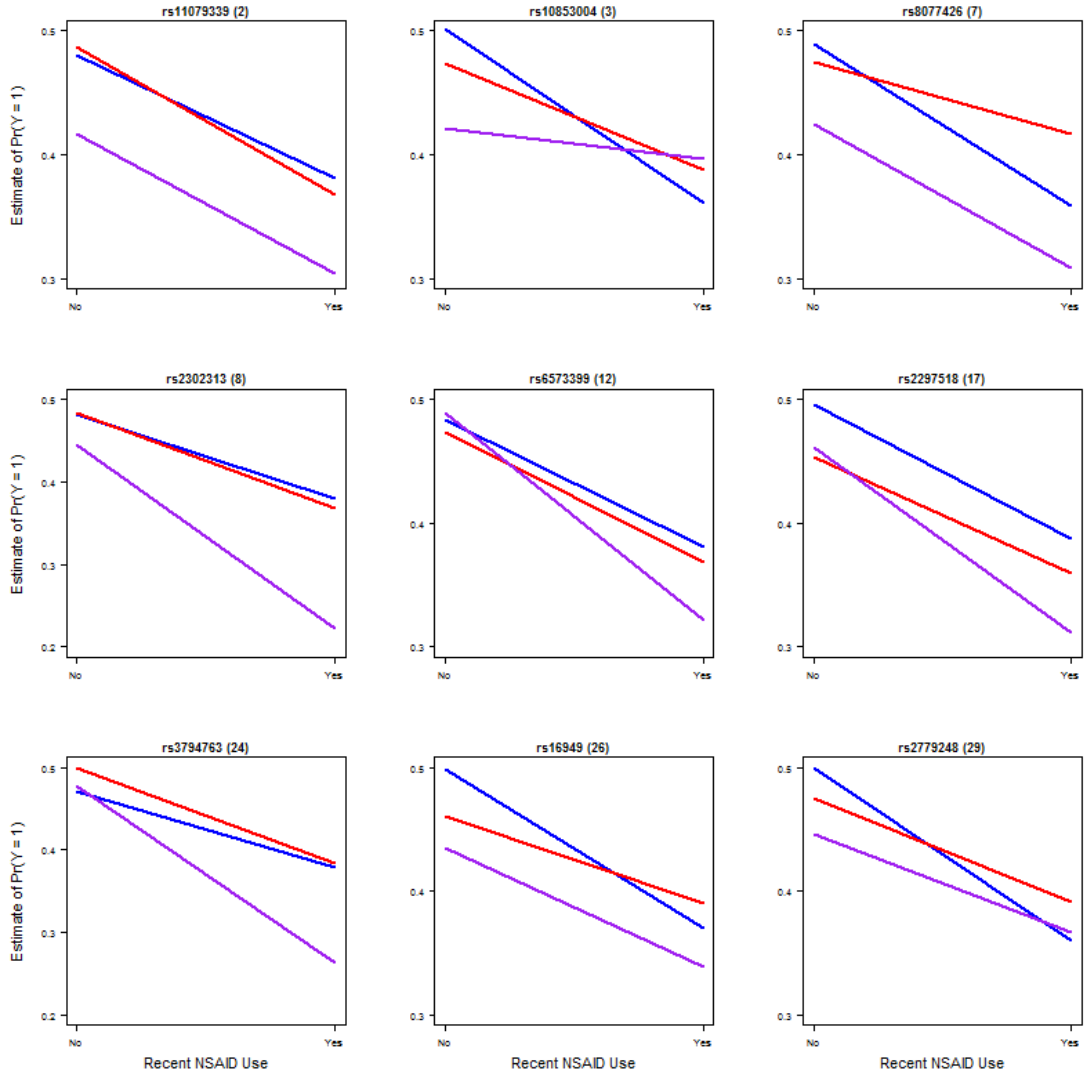

Fig. 4.12: Relationships Between Recent NSAID Use and Risk of Colon Cancer, Stratified by the Levels of Genotype, Amongst the 9 SNPs Determined to Possess the Strongest Association Signal Within GEM upon Candidate Pattern $L_{A_3} = (G_j \in \{0, 1\}) \wedge (E = 0)$. Blue Curves Correspond to Genotype AA ($G_j = 0$), Red Curves to Genotype Aa ($G_j = 1$), and Purple Curves to Genotype aa ($G_j = 2$), for a SNP with Respective Major and Minor Alleles A and a. The Genome SNP ID and SNP Index (in Parentheses) Are Shown above Each Plot.

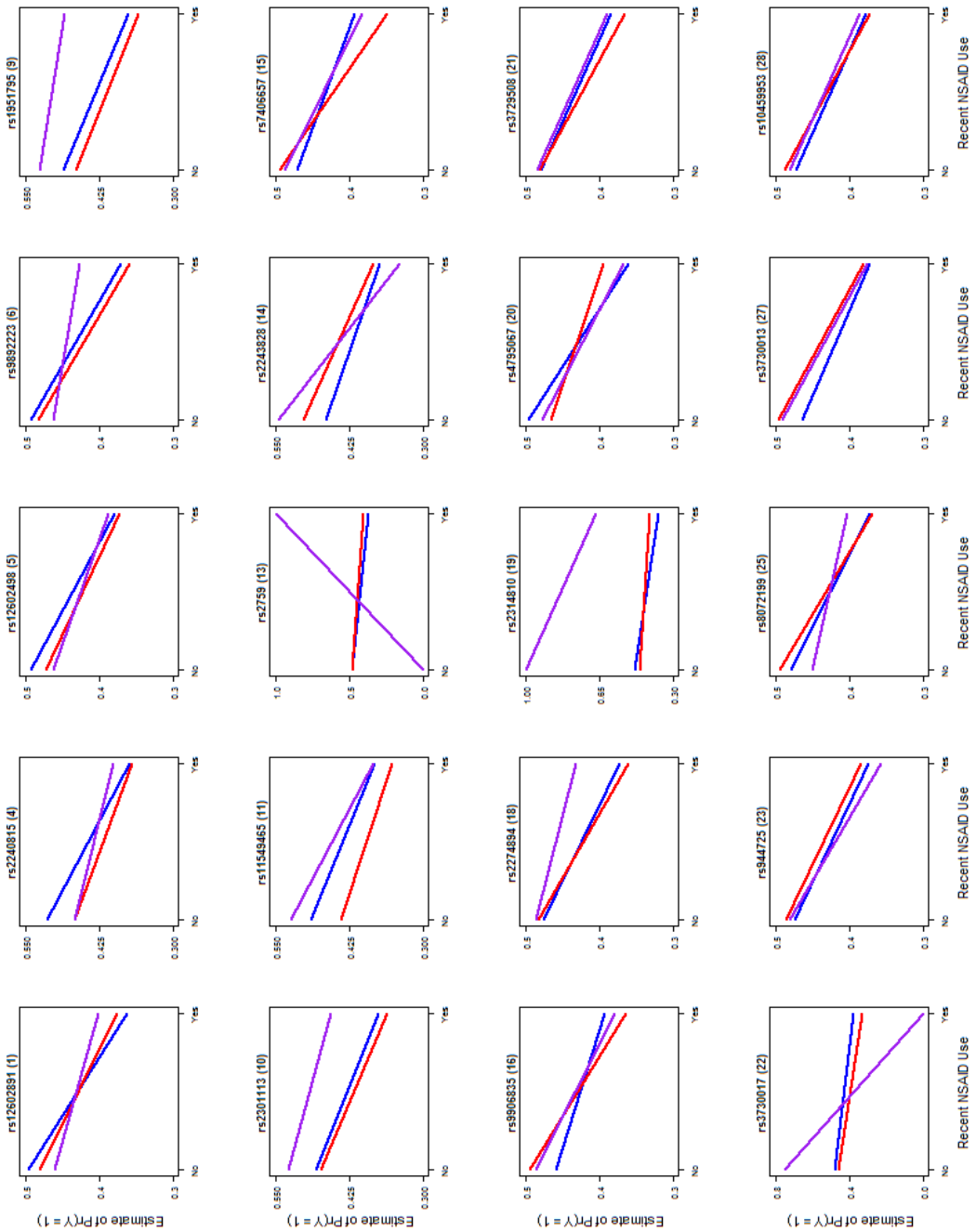

Fig. 4.13: Relationships Between Recent NSAID Use and Risk of Colon Cancer, Stratified by the Levels of Genotype, Amongst the 20 SNPs Determined to Possess the Strongest Association Signal Within GEM upon Candidate Pattern $L_{A_4} = (G_j \in \{0, 1\}) \wedge (E = 1)$. Blue Curves Correspond to Genotype AA ($G_j = 0$), Red Curves to Genotype Aa ($G_j = 1$), and Purple Curves to Genotype aa ($G_j = 2$), for a SNP with Respective Major and Minor Alleles A and a. The Genome SNP ID and SNP Index (in Parentheses) Are Shown above Each Plot.

Table 4.15: Summary of SNP Loci Depicting Cross-Interaction with Recent NSAID Use or Recent Cigarette Consumption in Risk Towards Colon or Rectal Cancer, among Loci Determined to Exhibit Statistically Significant GxE Interaction at the 5% FWER Level by GEM†.

Recent NSAID Use and Colon Cancer				
	SNP	Adjusted P-value	Stratified Genotype Odds Ratios	
Gene	ID	Main Effect in G	$E = 0$	$E = 1$
EPX	1	1.00	(0.94, 0.92, 0.86)	(1.06, 1.11, 1.18)
	3	0.46	(0.89, 0.81, 0.72)	(1.12, 1.04, 1.16)
NOS2A	15	1.00	(1.10, 0.97, 1.07)	(0.82, 1.17, 0.96)
	16	0.87	(1.16, 0.96, 1.11)	(0.88, 1.06, 0.94)
	20	1.00	(0.89, 1.05, 0.93)	(1.15, 0.89, 1.03)
	25	1.00	(1.06, 0.84, 0.89)	(0.99, 1.15, 1.14)
Recent NSAID Use and Rectal Cancer				
	SNP	Adjusted P-value	Stratified Genotype Odds Ratios	
Gene	ID	Main Effect in G	$E = 0$	$E = 1$
EPX	5	1.00	(0.94, 1.14, 1.07)	(1.05, 0.92, 0.97)
NOS2A	17	1.00	(0.98, 0.81, 0.80)	(1.14, 1.15, 1.30)
	22	0.56	(1.02, NC, NC)	(0.75, NC, NC)
	27	1.00	(1.00, 0.87, 0.87)	(1.03, 1.12, 1.16)
Recent Cigarette Consumption and Colon Cancer				
	SNP	Adjusted P-value	Stratified Genotype Odds Ratios	
Gene	ID	Main Effect in G	$E = 0$	$E = 1$
NOS2A	15	1.00	(1.06, 1.19, 1.27)	(0.79, 0.61, 0.48)
	21	1.00	(0.96, 0.93, 0.89)	(1.06, 1.45, 1.54)
	23	1.00	(1.11, 1.04, 1.16)	(0.87, 0.68, 0.59)
	27	0.33	(1.20, 0.95, 1.14)	(0.79, 1.16, 0.92)
	28	1.00	(0.99, 0.92, 0.91)	(1.28, 1.14, 1.47)
Recent Cigarette Consumption and Rectal Cancer				
	SNP	Adjusted P-value	Stratified Genotype Odds Ratios	
Gene	ID	Main Effect in G	$E = 0$	$E = 1$
HIF1A	9	1.00	(1.12, 1.19, 1.33)	(0.60, 0.79, 0.47)
	10	1.00	(1.01, 1.27, 1.28)	(0.68, 0.79, 0.54)
NOS2A	16	0.11	(0.69, 1.17, 0.81)	(1.59, 0.96, 1.53)
	18	0.50	(1.05, 1.29, 1.36)	(0.84, 0.76, 0.64)
	22	0.56	(1.07, NC, NC)	(0.52, NC, NC)
	29	1.00	(1.08, 1.10, 1.19)	(0.65, 0.89, 0.58)

†Statistically significant GxE interaction is based upon the maxT adjusted p-value, $\min\{\tilde{p}^*_{j1\sigma}, \ldots, \tilde{p}^*_{j8\sigma}\}$; Adjusted p-value for assessing the main effect in G is Bonferroni corrected, for having tested the null hypothesis of no genotype-phenotype association upon each of the dominant and recessive GMIs; For a SNP locus with major/minor alleles A/a, the vector of odds ratios = Odds of cancer, (comparing subjects with genotype Aa to subjects with genotype AA, comparing subjects with genotype aa to subjects with genotype Aa, comparing subjects with genotype aa to subjects with genotype AA); NC = not calculable.

1.0 (perfect agreement), it would seem that the consensus in disease risk is maximized within the subjects represented by candidate pattern $L_{A_3} = (G_j \in \{0, 1\}) \wedge (E = 0)$, which is precisely what GEM has suggested to be the circumstance. Although some of the plots within each figure may

appear to contradict the candidate pattern for which GEM has suggested most associated with risk of colon cancer (e.g., candidate pattern $L_{A_4} = (G_j \in \{0,1\}) \wedge (E = 1)$ upon SNP rs3794763 (24) of the former figure seems more appropriate than that of L_{A_3}; candidate pattern L_{A_3} upon SNP rs3729508 (21) of the latter figure seems more appropriate than that of L_{A_4}), the statistical significance governing the tests of null hypotheses over the candidate patterns L_{A_3} and L_{A_4} is roughly the same, across all 29 tSNP loci. This notion, coupled with the affiliated standard errors in estimating the adjusted p-values $\tilde{p}_{jl\mu}$ (4.14) with those of $\tilde{p}^*_{jl\mu}$ (4.19), can explain this phenomenon.

Table 4.16 summarizes the statistically significant GxE interactions between recent NSAID use and genotype at each of the 29 genetic markers in their respective effect towards risk of rectal cancer, after multiplicity correction at the SNP locus level by way of GEM ($\tilde{p}^*_{jl\sigma} \leq 0.02$, some $l \leq 8$, for all $j = 1, \ldots, 29$). Furthermore, after multiple testing correction at the gene level, the genotypes at all but four markers (i.e., rs9906835 ($j = 16$), rs3729508 (21), rs944725 (23), and rs10459953 (28)) statistically significantly interact with recent NSAID use status in their effect towards risk of rectal cancer ($\tilde{p}_{jl\mu} < 0.05$, some $l \leq 8$, for all $j \notin \{16, 21, 23, 28\}$). However, with regard to the aforementioned four markers, whereas these GxE interactions are statistically significant at the SNP level adjustment using GEM ($\tilde{p}^*_{jl\sigma} \leq 0.02$), the corresponding GxE interactions using the LRTI method are not statistically significant ($p \geq 0.17$). Finally, note that the pACT algorithm failed to converge in performing the appropriate multiple testing correction upon the genes *HIF1A* and *NOS2A*, indicating that this MHT approach may not be well suited for tests involving GxE interaction.

After applying GEM, two (2) markers showed their respective strongest association with risk of rectal cancer upon candidate pattern $L_{A_1} = (G_j = 0) \wedge (E = 0)$, one (1) marker upon candidate pattern $L_{A_2} = (G_j = 0) \wedge (E = 1)$, seven (7) markers upon the candidate pattern $L_{A_3} = (G_j \in \{0,1\}) \wedge (E = 0)$, thirteen (13) markers upon candidate pattern $L_{A_4} = (G_j \in \{0,1\}) \wedge (E = 1)$, one marker upon candidate pattern $L_{A_5} = (G_j \in \{1,2\}) \wedge (E = 0)$, four (4) markers upon candidate pattern $L_{A_6} = (G_j \in \{1,2\}) \wedge (E = 1)$, and one marker upon the candidate pattern $L_{A_7} = (G_j = 2) \wedge (E = 0)$. Figures 4.14 and 4.15 display the relationships between recent NSAID use and risk of rectal cancer, stratified by genotype, where the former figure captures the nine SNPs having demonstrated strongest association with risk of rectal cancer upon either of the candidate patterns L_{A_1} and L_{A_3}. These plots indicate that recent NSAID use may be an effect modifier for the relationship between genotype and risk of rectal cancer upon some SNP loci. For example, consider SNP rs16949 (the

plot within the lower right panel of the former figure). Among non-recent NSAID users, the odds of rectal cancer for subjects with genotype AA are 1.00 and 1.32 times those of subjects with respective genotypes Aa and aa; and, the odds of rectal cancer for subjects with genotype Aa is 1.32 times that of subjects with genotype aa. Among recent NSAID users, these odds ratios are 0.91, 1.02, and 1.12, respectively. Insofar as these genotype-phenotype odds ratios appear to differ between exposure status of recent NSAID use, said environmental factor could be an effect modifier for the

Table 4.16: Statistically Significant Interactions Between Recent Use of NSAIDs and the Genes *EPX*, *MPO*, *HIF1A*, and *NOS2A* in Their Effect Towards Risk of Rectal Cancer, at the 5% FWER Level as Determined by GEM†.

Gene	SNP ID	*P*-value LRTI Raw (GMI)	*P*-value LRTI pACT	*P*-value GEM $\min_{1\le l\le 8}\{\tilde{p}^*_{jl\sigma}\}$ (l)	*P*-value GEM $\tilde{p}^*_{jl\mu}$	Odds Ratio (95% CI)
EPX	1	0.05 (REC)	0.44	0.003 (7)	0.015	1.75 (1.27, 2.40)
	2	0.02 (DOM)	0.17	0.004 (4)	0.022	0.70 (0.58, 0.86)
	3	0.30 (DOM)	0.94	0.003 (4)	0.014	0.69 (0.56, 0.85)
	4	0.33 (DOM)	0.94	0.008 (4)	0.037	0.70 (0.57, 0.87)
	5	0.65 (REC)	1.00	0.009 (4)	0.046	0.72 (0.58, 0.88)
	6	0.35 (DOM)	0.95	*0.011* (4)	0.047	0.71 (0.57, 0.88)
	7	0.25 (REC)	0.92	0.003 (4)	0.015	0.70 (0.57, 0.85)
	8	0.48 (DOM)	0.98	< 0.001 (1)	< 0.001	1.54 (1.26, 1.88)
HIF1A	9	0.15 (DOM)	0.54	0.001 (4)	0.005	0.69 (0.56, 0.84)
	10	0.50 (DOM)	0.96	0.002 (4)	0.005	0.69 (0.56, 0.85)
	11	0.01 (REC)	0.07	< 0.001 (3)	0.002	1.48 (1.21, 1.81)
	12	0.62 (DOM)	1.00	0.002 (3)	0.005	1.44 (1.18, 1.76)
MPO	13	0.48 (DOM)	DNC	< 0.001 (1)	0.002	1.44 (1.18, 1.76)
	14	0.09 (REC)	DNC	< 0.001 (3)	0.001	1.49 (1.22, 1.82)
NOS2A	15	0.46 (DOM)	DNC	0.004 (4)	0.036	0.70 (0.57, 0.86)
	16	0.76 (DOM)	DNC	0.008 (6)	0.064	0.69 (0.54, 0.86)
	17	0.39 (DOM)	DNC	0.001 (3)	0.017	1.45 (1.18, 1.77)
	18	0.22 (DOM)	DNC	0.004 (4)	0.039	0.70 (0.57, 0.86)
	19	0.41 (DOM)	DNC	< 0.001 (2)	0.007	0.67 (0.54, 0.82)
	20	0.03 (DOM)	DNC	0.001 (3)	0.011	1.46 (1.19, 1.77)
	21	0.18 (DOM)	DNC	0.006 (5)	0.055	1.41 (1.15, 1.73)
	22	0.42 (DOM)	DNC	0.001 (3)	0.019	1.45 (1.18, 1.77)
	23	0.33 (REC)	DNC	*0.021* (6)	0.165	0.71 (0.56, 0.89)
	24	0.91 (REC)	DNC	0.002 (4)	0.025	0.69 (0.57, 0.85)
	25	0.11 (REC)	DNC	0.002 (6)	0.018	0.66 (0.53, 0.83)
	26	0.57 (DOM)	DNC	0.001 (3)	0.012	1.46 (1.19, 1.78)
	27	0.40 (REC)	DNC	0.003 (4)	0.030	0.69 (0.56, 0.85)
	28	0.17 (REC)	DNC	*0.021* (6)	0.159	0.71 (0.56, 0.89)
	29	0.43 (DOM)	DNC	0.005 (4)	0.046	0.70 (0.57, 0.86)

†Odds ratio = Odds of colon cancer for individuals over candidate pattern L_{A_l}, compared to that for individuals over candidate pattern L_{B_l}; Raw p-value for LRTI is unadjusted for MHT, where GMI is the genetic model yielding the largest LRT statistic for this approach; 95% Fisher's exact-based confidence intervals are uncorrected for multiple comparisons; DNC = pACT algorithm did not converge; Italicized p-values are exact, based upon implementation of Algorithm 4.2.

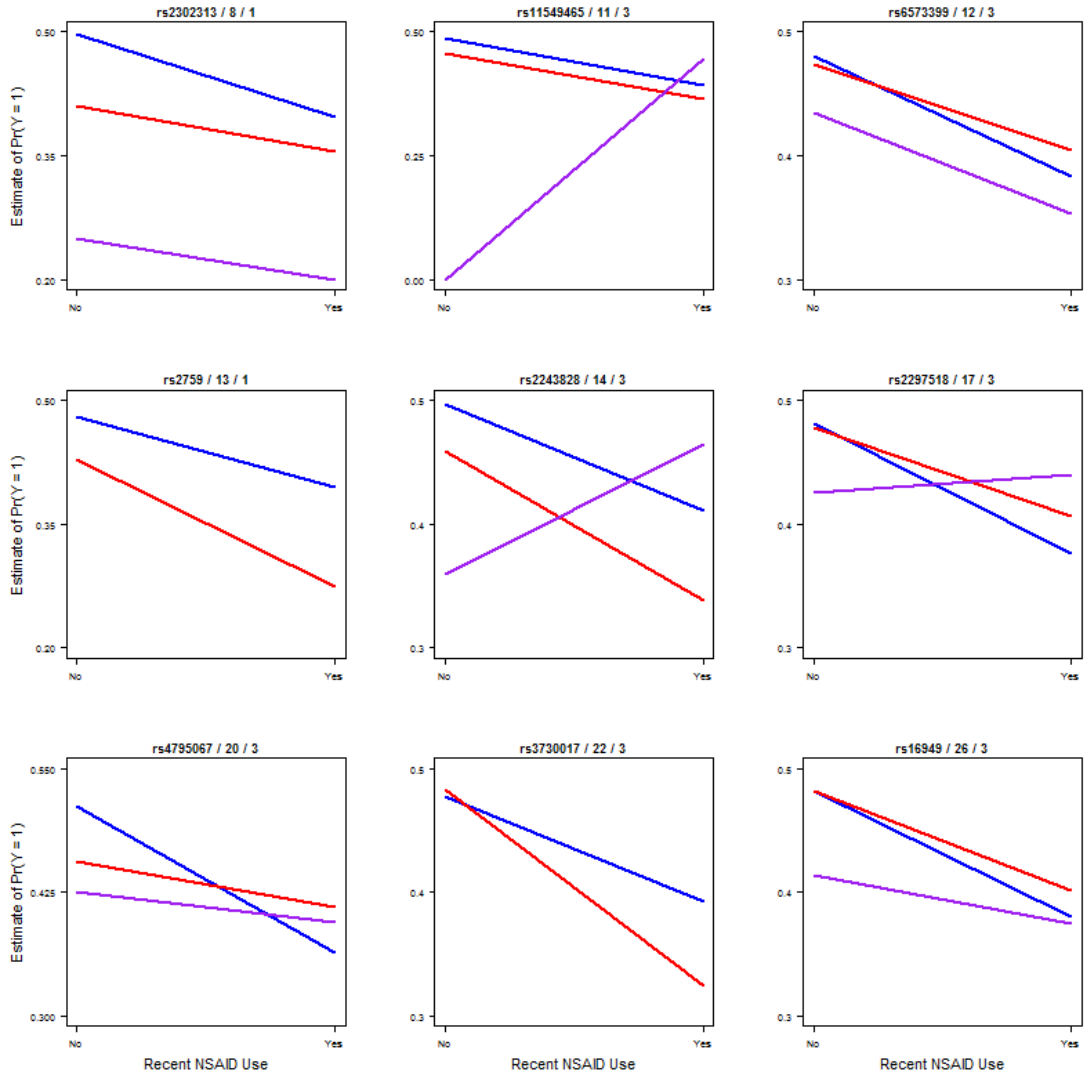

Fig. 4.14: Relationships Between Recent NSAID Use and Risk of Rectal Cancer, Stratified by the Levels of Genotype, Amongst the 9 SNPs Determined to Possess the Strongest Association Signal Within GEM upon Either of the Candidate Patterns $L_{A_1} = (G_j = 0) \wedge (E = 0)$ and $L_{A_3} = (G_j \in \{0, 1\}) \wedge (E = 0)$. Blue Curves Correspond to Genotype AA $(G_j = 0)$, Red Curves to Genotype Aa $(G_j = 1)$, and Purple Curves to Genotype aa $(G_j = 2)$, for a SNP with Respective Major and Minor Alleles A and a. The Genome SNP ID / SNP Index / Candidate Pattern Index Are Shown above Each Plot. The Missing Purple Line Within Each of the Two Appropriate Plots Is due to Data Sparsity.

genotype-phenotype relationship. Furthermore, these plots indicate that genotype may be an effect modifier for the relationship between recent NSAID use and risk of rectal cancer upon some SNP loci. To illustrate, consider again SNP rs16949. Upon subjects with genotype AA at this locus, these data indicate that the odds of rectal cancer among non-recent NSAID users is 1.52 times that of recent NSAID users; upon subjects with genotype Aa at this locus, this odds ratio is 1.39; and,

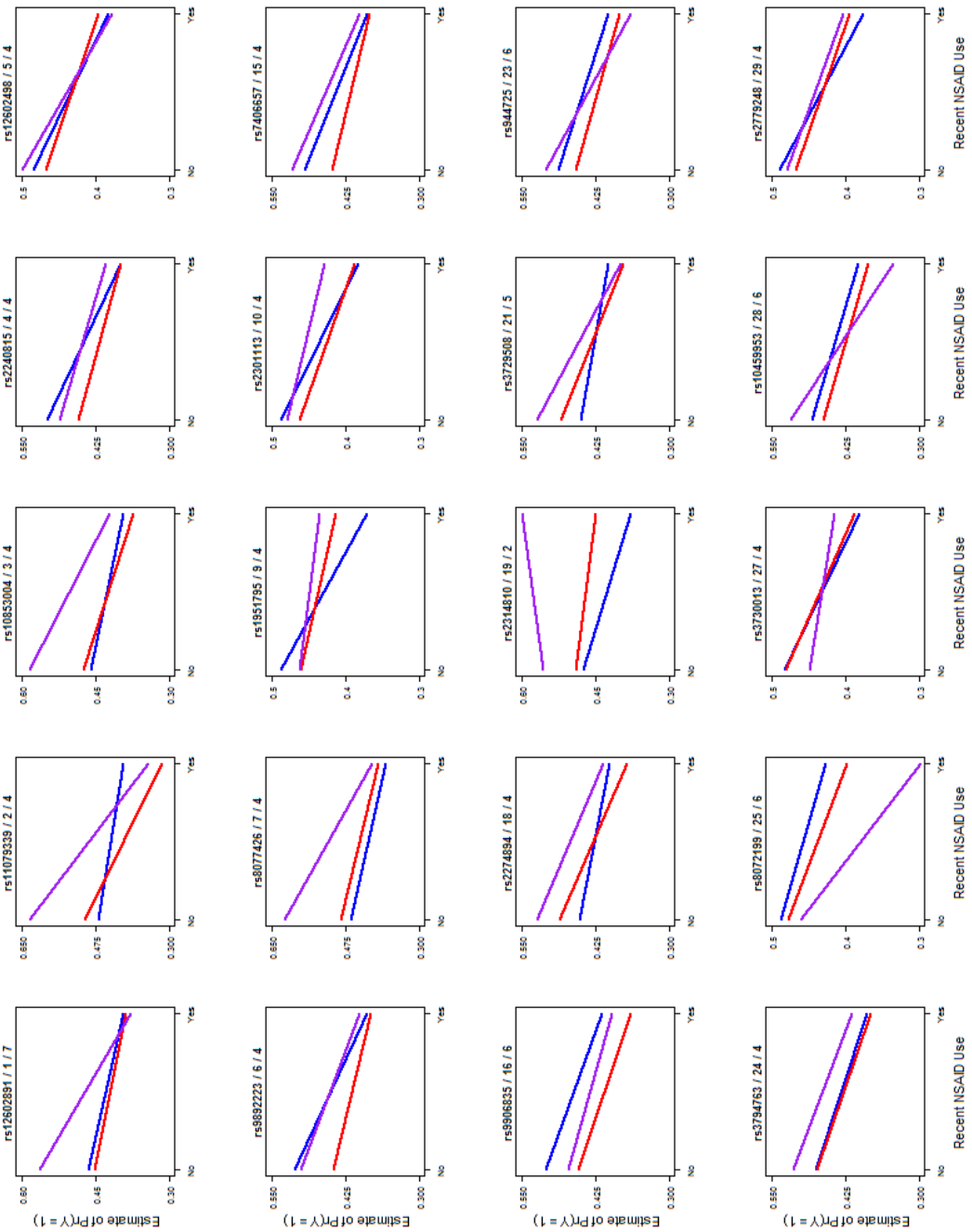

Fig. 4.15: Relationships Between Recent NSAID Use and Risk of Rectal Cancer, Stratified by the Levels of Genotype, Amongst the 20 SNPs Determined to Possess the Strongest Association Signal Within GEM Amongst the Candidate Patterns L_{A_2}, L_{A_4}, L_{A_5}, L_{A_6}, and L_{A_7}. Blue Curves Correspond to Genotype AA ($G_j = 0$), Red Curves to Genotype Aa ($G_j = 1$), and Purple Curves to Genotype aa ($G_j = 2$), for a SNP with Respective Major and Minor Alleles A and a. The Genome SNP ID / SNP Index / Candidate Pattern Index Are Shown above Each Plot.

upon subjects with genotype aa at this locus, this odds ratio is 1.18. Since these odds ratios differ across genotype levels, this suggests that the relationship between recent NSAID use and risk of rectal cancer may be modified by genotype at this SNP locus. Finally, several SNP loci (e.g., SNPs rs2297518 and rs3730017 upon the former figure, and SNP rs12602498 upon the latter figure) demonstrate cross-interaction (see Table 4.15 for complete list of cross-interactions upon these data). To illustrate, consider SNP rs2297518. Among non-recent NSAID users, the odds of rectal cancer for subjects with genotype Aa is 0.98 times that of subjects with genotype AA, this odds ratio is 0.81 comparing subjects with genotype aa to subjects with genotype Aa, and this odds ratio is 0.80 comparing subjects with genotype aa to subjects with genotype AA. Conversely, among recent NSAID users, these odds ratios are 1.14, 1.15, and 1.30, respectively. When comparing the two strata of recent NSAID use, it seems that the within-NSAID strata genotype effects are in opposite directions across the genotype levels at this locus, which suggests cross-interaction is at play. Moreover, we note that 17.2% (5/29) of the GxE interactions for these data show the cross-interaction pattern.

Table 4.17 summarizes the statistically significant GxE interactions between recent cigarette consumption and genotype at each of 10 genetic markers in their respective effect towards risk of colon cancer, after multiplicity correction at the SNP locus level by way of GEM (exact $\tilde{p}_{jl\sigma} \leq 0.054$, for some $l \leq 8$ and $j \in \{10, 14, 15, 17, 18, 21, 23, 26, 27, 28\}$); the statistically significant GxE interactions between recent cigarette consumption and genotype at each of 12 genetic markers in their respective effect towards risk of rectal cancer, after multiplicity correction at the SNP locus level by way of GEM (exact $\tilde{p}_{jl\sigma} \leq 0.052$, for some $l \leq 8$ and $j \in \{3, 8, 9, 10, 11, 12, 14, 16, 18, 22, 25, 29\}$). As with the observations made upon Tables 4.14 and 4.16 for the recent NSAID use environmental factor, here for the most part, these data indicate the statistical power of GEM to exceed the conventional LRTI approach in detecting GxE interaction upon the recent cigarette consumption environmental factor. In fact, for the GxE interactions depicted within the table for rectal cancer risk, whereas each of the 12 markers results in a statistically significant interaction using GEM (adjustment at the SNP locus level), only a handful (i.e., SNPs with ID 9, 10, 14, 18, 29) suggest statistically significant GxE interaction applying the LRTI approach.

After applying GEM, one marker showed its strongest association with risk of colon cancer upon candidate pattern $L_{A_1} = (G_j = 0) \wedge (E = 0)$, three markers upon candidate pattern $L_{A_2} = (G_j = 0) \wedge (E = 1)$, one marker upon the candidate pattern $L_{A_3} = (G_j \in \{0, 1\}) \wedge (E = 0)$, one

marker upon candidate pattern $L_{A_4} = (G_j \in \{0,1\}) \wedge (E = 1)$, three markers upon candidate pattern $L_{A_6} = (G_j \in \{1,2\}) \wedge (E = 1)$, and one marker upon the candidate pattern $L_{A_8} = (G_j = 2) \wedge (E = 1)$. Also, seven markers showed their respective strongest association with risk of rectal cancer upon candidate pattern $L_{A_2} = (G_j = 0) \wedge (E = 1)$, two markers upon candidate pattern $L_{A_3} = (G_j \in \{0,1\}) \wedge (E = 0)$, one marker upon candidate pattern $L_{A_4} = (G_j \in \{0,1\}) \wedge (E = 1)$, and two markers upon candidate pattern $L_{A_5} = (G_j \in \{1,2\}) \wedge (E = 0)$.

Table 4.17: Statistically Significant Interactions Between Recent Consumption of Cigarettes and the Genes EPX, MPO, $HIF1A$, and $NOS2A$ in Their Effect Towards Risk of Colon or Rectal Cancer, at the 5% FWER Level as Determined by GEM†.

Recent Consumption of Cigarettes and Colon Cancer						
Gene	SNP ID	*P*-value LRTI Raw (GMI)	pACT	*P*-value GEM $\min_{1\le l\le 8}\{\tilde{p}_{jl\sigma}\}$ (l)	$\tilde{p}^*_{jl\mu}$	Odds Ratio (95% CI)
HIF1A	10	0.55 (DOM)	1.00	*0.047* (3)	0.110	0.81 (0.69, 0.95)
MPO	14	0.23 (DOM)	DNC	*0.049* (6)	0.067	1.42 (1.09, 1.85)
NOS2A	15	0.03 (DOM)	DNC	*0.040* (2)	0.291	1.35 (1.09, 1.68)
	17	0.15 (DOM)	DNC	*0.019* (2)	0.162	1.36 (1.11, 1.67)
	18	0.02 (DOM)	DNC	*0.026* (6)	0.185	1.36 (1.10, 1.67)
	21	0.03 (REC)	DNC	*0.054* (8)	0.358	1.63 (1.12, 2.37)
	23	0.03 (REC)	DNC	*0.041* (4)	0.287	1.29 (1.08, 1.56)
	26	0.11 (DOM)	DNC	*0.030* (2)	0.231	1.36 (1.10, 1.68)
	27	0.03 (DOM)	DNC	*0.036* (1)	0.260	0.82 (0.71, 0.94)
	28	0.08 (DOM)	DNC	*0.050* (6)	0.328	1.33 (1.08, 1.64)
Recent Consumption of Cigarettes and Rectal Cancer						
Gene	SNP ID	*P*-value LRTI Raw (GMI)	pACT	*P*-value GEM $\min_{1\le l\le 8}\{\tilde{p}_{jl\sigma}\}$ (l)	$\tilde{p}^*_{jl\mu}$	Odds Ratio (95% CI)
EPX	3	0.67 (REC)	1.00	*0.045* (3)	0.192	0.74 (0.59, 0.92)
	8	0.36 (DOM)	0.97	*0.018* (5)	0.114	0.66 (0.50, 0.88)
HIF1A	9	0.01 (DOM)	0.05	*0.009* (2)	0.023	1.67 (1.21, 2.32)
	10	0.06 (DOM)	0.30	*0.030* (2)	0.069	1.62 (1.15, 2.28)
	11	0.34 (DOM)	0.55	*0.050* (2)	0.140	1.44 (1.08, 1.91)
	12	0.08 (DOM)	0.33	*0.043* (2)	0.111	1.49 (1.11, 2.02)
MPO	14	0.03 (DOM)	DNC	*0.003* (2)	0.004	1.76 (1.27, 2.44)
NOS2A	16	0.003 (DOM)	DNC	*0.002* (5)	0.021	0.70 (0.58, 0.85)
	18	0.06 (REC)	DNC	*0.037* (3)	0.256	0.74 (0.60, 0.92)
	22	0.19 (DOM)	DNC	*0.052* (2)	0.501	1.39 (1.06, 1.81)
	25	0.07 (REC)	DNC	*0.016* (4)	0.128	1.54 (1.16, 2.05)
	29	0.04 (DOM)	DNC	*0.036* (2)	0.263	1.75 (1.16, 2.65)

†Odds ratio = Odds of colon cancer for individuals over candidate pattern L_{A_l}, compared to that for individuals over candidate pattern L_{B_l}; Raw p-value for LRTI is unadjusted for MHT, where GMI is the genetic model yielding the largest LRT statistic for this approach; 95% Fisher's exact-based confidence intervals are uncorrected for multiple comparisons; Statistically significant interactions based upon the maxT adjusted p-value $\tilde{p}^*_{jl\sigma}$; DNC = pACT algorithm did not converge; Italicized p-values are exact, based upon implementation of Algorithm 4.2.

Figures 4.16 and 4.17 display the relationships between recent cigarette consumption with risk of colon cancer (former figure) and rectal cancer (latter figure), upon the markers determined to exhibit statistically significant GxE interaction by GEM (SNP locus level multiple testing adjustment). Note that each of the four SNPs within the *HIF1A* gene exhibit their respective strongest association with risk of rectal cancer upon candidate pattern $L_{A_2} = (G_j = 0) \wedge (E = 1)$ of GEM (see Table 4.17). Interestingly, upon examination of the latter figure, we see that three (SNP IDs 9, 10, and 12) of the four markers within this gene show similar patterns in their respective relationships between recent cigarette consumption and risk of rectal cancer, across each of the three genotype levels at these loci. To illustrate, upon subjects with genotype AA, the odds of rectal cancer among recent cigarette consumers is 1.75, 1.65, and 1.53 times that of non-recent cigarette consumers for the respective SNP IDs 9, 10, and 12; these odds ratios upon subjects with genotype Aa are 0.93, 1.13, and 1.02, for the respective SNP IDs 9, 10, and 12; and these odds ratios upon subjects with genotype aa are 0.63, 0.69, and 0.49, for the respective SNP IDs 9, 10, and 12. Finally, we note that 50% of the GxE interactions (5/10 for colon; 6/12 for rectal) for these data show the cross-interaction pattern – see Table 4.15 for the summary of the SNP loci showing cross-interaction with the recent cigarette consumption environmental factor.

4.11 Conclusions and Future Directions

Within this chapter we have proposed adapting the SNP-SNP (gene-gene) interaction testing framework of [109] (LPCV) to tests of gene-environment interaction. Upon a given genetic marker (a SNP) and a categorical environmental factor, the goal of the LPCV approach in this context is the simultaneous assessment of a main genetic effect, a main environmental effect, and GxE interaction, without the requirement of specifying the genetic model of inheritance. This is carried out by way of the MHT of the null hypothesis of no association between a pair of competitive candidate patterns – formulated in terms of strata upon the genetic and environmental factors, by way of the random variable X_j (4.2) – and a binary response. A chi-square test statistic is computed upon testing each of the null hypotheses. Correction for MHT is performed by way of referring to the test statistics null distribution for the maximum of these chi-square test statistics. Whereas the MHT correction of [109] is based upon an asymptotic MVN approximation to the test statistics null distribution, our approach to the MHT correction refrains from making any approximations to said distribution and is based upon the permutation null distribution of the maximum chi-square test statistic.

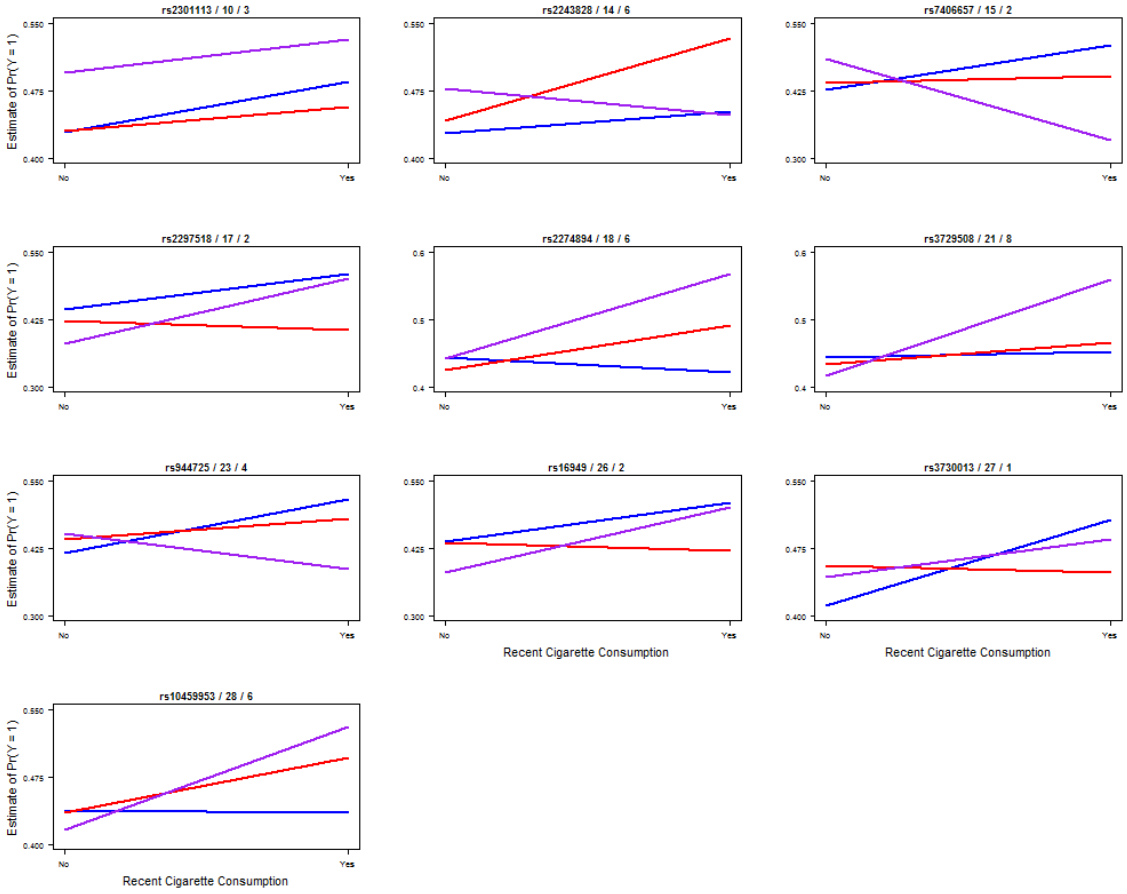

Fig. 4.16: Relationships Between Recent Consumption of Cigarettes and Risk of Colon Cancer, Stratified by the Levels of Genotype, for the 10 SNPs Determined to Possess Statistically Significant GxE Interaction Using GEM. Blue Curves Correspond to Genotype AA ($G_j = 0$), Red Curves to Genotype Aa ($G_j = 1$), and Purple Curves to Genotype aa ($G_j = 2$), for a SNP with Respective Major and Minor Alleles A and a. The Genome SNP ID / SNP Index / Candidate Pattern Index Are Shown above Each Plot.

Our simulation results upon the RGEM and NLRT competitive methods suggest that such a correction – to the multiple hypothesis testing problem invoked from the *fishing for associations* approach of these methods – is necessary for the proper reporting of Type I errors. Furthermore, our results suggest that control of the FWER at the 5% level to not be realistic for the competitive methods, BLRT, BRGEM, PCT, and GLRT, each of which is based upon an asymptotic test statistics null distribution for its control of the FWER, as illustrated empirically by the observed FWERs presented within Table 4.7. This notion is particularly true upon SNP loci possessing a rare minor allele frequency (say not greater than 5%) within the population and whose allele frequencies adhere to Hardy-Weinberg equilibrium within the population, as illustrated by Figure 4.4.

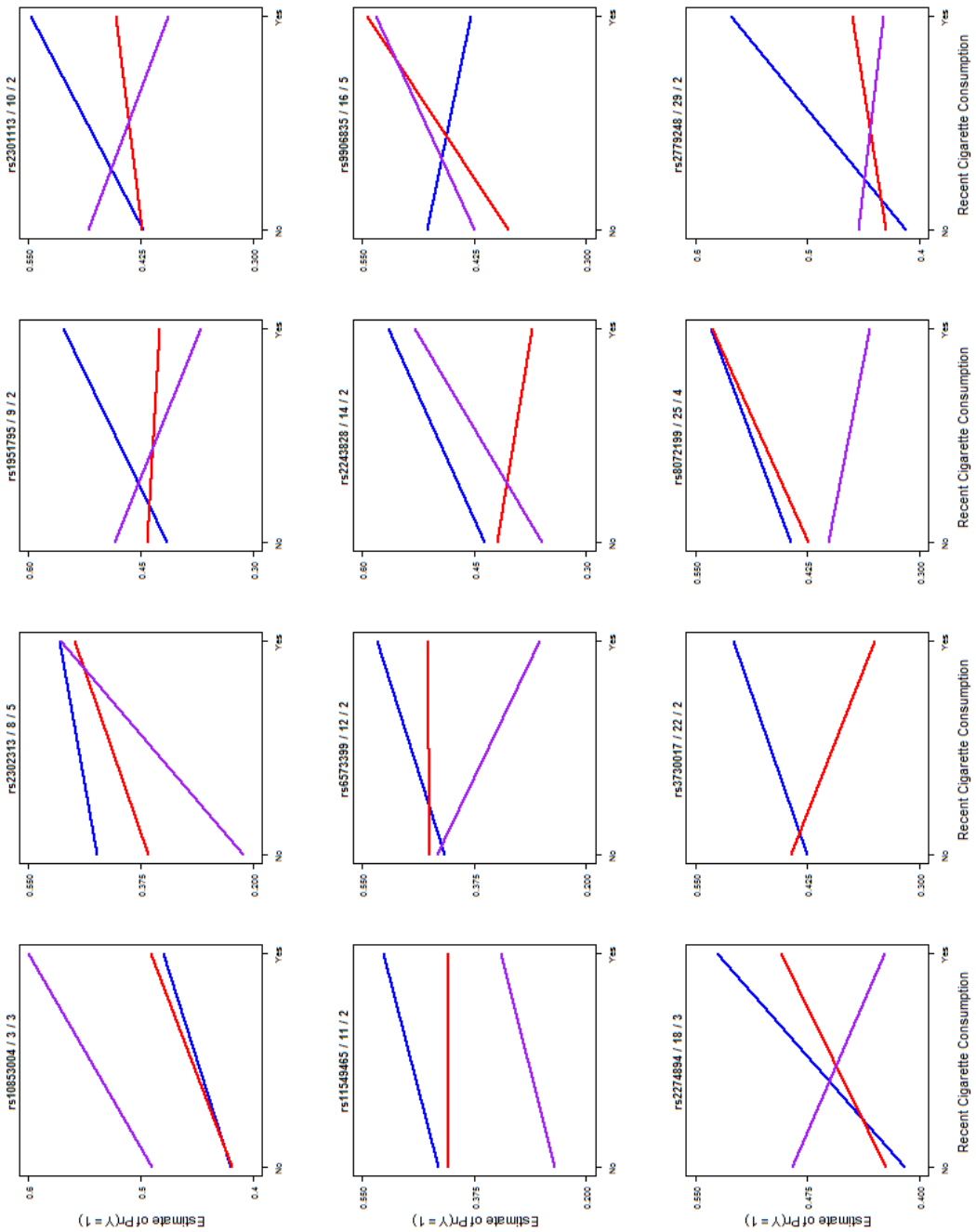

Fig. 4.17: Relationships Between Recent Consumption of Cigarettes and Risk of Rectal Cancer, Stratified by the Levels of Genotype, for the 12 SNPs Determined to Possess Statistically Significant GxE Interaction Using GEM. Blue Curves Correspond to Genotype AA ($G_j = 0$), Red Curves to Genotype Aa ($G_j = 1$), and Purple Curves to Genotype aa ($G_j = 2$), for a SNP with Respective Major and Minor Alleles A and a. The Genome SNP ID / SNP Index / Candidate Pattern Index Are Shown above Each Plot. The Plot Missing a Purple Line Is due to Data Sparsity.

Under $\mathcal{H}_0$, there are two apparent approaches to controlling the FWER over the maximum test statistic for GEM: approximate the test statistics null distribution, by adapting an appropriate asymptotic framework; or, conditional upon the data, work upon the exact conditional (i.e., permutation) null distribution of said test statistic. However, applying an asymptotic approximation to the true null distribution for the maximum chi-square test statistic, say, by way of the MVN theoretical framework presented within [109], is suspect and can lead to inaccurate control in the FWER. This notion is true, even upon large samples, because sparse cell counts (in reference to Table 4.2) can occur upon SNPs with low population minor allele frequencies. Moreover, the MVN framework for the maximum chi-square test statistic (4.11), as presented within [109], lacks the ability to correct the MHT problem of assessing GxE interaction upon multiple genetic markers (i.e., upon circumstances for which $m > 1$). Apparently, there are two approaches for adapting the MVN framework when assessing GxE interaction upon multiple genetic markers: (a) modify the MVN framework to a single $3^m\varepsilon$-level categorical variable which summarizes the random variables $G_1, \ldots, G_m$ and E (e.g., use the random variable X defined within §4.5.2); (b) or, adapt an MTP which corrects for the multiple implementation of the LPCV approach across several genetic markers (e.g., Bonferroni). However, there are problems affiliated with each of these approaches. For the former approach (a), there are two problems. First, the notion of sparse cell counts (referring to Table 4.4) is exacerbated over the single genetic marker implementation of the LPCV approach. For the fixed numbers of cases and controls, many of these categories are likely to contain few (if any) observations. Hence, the integrity of the asymptotic assumption governing the maximum chi-square test statistic across the loci becomes increasingly suspect for increasing values in m. The second problem is computational in nature. As the value in m increases, the number of hypothesis tests encompassing $\mathcal{H}_0$ also increases. Within such a setting, integration over the assumed asymptotic PDF for the maximum chi-square test statistic across the loci becomes computationally prohibitive [109]. A problem with the latter approach (b) is that it lacks accounting for the joint distribution of the maximum test statistics across the genetic markers. As a consequence, this can result in a conservative MHT correction. For example, to account for the MHT problem, induced from application of the LPCV approach to multiple SNP-SNP pairings, [109] applied the conservative method of [177] to control the false discovery rate (FDR). The problem with this approach in controlling the FDR, is that it accounts for an arbitrary dependence structure of the maximum test statistics across the loci [60], thereby failing to incorporate the true underlying correlation of the data. As a result, one can incur

a loss in statistical power.

On the other hand, our permutation based approach to the LPCV induced multiple testing problem, refrains from making any approximations to the null distribution for the maximum chi-square test statistic and therefore results in proper control of the FWER. Moreover, our approach incorporates MHT correction upon the permutation null distribution of the maximum chi-square test statistic across multiple genetic markers, thereby accounting for the joint distribution of the maximum test statistic (4.13) across the loci. As a result, there is no need to implement a secondary MTP (to that of the maxT MTP) to control for MHT across the genetic markers. This can result in increased statistical power to detect true associations, while properly controlling the FWER.

In terms of statistical power (where control of the FWER is at the 5% level), with the exception of two simulation conditions (§4.7) – each of which, nevertheless, resulted in low statistical power amongst all competitive methods chosen for this investigation – our GEM approach outperformed all competitive methods chosen for this investigation, when the true data generating model involved a main genetic effect, a main environmental effect, or GxE interaction. Moreover, based upon the simulation conditions chosen within our investigation, this notion appears to be prevalent whenever the number of sampled genetic markers, m, assumes a value not exceeding 10. This makes our GEM approach ideal for assessing GxE interaction upon a small sample of SNP markers. Although the relative performance of GEM to competing methods seemed to vary upon the simulation conditions involving cross-interaction (§4.8), its performance was essentially on par with the competing methods. Moreover, upon the circumstances in which the relative statistical power of GEM seemed to suffer (i.e., $\beta_e = -0.25$), could be attributed to the competing methods possessing increasing power to detect the marginal effects in both the genetic factor (β_g) and the environmental factor (β_e) – in addition to effect of GxE interaction – particularly, for the multiple logistic regression model of the BLRT method, comprised of the predictors G_{j1} (dominant genetic model), E, and the appropriate term for their interaction effect.

We have proposed several tools for addressing the computational problem unfolding from adapting GEM in practice. Algorithm 4.1 is an efficient computational tool for utility in sampling from the permutation null distribution of the maximum test statistic for GEM. We utilized this algorithm to rapidly carry out $R = 10$K permutation upon the columns of $\mathbf{GE}^*$ for each of the simulated data sets of our simulation study of §4.7, for example, and we used the algorithm to efficiently generate $R = 100$K permutations upon the columns of $\mathbf{GE}^*$ for the application of GEM to the colon

and rectal data sets. In fact, these latter permutations upon the live data sets required at most 1.5 minutes to complete (maximum required time for GEM occurred upon the colon cancer data integrated with the recent NSAID use environmental factor), when applying our proposed *GEM* `R` package code upon the desktop computer summarized within Table 2.3. Moreover, in the case of assessing GxE interaction upon a single genetic marker and a binary environmental factor, we have proposed a network algorithm (NA) approach (Algorithm 4.2) which produces exact conditional maxT adjusted p-values. Without the uncertainty associated with simulating a null distribution, this approach provides the highest accuracy in the control of the FWER over the permutation null distribution of $Z^2_{\max}$. As a future research endeavor, this NA approach could be extended to include investigations related to statistical power for GEM.

We have demonstrated application of GEM to real case-control data involving cases upon each of colon and rectal cancer. Our GEM approach detected highly statistically significant GxE interactions between each of the 29 SNP markers of the candidate pathway and recent NSAID use in their synergistic effect towards risk of colon and rectal cancer, whereas the LRTI approach – used as a model to the conventional logistic regression modeling approach in detecting GxE interaction – failed to detect a single statistically significant GxE interaction at the 5% level in the FWER. Although we have yet to replicate these significant GEM findings, the results adhere with our simulation findings for increased statistical power to detect GxE interaction through application of GEM when compared to the conventional logistic regression modeling approach to detect this type of interaction. Aside from replicating these findings, the next step in the analysis of these data might be to obtain a collective measure of risk for the SNPs comprising the candidate pathway. This could be carried out by using the *combined statistic* approach of [178], or the *adaptive rank truncated product* statistic approach of [179].

We recognize that there are several limitations to our approach. First, because of the sheer magnitude in the number of hypothesis tests conducted across the sampled SNP loci (i.e., $m \times q$), control of the FWER is anticipated to possess lower statistical power when compared to control of alternative Type I error rates, such as the FDR [60,180]. This notion is of particular interest, because genome-wide interaction studies (GWIS) are beginning to emerge [26]. One of the greatest challenges for this approach to gene-environment interaction discovery is that of statistical power [3,181]. In this regard, when assessing GxE interaction upon a large number of sampled genetic markers (say, $m \geq$ 1K – e.g., GWIS) using GEM, control of the FDR may be more appropriate. In such circumstances,

one most surely desires accountability of the correlation in the data (and, so also the test statistics) under the complete null hypothesis, and apparently resampling seems to be the most lucid approach to implicitly accomplish this feat. This could be carried out by adapting the resampling procedures of [182,183] to GEM. Second, in parallel with the novel approaches of e.g., [3,27,109][15] – assessing GxG or GxE interaction within the context of a population-based study design (e.g., case-control) – our approach lacks accountability for confounding factor(s). Confounding occurs whenever the response-explanatory association is over- or under-estimated because of the relation of an underlying variable (i.e., the confounder) with both the explanatory variable and the response variable. Adjustment for the confounding variable(s) within the GEM model can remove its effects. This could be carried out by application of the methodology within [184], as follows. If Z is a confounding variable for the Y-W_l association, some $l = 1, \ldots, q$, we fit a generalized linear model (GLM) to [a function of] $E(Y) = \mu$ and Z, say

$$g(\mu_i) = \beta_0 + \beta_1 Z_i,$$

where g is some appropriate link function (e.g., logit), and where μ_i and Z_i are the respective expected value of Y and measurement of Z for subject i, $i = 1, \ldots, n$. The MLEs for β_0 and β_1 are calculated and the residual, ϵ_i, is estimated by

$$\hat{\epsilon}_i = Y_i - \hat{\mu}_i = Y_i - g^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1 Z_i\right),$$

where Y_i is the phenotype for subject i. The confounder Z could then be accounted for within GEM, by using $\hat{\epsilon}_i$ in lieu of Y_i when conducting the chi-square tests of §4.3. Third, in conjunction with the inference conducted within §3.2.3 for the CATT statistic, there is no compelling reason to assume that the test statistics, $Z_{jA_1}, \ldots, Z_{jA_q}$ (4.10) are identically distributed under $\mathcal{H}_0$, $j = 1, \ldots, m$, for which the maxT MTP for GEM could be unbalanced in its multiplicity adjustment. To illustrate, Figure 4.18 displays the natural logarithm of the ratio of the exact unconditional probability of Type I error for the test statistics upon candidate patterns L_{A_1} and L_{A_6}, for balanced and unbalanced case-control samples and various parameterizations of the ordered pair (π_G, π_E). If the distributions of the test statistics Z_{jA_1} and Z_{jA_6} are identical under $\mathcal{H}_0$, each curve depicted within each of the panel plots of this figure would lie upon the horizontal reference line (light dashed black line).

[15] Interestingly, to this author's review, each of these cited articles fails to acknowledge the notion of confounding factors. Yet, these population-based approaches are susceptible to distortions in associations of interest brought about by unaccounted confounding factors.

However, the panel plots within this figure suggests that the distributions of Z_{jA_1} and Z_{jA_6} are not identical under $\mathcal{H}_0$, as seen by the deviation of each curve from the reference line. Hence, the maxT MTP could be unbalanced in its control of the FWER. A plausible resolution to this potential problem is to adapt a minP MTP approach to GEM, in an analogous manner to that conducted upon the CATT statistic within Chapter 3 of this manuscript. Finally, as argued within Proposition A.9, GEM can be assumed to control the FWER only in the weak sense (in the circumstance for which a binary environmental factor has been sampled). However, because all of the simulations conducted within this chapter (§4.7–§4.9) suggest that GEM controls the FWER at the 5% level under a variety of conditions for which a partial null hypothesis holds, we are optimistic in GEMs ability to control the FWER (at the 5% level) in practice.

We have demonstrated application of our approach within the context of a binary environmental factor, and have assumed that the heterozygote genotype (i.e., $G_j = 1$) to be grouped with one of the homozygote genotypes for each of the genetic factors. In practice, the investigator may be interested in an alternative type of categorical environmental factor, such as a nominal or ordinal multinomial exposure, and/or alternative assumptions governing the genotypes for the genetic factor. Our approach can naturally be extended to accommodate variations in each of the environmental or genetic factors. For example, the candidate patterns outlined within Table 4.1 can be used to apply GEM to a three (or, more) qualitative environmental factor. The statistical power of these extensions would depend upon the underlying data distributions, but – pursuant to our simulation results (§4.7.4) – we would expect similar increases in power for GEM over competing approaches to detect GxE interaction.

In conclusion, gene-environment interactions are worth studying for a number of reasons (§1.2), as they can lead to a better understanding of the complete etiology of disease, inclusive of both distinct and interacting pathways comprised of genetic and environmental factors. In some circumstances, interactions between genetic and environmental factors are believed to exhibit a greater effect than either of the accompanying main effects upon the factors [53,185]. Failing to account for the presence of GxE interaction (e.g., the GWAS approach of assessing solely for main genetic effects) can result in spurious conclusions about the etiology of complex disease, and often attributed as a reason of discordant study findings [53]. Many GWAS, either currently underway or completed, have been conducted on samples with large amounts of existing environmental data (see e.g., [24,59,186,187,188]). Hence, additional testing for interactions to identify novel genetic mark-

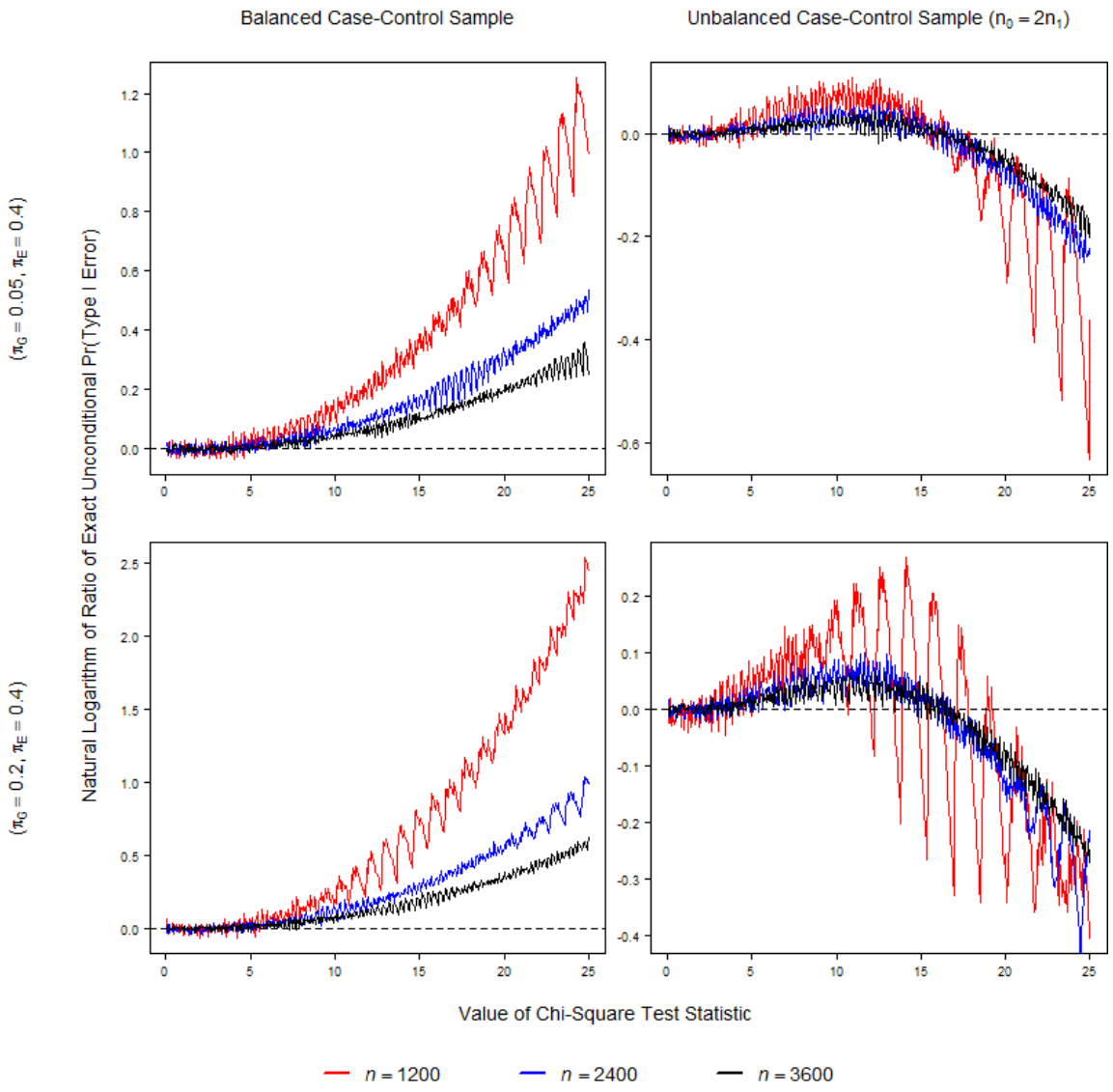

Fig. 4.18: Plot of the Ratio (Natural Logarithm Thereof) of the Exact Unconditional Probability of Type I Error for the Test Statistics upon Candidate Patterns L_{A_1} and L_{A_6} of GEM, for a Binary Environmental Factor with Population Prevalence of Exposure, $\pi_E = \Pr(E = 1) = 0.4$, and a SNP Marker Adhering to Population Hardy-Weinberg Equilibrium with Minor Allele Frequency (π_G) 0.05 (Upper Panel Plots) / 0.20 (Lower Panel Plots). Balanced/Unbalanced Case-Control Samples Depicted Within the Left/Right Panel Plots.

ers beyond those that would be detected by main genetic effect testing alone within these studies, would seem to be a practical cost effective approach. The evidence for GxE interaction upon complex diseases within the literature is compelling (§1.2), as is the argument that failure to model GxE interactions in genetic studies will result in missing potentially important loci which demonstrate

interactions, particularly those coupled with an environmental factor showing a cross-interaction pattern of association. Based upon the analysis conducted within this chapter, our GEM approach appears to possess the potential to increase the yield of genetic association studies, by identifying important loci that synergistically work in concert with an environmental factor to influence risk of a complex disease.

CHAPTER 5

SUMMARY

The mapping of the human genome and the completion of the Human HapMap project over the past decade have significantly altered how research is conducted with respect to the genetic epidemiology of human disease. Study designs and analytic approaches have evolved rapidly from investigations involving relatively few targeted candidate genes to hypothesis-free genome-wide association studies, where thousands – and now even millions – of single molecular mutations are simultaneously analyzed to identify regions of the genome that may influence disease. As laboratory techniques continue to improve and costs decrease, the volume of genetic data will inexorably rise, and robust tools for data management, statistical analysis, and computation will likewise need to keep pace.

This project has focused attention on several analytic and computational problems arising from these new technologies and study designs. Within Chapter 2, we proposed two data management techniques and a parallel processing algorithm (named GPER), whose collective aim is to accelerate simulation of the permutation null distributions for the maxT and minP MTPs upon GWAS data. Our approach presents a significant improvement in computational performance over that of the widely utilized GWAS PLINK software, and is on par with the fastest alternative methods (e.g., PRESTO, PERMORY). However, unlike these methods – which utilize the CPU of the personal computer upon a purely serial-based computing algorithm – our approach is novel insofar as we offload the computational burden for the maxT and minP MTPs to the GPU of the personal computer, and employ a parallel processing approach to accelerate the computational performance.

In Chapter 3 we extended these computational and data management tools, and proposed tools which enhance the statistical analysis governing the Cochran-Armitage trend test (CATT) statistic upon GWAS data. In practice, these proposed enhancements introduce a rather profuse computational problem. We leveraged upon the GPU basis of the GPER algorithm and proposed a parallel processing approach to tackle this computational problem. Insofar as our approach is based upon the minP MTP, implementation of the tools developed within Chapter 3 lead to proper control over the FWER – in the strong sense – while simultaneously preserving high statistical power for control over the nominal Type I error rate. We have demonstrated – through simulation and through

the analysis of a GWAS of Bipolar Disorder – we can attain a considerable boost in statistical power through applying our proposed test statistics null distribution for the CATT statistic within the minP MTP, when compared to utilizing the asymptotic null distribution within the maxT MTP.

In Chapter 4 we extended the utility of the maxT MTP, adapting its control over the FWER when detecting markers involved in gene-environment interactions. We have proposed several tools for addressing the computational problems arising from adapting GEM in practice, including a data management tool analogous to that proposed within Chapter 2 for GWAS data. In the case of assessing a GxE interaction upon a single genetic marker and a binary environmental factor, we proposed a network algorithm (NA) approach which produces exact conditional maxT adjusted p-values. Without the uncertainty associated with simulating a null distribution, this approach provides the highest accuracy in the control of the FWER over the permutation null distribution of the appropriate maximum test statistic. Because our NA operates upon the joint distribution of several test statistics (in contrast, typically an NA involves the distribution of a single test statistic), it could be used as a model for the implementation of an exact approach to the maxT MTP upon GWAS data.

We recognize that there are limitations to some of our proposed tools. For example, calculation of the appropriate p-values under the proposed exact unconditional null distribution for the CATT statistic (Chapter 3), depends upon the nuisance parameter vector $\boldsymbol{\theta}$ (see §3.5.2). We have proposed estimating the parameter elements of this vector at their respective MLEs under the complete null hypothesis. In doing this, our computed pointwise p-values are called bootstrap p-values and are approximate (i.e., not exact). However, this approach seems tenable, insofar as the calculation of the Cochran-Armitage trend test statistic at a particular marker locus (2.5) itself involves estimating a nuisance parameter at its MLE under the complete null hypothesis (for details, see page 150 of [62]). Although future research – beyond the body of research comprising this Dissertation – is needed for developing a methodology to transform these approximate p-values to their exact counterparts, based upon the simulation results presented within §3.6.1 we are optimistic that the approximations are sufficient for accurate strong control of the FWER in a GWAS. This notion holds particularly true to the minP MTP, as any discrepancies between the approximate and exact p-values will result in an unbalanced multiplicity adjustment, as a worst case scenario, and should not compromise the overall control in the FWER for this MTP. Moreover, recent research investigating the distributional properties of bootstrap p-values, particularly within the realm of discrete data, suggests the accuracy

of these p-values to be quite remarkable (see e.g., [140]), which is in direct agreement with the results we obtained within our simulation (§3.6.1).

There are also several limitations to our proposed GEM method, as described within §4.11. First, for large marker panels (i.e., large m), control of the FWER may require an increase in the number of reported Type II errors when compared to control of alternative Type I error rates, such as the FDR. This is of particular interest, because genome-wide interaction studies (GWIS) are now becoming more common [26]. Second, since our GEM method does not as yet allow for inclusion of additional covariates, it can be susceptible to confounding. Within §4.11 we suggested a procedure to control for other factors within GEM, based upon fitting the residuals of a GLM as the response variable for GEM. Finally, as argued within Proposition A.9, GEM can be assumed to control the FWER only in the weak sense. Although we demonstrated this in the context of a binary environmental factor, we can easily extend this approach to an environmental factor with three or more levels. However, because each of the simulations conducted within Chapter 4 (see §4.7–§4.9) suggests that GEM controls the FWER at the 5% level under a variety of conditions for which a partial null hypothesis holds, our preliminary results suggest that GEM could control the FWER (at the 5% level) in practice. This will be a focus of further study.

REFERENCES

[1] J. Hardy and A. Singleton, "Genomewide association studies and human disease," *N Engl J Med*, vol. 360, pp. 1759–1768, 2009.

[2] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, pp. 1516–1517, 1996.

[3] C. Murcray, J. Lewinger, and W. Gauderman, "Gene-environment interaction in genome-wide association studies," *Am J Epidemiol*, vol. 169, no. 2, pp. 219–226, 2009.

[4] E. Hauser, R. Watanabe, D. Duren, et al., "Ordered subset analysis in genetic linkage mapping of complex traits," *Genet Epidemiol*, vol. 27, pp. 53–63, 2004.

[5] T. Pearson and T. Manolio, "How to interpret a genome-wide association study," *JAMA*, vol. 299, no. 11, pp. 1335–1344, 2008.

[6] F. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from large-scale biology," *Science*, vol. 300, no. 5617, pp. 286–290, 2003.

[7] A. Goldstein and N. Andrieu, "Detection of interaction involving identified genes: available study designs," *Monogr Natl Cancer Inst*, vol. 26, pp. 49–54, 1999.

[8] W. Gauderman, J. Witte, and D. Thomas, "Family-based association studies," *Monogr Natl Cancer Inst*, vol. 26, pp. 31–37, 1999.

[9] E. Lander, "The new genomics: global views of biology," *Science*, vol. 274, pp. 536–539, 1996.

[10] International HapMap Consortium, "The international HapMap project," *Nature*, vol. 426, pp. 789–796, 2003.

[11] R. McPherson, A. Pertsemlidis, N. Kavaslar, et al., "A common allele on chromosome 9 associated with coronary heart disease," *Science*, vol. 316, pp. 1488–1491, 2007.

[12] A. Helgadottir, G. Manolescu, S. Gretarsdottir, et al., "A common variant on chromosome 9p21 affects the risk of myocardial infarction," *Science*, vol. 316, pp. 1491–1493, 2007.

[13] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature* vol. 447, pp. 661–678, 2007.

[14] R. Saxena, B. Voight, V. Lyssenko, et al., "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels," *Science*, vol. 316, pp. 1331–1336, 2007.

[15] E. Zeggini, M. Weedon, C. Lindgren, et al., "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes," *Science*, vol. 316, pp. 1336–1341, 2007.

[16] A. Singleton, J. Hardy, B. Traynor, et al., "Towards a complete resolution of the genetic architecture of disease," *Trends Genet*, vol. 26, pp. 438–442, 2010.

[17] W. Cochran, "Some methods for strengthening the common chi-square tests," *Biometrics*, vol. 10, no. 4, pp. 417–451, 1954.

[18] P. Armitage, "Tests for linear trends in proportions and frequencies," *Biometrics*, vol. 11, no. 3, pp. 375–386, 1955.

[19] M. Guedj, J. Wojcjk, E. Della-Chiesa, et al., "A fast, unbiased and exact allelic test for case-control association studies," *Hum Hered*, vol. 61, pp. 210–221, 2006.

[20] K. Conneely and M. Boehnke, "So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests," *Am J Hum Genet*, vol. 81, pp. 1158–1168, 2007.

[21] B. Han, H. Kang, and E. Eskin, "Rapid and accurate multiple testing correction and power estimation for millions of correlated markers," *PLoS Genet*, vol. 5, no. 4, pp. e1000456, 2009.

[22] R. Pahl and H. Schäfer, "PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing," *Bioinformatics*, vol. 26, no. 17, pp. 2093–2100, 2010.

[23] M. Iles, "What can genome-wide association studies tell us about the genetics of common disease?," *PLoS Genet*, vol. 4, no. 2, e33, 2008.

[24] D. Hunter, P. Kraft, K. Jacobs, et al., "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer," *Nat Genet*, vol. 39, pp. 870–874, 2007.

[25] S. Stacey, A. Manolescu, P. Sulem, et al., "Common variants on chromosomes 2q35 and 16q12 confer suceptibility to estrogen receptor-positive breast cancer," *Nat Genet*, vol. 39, pp. 870–874, 2007.

[26] C. Ober and D. Vercelli, "Gene-environment interactions in human disease: nuisance or opportunity?," *Trends Genet*, vol. 27, no. 3, pp. 107–115, 2011.

[27] C. Murcray, J. Lewinger, D. Conti, et al., "Sample size requirements to detect gene-environment interactions in genome-wide association studies," *Genet Epidemiol*, vol. 35, pp. 201–210, 2011.

[28] B. Maher, "Personal genomes: the case of the missing heritability," *Nature*, vol. 456, pp. 18–21, 2008.

[29] T. Manolio, F. Collins, N. Cox, et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, 2009.

[30] M. McCarthy, G. Abecasis, L. Cardon, et al., "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nature Rev Genet*, vol. 9, pp. 356–369, 2008.

[31] J. Barrett, "Measuring the effects of genes and environment on complex traits," *Methods Mol Med*, vol. 141, pp. 55–69, 2008.

[32] M. Boks, M. Schipper, C. Schubart, et al., "Investigating gene environment interaction in complex diseases: increasing power by selective sampling for environmental exposure," *Int J Epidemiol*, vol. 36, pp. 1363–1369, 2007.

[33] M. Chamberlain, P. Baird, M. Dirani, et al., "Unraveling a complex genetic disease: age-related macular degeneration," *Surv Ophthalmol*, vol. 51, pp. 576–586, 2006.

[34] M. Blumenthal, "The role of genetics in the development of asthma and atopy," *Curr Opin Allergy Clin Immunol*, vol. 5, pp. 141–145, 2005.

[35] C. van der Zwaluw and R. Engels, "Gene-environment interactions and alcohol use and dependence: current status and future challenges," *Addiction*, vol. 104, no. 6, pp. 907–914, 2009.

[36] J. van Os, B. Rutten, and R. Poulton, "Gene-environment interactions in schizophrenia: review of epidemiological findings and future directions," *Schizophr Bull*, vol. 34, no. 6, pp. 1066–1082, 2008.

[37] C. Andreasen and G. Andersen, "Gene-environment interactions and obesity-further aspects of genomewide association studies," *Nutrition*, vol. 25, no. 10, pp. 998–1003, 2009.

[38] A. Hirvonen, "Gene-environment interactions in chronic pulmonary diseases," *Mutat Res*, vol. 667, no. 1, pp. 132–141, 2009.

[39] M. Andreassi, "Metabolic syndrome, diabetes and atherosclerosis: influence of gene-environment interaction," *Mutat Res*, vol. 667, no. 1, pp. 35–43, 2009.

[40] J. Ordovas and J. Shen, "Gene-environment interactions and susceptibility to metabolic syndrome and other chronic diseases," *J Periodontol*, vol. 79, no. 8, pp. 1508–1513, 2008.

[41] L. Qi, F. Hu, and G. Hu, "Genes, environment, and interactions in prevention of type 2 diabetes: a focus on physical activity and lifestyle changes," *Cur Mol Med*, vol. 8, no. 6, pp. 519–532, 2008.

[42] S. Visvikis-Siest and G .Siest, "The STANISLAS Cohort: a 10-year follow-up of supposed healthy families. Gene-environment interactions, reference values and evaluation of biomarkers in prevention of cardiovascular diseases," *Clin Chem Lab Med*, vol. 46, no. 6, pp. 733–747, 2008.

[43] A. Alcais, L. Abel, and J. Casanova, "Human genetics of infectious diseases: between proof of principle and paradigm," *J Clin Invest*, vol. 119, no. 9, pp. 2506–2514, 2009.

[44] A. Hill, "The genomics and genetics of human infectious disease susceptibility," *Annu Rev Genomics Human Genet*, vol. 2, pp. 373–400, 2001.

[45] A. Reddy and S. Kleeberger, "Genetic polymorphisms associated with acute lung injury," *Pharmacogenomics*, vol. 10, no. 9, pp. 1527–1539, 2009.

[46] L. Le Marchand and L. Wilens, "Design considerations for genomic association studies: importance of gene-environment interactions," *Cancer Epidemiol Biomarkers Prev*, vol. 17, pp. 263–267, 2008.

[47] L. Tiret, L. Abel, and R. Rakotovao, "Effect of ignoring genotype-environment interaction on segregation analysis of quantitative traits," *Genet Epidemiol*, vol. 10, pp. 581–586, 1993.

[48] L. Eaves, "The resolution of genotype x environment interaction in segregation analysis of nuclear families," *Genet Epidemiol*, vol. 1, pp. 215–228, 1984.

[49] N. Andrieu and A. Goldstein, "Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods," *Epidemiol Rev*, vol. 20, pp. 137–147, 1998.

[50] R. Ottman, "An epidemiologic approach to gene-environment interaction," *Genet Epidemiol*, vol. 11, pp. 75–86, 1990.

[51] W. Gauderman and C. Faucett, "Detection of gene-environment interactions in joint segregation and linkage studies," *Am J Hum Genet*, vol. 61, pp. 1189–1199, 1997.

[52] K. Rothman, S. Greenland, and A. Walker, "Concepts of interaction," *Am J Epidemiol*, vol. 112, no. 4, pp. 467–470, 1980.

[53] E. Flowers, E. Froelicher, and B. Aouizerat, "Gene-environment interactions in cardiovascular disease," *Eur J Cardiovasc Nurs*, In Press, p. 7, 2011.

[54] K. Rothman, "The estimation of synergy or antagonism," *Am J Epidemiol*, vol. 103, no. 5, pp. 506–511, 1976.

[55] E. van Aalst-Cohen, A. Jansen, M. Tanck, et al., "Diagnosing familial hypercholesterolaemia: the relevance of genetic testing," *Eur Heart J*, vol. 27, pp. 2240–2246, 2006.

[56] D. Calhoun, D. Jones, S. Textor, et al., "Resistant hypertension: diagnosis, evaluation, and treatment : a scientific statement from the American Heart Association Professional Education Committee of the Council for High Blood Pressure Research," *Hypertension*, vol. 51, pp. 1403–1419, 2008.

[57] J. Howsonn, B. Barratt, J. Todd, et al., "Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci," *Genet Epidemiol*, vol. 29, pp. 51–67, 2005.

[58] C. Aslund, N. Nordquist, E. Comasco, et al., "Maltreatment, MAOA, and delinquency: sex differences in geneenvironment interaction in a large population-based cohort of adolescents," *Behav Genet*, vol. 41, pp. 262–272, 2011.

[59] J. Herbeck, G. Gottlieb, C. Winkler, et al., "Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS," *J Infect Dis*, vol. 201, pp. 618–626, 2010.

[60] S. Dudoit, J. Shaffer, and J. Boldrick, "Multiple hypothesis testing in microarray experiments," *Stat Sci*, vol. 18, no. 1, pp. 71–103, 2003.

[61] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J Roy Stat Soc B*, vol. 57, no. 1, pp. 289–300, 1995.

[62] P. Westfall and S. Young, *Resampling-based multiple testing.* New York: Wiley, 1993.

[63] S. Purcell, B. Neale, K. Todd-Brown, et al., "PLINK: a toolset for whole-genome association and population-based linkage analysis," *Am J Hum Genet*, vol. 81, pp. 559–575, 2007.

[64] G. Almasi and A. Gottlieb, *Highly Parallel Computing.* Redwood City: Benjamin-Cummings publishers, 1989.

[65] E. Lee. (2006, Jan. 10). *The problem with threads* [Online]. Available: `http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-1.pdf`.

[66] M. Herlihy and N. Shavit, *The Art of Multiprocessor Programming.* Burlington: Morgan Kaufmann, 2008.

[67] S. Adve, V. Adve, G. Agha, et al., *Parallel Computing Research at Illinois.* University of Illinois at Urbana-Champaign, 2008.

[68] NVIDIA Corporation. (2011, May 6). *NVIDIA CUDA C Best Practices Guide, version 4.0* [Online]. Available: `http://www.nvidia.com/object/cuda_get.html`.

[69] NVIDIA Corporation. (2011, May 6). *NVIDIA CUDA C Programming Guide, version 4.0* [Online]. Available: `http://www.nvidia.com/object/cuda_get.html`.

[70] NVIDIA Corporation, "Scalable parallel programming with CUDA," *ACM Queue*, vol. 6, no. 2, pp. 40–53, 2008.

[71] M. Harris. (2007, Feb. 14). *Parallel prefix sum (scan) with CUDA* [Online]. Available: `http://www.nvidia.com/object/cuda_get.html`.

[72] M. Harris. (2011, May 6). *Optimizing parallel reduction with CUDA* [Online]. Available: `http://www.nvidia.com/object/cuda_get.html`.

[73] S. Sherry and K. Sirotkin, "dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation," *Genome Research*, vol. 9, pp. 677–679, 1999.

[74] K. Manly, D. Nettleton, and J.T. Hwang, "Genomics, prior probability, and statistical tests of multiple hypotheses," *Genome Res*, vol. 14, pp. 997–1001, 2004.

[75] F. Dudbridge and B. Koeleman, "Rank truncated product of p-values, with application to genomewide association scans," *Genet Epidemiol*, vol. 25, pp. 360–366, 2003.

[76] F. Dudbridge and A. Gusnanto, "Estimation of significance thresholds for genomewide association scans," *Genet Epidemiol*, vol. 32, pp. 227–234, 2008.

[77] Z. Wei, W. Sun, K. Wang, et al., "Multiple Testing in Genome-Wide association studies via hidden markov models," *Bioinformatics*, vol. 25, pp. 2802–2808, 2009.

[78] S. Grant, K. Wang, H. Zhang, et al., "A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24," *J Pediatr*, vol. 155, no. 6, pp. 909–913, 2009.

[79] W. Satake, Y. Nakabayashi, I. Mizuta, et al., "Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease," *Nat Genet*, vol. 41, no. 12, pp. 1303–1307, 2009.

[80] E. Hattori, T. Toyota, Y. Ishitsuka, et al., "Preliminary genome-wide association study of bipolar disorder in the Japanese population," *Am J Med Genet Part B*, vol. 150B, no. 8, pp. 1110–1117, 2009.

[81] C. McGrath, S. Glatt, P. Sklar, et al., "Evidence for genetic association of RORB with bipolar disorder," *BMC Psychiatry*, vol. 9, no. 70, 2009.

[82] G. Petersen, L. Amundadottir, C. Fuchs, et al., "A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33," *Nat Genet*, vol. 42, no. 3, pp. 224–228, 2010.

[83] L. Shen, K. Sungeun, S. Risacher, et al., "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort," *Neuroimage*, vol. 53, no. 3, pp. 1051–1063, 2010.

[84] L. Lange, D. Croteau, A. Marvelle, et al., "Genome-wide association study of homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR effect in young adults," *Hum Mol Genet*, vol. 19, no. 10, pp. 2050–2058, 2010.

[85] B. Browning, "PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted p-values via permutation for one and two-stage genetic association studies," *BMC Bioinformatics*, vol. 9, no. 309, 2008.

[86] D. Nyholt, "A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other," *Am J Hum Genet*, vol. 74, pp. 765–769, 2004.

[87] G. Kimmel and R. Shamir, "A fast method for computing high-significance disease association in large population-based studies," *Am J Hum Genet*, vol. 79, pp. 481–492, 2006.

[88] X. Gao, J. Starmer, and E. Martin, "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms," *Genet Epidemiol*, vol. 32, pp. 361–369, 2008.

[89] J. Cheverud, "A simple correction for multiple comparisons in interval mapping genome scans," *Heredity*, vol. 87, pp. 52–58, 2001.

[90] J. Li and L. Ji, "Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix," *Heredity*, vol. 95, pp. 221–227, 2005.

[91] V. Moskvina and K. Schmidt, "On multiple-testing correction in genome-wide association studies," *Genet Epidemiol*, vol. 32, pp. 567–573, 2008.

[92] I. Pe'er, R. Yelensky, D. Altshuler, et al., "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants," *Genet Epidemiol*, vol. 32, pp. 381–385, 2008.

[93] S. Seaman and B. Müller-Myhsok, "Rapid simulation of p-values for product methods and multiple-testing adjustment in association studies," *Am J Hum Genet*, vol. 76, pp. 399–408, 2005.

[94] D. Lin, "An efficient monte carlo approach to assessing statistical significance in genomic studies," *Bioinformatics*, vol. 21, no. 6, pp. 781–787, 2005.

[95] D. Salyakina, S. Seaman, B. Browning, et al., "Evaluation of Nyholt's procedure for multiple testing correction," *Hum Hered*, vol. 60, pp. 19-25, 2005.

[96] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, 2nd ed. New York, NY: Springer, 2009.

[97] C. Lewis, "Genetic association studies: design, analysis and interpretation," *Brief Bioinform*, vol. 3, no. 2, pp. 146–152, 2002.

[98] P. Sasieni, "From genotypes to genes: doubling the sample size," *Biometrics*, vol. 53, pp. 1253–1261, 1997.

[99] E. Marubini and M. Valsecchi, *Analysing survival data from clinical trial and observational studies*. West Sussex, England: John Wiley & Sons, Inc., 1995.

[100] M. Matsumoto and T. Nishimura, "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM T Model Comput S*, vol. 8, pp. 3-30, 1998.

[101] K.E. Batcher, "Sorting Networks and their Applications," *Proc. AFIPS Spring Joint Comput. Conf.*, vol. 32, pp. 307–314, 1968.

[102] D.E. Knuth, *The Art of Computer Programming, Vol. 3 - Sorting and Searching*. Reading: Addison-Wesley, 1973.

[103] R Development Core Team. (2011, Dec. 22). *R: A Language and Environment for Statistical Computing* [Online]. Available: `http://www.R-project.org/`.

[104] C. Laurie, K. Doheny, D. Mirel, et al., "Quality control and quality assurance in genotypic data for genome-wide association studies," *Genet Epidemiol*, vol. 34, pp. 591–602, 2010.

[105] E. Smith, C. Bloss, J. Badner, et al., "Genome-wide association study of bipolar disorder in European American and African American individuals," *Mol Psychiatr*, vol. 14, no. 8, pp. 755–763, 2009.

[106] E. Smith, D. Koller, C. Panganiban, et al. , "Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes," *PLoS Genet*, vol. 7, no. 6, 2011.

[107] M. Suchard, Q. Wang, C. Chan, et al., "Understanding GPU programming for statistical computation: studies in massively parallel massive mixtures," *J Comput Graph Statist*, vol. 19, no. 2, pp. 419–438, 2010.

[108] K. Kohlhoff, M. Sosnick, W. Hsu, et al., "CAMPAIGN: an open-source library of GPU-accelerated data clustering algorithms," *Bioinformatics*, vol. 27, no. 16, pp. 2321–2322, 2011.

[109] A. Boulesteix, C. Stobl, S. Weidinger, et al., "Multiple testing for SNP-SNP interactions," *Stat Appl Genet Molec Biol*, vol. 6, no. 1, pp. 1–22, 2007.

[110] R. Gentleman(ed), V. Carey(ed), W. Huber(ed), et al., *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York City: Springer Science + Business Media, LLC, 2005.

[111] B. Efron, R. Tibshirani, V. Goss, et al., "Microarrays and their use in a comparative experiment," Dept. of Stat., Stanford Univ., Tech. Rep. 2000-37B/213, 2000.

[112] V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc of the National Academy of Sciences, USA*, vol. 98, no. 9, pp. 5116–5121, 2001.

[113] J. Wigginton, D. Cutler, and G. Abecasis, "A note on exact tests of hardy-weinberg equilibrium," *Am J Human Genet*, vol. 76, pp. 887–883, 2005.

[114] L. Hartwell, L. Hood, M. Goldberg, et al., *Genetics: From Genes to Genomes*, 3rd ed. Boston: McGraw Hill, 2008.

[115] NCI-NHGRI Working Group on Replication in Association Studies, "Replicating genotype-phenotype associations," *Nature*, vol. 447, pp. 655–660, 2007.

[116] J. Millstein, D. Conti, F. Gilliland, et al., "A testing framework for identifying susceptibility genes in the presence of epistasis," *Am J Hum Genet*, vol. 78, pp. 15–27, 2006.

[117] S. Kuida and D. Beier, "Genetic localization of interacting modifiers affecting severity in a murine model of polycystic kidney disease," *Genome Res*, vol. 10, pp. 49–54, 2000.

[118] C. Naber, J. Husing, U. Wolfhard, et al., "Interaction of the *ACE D* allele and the *GNB3 825T* allele in myocardial infarction," *Hypertension*, vol. 36, pp. 986–989, 2000.

[119] S. Williams, J. Addy, J. Phillips 3rd, et al., "Combinations of variations in multiple genes are associated with hypertension," *Hypertension*, vol. 36, pp. 2–6, 2000.

[120] W. Hsueh, S. Cole, A. Shuldiner, et al., "Interactions between variants in the β_3-adrenergic receptor and peroxisome proliferatoractivated receptor-γ2 genes and obesity," *Diabetes Care*, vol. 24, pp. 672–677, 2001.

[121] J. Kim, S. Sen, C. Avery, et al., "Genetic analysis of a new mouse model for non-insulin-dependent diabetes," *Genomics*, vol. 74, pp. 273–286, 2001.

[122] N. Tripodis, A. Hart, R. Fijneman, et al., "Complexity of lung cancer modifiers: mapping of thirty genes and twenty-five interactions in half of the mouse genome," *J Natl Cancer Inst*, vol. 93, pp. 1484–1491, 2001.

[123] O. Ukkola, L. Perusse, Y. Chagnon, et al., "Interactions among the glucocorticoid receptor, lipoprotein lipase and adrenergic receptor genes and abdominal fat in the Quebec Family Study," *Int J Obes Relat Metab Disord*, vol. 25, pp. 1332–1339, 2011.

[124] C. Barlassina, C. Lanzani, P. Manunta, et al., "Genetics of essential hypertension: from families to genes," *J Am Soc Nephrol Suppl 3*, vol. 13, pp. S155–S164, 2002.

[125] M. De Miglio, R. Pascale, M. Simile, et al., "Polygenic control of hepatocarcinogenesis in Copenhagen x F344 rats," *Int J Cancer*, vol. 111, pp. 9–16, 2004.

[126] E. Yanchina, T. Ivchik, E. Shvarts, et al., "Gene-gene interactions between glutathione-s transferase M1 and matrix metalloproteinase 9 in the formation of hereditary predisposition to chronic obstructive pulmonary disease," *Bull Exp Biol Med*, vol. 137, pp. 64–66, 2004.

[127] P. Yang, W.Bamlet, J. Ebbert, et al., "Glutathione pathway genes and lung cancer risk in young and old populations," *Carcinogenesis*, vol. 25, pp. 1935–1944, 2004.

[128] C. Aston, D. Ralph, D. Lalo, et al., "Oligogenic combinations associated with breast cancer risk in women under 53 years of age," *Hum Genet*, vol. 116, pp. 208–221, 2005.

[129] C. Dong, W. Li, and R. Price, "Interaction between obesity susceptibility loci in chromosome regions 2p25-p24 and 13q13-q21," *Eur J Human Genet*, vol. 13, pp. 102–108, 2005.

[130] V. Roldan, R. Gonzalez-Conejero, F. Martin, et al., "Five prothrombotic polymorphisms and the prevalence of premature myocardial infarction," *Haematologica*, vol. 90, pp. 421–423, 2005.

[131] A. Balmain and C. Harris, "Carcinogenesis in mouse and human cells: parallels and paradoxes," *Carcinogenesis*, vol. 21, pp. 371–377, 2000.

[132] J. Staessen, J. Wang, E. Brand, et al., "Effects of three candidate genes on prevalence and incidence of hypertension in a Caucasian population," *J Hypertens*, vol. 19, pp. 1349–1358, 2001.

[133] R. Culverhouse, "A perspective on epistasis: limits of models displaying no main effects," *Am J Hum Genet*, vol. 70, pp. 461–471, 2002.

[134] D. Vercelli, "Learning from discrepancies: CD14 polymorphisms, atopy and the endotoxin switch," *Clin Exp Allergy*, vol. 33, pp. 153–155, 2003.

[135] G. Koppelman, "Gene by environment interaction in asthma," *Curr Allergy Asthma Rep*, vol. 6, pp. 103–111, 2006.

[136] F. Martinez, "CD14, endotoxin, and asthma risk: actions and interactions," *Proc Am Thorac Soc*, vol. 4, pp. 221–225, 2007.

[137] J. Ioannidis, "Why most published research findings are false," *PLoS Med*, vol. 2, no. 8, pp. 696–701, 2005.

[138] R. Rohlfs and B. Weir, "Distributions of hardy-weinberg equilibrium test statistics," *Genetics*, vol. 180, pp. 1609–1616, 2008.

[139] J. Wakefield, "Bayesian methods for examining hardy-weinberg equilibrium," *Biometrics*, vol. 66, no. 1, pp. 257–265, 2010.

[140] C. Lloyd. (2008, Oct. 1). Bootstrap p-values in discrete models: asymptotic and non-asymptotic effects. *MBS Working paper* [Online]. Available: `http://works.bepress.com/chris_lloyd/15/`.

[141] C. Mehta, N. Patel, and P. Senchaudhuri, "Exact power and sample-size computations for the Cochran-Armitage trend test," *Biometrics*, vol. 54, pp. 1615–1621, 1998.

[142] C. Corcoran, C. Mehta, and P. Senchaudhuri, "Power comparisons for tests of trend in dose-response studies," *Statist Med*, vol. 19, pp. 3037–3050, 2000.

[143] F. Requena and N. Ciudad, "A major improvement to the network algorithm for Fisher's Exact Test in $2 \times c$ contingency tables," *Comput Stat Data An*, vol. 51, pp. 490–498, 2006.

[144] Psychiatric GWAS Consortium Coordinating Committee, "Genomewide association studies: history, rationale, and prospects for psychiatric disorders," *Am J Psychiatry*, vol. 166, no. 5, pp. 540–556, 2009.

[145] H. Ito, K. Matsuo, N. Hamajima, et al., "Gene-environment interactions between the smoking habit and polymorphisms in the DNA repair genes, APE1 Asp148Glu and XRCC1 Arg399Gln, in Japanese lung cancer risk," *Carcinogenesis*, vol. 25, no. 8, pp. 1395–1401, 2004.

[146] M. Stern, L. Johnson, D. Bell, et al., "XPD codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk," *Cancer Epidemiol Biomarkers Prev*, vol. 11, no. 10, pp. 1004–1011, 2002.

[147] D. Thomas, "Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies," *Annu Rev Public Health*, vol. 31, pp. 21–36, 2010.

[148] K. Pollard and M. van der Laan, "Choice of a null distribution in resampling-based multiple testing," *J Stat Plan Infer*, vol. 125, no. 1, pp. 85–100, 2004.

[149] S. Dudoit and M. van der Laan, *Multiple testing procedures with applications to genomics*. New York: Springer, 2008.

[150] P. Bůžková, T. Lumley, and K. Rice, "Permutation and parametric bootstrap tests for genegene and geneenvironment interactions," *Ann Hum Genet*, vol. 75, pp. 36–45, 2011.

[151] M. Khoury and W. Flanders, "Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls!," *Am J Epidemiol*, vol. 144, no. 3, pp. 207–213, 1996.

[152] W. Gauderman, "Sample size requirements for association studies of gene-gene interaction," *Am J Epidemiol*, vol. 155, no. 5, pp. 478–484, 2002.

[153] W. Piegorsch, C. Weinberg, and J. Taylor, "Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies," *Statist Med*, vol. 13, no. 2, pp. 153–162, 1994.

[154] C. Kooperberg and M. LeBlanc, "Increasing the power of identifying gene-gene interactions in genome-wide association studies," *Genet Epidemiol*, vol. 32, pp. 255–263, 2008.

[155] I. Ruczinski and C. Kooperberg, "Logic regression," *J Comput Graph Statist*, vol. 12, pp. 475–511, 2003.

[156] I. Ruczinski, C. Kooperberg, and M. LeBlanc, "Exploring interactions in high dimensional genomic data: an overview of logic regression, with applications," *J Multivariate Anal*, vol. 90, pp. 178–195, 2004.

[157] C. Kooperberg and I. Ruczinski, "Identifying interacting SNPs using Monte Carlo logic regression," *Genet Epidemiol*, vol. 28, pp. 157–170, 2005.

[158] H. Schwender and K. Ickstadt, "Identification of SNP interactions using logic regression," *Biostatistics*, vol. 9, no. 1, pp. 187–198, 2008.

[159] A. Bureau, J. Dupuis, and K. Falls, "Identifying SNPs predictive if phenotype using random forests," *Genet Epidemiol*, vol. 28, pp. 171–182, 2005.

[160] L. Chen, G. Yu, C. Langefeld, et al., "Comparative analysis of methods for detecting interacting loci," *BMC Genomics*, vol. 12, no. 344, 2011.

[161] D. Clayton and P. McKeigue, "Epidemiological methods for studying genes and environmental factors in complex diseases," *Lancet*, vol. 358, pp. 1356–1360, 2001.

[162] P. Kraft, Y. Yen, D. Stram, et al., "Exploiting gene-environment interaction to detect genetic associations," *Hum Hered*, vol. 63, pp. 111–119, 2007.

[163] H. Selinger-Leneman, E. Genin, J. Norris, et al., "Does accounting for gene-environment (GxE) interaction increase the power to detect the effect of a gene in a multifactorial disease?," *Genet Epidemiol*, vol. 24, pp. 200–207, 2003.

[164] M. Pepe, H. Janes, G. Longton, et al., "Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker," *Am J Epidemiol*, vol. 159, pp.882–890, 2004.

[165] P. McCullagh and J. Nelder, *Generalized linear models*, 2nd ed. Boca Raton: Chapman & Hall/CRC, 1989.

[166] C. Mehta and N. Patel, "A network algorithm for the exact treatment of the $2 \times k$ contingency table," *Commun Stat B-Simul*, vol. 9, no. 6, pp. 649–664, 1980.

[167] C. Mehta and N. Patel, "A network algorithm for performing Fisher's Exact Test in $r \times c$ contingency tables," *J Am Stat Assoc*, vol. 78, pp. 427–434, 1983.

[168] C. Mehta, N. Patel, and P. Senchaudhuri, "Exact stratified linear rank tests for binary data," *Computing Science and Statistics: Proc of the 23rd Symp on the Interface*, 1991, pp. 200–207.

[169] A. Agresti, C. Mehta, and N. Patel, "Exact inference for contingency tables with ordered categories," *J Am Stat Assoc*, vol. 85, pp. 453–458, 1990.

[170] K. Schneiter, N. Laird, and C. Corcoran, "Exact family-based association tests for biallelic data," *Genet Epidemiol*, vol. 29, pp. 185–194, 2005.

[171] J. Marchini, P. Donelly, and L. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases," *Nat Genet*, vol. 37, pp. 413–417, 2005.

[172] M. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, vol. 9, no. 1, pp. 30–50, 2008.

[173] C. Tsai, J. Hwang, L. Lai, et al., "Interaction of gender, hypertension, and the angiotensinogen gene haplotypes on the risk of coronary artery disease in a large angiographic cohort," *Atherosclerosis*, vol. 203, pp. 249–256, 2009.

[174] M. Slattery, J. Potter, B. Caan, et al., "Energy balance and colon cancer–beyond physical activity," *Cancer Res*, vol. 57, no. 1, pp. 75–80, 1997.

[175] M. Slattery, S. Edwards, K. Curtin, et al., "Physical activity and colorectal cancer," *Am J Epidemiol*, vol. 158, no. 3, pp. 214–224, 2003.

[176] M. Slattery, A. Lundgreen, B. Welbourn, et al., "Oxidative balance and colon and rectal cancer: interaction of lifestyle factors and genes," *Carcinogenesis*, to be published.

[177] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Ann Stat*, vol. 29, pp. 1165–1188, 2001.

[178] O. De la Cruz, X. Wen, B. Ke, et al., "Gene, region and pathway level analyses in whole-genome studies," *Genet Epidemiol*, vol. 34, pp. 222–231, 2010.

[179] K. Yu, Q. Li, A. Bergen, et al., "Pathway analysis by adaptive combination of p-values," *Genet Epidemiol*, vol. 33, pp. 700–709, 2009.

[180] S. Musani, D. Shriner, N. Liu, et al., "Detection of gene x gene interactions in genome-wide association studies of human population data," *Hum Hered*, vol. 63, pp. 67–84, 2007.

[181] M. Lanktree and R. Hegele, "Gene-gene and gene-environment interactions: new insights into the prevention, detection and management of coronary artery disease," *Genome Med*, vol. 1, no. 28, pp. 1–11, 2009, doi:10.1186/gm28.

[182] D. Yekutieli and Y. Benjamini, "Resampling-based false discovery rate controlling multiple testing procedures for correlated test statistics," *J Stat Plan Infer*, vol. 82, pp. 171–196, 1999.

[183] J. Romano, A. Shaikh, and M. Wolf, "Control of the false discovery rate under dependence using the bootstrap and subsampling," *Test*, vol. 17, pp. 417–442, 2008.

[184] K. Lunetta, S. Faraone, J. Biederman, et al., "Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions," *Am J Hum Genet*, vol. 66, pp. 605–614, 2000.

[185] D. Hunter, "Gene-environment interactions in human diseases," *Nat Rev Genet*, vol. 6, no. 4, pp. 287–298, 2005.

[186] M. Ising, S. Lucae, E. Binder, et al., "A genomewide association study points to multiple loci that predict antidepressant drug treatment outcome in depression," *Arch Gen Psychiatry*, vol. 66, pp. 966–975, 2009.

[187] E. van den Oord, P. Kuo, A. Hartmann, et al., "Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism," *Arch Gen Psychiatry*, vol. 65, pp. 1062–1071, 2007.

[188] L. Scott, K. Mohlke, L. Bonnycastle, et al., "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants," *Science*, vol. 316, pp. 1341–1345, 2007.

APPENDICES

APPENDIX A

PROPOSITIONS

Proposition A.1. *For $j = 1, \ldots, m$, we consider testing the null hypothesis of no association between Y and G_j ($H_0^{(j)}$), against the two-sided alternative hypothesis ($H_a^{(j)}$). To test these hypotheses within a GWAS, a commonly employed test statistic is based upon the Cochran-Armitage trend test (CATT), whose form is given by (2.5). Taking $(v_0, v_1, v_2) = (t, t+1, t+2)$, for some real number t, the CATT statistic can be used to test $H_0^{(j)}$ against $H_a^{(j)}$ under the additive GMI. From a parametric modeling perspective, the additive GMI satisfies the simple logistic regression model, specified by (2.3). That is, if $\pi_{jk} = \Pr(Y = 1 | G_j = k)$, for each $k \in \mathcal{G}$, the additive GMI assumes the behavior in the π_{jk} at locus j satisfies the model*

$$\log\left(Odds\left(\pi_{jk}\right)\right) = \beta_{0j} + \beta_{1j}k, \tag{A.1}$$

for some unknown parameters β_{0j} and β_{1j}. In terms of this model, $H_0^{(j)}$ and $H_a^{(j)}$ can be expressed by

$$\begin{aligned} H_0^{(j)} &: \beta_{1j} = 0 \\ H_a^{(j)} &: \beta_{1j} \neq 0. \end{aligned} \tag{A.2}$$

Under $H_0^{(j)}$, it follows that the CATT statistic is equivalent to Rao's Score test statistic in testing the hypotheses given by (A.2) upon the model (A.1).□

Proof: First, note that under the additive GMI, the CATT statistic (2.5) can be written as

$$T_j = \frac{n\left(n\left(n_{j02} - n_{j00}\right) - n_0\left(n_{j2} - n_{j0}\right)\right)^2}{\left(n_0\right)\left(n - n_0\right)\left(4n_{j0}n_{j2} + n_{j0}n_{j1} + n_{j1}n_{j2}\right)}, \tag{A.3}$$

where for each $c = 0, 1$ and $k \in \mathcal{G}$, the values of n_{jck} and n_{jk} are as depicted within Table 2.1. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ denote the vector of random responses for the binary trait among the n sampled participants. Further, conditional on X_j, denote the joint PMF of $\mathbf{Y}$ at realization $\mathbf{y}$ by

$f_{\mathbf{Y}}(\mathbf{y}; \pi_{j0}, \pi_{j1}, \pi_{j2})$. Since the responses are assumed independent, the likelihood function is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \pi_{j0}, \pi_{j1}, \pi_{j2}) &= \prod_{k=0}^{2} \prod_{i:y_i \in \mathbf{y}, g_{ji}=k} \Pr\left(Y_i = y_i | G_j = g_{ji}\right) \\ &= \prod_{k=0}^{2} \prod_{i:y_i \in \mathbf{y}, g_{ji}=k} \pi_{jk}^{y_i} \left(1 - \pi_{jk}\right)^{1-y_i}, \end{aligned} \tag{A.4}$$

where g_{ji} is as defined by (2.2). Taking the natural logarithm, the log-likelihood function – denoted as $l\left(\pi_{j0}, \pi_{j1}, \pi_{j2}; \mathbf{y}\right)$ – is given by

$$\begin{aligned} l\left(\pi_{j0}, \pi_{j1}, \pi_{j2}; \mathbf{y}\right) &= \sum_{k=0}^{2} \sum_{i:y_i \in \mathbf{y}, g_{ji}=k} \left[y_i \log\left(\pi_{jk}\right) + \left(1 - y_i\right) \log\left(1 - \pi_{jk}\right)\right] \\ &= \sum_{k=0}^{2} \sum_{i:y_i \in \mathbf{y}, g_{ji}=k} \left[y_i \operatorname{logit}\left(\pi_{jk}\right) + \log\left(1 - \pi_{jk}\right)\right] \\ &\overset{*}{=} \sum_{i=1}^{n} \left[y_i \left(\beta_{0j} + g_{ji}\beta_{1j}\right) - \log\left(1 + \exp\left\{\beta_{0j} + g_{ji}\beta_{1j}\right\}\right)\right] \\ &= l\left(\beta_{0j}, \beta_{1j}; \mathbf{y}\right), \end{aligned} \tag{A.5}$$

where $\operatorname{logit}(\cdot) = \log\left(\operatorname{Odds}(\cdot)\right)$ and where $(\overset{*}{=})$ holds by (A.1). Taking partial derivatives of (A.5) with respect to β_{kj}, $k \in \{0, 1\}$, we get

$$\begin{aligned} \frac{\partial l\left(\beta_{0j}, \beta_{1j}; \mathbf{y}\right)}{\partial \beta_{0j}} &= \sum_{i=1}^{n} \left[y_i - \operatorname{expit}\left(\beta_{0j} + g_{ji}\beta_{1j}\right)\right] \\ \frac{\partial l\left(\beta_{0j}, \beta_{1j}; \mathbf{y}\right)}{\partial \beta_{1j}} &= \sum_{i=1}^{n} \left[y_i g_{ji} - g_{ji} \operatorname{expit}\left(\beta_{0j} + g_{ji}\beta_{1j}\right)\right] \end{aligned}, \tag{A.6}$$

where $\operatorname{expit}(\cdot) = \exp(\cdot)\left(1 + \exp\{\cdot\}\right)^{-1}$ (the inverse function of $\operatorname{logit}(\cdot)$). Let $\mathbf{U}_j = \left(U_0(\boldsymbol{\beta}_j), U_1(\boldsymbol{\beta}_j)\right)'$ denote the efficient score for $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j})'$. For $k \in \{0, 1\}$, by (A.6) it holds

$$U_k\left(\boldsymbol{\beta}_j\right) = \frac{\partial l\left(\beta_{0j}, \beta_{1j}; \mathbf{y}\right)}{\partial \beta_{kj}} = \sum_{i=1}^{n} \left[g_{ji}^{k} \left(y_i - \operatorname{expit}\{\beta_{0j} + g_{ji}\beta_{1j}\}\right)\right]. \tag{A.7}$$

Now, under the null hypothesis of (A.2), in accordance with (A.6) it follows that

$$\frac{\partial l\left(\beta_{0j}, \beta_{1j}; \mathbf{y}\right)}{\partial \beta_{0j}} = 0 \qquad \Longleftrightarrow \qquad \tilde{\beta}_{0j} = \operatorname{logit}\left(\frac{n - n_0}{n}\right), \tag{A.8}$$

where $\tilde{\beta}_{0j}$ denotes the maximum likelihood estimate of β_{0j} under said null hypothesis. Hence, under the null hypothesis of (A.2), evaluating (A.7) at $\boldsymbol{\beta}_j = (\tilde{\beta}_{0j}, 0)'$, we get $U_0\left((\tilde{\beta}_{0j}, 0)'\right) = 0$ and

$$\begin{aligned} U_1\left((\tilde{\beta}_{0j}, 0)'\right) &= \sum_{i=1}^{n} g_{ji}\left(y_i - \frac{n - n_0}{n}\right) \\ &= \frac{n\left(n_{j11} + 2n_{j12}\right) - \left(n - n_0\right)\left(n_{j11} + n_{j01} + 2n_{j12} + 2n_{j02}\right)}{n} \\ &= \frac{n_0\left(n_{j12} - n_{j10}\right) + \left(n - n_0\right)\left(n_{j00} - n_{j02}\right)}{n}. \end{aligned} \tag{A.9}$$

Denote the observed Fisher's information matrix for $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j})'$ by $\mathbf{I}(\boldsymbol{\beta}_j)$. Also, for each $s, t \in \{1, 2\}$, denote the $(s,t)^{\text{th}}$ element of $\mathbf{I}(\boldsymbol{\beta}_j)$ by $\left[\mathbf{I}(\boldsymbol{\beta}_j)\right]_{(s,t)}$. Under the null hypothesis of (A.2), a consistent estimator of $\left[\mathbf{I}(\boldsymbol{\beta}_j)\right]_{(s+1,t+1)}$ is [99]

$$\left[\mathbf{I}\left((\tilde{\beta}_{0j}, 0)'\right)\right]_{(s+1,t+1)} = -\frac{\partial U_s(\boldsymbol{\beta}_j)}{\partial \beta_{tj}}\Big|_{\beta_{0j}=\tilde{\beta}_{0j}, \beta_{1j}=0} \quad \forall s, t \in \{0, 1\}. \tag{A.10}$$

We find

$$-\frac{\partial U_s(\boldsymbol{\beta}_j)}{\partial \beta_{tj}} = \sum_{i=1}^{n}\left[(g_{ji})^{I(s\neq t)+2I(s=t=1)}\left(\frac{\operatorname{expit}\left\{\beta_{0j} + g_{ji}\beta_{1j}\right\}}{1 + \exp\left\{\beta_{0j} + g_{ji}\beta_{1j}\right\}}\right)\right] \quad \forall s, t \in \{0, 1\}, \tag{A.11}$$

where recall $I(\cdot)$ is the indicator function, whose returned value is: one if its argument, namely $(\cdot)$, is true; and zero otherwise. Evaluating (A.11) at $\boldsymbol{\beta}_j = (\tilde{\beta}_{0j}, 0)'$, expression (A.10) reduces to

$$\left[\mathbf{I}\left((\tilde{\beta}_{0j}, 0)'\right)\right]_{(s+1,t+1)} = \begin{cases} n\left(\dfrac{n - n_0}{n}\right)\left(\dfrac{n_0}{n}\right), & \text{if } s = t = 0 \\ \left(\dfrac{n - n_0}{n}\right)\left(\dfrac{n_0}{n}\right)\left(n_{j1} + 2n_{j2}\right), & \text{if } s \neq t \\ \left(\dfrac{n - n_0}{n}\right)\left(\dfrac{n_0}{n}\right)\left(n_{j1} + 4n_{j2}\right), & \text{if } s = t = 1. \end{cases} \tag{A.12}$$

Now, by (A.12), it holds

$$\begin{aligned} \det\left(\mathbf{I}\left((\tilde{\beta}_{0j}, 0)'\right)\right) &= \left(\frac{n - n_0}{n}\right)^2\left(\frac{n_0}{n}\right)^2\left(n(n_{j1} + 4n_{j2}) - \left(n_{j1} + 2n_{j2}\right)^2\right) \\ &= \left(\frac{n_0(n - n_0)}{n^2}\right)^2\left(4n_{j0}n_{j2} + n_{j0}n_{j1} + n_{j1}n_{j2}\right). \end{aligned} \tag{A.13}$$

Denote the inverse of $\mathbf{I}(\boldsymbol{\beta}_j)$ by $\mathbf{I}^{-1}(\boldsymbol{\beta}_j)$. It follows that the $(2,2)^{\text{th}}$ element of $\mathbf{I}^{-1}((\tilde{\beta}_{0j},0)')$, denoted as $\left[\mathbf{I}^{-1}((\tilde{\beta}_{0j},0)')\right]_{(2,2)}$, is given by

$$\begin{aligned}\left[\mathbf{I}^{-1}((\tilde{\beta}_{0j},0)')\right]_{(2,2)} &= \left(\frac{1}{\det\left(\mathbf{I}\left((\tilde{\beta}_{0j},0)'\right)\right)}\right)\left[\mathbf{I}\left((\tilde{\beta}_{0j},0)'\right)\right]_{(0,0)} \\ &= \left(\frac{n^3}{n_0(n-n_0)}\right)\left(4n_{j0}n_{j2}+n_{j0}n_{j1}+n_{j1}n_{j2}\right)^{-1}. \qquad \text{(A.14)}\end{aligned}$$

Hence, under the null hypothesis of (A.2), Rao's Score test statistic, Q_j, is given by

$$\begin{aligned}Q_j &= \mathbf{U}_j'\mathbf{I}^{-1}(\boldsymbol{\beta}_j)\mathbf{U}_j\Big|_{\beta_{0j}=\tilde{\beta}_{0j},\beta_{1j}=0} \\ &= \left[0 \quad U_1\left((\tilde{\beta}_{0j},0)'\right)\right]\mathbf{I}^{-1}\left((\tilde{\beta}_{0j},0)'\right)\left(0,U_1\left((\tilde{\beta}_{0j},0)'\right)\right)' \\ &= \left(U_1\left((\tilde{\beta}_{0j},0)'\right)\right)^2\left[\mathbf{I}^{-1}((\tilde{\beta}_{0j},0)')\right]_{(2,2)} \\ &\overset{*}{=} \frac{n\left[n_0\left(n_{j12}-n_{j10}\right)+\left(n-n_0\right)\left(n_{j00}-n_{j02}\right)\right]^2}{n_0(n-n_0)\left(4n_{j0}n_{j2}+n_{j0}n_{j1}+n_{j1}n_{j2}\right)} \\ &\overset{**}{=} T_j \qquad \left(\text{expression (A.3) computed under } H_0^{(j)}\right), \qquad \text{(A.15)}\end{aligned}$$

where $\left(\overset{*}{=}\right)$ holds by (A.9) and (A.14), and $\left(\overset{**}{=}\right)$ holds by trivial algebra. But, showing that $Q_j = T_j$ under $H_0^{(j)}$ of (A.2) is precisely what we needed to demonstrate. Therefore, under $H_0^{(j)}$, the Cochran-Armitage trend test statistic under the additive GMI is equivalent to Rao's Score test statistic in testing the hypotheses (A.2) upon the logistic regression model (A.1). ■

Proposition A.2. *Let $\mathbf{G}_t^*$ denote the t^{th} row of the genotype matrix $\mathbf{G}^*$, $t = 1, \ldots, m$, where $\mathbf{G}^*$ is the matrix defined within §2.3.2. Here, for each $s = 1, \ldots, m'$, where $m' = m/\rho$ some $\rho \geq 1$, and each $i = 1, \ldots, n$, we consider*

$$g_{si}^{(*\rho)} = \sum_{j=(s-1)\rho+1}^{s\rho} 4^{j-(s-1)\rho-1} g_{ji}^*, \tag{A.16}$$

where g_{ji}^* is the $(j,i)^{th}$ element of $\mathbf{G}^*$. We assign missing genotype values to the numerical value of three (3), so that $g_{ji}^* \in \mathcal{G}^* = \mathcal{G} \cup \{3\}$. Then, it holds that each possible value of $g_{si}^{(*\rho)}$, namely $g_{si}^{(*\rho)} = 0, 1, \ldots, 4^\rho - 1$, corresponds to a unique specification of the vector $\left(g_{\{(s-1)\rho+1\}i}^*, \ldots, g_{\{s\rho\}i}^*\right)$. □

Proof: The proof is by mathematical induction with respect to ρ. To establish the basis for induction, suppose $\rho = 1$. It follows by (A.16) that

$$g_{si}^{(*\rho)} = g_{si}^* \in \mathcal{G}^*,$$

for all $s = 1, \ldots, m$. Clearly, each unique $g_{si}^{(*\rho)} \in \mathcal{G}^*$ corresponds to a unique specification of the [singleton] vector (g_{si}^*), so that the basis for induction holds.

Next, to establish the induction step, suppose that each possible value of $g_{si}^{(*\rho)}$, namely $g_{si}^{(*\rho)} = 0, 1, \ldots, 4^\rho - 1$, corresponds to a unique specification of the vector $\left(g_{\{(s-1)\rho+1\}i}^*, \ldots, g_{\{s\rho\}i}^*\right)$, for some $\rho \in \mathbb{N}$, $\rho > 1$, and all $s = 1, \ldots, m'$. We need to show that this result holds for $\rho+1 \in \mathbb{N}$. Here, for $\rho > 1$, let $g_{si}^{(*\rho)} \in \{0, 1, \ldots, 4^\rho - 1\}$, correspond to the unique specification of $\left(g_{\{(s-1)\rho+1\}i}^*, \ldots, g_{\{s\rho\}i}^*\right)$, for some $s = 1, \ldots, m' - 1$. It holds,

$$\begin{aligned} g_{si}^{(*\{\rho+1\})} &= \sum_{j=(s-1)(\rho+1)+1}^{s(\rho+1)} 4^{j-(s-1)(\rho+1)-1} g_{ji}^* \\ &= \sum_{k=(s-1)\rho+1}^{s\rho+1} 4^{k-(s-1)\rho-1} g_{ki}^* \\ &= g_{si}^{(*\rho)} + 4^\rho g_{\{s\rho+1\}i}^*, \end{aligned}$$

where it is assumed that $s\rho + 1 \leq m$. Here, the element $g_{si}^{(*\{\rho+1\})}$ corresponds to the vector $\left(g_{\{(s-1)\rho+1\}i}^*, \ldots, g_{\{s\rho\}i}^*, g_{\{s\rho+1\}i}^*\right)$. Note that each $g_{\{s\rho+1\}i}^* \in \mathcal{G}^*$ yields a unique value for $g_{si}^{(*\{\rho+1\})}$, since $0 \leq g_{si}^{(*\rho)} < 4^\rho$. The result immediately follows, since the value of $g_{si}^{(*\rho)}$ corresponds to a unique specification of the vector $\left(g_{\{(s-1)\rho+1\}i}^*, \ldots, g_{\{s\rho\}i}^*\right)$. This establishes the induction step.

Therefore, by mathematical induction it holds that each possible value of $g_{si}^{(*\rho)}$, corresponds to a unique specification of the vector $\left(g^*_{\{(s-1)\rho+1\}i}, \ldots, g^*_{\{s\rho\}i}\right)$, for all $s = 1, \ldots, m'$. ■

Proposition A.3. *Here, for each $s = 1, \ldots, m'$, where $m' = m/\rho$, for some $\rho \geq 1$, let $g_{si}^{(*\rho)}$ be given by (A.16), for all $i = 1, \ldots, n$. For every $h = 1, \ldots, \rho - 1$, it holds*

$$\left\lfloor \frac{\lfloor g_{si}^{(*\rho)}/4^{h-1} \rfloor + 4 - k}{4} \right\rfloor = \left\lfloor \frac{\lfloor g_{si}^{(*\rho)}/4^{h-1} \rfloor + 7 - k}{4} \right\rfloor \quad \Longleftrightarrow \quad g^*_{\{(s-1)\rho+h\}i} = k, \tag{A.17}$$

for all $k \in \mathcal{G}^* = \mathcal{G} \cup \{3\}$. Furthermore, for all $k \in \mathcal{G}^*$, it holds

$$\left\lfloor \frac{g_{si}^{(*\rho)}}{4^{\rho-1}} \right\rfloor = k \quad \Longleftrightarrow \quad g^*_{\{s\rho\}i} = k. \square \tag{A.18}$$

Proof: First, note that for each $h = 1, \ldots, \rho$ and all $i = 1, \ldots, n$, it holds

$$\left(4^{1-h}\right) g_{si}^{(*\rho)} = g^*_{\{(s-1)\rho+h\}i} + c_1 + c_2, \tag{A.19}$$

where

$$c_1 = \sum_{j=(s-1)\rho+1}^{(s-1)\rho+h-1} 4^{j-(s-1)\rho-h} g^*_{ji} \quad \text{and} \quad c_2 = \sum_{j=(s-1)\rho+h+1}^{\rho s} 4^{j-(s-1)\rho-h} g^*_{ji}.$$

Further, since $g^*_{ji} \in \mathcal{G}^*$, it follows that

$$\begin{aligned} c_1 &\leq 3 \sum_{j=1}^{h-1} 4^{-j} \qquad \text{(partial geometric series)} \\ &= 3 \left(\frac{1 - (1/4)^h}{3/4} - 1 \right) I(h > 1) \\ &< 1, \end{aligned}$$

for all $h \in \mathbb{N}$. Hence, by (A.19), it holds

$$\lfloor g_{si}^{(*\rho)}/4^{h-1} \rfloor = x^*_{\{(s-1)\rho+h\}i} + c_2. \tag{A.20}$$

Consider $h = 1, \ldots, \rho - 1$. Suppose that the premise of (A.17) holds for all $k \in \mathcal{G}^*$. Note that $c_2/4 = c_2^*$, where $c_2^* \in \mathbb{N} \cup \{0\}$. Thus, for each $k \in \mathcal{G}^*$, by (A.20) we have

$$\left\lfloor \frac{\lfloor g_{si}^{(*\rho)}/4^{h-1} \rfloor + 4 - k}{4} \right\rfloor = \left\lfloor \frac{\lfloor g_{si}^{(*\rho)}/4^{h-1} \rfloor + 7 - k}{4} \right\rfloor$$

$$\implies \quad \left\lfloor \frac{g^*_{\{(s-1)\rho+h\}i} - k}{4} + c_2^* + 1 \right\rfloor = \left\lfloor \frac{g^*_{\{(s-1)\rho+h\}i} + 3 - k}{4} + c_2^* + 1 \right\rfloor$$

$$\implies \quad g^*_{\{(s-1)\rho+h\}i} = k,$$

for which the conclusion of (A.17) holds. Conversely, suppose that the conclusion of (A.17) holds for all $k \in \mathcal{G}^*$. Again, we note that $c_2/4 = c_2^*$, where $c_2^* \in \mathbb{N} \cup \{0\}$. Thus, for each $k \in \mathcal{G}^*$, by (A.20) we have

$$\left\lfloor \frac{\lfloor g_{si}^{(*\rho)}/4^{h-1} \rfloor + 4 - k}{4} \right\rfloor = c_2^* + 1 = \left\lfloor \frac{\lfloor g_{si}^{(*\rho)}/4^{h-1} \rfloor + 7 - k}{4} \right\rfloor,$$

for which the premise of (A.17) holds. Therefore, (A.17) holds for all $k \in \mathcal{G}^*$ and all $h = 1, \ldots, \rho - 1$.

Finally, suppose that $h = \rho$. It follows that $c_2 = 0$, for which (A.20) provides that

$$\lfloor g_{si}^{(*\rho)}/4^{\rho-1} \rfloor = g^*_{\{s\rho\}i}.$$

The result of (A.18) immediately follows for all $k \in \mathcal{G}^*$. ■

Corollary A.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be the function defined by*

$$f(x) = \left\lfloor \frac{x}{4} \right\rfloor, \tag{A.21}$$

and for $k \in \mathbb{N} \cup \{0\}$ let $f^k(x)$ denote the k-fold iterated function over f, where $f^0(x) = x$. Let $s = 1, \ldots, m'$ and $i = 1, \ldots, n$, each be chosen arbitrarily. For all $h = 1, \ldots, \rho$, it holds

$$f^{h-1}\left(g_{si}^{(*\rho)}\right) = \left\lfloor \frac{g_{si}^{(*\rho)}}{4^{h-1}} \right\rfloor . \tag{A.22}$$

□

Proof: The result clearly holds for $h = 1$. So, suppose that $h > 1$. Here,

$$\begin{aligned}
f^{h-1}\left(g_{si}^{(*\rho)}\right) &= f^{h-2}\left(f\left(g_{si}^{(*\rho)}\right)\right) \\
&= f^{h-2}\left(f\left(\sum_{j=(s-1)\rho+1}^{s\rho} 4^{j-(s-1)\rho-1} g_{ji}^{*}\right)\right) \\
&= f^{h-2}\left(g_{\{(s-1)\rho+2\}i}^{*} + \sum_{j=(s-1)\rho+3}^{s\rho} 4^{j-(s-1)\rho-2} g_{ji}^{*}\right) \\
&= f^{h-3}\left(g_{\{(s-1)\rho+3\}i}^{*} + \sum_{j=(s-1)\rho+4}^{s\rho} 4^{j-(s-1)\rho-3} g_{ji}^{*}\right) \\
&\vdots \\
&= f^{h-(h-1)}\left(g_{\{(s-1)\rho+h-1\}i}^{*} + \sum_{j=(s-1)\rho+h}^{s\rho} 4^{j-(s-1)\rho-h+1} g_{ji}^{*}\right) \\
&= g_{\{(s-1)\rho+h\}i}^{*} + \sum_{j=(s-1)\rho+h+1}^{s\rho} 4^{j-(s-1)\rho-h} g_{ji}^{*} \\
&= \left\lfloor \frac{g_{si}^{(*\rho)}}{4^{h-1}} \right\rfloor,
\end{aligned}$$

by (A.20). Therefore, $f^{h-1}\left(g_{si}^{(*\rho)}\right) = \left\lfloor \frac{g_{si}^{(*\rho)}}{4^{h-1}} \right\rfloor$, for all $h = 1, \ldots, \rho$. ■

Proposition A.4. *For each $k \in \{0,1,2\} = \mathcal{G}$ and each $j = 1,\ldots,m$, let $\pi_{jk} = \Pr(Y = 1|G_j = k)$ be as previously defined within §2.2.2. In terms of these π_{jk}, the null hypothesis of no association between G_j and Y, $H_0^{(j)}$, can be expressed by*

$$H_0^{(j)} : \pi_{j0} = \pi_{j1} = \pi_{j2} \ \forall j = 1,\ldots,m. \tag{A.23}$$

For each $k \in \mathcal{G}$ let G_{j1k} and G_{j0k}, denote the respective random numbers of cases and controls carrying k copies of the minor allele at locus j. Under the null hypothesis $H_0^{(j)}$, for each $y \in \{0,1\} = \mathcal{Y}$, it follows that

$$\mathbf{G}_{jy} = (G_{jy0}, G_{jy1}, G_{jy2}) \sim \textit{Multinomial}\left(n_y, \boldsymbol{\pi}_j = (\pi_{0j}, \pi_{1j}, \pi_{2j})\right), \tag{A.24}$$

where $\pi_{kj} = \Pr(G_j = k)$ for all $k \in \mathcal{G}$.□

Proof: First, note that the vector $\mathbf{G}_{jy}$ follows a multinomial distribution, each $y \in \mathcal{Y}$. We need to show that $\Pr(G_j = k|Y = y) = \pi_{kj}$, for all $k \in \mathcal{G}$ and $y \in \mathcal{Y}$. Here, for each $y \in \mathcal{Y}$, let $\pi_y^* = \Pr(Y = y)$. Under the null hypothesis (A.23), for each $y \in \mathcal{Y}$ and $k \in \mathcal{G}$, we have

$$\pi_y^* = \Pr(Y = y|G_j = k) = \frac{\Pr(G_j = k|Y = y)\,\pi_y^*}{\Pr(G_j = k)},$$

which implies that $\Pr(G_j = k|Y = y) = \pi_{kj}$ and the result is established. Therefore, under the null hypothesis (A.23),

$$\mathbf{G}_{jy} = (G_{jy0}, G_{jy1}, G_{jy2}) \sim \text{Multinomial}\left(n_y, \boldsymbol{\pi}_j = (\pi_{0j}, \pi_{1j}, \pi_{2j})\right),$$

where $\pi_{kj} = \Pr(G_j = k)$ for all $k \in \mathcal{G}$. ■

Proposition A.5. *Consider SNP locus j with respective major and minor alleles, A and a. Furthermore, let π_j^{AA}, π_j^{Aa}, and π_j^{aa} denote the respective population frequencies for genotypes AA, Aa, and aa at the locus. If f_j denotes the inbreeding coefficient at locus j, we consider modeling the vector of parameters $\left(\pi_j^{aa}, \pi_j^{Aa}, \pi_j^{AA}\right)$ by*

$$\begin{aligned} \pi_j^{aa} &= \pi_j^2 + \pi_j\left(1-\pi_j\right) f_j, \\ \pi_j^{Aa} &= 2\pi_j\left(1-\pi_j\right)\left(1-f_j\right), \text{ and} \\ \pi_j^{AA} &= \left(1-\pi_j\right)^2 + \pi_j\left(1-\pi_j\right) f_j, \end{aligned} \tag{A.25}$$

where recall, π_j is the population frequency for allele a. It holds that

$$\frac{\pi_j}{\pi_j - 1} \leq f_j \leq 1._{\square} \tag{A.26}$$

Proof: First, we note that $\pi_j^k \in [0,1]$, for all $k \in \{aa, Aa, AA\}$. With a little algebra, by (A.25) we can demonstrate that this requires

$$f_j \in \left[\max\left\{1 - \frac{1}{2\pi_j\left(1-\pi_j\right)}, \frac{\pi_j - 1}{\pi_j}, \frac{\pi_j}{\pi_j - 1}\right\}, \min\left\{1, \frac{2-\pi_j}{1-\pi_j}, \frac{1+\pi_j}{\pi_j}\right\}\right].$$

For every $\pi_j \in [0, 0.5]$, it holds that

$$\min\left\{1, \frac{2-\pi_j}{1-\pi_j}, \frac{1+\pi_j}{\pi_j}\right\} = 1.$$

Also, since $\pi_j \leq 0.5$, it follows that

$$\max\left\{1 - \frac{1}{2\pi_j\left(1-\pi_j\right)}, \frac{\pi_j - 1}{\pi_j}, \frac{\pi_j}{\pi_j - 1}\right\} = \frac{\pi_j}{\pi_j - 1}.$$

Therefore, (A.26) holds. ■

Proposition A.6. *For SNP locus j with respective major and minor alleles, A and a, let π_j^{aa} and π_j^{Aa} denote the respective population proportions of homozygotes for the minor allele and heterozygotes. Let $Q_{0j}^*\left(\pi_j^{aa}, \pi_j^{Aa}, n_0, n_1\right)$ denote the unconditional distribution of the CATT statistic under $H_0^{(j)}$, as previously defined within §3.4. If Γ denotes the support of Q_{0j}^*, then it follows that the number of elements comprising Γ, $n(\Gamma)$, is given by*

$$n(\Gamma) = \binom{n_0+2}{2}\binom{n_1+2}{2}. \square \tag{A.27}$$

Proof: In terms of the notation provided upon Table 2.1, for each $y \in \{0,1\} = \mathcal{Y}$, let Γ_y be defined by

$$\Gamma_y = \left\{(n_{jy0}, n_{jy1}, n_{jy2}) : \sum_{k=0}^{2} n_{jyk} = n_y\right\}.$$

Now, since cases and controls are assumed unrelated, it follows that

$$n(\Gamma) = n(\Gamma_0)n(\Gamma_1).$$

Here, consider $y \in \mathcal{Y}$ arbitrarily. Note that $n_{jy0} = 0, \ldots, n_y$, $n_{jy1} = 0, \ldots, n_y - n_{jy0}$, and $n_{jy2} = n_y - n_{jy0} - n_{jy1}$, for which we have

$$\begin{aligned} n(\Gamma_y) &= \sum_{n_{jy0}=0}^{n_y} \sum_{n_{jy1}=0}^{n_y - n_{jy0}} (1) \\ &= \sum_{n_{jy0}=0}^{n_y} (n_y - n_{jy0} + 1) \\ &= (n_y+1)^2 - \sum_{n_{jy0}=1}^{n_y} n_{jy0} \\ &= (n_y+1)^2 - \frac{n_y(n_y+1)}{2} \\ &= \binom{n_y+2}{2}. \end{aligned}$$

Therefore,

$$n(\Gamma) = n(\Gamma_0)n(\Gamma_1) = \binom{n_0+2}{2}\binom{n_1+2}{2}. \quad \blacksquare$$

Proposition A.7. *For each $k \in \mathcal{G}$ and each $j = 1, \ldots, m$, let $\pi_{kj} = \Pr(G_j = k)$ be as previously defined within the conjecture of Proposition A.4. If π_j denotes the population minor allele frequency at locus j and if the genotype frequencies at said locus adhere to Hardy-Weinberg equilibrium (HWE) within the population, then under the null hypothesis of no genotype-phenotype association $\left(H_0^{(j)}\right)$, the maximum likelihood estimator of π_j, $\hat{\pi}_j$, is given by*

$$\hat{\pi}_j = \frac{2n_{j2} + n_{j1}}{2n}, \tag{A.28}$$

where n_{jk}, $k \in \mathcal{G}$ are as defined within Table 2.1. Further,

$$E(\hat{\pi}_j) = \pi_j, \qquad \text{and} \qquad Var(\hat{\pi}_j) = \frac{\pi_j(1-\pi_j)}{2n}. \tag{A.29}$$

□

Proof: Let the random vector $\mathbf{G}_j = (G_{j1}, \ldots, G_{jn})$ denote the j^{th} row of the genotype matrix $\mathbf{G}$, $j = 1, \ldots, m$, where $\mathbf{G}$ is the matrix defined within §2.2. Note that $\mathbf{G}_j$ corresponds to the random values of G_j upon SNP locus j for the n-size sample of cases and controls. Here, under $H_0^{(j)}$ we denote the joint PMF of $\mathbf{G}_j$ at realization $\mathbf{g} = (g_{j1}, \ldots, g_{jn})$ by $f_{\mathbf{G}_j}(\mathbf{g}; \pi_{0j}, \pi_{1j}, \pi_{2j})$, where g_{ji} is as defined by (2.2). Under said null hypothesis, the likelihood function for the random sample is given by

$$f_{\mathbf{G}_j}(\mathbf{g}; \pi_{0j}, \pi_{1j}, \pi_{2j}) = \prod_{i=1}^{n} \prod_{k \in \mathcal{G}} \pi_{kj}^{I(g_{ji}=k)}. \tag{A.30}$$

Taking the natural logarithm, the log-likelihood function – denoted as $l(\pi_{0j}, \pi_{1j}, \pi_{2j}; \mathbf{g})$ – under the HWE model is given by

$$\begin{aligned} l(\pi_{0j}, \pi_{1j}, \pi_{2j}; \mathbf{g}) &= \sum_{i=1}^{n} [2I(g_{ji}=0)\log(1-\pi_j) + I(g_{ji}=1)\log(2\pi_j(1-\pi_j)) + 2I(g_{ji}=2)\log(\pi_j)] \\ &= 2n_{j0}\log(1-\pi_j) + n_{j1}\log(2\pi_j(1-\pi_j)) + 2n_{j2}\log(\pi_j). \end{aligned} \tag{A.31}$$

Taking the derivative of this expression with respect to π_j, we get

$$\frac{dl(\pi_{0j}, \pi_{1j}, \pi_{2j}; \mathbf{g})}{d\pi_j} = \frac{n_{j1} + 2n_{j2}}{\pi_j} - \frac{n_{j1} + 2n_{j0}}{1-\pi_j}.$$

Setting this expression equal to zero and solving for the critical value(s) of π_j, we find that

$$\hat{\pi}_j = \frac{2n_{j2} + n_{j1}}{2n}.$$

Since the second derivative of (A.31) is strictly negative for all n_{0j}, n_{1j}, n_{2j}, and $\pi_j \in (0,1)$, it follows that $\hat{\pi}_j$ indeed maximizes the likelihood function. Also, under $H_0^{(j)}$, for each $y \in \mathcal{Y}$ and $k \in \mathcal{G}$ we have $G_{jyk} \sim \text{Binomial}(n_y, \pi_{kj})$, where G_{jyk} is as defined within the conjecture of Proposition A.4. Since the random vectors $\mathbf{G}_{j0}$ and $\mathbf{G}_{j1}$ (also defined within said Proposition) are mutually independent, under $H_0^{(j)}$ it follows that

$$G_{j0k} + G_{j1k} \sim \text{Binomial}(n, \pi_{kj}) \;\; \forall k \in \mathcal{G}.$$

Hence, assuming the HWE model under $H_0^{(j)}$, we have

$$\begin{aligned} E(\hat{\pi}_j) &= E\left(\frac{2(G_{j02} + G_{j12}) + (G_{j01} + G_{j11})}{2n}\right) \\ &= \frac{2\pi_{2j} + \pi_{1j}}{2} \\ &= \frac{2\pi_j^2 + 2\pi_j(1-\pi_j)}{2} \\ &= \pi_j, \end{aligned}$$

as desired. Finally, under $H_0^{(j)}$ and the HWE model

$$\begin{aligned} Var(\hat{\pi}_j) &= Var\left(\frac{2(G_{j02} + G_{j12}) + (G_{j01} + G_{j11})}{2n}\right) \\ &= \frac{4\pi_{2j}(1-\pi_{2j}) + \pi_{1j}(1-\pi_{1j})}{4n} + \left(\frac{4}{4n^2}\right)\sum_{y\in\mathcal{Y}} Cov(G_{jy2}, G_{jy1}) \\ &= \frac{4\pi_j^2(1-\pi_j^2) + 2\pi_j(1-\pi_j) - 4\pi_j^2(1-\pi_j)^2 - 8\pi_j^3(1-\pi_j)}{4n} \\ &= \frac{\pi_j(1-\pi_j)}{2n}, \end{aligned}$$

as required. Therefore, under $H_0^{(j)}$ and HWE at locus j, the MLE for π_j is given by (A.28); and, $E(\hat{\pi}_j)$ and $Var(\hat{\pi}_j)$ are given by (A.29). ■

Proposition A.8. *We consider a SNP locus whose true penetrance is given by the pure interaction model*

$$\text{logit}\left(\Pr\left(Y=1|E,G\right)\right)=\beta_0+\gamma GE, \tag{A.32}$$

where Y is an indicator of disease, G is considered the dominant genotype coding at the locus, and E is an indicator for exposure to a binary environmental factor. That is, this locus is assumed to only affect disease risk among exposed individuals within the population, and disease risk is assumed the same between the two exposure groups among non-carriers of the risk allele (measured by way of the parameter β_0). If $\pi_{E|G=1}$ denotes the prevalence of exposure within the population, among individuals carrying at least one copy of the risk allele at the locus, then a main genetic effect may or may not be detectible amongst an n-size sample of cases and controls, dependent upon the magnitude in the value for each of the parameters $\pi_{E|G=1}$ and γ.$\square$

Proof: Here, an association will exist between the genotype coding at this locus and the phenotype, whenever

$$\Pr\left(Y=1|G=1\right)=\Pr\left(Y=1|G=0\right)+\delta,$$

for some $\delta\neq 0$; no association between these variables exists whenever $\delta=0$. Furthermore, for a given n-size sample of cases and controls, all else being equal (e.g., the proportion of cases amidst the [assumed fixed] n-size sample remains unchanged), the statistical power to detect a genotype-phenotype association at the locus increases as δ traverses further away – positive or negative – from the null value of zero. Now, $\Pr\left(Y=1|G=0\right)=\text{expit}\left(\beta_0\right)$, where $\text{expit}(\cdot)=\exp(\cdot)(1+\exp(\cdot))^{-1}$. Also,

$$\begin{aligned}\Pr\left(Y=1|G=1\right)&=\frac{\sum_{x\in\{0,1\}}\Pr\left(Y=1,G=1,E=x\right)}{\Pr(G=1)}\\&=\sum_{x\in\{0,1\}}\Pr\left(Y=1|G=1,E=x\right)\Pr\left(E=x|G=1\right)\\&=\Pr\left(Y=1|G=0\right)+\pi_{E|G=1}\left(\text{expit}\left(\beta_0+\gamma\right)-\text{expit}\left(\beta_0\right)\right),\end{aligned}$$

which provides that $\delta=\pi_{E|G=1}\left(\text{expit}\left(\beta_0+\gamma\right)-\text{expit}\left(\beta_0\right)\right)$. But, the degree to which δ deviates – positive or negative – from the null value of zero, depends upon the magnitude of the parameters $\pi_{E|G=1}$ and γ, which establishes the desired result. ■

Proposition A.9.

We consider testing the q-fold collection of null hypotheses $\{H_0^{(j,l)}\}_{l=1,\ldots,q}$ for GEM, for some $j = 1, \ldots, m$, where $H_0^{(j,l)}$ is as defined within §4.3. If $\mathbf{P}_j = \left(P_{j1}^*, \ldots, P_{jq}^*\right)$ denotes the vector of unadjusted p-values corresponding with the test statistics $Z_{jA_1}^*, \ldots, Z_{jA_q}^*$ – where $Z_{jA_l}^*$ is given by (4.10) with (4.7) substituted in lieu of (4.8) therein, for all $l = 1, \ldots, q$ – then in accordance with Condition 2.1 of [62]:

> The distribution of $\mathbf{P}_j$ is said to have the *subset pivotality* property if the joint distribution of the subvector $\{P_{jl}^* : l \in K\}$ is identical under the restrictions $\cap_{l \in K} H_0^{(j,l)}$ and $\mathcal{H}_0^{(j)} = \cap_{l=1}^{q} H_0^{(j,l)}$, for all subsets $K = \{l_1, \ldots, l_i\}$ of true null hypotheses.

Subset pivotality is important for several reasons. First, it is convenient, as it allows for resampling to be performed under the complete null hypothesis ($\mathcal{H}_0^{(j)}$), rather than under partial null hypotheses. Second, when the condition holds, strong control of the family-wise Type I error rate (FWER) results. On the other hand, when the condition fails, resampling under $\mathcal{H}_0^{(j)}$ can be assumed to control the FWER only in the weak sense [62]. Here, we consider a binary environmental factor and demonstrate that GEM fails adherence to the subset pivotality condition.

The Joint Distribution of a Pair of Standardized Test Statistics

Here, we assume all notation as previously defined within §4.1–§4.5, and we consider assigning the set $K = \{r, s\}$, for some $r \neq s = 1, \ldots, q$. To illustrate GEMs failure in adherence with the property of subset pivotality, it suffices to show that the joint distribution of the test statistics $Z_{jA_r}^*$ and $Z_{jA_s}^*$ is different under the partial null hypothesis $\tilde{\mathcal{H}}_0^{\mathrm{p}} = \{H_0^{(j,r)}, H_0^{(j,s)}\}$ and the complete null hypothesis at locus j, $\mathcal{H}_0^{(j)}$. To show this, we will demonstrate that the covariance between said test statistics is different under $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ and $\mathcal{H}_0^{(j)}$.

Indeed, suppose $H_0^{(j,r)}$ and $H_0^{(j,s)}$ hold true. If θ_{rs} denotes the product of the standard errors for the statistics T_{jA_r} and T_{jA_s} (where these statistics are given by (4.6)) under $H_0^{(j,r)}$ and $H_0^{(j,s)}$, since the random vectors $\mathbf{X}_{j0}$ and $\mathbf{X}_{j1}$ are independent (holds true, since cases and controls are

assumed unrelated), under said null hypotheses it holds

$$\begin{aligned}
\theta_{rs} Cov\left(Z^*_{jA_r}, Z^*_{jA_s}\right) &= E\left(T_{jA_r} T_{jA_s}\right) - E\left(T_{jA_r}\right) E\left(T_{jA_s}\right) \\
&= E\left\{\left(\hat{\pi}_{jA_r1} - \hat{\pi}_{jA_r0}\right)\left(\hat{\pi}_{jA_s1} - \hat{\pi}_{jA_s0}\right)\right\} \\
&= \sum_{y \in \{0,1\} = \mathcal{Y}} E\left(\hat{\pi}_{jA_r y} \hat{\pi}_{jA_s y}\right) - \sum_{y, y' \in \mathcal{Y}: y' \neq y} E\left(\hat{\pi}_{jA_r y}\right) E\left(\hat{\pi}_{jA_s y'}\right) \\
&= \sum_{y \in \mathcal{Y}} \left\{E\left(\hat{\pi}_{jA_r y} \hat{\pi}_{jA_s y}\right) - E\left(\hat{\pi}_{jA_r y}\right) E\left(\hat{\pi}_{jA_s y}\right)\right\}. \qquad \text{(A.33)}
\end{aligned}$$

Now,

$$\begin{aligned}
E\left(\hat{\pi}_{jA_r1} \hat{\pi}_{jA_s1}\right) &= \left(\frac{1}{n_1}\right)^2 E\left(\sum_{h \in A_r} X_{j1h} \sum_{k \in A_s} X_{j1k}\right) \\
&= \left(\frac{1}{n_1}\right)^2 \left\{\sum_{h \in A^*_r} \left[Var\left(X_{j1h}\right) + \left(E\left(X_{j1h}\right)\right)^2\right] + \right. \\
&\qquad \left. \sum_{\substack{h \in A_r, \\ k \in A_s \setminus A^*_r}} \left[Cov\left(X_{j1h}, X_{j1k}\right) + E\left(X_{j1h}\right) E\left(X_{j1k}\right)\right]\right\} \\
&= \left(\frac{1}{n_1}\right)\left(\sum_{h \in A^*_r} \pi_{j1h} - \sum_{\substack{h \in A_r, \\ k \in A_s}} \pi_{j1h} \pi_{j1k}\right) + E\left(\hat{\pi}_{jA_r1}\right) E\left(\hat{\pi}_{jA_s1}\right),
\end{aligned}$$

where $A^*_r = (A_r \cap A_s)$. Similarly,

$$E\left(\hat{\pi}_{jA_r0} \hat{\pi}_{jA_s0}\right) = \left(\frac{1}{n_0}\right)\left(\sum_{h \in A^*_r} \pi_{j0h} - \sum_{\substack{h \in A_r, \\ k \in A_s}} \pi_{j0h} \pi_{j0k}\right) + E\left(\hat{\pi}_{jA_r0}\right) E\left(\hat{\pi}_{jA_s0}\right).$$

Therefore, by (A.33), under $H_0^{(j,r)}$ and $H_0^{(j,s)}$, we have

$$Cov\left(Z^*_{jA_r}, Z^*_{jA_s}\right) = \frac{n_0 n_1 \sum_{y \in \mathcal{Y}} \left\{\left(\frac{1}{n_y}\right)\left(\sum_{h \in A^*_r} \pi_{jyh} - \pi_{jA_r1} \pi_{jA_s1}\right)\right\}}{(n_0 + n_1)\sqrt{\pi_{jA_r1} \pi_{jA_s1} \left(1 - \pi_{jA_r1}\right)\left(1 - \pi_{jA_s1}\right)}}. \qquad \text{(A.34)}$$

GEM Fails Adherence to the Property of Subset Pivotality

For clarity in discussion, here we consider a balanced case-control study (i.e., $n_0 = n_1$), for which under $H_0^{(j,r)}$ and $H_0^{(j,s)}$, (A.34) reduces to

$$Cov\left(Z^*_{jA_r}, Z^*_{jA_s}\right) = \frac{\sum_{y\in\mathcal{Y}}\sum_{h\in A_r^*}\pi_{jyh} - 2\pi_{jA_r1}\pi_{jA_s1}}{2\sqrt{\pi_{jA_r1}\pi_{jA_s1}\left(1-\pi_{jA_r1}\right)\left(1-\pi_{jA_s1}\right)}}. \tag{A.35}$$

To illustrate GEMs failure in the subset pivotality property, it suffices to show that (A.35) differs between the partial null $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ and the complete null hypothesis at locus j, $\mathcal{H}_0^{(j)}$, for a single specification in $\boldsymbol{\pi}_1$ – given the specification in said probability vector, we can then define the respective null hypotheses, $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ and $\mathcal{H}_0^{(j)}$, each through a unique specification of the probability vector $\boldsymbol{\pi}_0$. With this notion in mind, we consider the following three candidate patterns [in particular] for GEM,

$$\begin{aligned} L_{A_1} = (G_j = 0) \wedge (E = 0) \quad &\Longleftrightarrow \quad A_1 = \{1\} \\ L_{A_3} = (G_j \in \{0,1\}) \wedge (E = 0) \quad &\Longleftrightarrow \quad A_3 = \{1,2\} \\ L_{A_9} = (G_j = 0) \wedge (E \in \{0,1\}) \quad &\Longleftrightarrow \quad A_9 = \{1,4\}, \end{aligned}$$

against their corresponding respective complements ($L_{B_l}, l = 1,3,9$), such that $q = 11$ (the number of candidate patterns consider by the GEM methodological development for a binary environmental factor), with the intention of assigning $K = \{3,9\}$. The initial two of these candidate patterns correspond to tests involving GxE interaction, while the final candidate pattern corresponds to the dominant GMI test for the main effect in G_j. These three candidate patterns are of particular interest, insofar as the intersection $A_3 \cap A_9 = A_1$ is not empty. To keep things simple, we consider assigning the elements of the vectors $\boldsymbol{\pi}_{j1}$ and $\boldsymbol{\pi}_{j0}$, such that

$$\begin{aligned} \boldsymbol{\pi}_{j1} &= (\pi_{j11}, \lambda, \lambda, \lambda, \lambda, \lambda) \\ \boldsymbol{\pi}_{j0} &= (\pi_{j11} + \delta, \lambda - \delta, \kappa, \lambda - \delta, \kappa, \kappa) \end{aligned}, \tag{A.36}$$

where the parameters $0 \leq \pi_{j11}, \lambda, \kappa$, and δ are chosen to satisfy the condition $\sum_k \pi_{jyk} = 1$, for each $y \in \mathcal{Y}$. Note that for any admissible choice of the vector of parameters $(\pi_{j11}, \lambda, \kappa, \delta)$, by inspection of (A.36), the null hypotheses $H_0^{(j,3)}$ and $H_0^{(j,9)}$ always hold true. Thus, by (A.35) – where, $r = 3$,

$s = 9$, and $A_3^* = (A_3 \cap A_9) = A_1 = \{1\}$ therein – we have

$$Cov\left(Z^*_{jA_3}, Z^*_{jA_9}\right) = \frac{2\pi_{j11} + \delta - 2\left(\pi_{j11} + \lambda\right)^2}{2\left(\pi_{j11} + \lambda\right)\left(1 - \pi_{j11} - \lambda\right)}.$$

Moreover, the complete null hypothesis at locus j, $\mathcal{H}_0^{(j)}$, holds if and only if $\kappa = \lambda$ and $\delta = 0$ within (A.36), for admissible π_{j11} and λ thereto. Thus, for given admissible values in π_{j11} and λ, observe

$$\begin{aligned} Cov\left(Z^*_{jA_3}, Z^*_{jA_9} | \mathcal{H}_0^{(j)}\right) &= \frac{2\pi_{j11} - 2\left(\pi_{j11} + \lambda\right)^2}{2\left(\pi_{j11} + \lambda\right)\left(1 - \pi_{j11} - \lambda\right)} \\ &\neq \frac{2\pi_{j11} + \delta - 2\left(\pi_{j11} + \lambda\right)^2}{2\left(\pi_{j11} + \lambda\right)\left(1 - \pi_{j11} - \lambda\right)} \\ &= Cov\left(Z^*_{jA_3}, Z^*_{jA_9} | \tilde{\mathcal{H}}_0^{\mathrm{p}}\right), \end{aligned}$$

for some admissible $\delta \neq 0$. Hence, for admissible π_{j11}, λ, and $\delta \neq 0$, such that $\tilde{\mathcal{H}}_0^{\mathrm{p}} = \{H_0^{(j,3)}, H_0^{(j,9)}\}$, since $Cov\left(Z^*_{jA_3}, Z^*_{jA_9} | \mathcal{H}_0^{(j)}\right)$ does not equal $Cov\left(Z^*_{jA_3}, Z^*_{jA_9} | \tilde{\mathcal{H}}_0^{\mathrm{p}}\right)$, the joint distribution of the test statistics $Z^*_{jA_3}$ and $Z^*_{jA_9}$ is not the same under $\tilde{\mathcal{H}}_0^{\mathrm{p}}$ and $\mathcal{H}_0^{(j)}$ – for example, $(\pi_{j11}, \lambda, \delta) = (0.20, 0.16, -0.15)$. Therefore, GEM fails adherence to the property of subset pivotality. ■

Proposition A.10. *We consider testing the q-fold collection of null hypotheses* $\{H_0^{(j,l)}\}_{l=1,\ldots,q}$ *for GEM upon a sampled binary environmental factor, for some* $j = 1, \ldots, m$*, where* $H_0^{(j,l)}$ *is as defined within §4.3. Then, the complete null hypothesis at the locus,* $\mathcal{H}_0^{(j)} = \cap_{l=1}^{q} H_0^{(j,l)}$*, is equivalent to (4.20).*□

Proof: We assume all notation as previously defined within §4.1–§4.5. We need to show that $\mathcal{H}_0^{(j)}$ is equivalent to $H_0 : \boldsymbol{\pi}_{j0} = \boldsymbol{\pi}_{j1}$. First, suppose that $\mathcal{H}_0^{(j)}$ holds true. We need to show that H_0 holds. Since $H_0^{(j,l)}$ holds, for all $l = 1, \ldots, q$, it follows that

(A.37)
$$\pi_{jA_l1} = \Pr(X_j \in A_l | Y = 1) = \frac{\sum_{x \in A_l} \Pr(X_j = x, Y = 1)}{\Pr(Y = 1)} = \frac{\sum_{x \in A_l} \Pr(X_j = x, Y = 0)}{\Pr(Y = 0)} = \pi_{jA_l0}.$$

This implies that

$$\Pr(X_j \in A_l, Y = y) = \Pr(X_j \in A_l) \Pr(Y = y), \tag{A.38}$$

for each $y \in \{0, 1\} = \mathcal{Y}$ and all $l = 1, \ldots, q$. Hence, for each $l \in \{1, 2, 7, 8\}$ and $y \in \mathcal{Y}$, we have

$$\Pr(X_j \in A_l = \{k\}, Y = y) = \Pr(X_j \in A_l = \{k\}) \Pr(Y = y), \tag{A.39}$$

for all $k \in \{1, 4, 3, 6\}$. Now, consider $l \in \{3, 4, 5, 6\}$ to be arbitrary. Then, for some $k \in \{1, 4, 3, 6\}$ and $h \in \{2, 5\}$, we have

$$\begin{aligned}\sum_{x \in \{k,h\}} \Pr(X_j = x, Y = y) &= \Pr(X_j \in \{k, h\}, Y = y) \\ &\overset{(A.38)}{=} \Pr(X_j \in \{k, h\}) \Pr(Y = y) \\ &\overset{(A.39)}{=} \Pr(X_j = k, Y = y) + \Pr(X_j = h) \Pr(Y = y).\end{aligned}$$

This result implies that

$$\Pr(X_j = h, Y = y) = \Pr(X_j = h) \Pr(Y = y), \tag{A.40}$$

for each $h \in \{2, 5\}$ and $y \in \mathcal{Y}$. In general, we find that (A.40) holds for all $h \in \mathcal{X}_2$ and $y \in \mathcal{Y}$. Hence, H_0 holds.

Conversely, suppose that H_0 holds true. We need to show that $\mathcal{H}_0^{(j)}$ holds. Since H_0 holds, it follows that the random variables Y and X_j are independent. Thus, for each $y \in \mathcal{Y}$ and $k \in \mathcal{X}_2$, we can define $\pi_{jk}^* = \pi_{jyk} = \Pr(X_j = k|Y = y)$. For any $k \in \mathcal{X}_2$, it holds

$$\begin{aligned}\Pr(X_j = k) &= \sum_{y\in\mathcal{Y}} \Pr(X_j = k|Y = y)\Pr(Y = y) \\ &\overset{H_0}{=} \pi_{jk}^* \sum_{y\in\mathcal{Y}} \Pr(Y = y) \\ &= \pi_{jk}^*.\end{aligned}$$

This expression implies that

$$\Pr(X_j = k, Y = y) = \Pr(X_j = k)\Pr(Y = y),$$

for all $k \in \mathcal{X}_2$ and $y \in \mathcal{Y}$. Thus, we have

$$\begin{aligned}\pi_{jA_l1} &= \sum_{k\in A_l} \Pr(X_j = k|Y = 1) \\ &= \sum_{k\in A_l} \pi_{jk}^* \\ &= \sum_{k\in A_l} \Pr(X_j = k|Y = 0) \\ &= \pi_{jA_l0},\end{aligned}$$

for all $l = 1, \ldots, q$, for which $\mathcal{H}_0^{(j)}$ holds. Therefore, the complete null hypothesis at the locus, $\mathcal{H}_0^{(j)} = \cap_{l=1}^q H_0^{(j,l)}$, is equivalent to (4.20). ■

APPENDIX B
CUDA KERNELS

B.1 The GPER Algorithm

B.1.1 Permutation

Having prepared the response vector $\mathbf{y}^*$ and genotype matrix $\mathbf{G}^{(*\rho)}$, as outlined within §2.3.1 and §2.3.2, we are poised to begin analyzing the data. Here, we describe a parallel processing approach to ascertaining random permutations of the columns upon $\mathbf{G}^{(*\rho)}$. This approach can essentially be implemented through the conglomeration of two components: (1) generate an n-sequence of unit-uniform (i.e., $U(0,1)$) random deviates, which we denote by $U_1, \ldots, U_n$. We concatenate this n-sequence, along with the sequence of their accompanying indices (i.e., the sequence $1, \ldots, n$ – representative of the column indices for $\mathbf{G}^{(*\rho)}$), to form a $2 \times n$ matrix (denoted $\mathbf{U}_n$); and (2) if the sequence $U_1, \ldots, U_n$ resides upon row t of $\mathbf{U}_n$, $t = 1, 2$, we order the columns upon this matrix in accordance to the values encompassing row t of the matrix. The elements upon row $2 - I(t = 2)$ of the ordered matrix $\mathbf{U}_n$ depict a random permutation of the column indices upon $\mathbf{G}^{(*\rho)}$. This latter component essentially reduces to sorting the sequence $U_1, \ldots, U_n$, while simultaneously tracking the indices of the accompanying elements during the sorting routine.

Each of these two components can be performed in a parallel manner within CUDA C, where it is noted that the latter component is dependent upon the former (i.e., the algorithm we implement for generation of the sequence $U_1, \ldots, U_n$ must complete its tasks prior to the column ordering of $\mathbf{U}_n$). Because of this dependency, we create CUDA C kernel(s) to carry out the specific task upon each of the two aforementioned components. Specifically, for the former component, we create a CUDA C kernel, denoted mMTK (shorthand for modified Mersenne Twister kernel); for the latter component, we create three CUDA C kernels, denoted respectively by BSK1 (BSK is shorthand for bitonic sort kernel), BSK2, and BSK3. We describe the details encompassing mMTK and the BSK's within the respective two sections which follow.

B.1.1.1 Parallel Random Number Generation

To generate the sequence $U_1, \ldots, U_n$ upon the GPU, we make use of the Mersenne Twister (MT) pseudorandom number generator (RNG), developed in 1998 by Makoto Matsumoto and Takuji Nishimura [100]. In particular, we modify the CUDA C MT kernel (MTK) developed by Victor Podlozhnyuk of the NVIDIA Corporation, the CUDA C programming code of which is provided as part of the CUDA toolkit (version 4.0, May 2011). The MT RNG possesses many important properties for random number generation, including: (a) a long period length, equal to the [colossal] value of $2^{19937}-1$ (a *Mersenne* prime number – which is where the name of the generator originates). As one will recall, the period of a random number generator essentially defines the number of [unique] random numbers generated in a sequence before the generator begins re-generating the sequence – the longer the period, the better the random number generator; (b) good distributional properties. The MT generates numbers with an almost $U(0,1)$ distribution; and (c) high performance and efficient use of memory. The speed, portability, and high quality of the MT RNG are desirable properties for random number generation. In fact, for many applications the MT RNG is the pseudorandom number generator of choice and is the default random number generator in the `R` software [103].

The host call for the CUDA C MTK – as provided within the CUDA toolkit (see §1.4.2 for review of the CUDA toolkit) – is `randomGPU<<< B, T >>>(d_Rand, nPerRng)`, where `d_Rand` is the returned two-dimensional matrix of $U(0,1)$ random numbers whose row and column dimensions are equal to `nPerRng` and `B`×`T`, respectively. Essentially, the t^{th} thread for this kernel, $t = 1, \ldots,$ `B`×`T`, generates the `nPerRng`-sequence of $U(0,1)$ random numbers pertaining to the t^{th} column of `d_Rand`.

Here, the row dimension of `d_Rand` could essentially be considered the GWAS sample size, n, while each of the columns for this matrix could represent a particular sequence $U_1, \ldots, U_n$ (e.g, to assist the BSK's in generation of a column permutation of $\mathbf{G}^{(*\rho)}$). However, this value in the row dimension of `d_Rand` will not suffice for the sorting kernels we develop, in general, because said kernels rely upon sorting sequences of length 2^p, some $p \in \mathbb{N}$. We propose utilizing a modified MTK (i.e., mMTK), whose primary intention is to assist the BSK's in generating permutations of the indices upon the columns of $\mathbf{G}^{(*\rho)}$. These modifications are: (a) implementation of a *seed* parameter for the MT RNG. This allows for clustering of GPER to multiple GPUs, and/or data partitioning of the GWAS data set; and (b) if the row dimension of `d_Rand` is 2^p, where p is some integer for which $n \leq 2^p$, we implement a parameter which informs mMTK to halt – upon each of the columns for `d_Rand` – random number generation of the sequence $U_1, \ldots, U_{2^p}$ at the random

number U_n. This modification to MTK is essential, because – as previously elucidated to – the BSK's rely upon sorting sequences of length 2^p, some $p \in \mathbb{N}$. Algorithm B.1 describes the procedure encompassing implementation of mMTK.

Algorithm B.1 Pseudorandom Number Generation

---

1. Allocate a matrix `d_Rand`, of size $2^{\mathtt{p}} \times \mathtt{B} \times \mathtt{T}$, within device memory, such that $n \leq 2^p$. Let $U_{s,t}$ denote the $(s+1, t+1)^{\text{th}}$ element of `d_Rand`.

2. Invoke mMTK, comprised of `B` blocks and `T` threads per block, within the host as follows:

   for $s = 0$ **to** $\mathtt{B} - 1$ **in parallel do**
   for $t = 0$ **to** $\mathtt{T} - 1$ **in parallel do**
   $h \leftarrow t + s \times \mathtt{T}$. {Which column of `d_Rand` does thread t work upon?}
   for $i = 0$ **to** $2^p - 1$ **do**
   if $i < n$ **then**
   $U_{i,h} \leftarrow$ MT RNG generated random deviate.
   else
   $U_{i,h} \leftarrow -1$.
   end if
   end for
   end for
   end for

3. In brief, thread t within block s of the kernel, $t = 0, \ldots, \mathtt{T} - 1$ and $s = 0, \ldots, \mathtt{B} - 1$, generates the sequence $\{U_i\}_{i=1,\ldots,2^p}$, upon the appropriate column of `d_Rand`, where

$$U_i = I(i \leq n)U(0,1) - I(i > n),$$

   for all $i = 1, \ldots, 2^p$. Our BSK's sort the sequence $U_1, \ldots, U_{2^p}$ into decreasing order, so that the elements $U_{n+1}, \ldots, U_{2^p}$ are assured to reside upon the final $(2^p - n)$ elements of the sorted sequence. Further details encompassing this notion are provided within Algorithm B.3 of the next section.

B.1.1.2 Parallel Bitonic Sort

To sort the sequence $U_1, \ldots, U_n$ upon the GPU, we make use of the bitonic sorting method. Bitonic sort is a member of the class of sorting algorithms called sorting networks, and is among the fastest algorithms in this class [101]. A sorting network is a special kind of sorting algorithm, in which the sequence of comparisons is not data-dependent [102]. Thus, sorting networks lend elegantly to the CUDA parallel framework. Overall, our BSK's for bitonic sort encompass a six-step algorithm (see Algorithm B.2), the implementation of which is guaranteed [101] to sort the sequence $U_1, \ldots, U_n$, into say decreasing order, where without loss of generality it is assumed that $n = 2^p$, for some $p \geq 3$. We begin the description of our bitonic sorting algorithm, by way of first defining what is meant by a bitonic sequence.

Definition – Bitonic Sequence:

A 0-1 n-sequence, say

$$a = \{a_1, \ldots, a_n\} \text{ with } a_i \in \{0, 1\},\ i = 1, \ldots, n,$$

is said to be a *bitonic sequence*, if it contains at most two changes between 0 and 1. That is, a is a bitonic sequence if there exists $k, l \in \{1, \ldots, n\}$, $k \leq l$, for which

$$\begin{aligned} &a_1, \ldots, a_k = 0, \quad a_{k+1}, \ldots, a_l = 1, \quad a_{l+1}, \ldots, a_n = 0, \text{ or} \\ &a_1, \ldots, a_k = 1, \quad a_{k+1}, \ldots, a_l = 0, \quad a_{l+1}, \ldots, a_n = 1. \end{aligned}$$

Some examples of 0-1 bitonic sequences for $n = 4$:

$$\underbrace{\{0,0,0,0\}}_{k=l=2}, \quad \underbrace{\{1,1,1,1\}}_{k=l=2}, \quad \underbrace{\{0,0,1,0\}}_{k=l-1=2}, \quad \underbrace{\{1,1,0,1\}}_{k=l-1=2}, \quad \underbrace{\{0,1,1,1\}}_{k=1,l=n}, \quad \underbrace{\{1,0,0,0\}}_{k=1,l=n}.$$

More generally, an n-sequence of real numbers, say

$$c = \{c_1, \ldots, c_n\},$$

is bitonic if it contains at most one local extrema. That is, c is bitonic if there exists $k = 1, \ldots, n$, for which

$$c_1 \oplus c_2 \oplus \cdots \oplus c_k \oplus' c_{k+1} \oplus' \cdots \oplus' c_n,$$

where $\oplus \in \{\leq, \geq\}$ and $\oplus' \in \{\leq, \geq\} \backslash \{\oplus\}$. Some examples of real valued bitonic sequences for $n = 4$:

$$\underbrace{\{1,2,3,4\}}_{k=n,\oplus=\leq}, \quad \underbrace{\{10,6,4,2\}}_{k=n,\oplus=\geq}, \quad \underbrace{\{1,5,3,2\}}_{k=2,\oplus=\leq}, \quad \underbrace{\{5,1,2,3\}}_{k=2,\oplus=\geq}, \quad \underbrace{\{1,2,4,3\}}_{k=3,\oplus=\leq}, \quad \underbrace{\{4,3,2,5\}}_{k=3,\oplus=\geq}. \square$$

Having defined a bitonic sequence, we proceed by describing an algorithm for bitonic sort.

Algorithm B.2 Bitonic Sort

Consider an arbitrary sequence of real numbers

$$c = \{c_0, c_1 \ldots, c_{n-1}\},$$

for which it is desired to sort the elements of this sequence, say in decreasing order, where without loss of generality it is assumed that $n = 2^p$, some $p \geq 3$. The following six-step algorithm applied to c yields the desired sorted n-sequence:

1. **for** $s = 0$ **to** $2^{p-1} - 1$ **do**

   if $c_{2s} \oplus c_{2s+1}$ **then**

   Swap the values c_{2s} and c_{2s+1}, such that

   (B.1)
   $$\oplus = \begin{cases} \leq, & \text{if } s = 2(k-1), \text{ for some } k \in \mathbb{N} \\ \geq, & \text{if } s = 2k-1, \text{ for some } k \in \mathbb{N}. \end{cases}$$

   end if

   end for

2. **for** $s = 0$ **to** $2^{p-1} - 1$ **do**

   if $s = 2(k-1)$, for some $k \in \mathbb{N}$ and $c_{2s} \oplus c_{2(s+1)}$ **then**

   Swap the values c_{2s} and $c_{2(s+1)}$, such that

   (B.2)
   $$\oplus = \begin{cases} \leq, & \text{if } \left\lfloor \frac{s}{2} \right\rfloor = 2(k-1), \text{ for some } k \in \mathbb{N} \\ \geq, & \text{if } \left\lfloor \frac{s}{2} \right\rfloor = 2k-1, \text{ for some } k \in \mathbb{N}, \end{cases}$$

   where $\lfloor \cdot \rfloor$ is the greatest integer function applied to the argument $(\cdot)$.

   else if $s = 2k - 1$, for some $k \in \mathbb{N}$ and $c_{2s-1} \oplus c_{2s+1}$ **then**

   Swap the values c_{2s-1} and c_{2s+1}, such that $\oplus$ is given by (B.2).

end if

end for

3. **for** $s = 0$ **to** $2^{p-1} - 1$ **do**

if $c_{2s} \oplus c_{2s+1}$ **then**

Swap the values c_{2s} and c_{2s+1}, such that $\oplus$ is given by (B.2).

end if

end for

4. Let $h = 1, \ldots, p-2$, denote the number of 'visits' to this step of the algorithm; let $d = 2^{h+1}$ represent the maximum stride between two elements to be compared within c.

5. **for** $s = 0$ **to** $2^{p-2-h} - 1$ **do**

for $w = 0$ **to** $h + 1$ **do**

for $z = 0$ **to** $2^w - 1$ **do**

for

$$t = 2d\left(s + \frac{z}{2^w}\right), 2d\left(s + \frac{z}{2^w}\right) + 1, \ldots, 2d\left(s + \frac{z}{2^w} + \frac{1}{2^{1+w}}\right) - 1$$

do

$u \leftarrow 2^{h+1-w} + t$.

if $c_t \oplus c_u$ **then**

Swap the values c_t and c_u, such that $\oplus$ is as defined by (B.1).

end if

end for

end for

end for

end for

6. **if** $h < p - 2$ **then**

Repeat step 4 of the algorithm.

else

Terminate the algorithm.

end if

In brief, the initial three steps of this algorithm create bitonic subsequences upon c, each of length eight elements. Moreover, if k indexes these subsequences (within any iteration of steps four thru six of the algorithm), the t^{th} cycle (iteration) upon steps four thru six of this algorithm merges the bitonic subsequences k' and $k'+1$ to form a new bitonic subsequence of length 2^{3+t}, such that $\lfloor k'/2 \rfloor = \lfloor (k'+1)/2 \rfloor$, where k' and $k'+1$ are each equal to some k and $t < p-2$. The final iteration upon steps four thru six of the algorithm yields the sorted sequence.

Note that implementation of Algorithm B.2 is readily carried out in a parallel manner upon the GPU, where each binary operator $\oplus$ essentially depicts a CUDA thread execution. Having outlined the algorithm, we are poised to describe our implementation of parallel bitonic sort upon the GPU. Prior to doing this, we provide a simple example for the application of Algorithm B.2 and provide a brief review of the bitwise AND operation.

Example: Consider the unsorted n-sequence of $n = 2^4 = 16$ positive integers,

$$\{10, 11, 7, 3, 6, 16, 14, 13, 5, 8, 15, 12, 9, 2, 4, 1\}, \tag{B.3}$$

for which it is desired to sort this sequence into decreasing order. Table B.1 summarizes the dynamics encompassing the sorting of the sequence (B.3) upon application of Algorithm B.2.$\square$

Table B.1: The Dynamics Entailing Application of Algorithm B.2 to the n-sequence (B.3).

Algorithm Step	(h, d, w)	Resulting Sequence
Unsorted Sequence		$\{10, 11, 7, 3, 6, 16, 14, 13, 5, 8, 15, 12, 9, 2, 4, 1\}$
1	–	$\{11, 10, 3, 7, 16, 6, 13, 14, 8, 5, 12, 15, 9, 2, 1, 4\}$
2	–	$\{11, 10, 3, 7, 13, 6, 16, 14, 12, 15, 8, 5, 1, 2, 9, 4\}$
3	–	$\{11, 10, 7, 3, 6, 13, 14, 16, 15, 12, 8, 5, 1, 2, 4, 9\}$
4-5	$(1, 4, 0)$	$\{11, 13, 14, 16, 6, 10, 7, 3, 1, 2, 4, 5, 15, 12, 8, 9\}$
5	$(1, 4, 1)$	$\{14, 16, 11, 13, 7, 10, 6, 3, 1, 2, 4, 5, 8, 9, 15, 12\}$
5-6	$(1, 4, 2)$	$\{16, 14, 13, 11, 10, 7, 6, 3, 1, 2, 4, 5, 8, 9, 12, 15\}$
4-5	$(2, 8, 0)$	$\{16, 14, 13, 11, 10, 9, 12, 15, 1, 2, 4, 5, 8, 7, 6, 3\}$
5	$(2, 8, 1)$	$\{16, 14, 13, 15, 10, 9, 12, 11, 8, 7, 6, 5, 1, 2, 4, 3\}$
5	$(2, 8, 2)$	$\{16, 15, 13, 14, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 1, 2\}$
5-6	$(2, 8, 3)$	$\{16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1\}$

Next, we provide a brief review of the bitwise AND operation. Here, if $z_{(10)}$ denotes some positive integer written in the decimal (base-10) number system, then we use $z_{(2)}$ to denote this

integer written in the binary (base-2) number system. Now, $z_{(10)}$ can be expressed as

$$z_{(10)} = \sum_{k=0}^{w-1} a_k 2^k, \tag{B.4}$$

where $a_k \in \{0, 1\}$, for all $k = 0, \ldots, w-1$, some $w \geq 1$. Also, in terms of the sequence, $a_0, \ldots, a_{w-1}$, $z_{(2)}$ is given by

$$z_{(2)} = (a_{w-1} \cdots a_0)_{(2)}\,, \tag{B.5}$$

and for $k = 0, \ldots, w-1$, a_k is called a bit. Hence, the relationship between the decimal and binary representations of some positive integer $z_{(10)}$ is given through expressions (B.4) and (B.5). For example, the binary representation of the integer $315_{(10)}$ is $100111011_{(2)}$.

Recall, the bitwise AND operation takes a binary representation for each of two integers $z_{(10),1}$ and $z_{(10),2}$, say $z_{(2),1}$ and $z_{(2),2}$, respectively, and performs the logical AND operation on each pair of corresponding bits thereof. For each pair of bits, the result is 1 if both bits are 1, and 0 otherwise. To illustrate the bitwise AND operation, consider the integers, $z_{(10),1} = 550$ and $z_{(10),2} = 300$. It is,

$$\begin{aligned}\left(z_{(10),1} = 550\right) \text{ AND } \left(z_{(10),2} = 300\right) &= \left(1000100110_{(2)}\right) \text{ AND } \left(0100101100_{(2)}\right) \\ &= 0000100100_{(2)} \\ &= 36_{(10)}.\end{aligned}$$

Now, our BSK1 kernel sorts a sequence of length 2^p, $p \geq 10$, into bitonic subsequences, where each subsequence is of length 2^{10}. Each bitonic subsequence is the result of a single thread block sorting routine. Each thread block sorting routine is performed within shared memory upon the device, resulting in exceptionally efficient pairwise element swapping (i.e., pairwise element exchange/sorting). However, since thread blocks cannot communicate amongst each other, we require some subsequent CUDA kernel to merge the bitonic subsequences together. The BSK2 and BSK3 kernels perform this task. Algorithm B.3 describes the methodology for implementation of parallel bitonic sort.

The host call to BSK1 is `bitonicSort1<<< B, T >>>(d_Rand, d_index, sml, col_offset)`, where: `d_Rand` is the returned matrix from implementation of Algorithm B.1; `d_index` is a matrix whose respective row and column dimensions equal those of the matrix `d_Rand`, where it is assumed

that $p \geq 10$ (recall, 2^p is the row dimension of `d_Rand`); `sml` is the CUDA *shared memory limit* for pairwise element comparisons within BSK1 (specifically, we assign `sml` to the value of $1024 = 2^{10}$); `col_offset` is a parameter used to offset the columns upon each of `d_Rand` and `d_index`; the number of blocks for BSK1, `B`, is equal to $\max\{1, 2^p/\texttt{sml}\}$; and `T` is the number of threads-per-block for BSK1, equal to $\min\{2^{p-1}, \texttt{sml}/2\}$.

Algorithm B.3 Parallel Bitonic Sort Implementation

1. Implement Algorithm B.1. Allocate a matrix `d_index`, of size equal to that of `d_Rand`, within device memory. Unless otherwise stated, the pairwise arguments for all bitwise AND operations to follow are assumed base-10 integers.

2. Let n_c denote the number of columns for the matrix `d_Rand` and let $p \in \mathbb{N}$, $p \geq 10$, satisfy $n \leq 2^p$.

   1: **for** $k = 0$ **to** $n_c - 1$ **do**
   2: Invoke BSK1, comprised of $\texttt{B} = 2^{p-10}$ blocks and $\texttt{T} = 2^9 = 512$ threads per block within the host as follows:
   3: **for** $s = 0$ **to** $\texttt{B} - 1$ **in parallel do**
   4: **for** $t = 0$ **to** $\texttt{T} - 1$ **in parallel do**
   5: Allocate shared memory object `s_val`, to warehouse elements from mMTK return. Allocate shared memory object `s_key`, to warehouse the column labels (keys) of $\mathbf{G}^{(*\rho)}$ to be ordered. Copy device memory to shared memory.
   6: `s_val[t]` $\leftarrow$ `d_Rand[k + (t + 2sT)(`n_c`)]`. {Load a $U(0,1)$ deviate.}
   7: `s_key[t]` $\leftarrow t + 2s\texttt{T}$. {Initialize a column label.}
   8: `s_val[t + T]` $\leftarrow$ `d_Rand[k + (t + 2sT + T)(`n_c`)]`.
   9: `s_key[t + T]` $\leftarrow t + 2s\texttt{T} + \texttt{T}$.
   10: $d \leftarrow 2$. {Initialize the stride for pairwise comparisons.}
   11: **while** $d < 2\texttt{T}$ **do**
   12: {Continue until sequence of keys is bitonic.}
   13: **if** t AND $(d/2)$ is not equal to zero **then**
   14: $c \leftarrow 1$. {Corresponding thread sorts in ascending order.}
   15: **else**
   16: $c \leftarrow 0$. {Corresponding thread sorts in descending order.}

17: **end if**

18: $h \leftarrow d/2$. {Initialize stride length between pairwise elements.}

19: **while** $h \geq 1$ **do**

20: {Continue until stride is of length one.}

21: Synchronize threads.

22: $l \leftarrow 2t-(t$ AND $(h-1))$. {Which elements in the sequence does the thread examine?}

23: **if** (`s_val[l]` $>$ `s_val[l + h]`) equals c **then**

24: {Does the thread swap comparative values?}

25: Swap the values `s_val[l]` and `s_val[l + h]`.

26: Swap the values `s_key[l]` and `s_key[l + h]`.

27: **end if**

28: $h \leftarrow h/2$. {Halve the stride between paired elements.}

29: **end while**

30: $d \leftarrow 2d$. {Double the number of elements comprising a bitonic subsequence.}

31: **end while**

32: $c \leftarrow s$ AND 1. {Determine the sorting order of the bitonic sequence.}

33: $h \leftarrow$ `T`. {Initialize the stride between paired elements.}

34: Repeat the while loop upon lines 19-29 above.

35: Synchronize threads and copy shared memory to device memory.

36: `d_index[k + (t + 2sT)(`n_c`)]` $\leftarrow$ `s_key[t]`.

37: `d_Rand[k + (t + 2sT)(`n_c`)]` $\leftarrow$ `s_val[t]`.

38: `d_index[k + (t + 2sT + T)(`n_c`)]` $\leftarrow$ `s_key[t + T]`.

39: `d_Rand[k + (t + 2sT + T)(`n_c`)]` $\leftarrow$ `s_val[t + T]`.

40: **end for**

41: **end for**

42: $bsl \leftarrow$ `2T` {Initialize bitonic subsequence length (bsl).}

43: **while** $bsl < 2^p$ **do**

44: {Continue until the bitonic sequence is of length 2^p.}

45: $hs \leftarrow bsl/2$ {Initialize host stride (hs).}

46: **while** $hs \geq 1$ **do**

{Continue until host stride is one.}
if $hs \geq 2\mathtt{T}$ **then**
    Invoke BSK2 comprised of $\mathtt{B}_1 = 2\mathtt{B}$ blocks and $\mathtt{T}_1 = \mathtt{T}/2$ threads per block within the host as follows:
    for $s = 0$ **to** $\mathtt{B}_1 - 1$ **in parallel do**
        for $t = 0$ **to** $\mathtt{T}_1 - 1$ **in parallel do**
            $l \leftarrow t + s\mathtt{T}_1$. {Which order does the thread sort?}
            if l AND $(bsl/2)$ is not equal to zero **then**
                $c \leftarrow 1$. {Sort elements in ascending order.}
            else
                $c \leftarrow 0$. {Sort elements in descending order.}
            end if
            $l \leftarrow 2l - (l$ AND $(hs - 1))$. {Which elements does thread load into the kernel?}
            $v_1 \leftarrow$ `d_Rand[k + (l)(`n_c`)]`. {Load $U(0, 1)$ deviate.}
            $v_2 \leftarrow$ `d_Rand[k + (l + hs)(`n_c`)]`.
            $k_1 \leftarrow$ `d_index[k + (l)(`n_c`)]`. {Load a column index element.}
            $k_2 \leftarrow$ `d_index[k + (l + hs)(`n_c`)]`.
            if $(v_1 > v_2)$ equals c **then**
                Swap the values of v_1 and v_2.
                Swap the values of k_1 and k_2.
            end if
            `d_Rand[k + (l)(`n_c`)]` $\leftarrow v_1$. {Copy ordered elements out to device memory.}
            `d_Rand[k + (l + hs)(`n_c`)]` $\leftarrow v_2$.
            `d_index[k + (l)(`n_c`)]` $\leftarrow k_1$.
            `d_index[k + (l + hs)(`n_c`)]` $\leftarrow k_2$.
        end for
    end for
else

74: Invoke BSK3 comprised of `B` blocks and `T` threads per block within the host as follows:

75: **for** $s = 0$ **to** `B` $- 1$ **in parallel do**

76: **for** $t = 0$ **to** `T` $- 1$ **in parallel do**

77: Allocate shared memory for $U(0, 1)$ deviates, denoted `s_val`. Allocated shared memory for column indices of $\mathbf{G}^{(*\rho)}$, denoted `s_key`.

78: $l \leftarrow t + 2s$`T`. {Which elements does the thread load into shared memory?}

79: `s_val[t]` $\leftarrow$ `d_Rand[k + (l)(`n_c`)]`. {Copy elements from device memory to shared memory.}

80: `s_val[t + T]` $\leftarrow$ `d_Rand[k + (l + T)(`n_c`)]`.

81: `s_key[t]` $\leftarrow$ `d_index[k + (l)(`n_c`)]`.

82: `s_key[t + T]` $\leftarrow$ `d_index[k + (l + T)(`n_c`)]`.

83: $l \leftarrow l - s$`T`. {Which order does thread sort?}

84: **if** l AND $(bsl/2)$ is not equal to zero **then**

85: $c \leftarrow 1$. {Sort elements into ascending order.}

86: **else**

87: $c \leftarrow 0$. {Sort elements into descending order.}

88: **end if**

89: $ds \leftarrow$ `T`. {Initialize device stride.}

90: **while** $ds \geq 1$ **do**

91: Synchronize threads.

92: $l \leftarrow 2t - (t$ AND $(ds - 1))$. {Which elements does thread compare?}

93: **if** (`s_val[l]` $>$ `s_val[l + ds]`) equals c **then**

94: Swap the elements `s_val[l]` and `s_val[l + ds]`.

95: Swap the elements `s_key[l]` and `s_key[l + ds]`.

96: **end if**

97: $ds \leftarrow ds/2$. {Halve the device stride.}

98: **end while**

99: Synchronize threads and copy shared memory to device memory.

100: $l \leftarrow t + 2s$`T`. {Which elements does thread copy out to device memory?}

101: `d_Rand[k + (l)(`n_c`)]` $\leftarrow$ `s_val[t]`. {Copy $U(0,1)$ deviate to device memory.}

102: `d_Rand[k + (l + T)(`n_c`)]` $\leftarrow$ `s_val[t + T]`.

103: `d_index[k + (l)(`n_c`)]` $\leftarrow$ `s_key[t]`. {Copy an index element to device memory.}

104: `d_index[k + (l + T)(`n_c`)]` $\leftarrow$ `s_key[t + T]`.

105: **end for**

106: **end for**

107: **end if**

108: $hs \leftarrow (hs)/2$. {Halve the host stride.}

109: **end while**

110: $bsl \leftarrow 2(bsl)$. {Double the length of bitonic subsequences.}

111: **end while**

112: **end for**

3. Insofar as the sequence $U_{0,k}, \ldots, U_{2^p-1,k}$, each $k = 0, \ldots, n_c - 1$, is sorted into decreasing order and the elements $U_{n,k}, \ldots, U_{2^p-1,k}$ (for the unsorted sequence $U_{0,k}, \ldots, U_{2^p-1,k}$) are initialized to the value of minus one (-1) (see Algorithm B.1), the initial n elements upon column k of the matrix `d_index` comprise a random permutation of the column indices for $\mathbf{G}^{(*\rho)}$. That is, one call to Algorithm B.1 and one pass through the first two steps of [this] Algorithm B.3, yields a total of n_c random permutations of the column indices for $\mathbf{G}^{(*\rho)}$.

B.1.2 Contingency Table Construction

Given a random permutation of the column indices for $\mathbf{G}^{(*\rho)}$, here we develop a parallel processing approach to generating, say, the control row upon the 2×3 table at SNP locus j. Insofar as the column margin for the table at locus j is fixed (i.e., permutation invariant), reconstruction of a single row of the table is sufficient for: reconstruction of the table, and computation of the CATT statistic under $H_0^{(j)}$ (2.5). To construct the control rows across the m tables, we create a CUDA C kernel, denoted CTK. In brief, given a random permutation of the column indices for $\mathbf{G}^{(*\rho)}$, the

threads upon block s of CTK, $s = 0, \ldots, \texttt{B} - 1$: decompress the random genotype data upon loci $s\rho + 1, \ldots, (s+1)\rho$, and construct the control rows for the 2×3 tables at these loci. Algorithm B.4 provides the details for implementation of CTK.

The host call for CTK is

```
controw_rand_perm<<< B, T >>> (d_compdata, d_index, index_offset,
                               n, cols_d_index, iters_perthread,
                               resid_iters, d_controw),
```
(B.6)

where: `d_compdata` is a $m' \times n$ matrix, where $m' = m/\rho$ and ρ are defined within §2.3.2; `d_index` is the returned matrix from implementation of Algorithm B.3; `index_offset` is a parameter to offset column reads upon `d_index`; `n` is the GWAS sample size; `cols_d_index` is the column dimension of the matrix `d_index`; `iters_perthread` depicts the number of data reads upon the columns of `d_compdata` each thread of the kernel will undergo; assuming the number of controls, n_0, is not a multiple of `T`, the parameter `resid_iters` represents the number of kernel threads which read-in the final (i.e., residual) control data upon the columns of `d_compdata`; `d_controw` is the returned $m \times 3$ matrix of control rows across the m tables for a given permutation of the columns upon $\mathbf{G}^{(*\rho)}$; `B`, the number of blocks for the kernel, is equal to the row dimension of `d_compdata`, m'; and `T` is the number of threads per block of the kernel.

Algorithm B.4 Contingency Table Construction

1. Copy the [compressed] genotype matrix $\mathbf{G}^{(*\rho)}$ to device memory as object `d_compdata`. Allocate device memory to warehouse the $m \times 3$ matrix of control rows, `d_controw`. If thread t within block s of CTK reads control data upon the $(s+1)^{\text{st}}$ row of `d_compdata`, $t = 0, \ldots, \texttt{T} - 1$ and $s = 0, \ldots, \texttt{B} - 1$, then said thread reads a total of $\lfloor n_0/\texttt{T} \rfloor + I\left(t < n_0 - (\texttt{T})\lfloor n_0/\texttt{T} \rfloor\right)$ elements from said row of `d_compdata`. Hence, let $\delta = \lfloor n_0/\texttt{T} \rfloor$ and let $r_t = n_0 - (\texttt{T})\lfloor n_0/\texttt{T} \rfloor$. If $r = 1, \ldots, R$ indexes the permutations of the columns upon $\mathbf{G}^{(*\rho)}$, then let $o = r - 1$ denote the column offset for reads upon the columns of `d_index`, where it is assumed that the column dimension of `d_index` is $R = n_c$, where n_c is as previously defined within Algorithm B.3.

2. Invoke the CTK, comprised of $\texttt{B} = m'$ blocks and, say $\texttt{T} = 64$ threads[1] per block, within the host as follows:

[1] Here, we assign two (2) warps (64 threads) per thread block, because each multiprocessor of the NVIDIA GeForce GTX 470 GPU is comprised of two warp schedulers [69].

1: **for** $s = 0$ **to** $\texttt{B} - 1$ **in parallel do**

2: **for** $t = 0$ **to** $\texttt{T} - 1$ **in parallel do**

3: Allocate a shared memory object, denoted `s_compdata`, to warehouse data reads from the device memory object `d_compdata`.

4: **for** $k \in \mathcal{G} = \{0, 1, 2\}$ **do**

5: Allocate a shared memory object, denoted `s_sum_k`, to generate the appropriate control rows for thread t of block s.

6: **end for**

7: **for** $d = 0$ **to** $\rho - 1$ **do**

8: **for** $k \in \mathcal{G}$ **do**

9: `s_sum_k[t + dT]` $\leftarrow 0$. {Initialize shared memory.}

10: **end for**

11: **end for**

12: **for** $d = 0$ **to** $\delta - 1$ **do**

13: `s_compdata[t]` $\leftarrow$ `d_compdata[d_index[o + (t + dT)(`n_c`)]+(s)(n)]` {device-to-shared memory copy.}

14: **for** $h = 0$ **to** $\rho - 2$ **do**

15: **for** $k \in \mathcal{G}$ **do**

16: **if** $\left\lfloor \frac{\texttt{s\_compdata[t]} + 4 - k}{4} \right\rfloor = \left\lfloor \frac{\texttt{s\_compdata[t]} + 7 - k}{4} \right\rfloor$ **then**

17: `s_sum_k[t + hT]` $\leftarrow$ `s_sum_k[t + hT]`$+1$. {Decompress data; $g^*_{\{s\rho+h+1\}i} = k$ (see Proposition A.3 and Corollary A.1), where $i \leq n_0$ corresponds to some column index upon the initial n_0 columns of $\mathbf{G}^{(*\rho)}$}.

18: **end if**

19: **end for**

20: `s_compdata[t]` $\leftarrow \lfloor$`s_compdata[t]`$/4\rfloor$. {Update pursuant to Corollary A.1.}

21: **end for**

22: **for** $k \in \mathcal{G}$ **do**

23: **if** `s_compdata[t]` $= k$ **then**

24: `s_sum_k[t + (`ρ `- 1)T]` $\leftarrow$ `s_sum_k[t + (`ρ `- 1)T]` $+ 1$. {Final comparison per Proposition A.3.}

25: **end if**

26: **end for**

27: **end for**

28: **if** $t < r_t$ **then**

29: `s_compdata[t]` $\leftarrow$ `d_compdata[d_index[o + (t + (d`ρ`)(T))(`n_c`)]+(s)(n)]`.

30: Repeat lines 14-27 above.

31: **end if**

32: Synchronize threads.

33: **for** $d = 0$ **to** $\rho - 1$ **do**

34: $h \leftarrow$ `T`$/2$. {Initialize parallel reduction}.

35: **if** $t < h$ **then**

36: **for** $k \in \mathcal{G}$ **do**

37: `s_sum_k[t + dT]` $\leftarrow$ `s_sum_k[t + dT]` $+$ `s_sum_k[t + h + dT]`.

38: **end for**

39: **end if**

40: **if** $h > 1$ **then**

41: $h \leftarrow h/2$. Synchronize threads. Proceed to line 35.

42: **end if**

43: **end for**

44: Synchronize threads.

45: **if** $t < 3\rho$ **then**

46: $k \leftarrow \lfloor t/\rho \rfloor$. {Which shared memory vector do we read from?}

47: $d \leftarrow t - \rho k$. {Which row upon the shared vector do we read from?}

48: `d_controw[t + 3`ρ`s]` $\leftarrow$ `s_sum_k[dT]`. {Copy shared memory to device memory.}

49: **end if**

50: **end for**

51: **end for**

3. By combining the initial column upon each of the shared memory objects `s_sum_k` within block s of the kernel, each $k \in \mathcal{G}$ and $s = 0, \ldots, m' - 1$, we have obtained the constructed control rows which correspond with the genotype data upon rows $s\rho + 1, \ldots, (s+1)\rho$ of $\mathbf{G}^{(*\rho)}$ for a random permutation of the columns upon this genotype matrix.

B.1.3 Test Statistic Computation

Having constructed the control rows of the 2×3 tables across the m loci for, say, the r^{th} permutation of the columns upon $\mathbf{G}^{(*\rho)}$, some $r = 1, \ldots, R$, we are poised to compute the test statistics $\{t_{j,r}\}_{j=1}^{m}$. For simplicity, we demonstrate the application of the maxT MTP, but note here that provided one has correctly identified the null distribution for T_j under $\mathcal{H}_0$, the minP MTP essentially entails the implementation of one additional step, namely the computation of $p_{j,r}$ under $\mathcal{H}_0$. To calculate the realization of T_j (2.5), $t_{j,r}$, all $j = 1, \ldots, m$, we create a CUDA C kernel, denoted TSK. In brief, each thread upon TSK computes $t_{j,r}$, some $j = 1, \ldots, m$. Algorithm B.5 outlines the details for the implementation of TSK.

The host call for TSK is `maxT_CATT<<< B, T >>> (d_controw, d_cmargin, d_calcTS, offset)`, where: `d_controw` is the returned $m \times 3$ matrix of control row from implementation of Algorithm B.4; `d_cmargin` is a $m \times 3$ matrix whose j^{th} row warehouses the column margin of the 2×3 table for locus j; `d_calcTS` is the m-vector of returned test statistics, $(t_{1,r}, \ldots, t_{m,r})$; `offset` is used to offset thread reads upon the rows of `d_controw`, provided that the value of m is not a multiple of the size of a CUDA warp (i.e., 32 threads); `B` is the number of blocks for TSK; and `T` is the number of threads per block for TSK.

Algorithm B.5 Test Statistic Calculation

---

1. Allocate the device memory object `d_cmargin`. Copy the elements comprising the column margins upon $\mathbf{G}^*$ to said device memory object. Allocate device memory to warehouse the m-vector of computed test statistics, `d_calcTS`.

2. Here, we consider implementation of TSK whose number of threads per block, `T`, satisfies `T` ≤ 512. Two calls to TSK are invoked: the first call, comprises `B` $= \lfloor m/512 \rfloor$ blocks of `T` $= 512$ threads each; and the second call, comprises `B` $= 1$ block of `T` $= m - (512)\lfloor m/512 \rfloor$ threads. Assign d, a parameter used to offset thread reads upon the rows of `d_controw`, to the value of $m -$ `T`.

3. Upon the initial call to TSK:

   for $s = 0$ **to** `B` $- 1$ **in parallel do**

   for $t = 0$ **to** `T` $- 1$ **in parallel do**

   Read the elements of row $512s + t + 1$ upon each of the matrices `d_cmargin` and `d_controw`. Compute $t_{\{512s+t+1\},r}$ in accordance with (A.3), the equivalent form of

T_j (2.5) under the additive GMI. Store the computed value to element $512s + t + 1$ of the vector `d_calcTS`.

end for

end for

4. Upon the second call to TSK:

for $t = 0$ **to** $\mathtt{T} - 1$ **in parallel do**

Read the elements of row $d + t + 1$ upon each of the matrices `d_cmargin` and `d_controw`. Compute $t_{\{d+t+1\},r}$ in accordance with (A.3). Store the computed value to element $d + t + 1$ of the vector `d_calcTS`.

end for

B.1.4 Locating the Maximum Test Statistic: Parallel Reduction

The final parallel component to GPER lies with locating the maximum test statistic amongst the collection $\{t_{j,r}\}_{j=1}^{m}$, some $r = 1, \ldots, R$ (see step 5 upon the GPER pseudocode §2.4). To do this, we will make use of a parallel processing technique called *reduction*. Reduction uses an algorithmic pattern that arises often in parallel computing: *balanced trees* [71]. The idea is to build a balanced binary tree upon the collection $\{t_{j,r}\}_{j=1}^{m}$ – where without loss of generality we [temporarily] assume $m = 2^p$, some $p \in \mathbb{N}$ – and sweep it to locate the maximum test statistic value. In the case of reduction for this set of m elements, a binary tree can be depicted as a $(\log_2(m)+1)$-level tree based structure of connected nodes, such that each disjoint pairing – of the $2^{m-(x+1)}$ total pairings – of adjacent nodes at level x of the tree (the pair of nodes in which we denote as parent nodes) extend and connect to a single common node upon level $x + 1$ of the tree (this node is denoted as the child node) by way of a pair of arcs, each $x = 0, \ldots, \log_2(m) - 1$. Figure B.1 displays an example of such a tree based structure for $m = 8$.

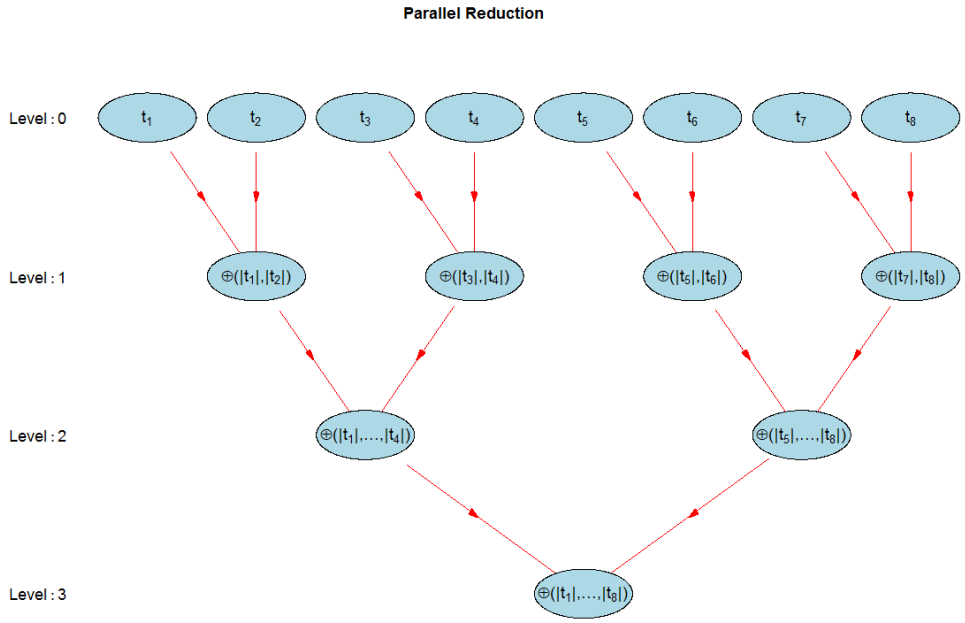

Fig. B.1: A Binary Tree of Connected Nodes for Parallel Reduction of the Elements, t_j, $j = 1, \ldots, 8$, Where t_j Represents a Realization of (2.5). Each Pair of Arrows (Arcs) Extends from Two Disjoint Parent Nodes at Level x of the Tree and Terminates upon a Common Child Node at Level $x + 1$ of the Tree, Some $x \in \{0, 1, 2\}$. The Child Node Warehouses the Resultant from Applying the Binary Operator $\oplus$ upon the Values Comprising Its Parent Nodes.

Here, the reduction begins by assigning the j^{th} parental node at the initial level ($x = 0$) of the tree to the value of $t_{j,r}$. We progress through the levels of the tree, by assigning to each child node the maximum value of its corresponding parent nodes. Reduction terminates upon the child node of level $\log_2(m)$, where it is noted that the assigned value for said child node is the desired sought maximum value (result) over the collection $\{t_{j,r}\}_{j=1}^{m}$.

This tree based structure is well suited to a CUDA C approach, because we can assign a thread to compute the maximum value upon each of the parental node pairings, irrespective of the level of the tree. Unfortunately, locating the maximum of the collection $\{t_{j,r}\}_{j=1}^{m}$ within CUDA C is not quite as simple as depicted above for two reasons: first, it is unlikely that m is exactly equal to 2^p, some $p \in \mathbb{N}$. Thus, we need to somehow modify this binary tree approach to incorporate collections $\{t_{j,r}\}_{j=1}^{m}$ for m taking any possible value over $\mathbb{N}$; and second, we need to use multiple thread blocks within the kernel we develop for reduction, so that all of the multiprocessors of the GPU are active [72]. However, each thread block has no means by which to broadcast its corresponding

result to other thread blocks of the kernel. Hence, we need to somehow modify this binary tree approach, so that thread blocks can communicate amongst each other.

First, consider the latter problem. If we could synchronize across thread blocks of the kernel (called *global synchronization*), such that the synchronization occurs upon all thread blocks completing their corresponding operations, then reduction could continue in a recursive manner. However, CUDA does not possess the ability for [device] global synchronization, because it is expensive to build into hardware for GPUs with high processor counts [72]. To resolve the problem, [72] suggests multiple kernel executions, because kernel execution acts as a [host] global synchronization point. Furthermore, kernel execution has negligible hardware overhead and low software overhead [72].

Here, to resolve the latter problem, we adopt the aforementioned suggestion of [72], and denote our reduction kernel by MTSK (shorthand for maximum test statistic kernel). In doing this, the former problem can be resolved as follows. We partition the collection of test statistics $\{t_{j,r}\}_{j=1}^{m}$ into blocks of size, say, 2^{10} elements each – here, we choose to invoke a maximum of 512 threads (12 warps) per thread block of MTSK. Within each thread block of MTSK, 2^{10} parental nodes are created upon the initial level of a binary tree, and reduction is carried out upon the tree, where threads evaluate particular parental node pairings. To this end, reduction has lead to each thread block comprising a maximum test statistic upon the child node at level ten of its corresponding binary tree, and a global synchronization point upon the host has been reached. These maximum test statistic values are aggregated, along with any test statistics not having been processed by some thread block. We partition these test statistic elements and invoke MTSK. This recursive process of partitioning test statistic elements and executing MTSK, continues until which time the execution of MTSK is comprised of a single thread block and all test statistics are processed within the thread block. Algorithm B.6 summarizes our implementation of MTSK.

The host call for MTSK is `max_list<<< B, T >>> (d_list1, d_list2, blockoffset, tsrem, d_maxT, permindex)`, where: `d_list1` is the list of test statistics to be processed upon some binary tree; `d_list2` is the list of test statistics which require processing upon subsequent calls to MTSK; `blockoffset` is a parameter to offset test statistic reads upon `d_list1`, so that test statistics are read into the proper thread block; `tsrem` represents the number of test statistics within `d_list1` which are not processed upon any thread blocks of MTSK; `d_maxT` is the returned maximum test statistic value for the collection $\{t_{j,r}\}_{j=1,\ldots,m}$; `permindex` is the value of r; `B` is the number of thread blocks; and `T` is the number of threads per block.

Algorithm B.6 Locating and Retrieving the Maximum Test Statistic

1. Let k index the invocations over MTSK (the value of k denotes the number of 'visits' to this step of the algorithm), and let d_k and r_k be defined by

$$d_k = \max\left\{l \in \mathbb{N}, l \leq 10 : \left\lfloor \frac{m_k}{2^l} \right\rfloor \in \mathbb{N}\right\}; \text{ and}$$
$$r_k = m_k - \left\lfloor \frac{m_k}{2^{d_k}} \right\rfloor \left(2^{d_k}\right),$$

   where $m_1 = m$. Essentially, the values of d_k and r_k represent the respective numbers of test statistics assigned to some thread block and test statistics not assigned to any thread block (i.e., $r_k > 0$ whenever the value of m_k is not divisible by a power of two).

2. If $k = r = 1$, then: (a) initialize device objects (vectors) `d_list1` and `d_list2`, each of length

$$m_1 - \left\lfloor \frac{m_1}{2^{d_1}} \right\rfloor \left(2^{d_1} - 1\right).$$

   These memory objects serve as data repositories for updated maximum test statistics upon successive calls to MTSK. Their lengths equal the number of test statistics which require processing upon the second call to MTSK; and (b) initialize the device object (vector) `d_maxT`, of length R, which will warehouse the maximum test statistics over the R column permutations of $\mathbf{G}^{(*\rho)}$.

3. Assign the parameters `tsrem` and `permindex` to the respective values of r_k and r. Invoke MTSK, with `B` $= \lfloor m_k/2^{d_k} \rfloor$ and `T` $= 2^{d_k - 1}$, as follows: if $k = 1$, then it is

```
max_list<<< B, T >>> (d_calcTS, d_list1, 2T, tsrem, d_maxT, permindex),
```

   where `d_calcTS` is the returned vector of calculated test statistics from implementation of Algorithm B.5; if $k = 2l$, for some $l \in \mathbb{N}$, then it is

```
max_list<<< B, T >>> (d_list1, d_list2, 2T, tsrem, d_maxT, permindex);
```

   otherwise, it is

```
max_list<<< B, T >>> (d_list2, d_list1, 2T, tsrem, d_maxT, permindex).
```

4. (Parallel reduction) Within each of the thread blocks, initialize a shared memory object (vector) of length `2T`, denoted `s_nodes`, the elements of which depict the parent nodes of the initial level of a binary tree.

1: **for** $s = 0$ **to** `B` $- 1$ **in parallel do**
2: **for** $t = 0$ **to** `T` $- 1$ **in parallel do**
3: **for** $c = 0$ **to** 1 **do**
4: `s_nodes[2t + c]` $\leftarrow$ `d_listx[2t + (s)(offset) + c]`, where `d_listx` is the first parameter within the call to MTSK (e.g., `d_calcTS` for $k = 1$, `d_list1` for $k = 2$, etc. – see step 3 above).
5: **end for**
6: $h \leftarrow$ `T`$/2$. Synchronize threads.
7: **if** $t < h$ **then**
8: `s_nodes[t]` $\leftarrow \max\{$`s_nodes[t]`$,$ `s_nodes[t + h]`$\}$.
9: **end if**
10: **if** $h > 1$ **then**
11: $h \leftarrow h/2$. Synchronize threads. Proceed to line 7.
12: **else if** `B` $= 1$ **and** $r_k = 0$ **then**
13: `d_maxT[r-1]` $\leftarrow$ `s_nodes[0]`. Terminate algorithm.
14: **else**
15: `d_listy[s]` $\leftarrow$ `s_nodes[0]`, where `d_listy` is the second parameter within the call to MTSK (i.e., `d_list1` for $k = 2l - 1$ and `d_list2` for $k = 2l$, where $l \in \mathbb{N}$ – see step 3 above).
16: **end if**
17: **if** $s =$ `B` $- 1$ **then**
18: **for** $c = 0$ **to** $r_k - 1$ **do**
19: `d_listy[B + c]` $\leftarrow$ `d_listx[2BT + c]`.
20: **end for**
21: $m_{k+1} \leftarrow r_k + \left\lfloor \dfrac{m_k}{2^{d_k}} \right\rfloor$.
22: Proceed to step 1 above.
23: **end if**
24: **end for**

25: **end for**

Table B.2 summarizes the dynamics in the values of m_k, d_k, and r_k, upon implementing Algorithm B.6 against $m = 769\,672$ SNP markers of a GWAS data set (see §2.6 for data set details).

Table B.2: The Dynamics of Algorithm B.6 Applied Against a GWAS Data Set Comprised of $m = 769\,672$ SNP Markers.

k	m_k	d_k	r_k	B	T
1	769 672	10	648	751	512
2	1399	10	375	1	512
3	376	8	120	1	128
4	121	6	57	1	32
5	58	5	26	1	16
6	27	4	11	1	8
7	12	3	4	1	4
8	5	2	1	1	2
9	2	1	0	1	1

B.2 Efficient Generation of the *P*-value Lookup Table

Here, we provide the underlying details for the CUDA kernels which encompass Algorithm 3.5. In particular, given: some [initial value] δ (essentially, can be thought of as δ_1 within Algorithm 3.4); the collection $\mathcal{O}$ of vectors $\boldsymbol{\theta} = \left(\pi^{aa}, \pi^{Aa}\right)$; the collection of upper interval endpoints for the CATT statistic, $\mathcal{T}$; and user specified precision ϵ for the estimates of $p^{\text{o}}_{u,w} = [\mathbf{P}]_{u,w}$, here we develop the CUDA kernels of Algorithm 3.5 which assist in generating $\mathbf{P}^{\epsilon}$.

Algorithm B.7 CUDA Kernel Pseudocode for Estimating the Pointwise *P*-value Lookup Table

---

1. (TURK1) For each $y \in \mathcal{Y} = \{0, 1\}$, let `d_gammay` warehouse the elements of Γ_y. Also, let `d_pi` warehouse those elements $\boldsymbol{\theta}_w$, such that $w \in \mathcal{W}$. We make a call to TURK1, for each $y \in \mathcal{Y}$, comprised of $\texttt{B} = n(\mathcal{W})$ blocks and $\texttt{T} = 256$ threads per block. Given the value in δ, the threads within block s, $s = 1, \ldots, \texttt{B} - 1$, will read over the elements within Γ_y and determine the value of $e(\Gamma(\boldsymbol{\theta}_{s+1}))$. Let $t_y = \lfloor n(\Gamma_y)/\texttt{T} \rfloor$ (i.e., the number of elements upon Γ_y processed by each thread within a given thread block) and for $t = 0, \ldots, \texttt{T} - 1$, let $r_y = I(t < n(\Gamma_y) - (\texttt{T})(t_y))$ (i.e., those threads which process 'residual' elements upon the collection Γ_y, whenever the number of elements for this collection is not divisible by T).

   1: **for** $y \in \mathcal{Y}$ **do**

2: Invoke TURK1 upon the host, whose pseudocode follows.
3: **for** $s = 0$ **to** $\texttt{B} - 1$ **in parallel do**
4: **for** $t = 0$ **to** $\texttt{T} - 1$ **in parallel do**
5: **if** $y = 0$ **and** $t = 0$ **then**
6: $e\left(\Gamma\left(\boldsymbol{\theta}_{s+1}\right)\right) \leftarrow 0$. {Initialize (3.14).}
7: **end if**
8: `s_sum[t]` $\leftarrow 0$. {Initialize shared memory.}
9: **for** $k = 0$ **to** $t_y - 1$ **do**
10: $\mathbf{z}_y \leftarrow$ `d_gammay[t + kT]` and $\boldsymbol{\theta} \leftarrow$ `d_pi[s]`.
11: **if** $h\left(\mathbf{z}_y|\boldsymbol{\theta}\right) \leq \delta$ **then**
12: `s_sum[t]` $\leftarrow$ `s_sum[t]` $+ h\left(\mathbf{z}_y|\boldsymbol{\theta}\right)$. {Increment appropriate sum upon (3.14).}
13: **end if**
14: **end for**
15: **if** $r_y = 1$ **then**
16: $\mathbf{z}_y \leftarrow$ `d_gammay[t + (`t_y`)(T)]` and $\boldsymbol{\theta} \leftarrow$ `d_pi[s]`
17: Repeat the conditional statement given by lines 11-13 above.
18: **end if**
19: $d \leftarrow \texttt{T}/2$. {Synchronize threads and prepare for reduction.}
20: **if** $t < d$ **then**
21: `s_sum[t]` $\leftarrow$ `s_sum[t]` $+$ `s_sum[t + d]`.
22: **end if**
23: **if** $d > 1$ **then**
24: $d \leftarrow d/2$; Synchronize threads; and, proceed to line 20.
25: **else if** $t = 0$ **then**
26: $e\left(\Gamma\left(\boldsymbol{\theta}_{s+1}\right)\right) \leftarrow e\left(\Gamma\left(\boldsymbol{\theta}_{s+1}\right)\right) +$ `s_sum[0]`.
27: **end if**
28: **end for**
29: **end for**
30: **end for**

2. (TURK2) Let $w \in \mathcal{W}$ and δ be given. For each $y \in \mathcal{Y}$, let `d_indy` be the device memory object,

such that the elements upon said object indicate those elements over the collection Γ_y which are also contained within the collection $\Gamma_y(\boldsymbol{\theta}_w)$. We make a call to TURK2, for each $y \in \mathcal{Y}$, comprised of `B` blocks and `T` threads per block, such that $\texttt{B} \times \texttt{T} = n(\Gamma_y)$. Given the value in δ, the threads over all blocks will: read over the elements within Γ_y; determine which of these elements are contained within the collection $\Gamma_y(\boldsymbol{\theta}_w)$; and, copy the appropriate elements from Γ_y to the collection $\Gamma_y(\boldsymbol{\theta}_w)$.

for $y \in \mathcal{Y}$ **do**

$n(\Gamma_y(\boldsymbol{\theta}_w)) \leftarrow 0$. {Initialize the number of elements contained within $\Gamma_y(\boldsymbol{\theta}_w)$.}

for $s = 0$ **to** $\texttt{B} - 1$ **in parallel do**

for $t = 0$ **to** $\texttt{T} - 1$ **in parallel do**

$h \leftarrow t + s\texttt{T}$. {Which element upon `d_gammay` does thread t examine?}

$\mathbf{z}_y \leftarrow$ `d_gammay[h]`.

if $h(\mathbf{z}_y|\boldsymbol{\theta}_w) > \delta$ **then**

`d_ind[h]` $\leftarrow 1$. {$\mathbf{z}_y \in \Gamma_y(\boldsymbol{\theta}_w)$.}

$n(\Gamma_y(\boldsymbol{\theta}_w)) \leftarrow n(\Gamma_y(\boldsymbol{\theta}_w)) + 1$.

else

`d_ind[h]` $\leftarrow 0$. {$\mathbf{z}_y \in \Gamma'_y(\boldsymbol{\theta}_w)$.}

end if

end for

end for

Allocate device memory object, `d_gammayw`, to warehouse the elements of $\Gamma_y(\boldsymbol{\theta}_w)$.

$h \leftarrow 0$.

for $d = 0$ **to** $n(\Gamma_y)$ **do**

if `d_ind[d]` $= 1$ **then**

`d_gammayw[h]` $\leftarrow$ `d_gammay[d]`. {Copy the appropriate element within `d_gammay[d]` to `d_gammayw[h]`.}

$h \leftarrow h + 1$.

end if

end for

end for

3. (TURK3) Here, given: $\boldsymbol{\theta}_w$; the collections $\Gamma_0(\boldsymbol{\theta}_w)$ and $\Gamma_1(\boldsymbol{\theta}_w)$; and $\tau_\iota \in \mathcal{T}$, we derive the value

$p_{\iota,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$. We make a call to TURK3, comprised of $\mathtt{B} = \min\left\{n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right), n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right)\right\}$ (for clarity, we assume $\mathtt{B} = n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right)$) blocks and $\mathtt{T} = 128$ threads per block. Let $t_1 = \left\lfloor n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right)/\mathtt{T}\right\rfloor$ (the number of elements upon the collection $\Gamma_1\left(\boldsymbol{\theta}_w\right)$ processed by each thread within a given thread block) and for $t = 0, \ldots, \mathtt{T}-1$, let $r_t = I(t < n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right) - (\mathtt{T})(t_1))$. Allocate device memory (vector) of length `B`, say `d_prob`, which will warehouse each blocks' contribution to the value $p_{\iota,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$. The pseudocode for TURK3 follows:

1: **for** $s = 0$ **to** $\mathtt{B} - 1$ **in parallel do**
2: **for** $t = 0$ **to** $\mathtt{T} - 1$ **in parallel do**
3: $\mathbf{z}_0 \leftarrow$ `d_gamma0w[s]`. {Load an element from $\Gamma_0\left(\boldsymbol{\theta}_w\right)$.}
4: `s_sum[t]` $\leftarrow 0$. {Initialize shared memory.}
5: **for** $d = 0$ **to** $t_1 - 1$ **do**
6: $\mathbf{z}_1 \leftarrow$ `d_gamma1w[t + dT]`. {Load an element from $\Gamma_1\left(\boldsymbol{\theta}_w\right)$.}
7: Under $\mathcal{H}_0$, compute the realization in the CATT statistic and denote it by $T\left(\mathbf{z}_0, \mathbf{z}_1\right)$.
8: **if** $T\left(\mathbf{z}_0, \mathbf{z}_1\right) \geq \tau_\iota$ **then**
9: `s_sum[t]` $\leftarrow$ `s_sum[t]` $+ g\left(\mathbf{z}_0, \mathbf{z}_1 | \boldsymbol{\theta}_w\right)$. {Contribute to p-value.}
10: **end if**
11: **end for**
12: **if** $r_t = 1$ **then**
13: $\mathbf{x}_1 \leftarrow$ `d_gamma1w[t + (`t_1`)(T)]`. {Load an element from $\Gamma_1\left(\boldsymbol{\theta}_w\right)$.}
14: Repeat lines 7-10 above.
15: **end if**
16: $h \leftarrow \mathtt{T}/2$. Synchronize threads. {Prepare for reduction.}
17: **if** $t < h$ **then**
18: `s_sum[t]` $\leftarrow$ `s_sum[t]` $+$ `s_sum[t + h]`.
19: **end if**
20: **if** $h > 1$ **then**
21: $h \leftarrow h/2$; synchronize threads; and, proceed to line 17.
22: **else if** $t = 0$ **then**
23: `d_prob[s]` $\leftarrow$ `s_sum[0]`. {Copy shared memory to device memory.}
24: **end if**

25: **end for**

26: **end for**

27: **for** $d = 1$ **to** $\mathtt{B} - 1$ **do**

28: $\mathtt{d\_prob[0]} \leftarrow \mathtt{d\_prob[0]} + \mathtt{d\_prob[d]}$.

29: **end for**

30: $p_{\iota,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right) \leftarrow \mathtt{d\_prob[0]}$.

4. (PPTK1) Here, for $w \in \mathcal{W}$, given the collection $\Gamma\left(\boldsymbol{\theta}_w\right)$ we compute the table probabilities $g\left(\mathbf{z}_0, \mathbf{z}_1 | \boldsymbol{\theta}_w\right)$ for all $\left(\mathbf{z}_0, \mathbf{z}_1\right) \in \Gamma\left(\boldsymbol{\theta}_w\right)$ and we evaluate (3.21) over said truncated unconditional reference set. As with TURK3, here we assume that $n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right) = \min\{n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right), n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right)\}$. Allocate two vectors within device memory, denoted `d_prob` and `d_lambda`, each of length $n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right)$ which will warehouse the respective returned values of $g\left(\mathbf{z}_0, \mathbf{z}_1 | \boldsymbol{\theta}_w\right)$ and (3.21), upon calling PPTK1. Our invocation of PPTK1 entails taking $\mathtt{B} = \lfloor n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right) / \mathtt{T} \rfloor$ blocks, where $\mathtt{T} = 512$ threads per block. For each $t = 0, \ldots, \mathtt{T}-1$, let $r_{t1} = I(t < n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right) - (\mathtt{T})(\mathtt{B}))$, denote: those threads which will process two elements upon the collection $\Gamma_1\left(\boldsymbol{\theta}_w\right)$ $(r_{t1} = 1)$; and, those threads which will process one element upon the collection $\Gamma_1\left(\boldsymbol{\theta}_w\right)$ $(r_{t1} = 0)$.

(PPTK2) Here, for each $\tau_u \in \mathcal{T}$, we evaluate (3.22) over the collections `d_prob` and `d_lambda` returned from PPTK1, where the PPTK2 returned values (i.e., $p_{u,w}\left(\Gamma\left(\boldsymbol{\theta}_w\right)\right)$, for all $u = 1, \ldots, n\left(\mathcal{T}\right)$) are to be contained within the allocated device memory object (vector) `d_Pw` (of length equal to that of $\mathcal{T}$). Said device memory object is to warehouse column w upon $\mathbf{P}^{\epsilon}$. Our invocation of PPTK2 entails taking $\mathtt{B}_1 = n\left(\mathcal{T}\right)$ blocks and $\mathtt{T}_1 = 64$ threads per block. For each $t = 0, \ldots, \mathtt{T} - 1$, let $t_2 = \lfloor n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right) / \mathtt{T}_1 \rfloor$, denote the number of elements upon the collection $\Gamma_1\left(\boldsymbol{\theta}_w\right)$ each thread (within each thread block) will process; and let $r_{t2} = I(t < n\left(\Gamma_1\left(\boldsymbol{\theta}_w\right)\right) - (\mathtt{T}_1)(\mathtt{B}_1))$ be an indicator for 'residual' thread processing over the collection $\Gamma_1\left(\boldsymbol{\theta}_w\right)$. The pseudocode for invocation of PPTK1 and PPTK2 follows:

1: **for** $d = 0$ **to** $n\left(\Gamma_0\left(\boldsymbol{\theta}_w\right)\right)$ **do**

2: Invoke PPTK1:

3: **for** $s = 0$ **to** $\mathtt{B} - 1$ **in parallel do**

4: **for** $t = 0$ **to** $\mathtt{T} - 1$ **in parallel do**

5: $\mathbf{z}_0 \leftarrow \mathtt{d\_gamma0w[d]}$.

6: $h \leftarrow t + s\mathtt{T}$.

7: $\mathbf{z}_1 \leftarrow$ `d_gamma1w[h]`.

8: Compute $T(\mathbf{z}_0, \mathbf{z}_1)$.

9: `d_lambda[h]` $\leftarrow$ evaluation of (3.21).

10: `d_prob[h]` $\leftarrow g(\mathbf{z}_0, \mathbf{z}_1|\boldsymbol{\theta}_w)$.

11: **if** $r_{t1} = 1$ **then**

12: $h \leftarrow t +$ `BT`.

13: Repeat lines 7-10 above.

14: **end if**

15: **end for**

16: **end for**

17: Invoke PPTK2:

18: **for** $s = 0$ **to** $\mathtt{B}_1 - 1$ **in parallel do**

19: **for** $t = 0$ **to** $\mathtt{T}_1 - 1$ **in parallel do**

20: `s_sum[t]` $\leftarrow 0$. {Initialize shared memory.}

21: **for** $h = 0$ **to** $t_2 - 1$ **do**

22: $c \leftarrow t + h\mathtt{T}_1$.

23: **if** `d_lambda[c]` $> s$ **then**

24: `s_sum[t]` $\leftarrow$ `s_sum[t]` + `d_prob[c]`.

25: **end if**

26: **end for**

27: **if** $r_{t2} = 1$ **then**

28: $c \leftarrow t + (t_2)(\mathtt{T}_1)$.

29: Repeat lines 23-25 above.

30: **end if**

31: $h \leftarrow \mathtt{T}_1/2$. Synchronize threads. {Prepare for reduction.}

32: **if** $t < h$ **then**

33: `s_sum[t]` $\leftarrow$ `s_sum[t]` + `s_sum[t + h]`.

34: **end if**

35: **if** $h > 1$ **then**

36: $h \leftarrow h/2$; synchronize threads; and, proceed to line 32.

37: **else if** $t = 0$ **then**

```
38:                 d_Pw[s] ← s_sum[t].
39:             end if
40:         end for
41:     end for
42: end for
```

APPENDIX C

R-PACKAGE FOR GEM

Here, we provide a brief description for the implementation of our proposed `R` package (tentatively denoted *GEM*). We assume that: a random sample of n_1 cases and n_0 controls has been obtained from the population of interest; a total of m tSNP loci have been sampled from amongst those within the human genome and genotyped across the n $(= n_0 + n_1)$ study subjects; and, data has been collected from each of the n study subjects upon either a binary or 3-level categorical environmental factor.

The package is essentially comprised of two functions, `GEM_2e` and `GEM_3e_flex`. The former function conducts the GEM methodology, as outlined within Chapter 4, upon a binary environmental factor. The latter function conducts the GEM methodology upon a 3-level environmental factor, and gives the user the task (i.e., flexibility) over the construction – both in the number of- and in the form of- – the candidate patterns. These functions each carry out Algorithm 4.1 and share a common set of 6 user input parameters (here, we use the prefix 'I' upon the following numbered list to signify that these are function inputs):

I1. The number of controls within the case-control sample, n_0.

I2. The number of study subjects within the sample, n.

I3. The number of tSNP loci, m.

I4. The data, consisting of an $m \times n$ matrix – rows correspond to tSNP loci and columns correspond to study subjects. The matrix is identical to that of $\mathbf{GE}^*$ (defined within §4.5.3), and should be ordered such that the initial/final n_0/n_1 columns comprise the control/case data for the realizations in the random variables X_j (4.2), $j = 1, \ldots, m$.

I5. The number of desired column permutations upon the matrix $\mathbf{GE}^*$, R.

I6. A vector, whose length is equal to the number of candidate genes under study, and whose i^{th} element equals the number of tSNPs sampled from the i^{th} candidate gene – the elements of this vector should sum to m.

In addition to the aforementioned 6 parameters, the function `GEM_3e_flex` requires 2 additional user input parameters:

I7. The $q \times 9$ [indicator] matrix, $\mathbf{I}$: as defined within step 1 of Algorithm 4.1, where $q = 16$ therein. This would allow for application of the GEM methodology as outlined within Chapter 4 upon a 3-level environmental factor; or, any variation of $\mathbf{I}$ the user desires, such that the $(l,k)^{\text{th}}$ matrix entry, denoted $[\mathbf{I}]_{(l,k)}$, is coded by

$$[\mathbf{I}]_{(l,k)} = I\left(k \in A_l\right) - I\left(k \in \left(A_l \cup B_l\right)'\right),$$

for all $l = 1, \ldots, q$ and $k = 1, \ldots, 9$.

I8. The row dimension of $\mathbf{I}$, q.

To illustrate application of the function `GEM_2e`, here we apply it against: the colon cancer data set (see §4.10.1), comprised of $n_1 = 1555$ cases of cancer and $n_0 = 1956$ healthy controls; the recent use of NSAIDs binary environmental factor; and, the 29 tSNPs upon the 4 candidate genes – 8 markers for the *EPX* gene, 4 markers for the *HIF1A* gene, 2 markers for the *MPO* gene, and 15 markers for the *NOS2A* gene. The data are assumed to reside within the ASCII file, `GEM_colon_nsaid.txt`, whose structure is 29 rows (the order of the SNPs are assumed in accordance with the aforementioned listing of candidate genes), each row with $n = 3511$ space-delimited columns. Each of the functions `GEM_2e` and `GEM_3e_flex` are accessed via the dynamic linked library (DLL), `GEM_DLL.dll` (this DLL file was compiled from C code and designed to be interfaced with `R`). The ASCII and DLL files are assumed to reside within the folder `C:\GEM_project\DLL`. The following `R` code changes the working directory and 'sources-in' the `DLL_test_main.R` file (see §D.2) – this file more-or-less prepares the `R` environment for use over the functions `GEM_2e` and `GEM_3e_flex`:

```
setwd('C:/GEM_project/DLL')
source('DLL_test_main.R')
```

The following `R` code reads in the contents of the ASCII file (`GEM_colon_nsaid.txt`) and calls the `GEM_2e` function, where $R = 100$K column permutations upon $\mathbf{GE}^*$ are requested – the return from the function call is stored within the `GEM_return` `R` object (list):

```
### INITIALIZE CPU TIME
#
beg.time.stamp = proc.time()[3]
#
```

```
### READ-IN THE CONTENTS OF THE ASCII FILE (STORE TO OBJECT x)
#
x = read.table('GEM_colon_nsaid.txt', header = F)
#
### NOW, MAKE THE CALL TO THE GEM_2e FUNCTION...
#
GEM_return = GEM_2e( n.controls        = 1956,
                     n.sample          = dim(x)[2],
                     n.snps            = dim(x)[1],
                     dat               = as.vector(t(as.matrix(x))),
                     n.perms           = 100000,
                     nsnps.per.geneset = c(8, 4, 2, 15) )
#
### FINALIZE CPU TIME
#
end.time.stamp = proc.time()[3]
#
### DISPLAY REQUIRED COMPUTATIONAL TIME FOR GEM_2e CALL (SECONDS)
#
end.time.stamp - beg.time.stamp
#
### UNLOAD THE .dll FILE
#
dyn.unload('GEM_DLL.dll')
```

The `GEM_return` `R` object is a list comprised of the following 6 elements (here, we use the prefix 'R' upon the following numbered list to signify that these are function returns – note that $m = 29$ and $q = 11$):

R1. A $m \times 10$ data frame summarizing the table margins for the 2×6 contingency table $\mathbf{X}_j$ (e.g., Table 4.5), $j = 1, \ldots, m$. The j^{th} row of the data frame essentially summarizes the random variable X_j for the j^{th} row of the ASCII file `GEM_colon_nsaid.txt`. Collectively, the values upon the initial 6 columns of the j^{th} row of the data frame represent the elements of the vector $\mathbf{c}_j$ (see (4.15)). The values upon the latter 4 columns of the j^{th} row of the data frame represent the respective values, $[\mathbf{M}]_{(j,8)}$, $[\mathbf{M}]_{(j,7)}$, $[\mathbf{M}]_{(j,9)}$, and $n_0 - [\mathbf{M}]_{(j,7)}$, where the $[\mathbf{M}]_{(j,k)}$ are as defined within step 1 of Algorithm 4.1.

R2. A $m \times q$ data frame, whose j^{th} row warehouses the standardized Wald-based test statistics, $Z_{jA_1}, \ldots, Z_{jA_q}$ (4.10), $j = 1, \ldots, m$.

R3. A $m \times q$ data frame, whose j^{th} row warehouses the estimated [locus-level adjusted] maxT permutation adjusted p-values, $\tilde{p}^*_{j1\sigma}, \ldots, \tilde{p}^*_{jq\sigma}$ (4.18), $j = 1, \ldots, m$.

R4. A $m \times q$ data frame, whose j^{th} row warehouses the estimated [gene-level adjusted] maxT permutation adjusted p-values, $\tilde{p}^*_{j1\mu}, \ldots, \tilde{p}^*_{jq\mu}$ (4.19), $j = 1, \ldots, m$.

R5. A $m \times q$ data frame, whose $(j, l)^{\text{th}}$ element warehouses the estimated log-odds ratio of colon cancer, comparing subjects satisfying candidate pattern A_l to subjects satisfying candidate pattern B_l, for all $j = 1, \ldots, m$ and $l = 1, \ldots, q$.

R6. A $m \times q$ data frame, whose $(j, l)^{\text{th}}$ element warehouses the standard error of the corresponding $(j, l)^{\text{th}}$ element of R5 above, for all $j = 1, \ldots, m$ and $l = 1, \ldots, q$.

APPENDIX D

SELECT PROGRAMMING CODE

D.1 Implementation of Algorithm 3.4

The following C program code reads in data from an ASCII file entitled, `input.txt`. The first line of the ASCII file must contain the numbers for each of cases and controls of the GWAS sample, where the corresponding data values within the file should be separated by a space. Each subsequent line of data within the ASCII file warehouses a particular realization of the elements upon the vector $\boldsymbol{\theta}_w = \left(\pi^{aa}, \pi^{Aa}\right)$, where w indexes the columns upon $\mathbf{P}$ (the pointwise p-value table). The values of π^{aa} and π^{Aa}, as listed within the ASCII file, should each: be separated by at least one space; and, be multiplied by 10^6 and stated as a counting (i.e., positive integer) value. The programming code follows:

```
// LOAD REQUIRED C HEADER FILES //
#include <stdlib.h>
#include <stdio.h>
#include <time.h>
#include <direct.h>
#include <process.h>
#include <conio.h>
#include <string.h>
#include <math.h>

// NATURAL LOGARITHM OF THE GAMMA FUNCTION -- NUMERICAL RECIPES //
double gammln(double xx)
{
  double x, y, tmp, ser;
  static double cof[6] = {76.18009172947146,
                          -86.50532032941677,
                          24.01409824083091,
                          -1.231739572450155,
                          0.1208650973866179e-2,
                          -0.5395239384953e-5
                         };
  int j;
  y = x = xx;
  tmp = x + 5.5;
  tmp -= (x + 0.5) * log(tmp);
  ser 0 =1.000000000190015;
  for(j = 0; j <= 5; j++) ser += cof[j] / ++y;
  return -tmp + log(2.5066282746310005 * ser / x);
}
```

```
// A FUNCTION TO SQUARE THE VALUE OF THE ARGUMENT x //
double SQR (double x)
{
  return x * x;
}

// LOCATE THE COLLECTION Gamma_y(theta_k) //
void probmargins( unsigned int n, unsigned int *row, double thresh, double *fact,
                  double *prob, unsigned long *num_elems_overall,
                  unsigned long *num_elems_new, char *row_ind_member,
                  unsigned long *row_ind_update, unsigned long *row_ind_old, double pi1,
                  double pi2, double *row_prob )
{
  unsigned long pos = 0, pos1 = 0, cnt = 0, pos2 = 0;
  double        egamthetak = 0.0, tmpprob;
  unsigned int  i, j;

  for(i = 0; i <= n; i++)
    for(j = 0; j <= (n - i); j++)
    {
      tmpprob  = (double) (fact[n] - fact[i] - fact[j] - fact[n - i - j]);
      tmpprob += (double) i * log( (double) pi1) + (double) j * log( (double) pi2) +
                 (double) (n - i - j) * log( (double) 1.0 - pi1 - pi2 );

      if(tmpprob >= thresh)
      {
        row[0 + 3 * pos] = i;
        row[1 + 3 * pos] = j;
        row[2 + 3 * pos] = n - i - j;
        row_prob[pos] = tmpprob;

        if(row_ind_member[cnt] == 0)
        {
          row_ind_member[cnt]++;
          row_ind_update[pos1] = pos;
          pos1++;
        }

        else
        {
          row_ind_old[pos2] = pos;
          pos2++;
        }

        pos++;
      }

      else
        egamthetak += (double) exp( (double) tmpprob);

      cnt++;
    }

  *num_elems_overall = pos;
  *num_elems_new     = pos1;
  *prob              = egamthetak;
}
```

```
// COMPUTE p_{i,k}(Gamma(theta_k))
void probtruncuncondrefset( unsigned int n1, unsigned int n2, unsigned int *row1,
                            unsigned int *row2, unsigned long pos1, unsigned long pos2,
                            unsigned long pos1_new, unsigned long pos2_new,
                            unsigned long *row1_update, unsigned long *row2_update,
                            unsigned long *row1_old, double test_stat_crit,
                            double var_common, double *pvalue, double *row1prob,
                            double *row2prob, double pval_init )
{
  unsigned int col_marg0, col_marg1, col_marg2;
  unsigned long i, j, deno;
  double TS_num, TS, tmpprob, PV;
  long num;

  PV = pval_init;
  for(i = 0; i < pos1_new; i++)
    for(j = 0; j < pos2; j++)
    {
      col_marg0 = (unsigned int) row1[0 + 3 * row1_update[i]] + row2[0 + 3 * j];
      col_marg1 = (unsigned int) row1[1 + 3 * row1_update[i]] + row2[1 + 3 * j];
      col_marg2 = (unsigned int) row1[2 + 3 * row1_update[i]] + row2[2 + 3 * j];

      deno = (unsigned long) 4 * col_marg0 * col_marg2 + col_marg0 * col_marg1 +
                             col_marg2 * col_marg1;

      if(deno)
      {
        num = n2 * (row1[0 + 3 * row1_update[i]] - row1[2 + 3 * row1_update[i]]) +
              n1 * (row2[2 + 3 * j] - row2[0 + 3 * j]);
        TS  = (double) var_common * SQR( (double) num) / deno;
      }
      else
        TS = 0.0;

      if( TS >= test_stat_crit )
        PV += (double) exp(row1prob[row1_update[i]] + row2prob[j]);

    }

  for(i = 0; i < (pos1 - pos1_new); i++)
    for(j = 0; j < pos2_new; j++)
    {
      col_marg0 = (unsigned int) row1[0 + 3 * row1_old[i]] +
                                 row2[0 + 3 * row2_update[j]];
      col_marg1 = (unsigned int) row1[1 + 3 * row1_old[i]] +
                                 row2[1 + 3 * row2_update[j]];
      col_marg2 = (unsigned int) row1[2 + 3 * row1_old[i]] +
                                 row2[2 + 3 * row2_update[j]];

      deno = (unsigned long) 4 * col_marg0 * col_marg2 + col_marg0 * col_marg1 +
                              col_marg2 * col_marg1;
```

```
      if(deno)
      {
        num = n2 * (row1[0 + 3 * row1_old[i]] - row1[2 + 3 * row1_old[i]]) +
              n1 * (row2[2 + 3 * row2_update[j]] - row2[0 + 3 * row2_update[j]]);
        TS  = (double) var_common * SQR( (double) num) / deno;
      }
      else
        TS = 0.0;

      if( TS >= test_stat_crit )
        PV += (double) exp(row1prob[row1_old[i]] + row2prob[row2_update[j]]);
    }

  *pvalue = (double) PV;
}

// MAIN SECTION //
int main(void)
{
  double *h_lfact, var_const, delta, tmpprob, egamthetak1, egamthetak2, t_iota, p_iotak;
  unsigned int i, j, n, n1, n2, *row1, *row2, tmp;
  unsigned long n_row1, n_row2, pos, pos1, pos2, pos1_new, pos2_new, *row1_ind_update,
                *row2_ind_update, *row1_ind_old, *row2_ind_old;
  FILE *rstream1, *wstream1;
  char filo[15], one[2];
  double pi1, pi2, prob, *row1_prob, *row2_prob, epsilon, p_asy, multiplier, dura,
         prob_old;
  char ind, *row1_ind_evermember, *row2_ind_evermember;
  clock_t start, finish;
  time_t curr = time(0);   // TIME STAMP

  // PREPARE HEADER ROW OF OUTPUT FILE //
  fopen_s(&wstream1, "results.txt", "w");
  fprintf_s(wstream1, "CASES CONTROLS PI1 PI2 ERROR_ROW1 ERROR_ROW2 NUM_ROW1 NUM_ROW2
                       P_VALUE_EST DELTA_S TIME\n");
  fclose(wstream1);

  // SAMPLE SIZE //
  // n_1 = cases; n_2 = controls
  fopen_s(&rstream1, "input.txt", "r");
  fscanf_s(rstream1, "%d", (unsigned int *) &n1);
  fscanf_s(rstream1, "%d", (unsigned int *) &n2);

  // LARGEST VALUE OF CA-TEST STATISTIC WITHIN THE PPT //
  t_iota = 40.0;

  // DESIRED RELATIVE ACCURACY OF P-VALUE ESTIMATES //
  epsilon = 0.9999;

  // ASYMP CHI-SQUARE P-VALUE FOR REALIZATION t_iota //
  p_asy = 2.54e-10;

  n = (unsigned int) n1 + n2;

  // COMMON COMPONENT OF VARIANCE FOR CA-TREND TEST STATISTIC //
  var_const = (double) n / (n1 * n2);
```

```
// NUMBER OF ELEMENTS FOR EACH ROW WITHIN THE UNCONDITIONAL REFERENCE SET //
n_row1 = (unsigned long) (n1 + 2) * (n1 + 1) / 2;
n_row2 = (unsigned long) (n2 + 2) * (n2 + 1) / 2;

// MEMORY ALLOCATION //
h_lfact             = (double *       ) malloc(   (n + 1) * sizeof(double)     );
row1                = (unsigned int * ) malloc(3 * n_row1 * sizeof(unsigned int) );
row2                = (unsigned int * ) malloc(3 * n_row2 * sizeof(unsigned int) );
row1_prob           = (double *       ) malloc(    n_row1 * sizeof(double)       );
row2_prob           = (double *       ) malloc(    n_row2 * sizeof(double)       );
row1_ind_evermember = (char *         ) malloc(    n_row1 * sizeof(char)         );
row2_ind_evermember = (char *         ) malloc(    n_row2 * sizeof(char)         );
row1_ind_update     = (unsigned long *) malloc(    n_row1 * sizeof(unsigned long));
row2_ind_update     = (unsigned long *) malloc(    n_row2 * sizeof(unsigned long));
row1_ind_old        = (unsigned long *) malloc(    n_row1 * sizeof(unsigned long));
row2_ind_old        = (unsigned long *) malloc(    n_row2 * sizeof(unsigned long));

// LOG-FACTORIALS
for(i = 0; i <= n; i++)  h_lfact[i] = (double) gammln( (double) (i + 1.0) );

// LOOP OVER THE ROWS OF ASCII INPUT FILE
for(j = 1; j <= 120; j++)
{
  system("CLS");
  printf("piset: %d\n\nIteration 1: %s", (unsigned int) j, ctime(&curr));

  // DELTA_1; pi^{aa} (pi1); pi^{Aa} (pi2)
  delta = (double) log( (double) (1.0 - epsilon) * p_asy / epsilon );
  fscanf_s(rstream1, "%d", (unsigned int *) &tmp);
  pi1 = (double) tmp / 1000000.0;

  fscanf_s(rstream1, "%d", (unsigned int *) &tmp);
  pi2 = (double) tmp / 1000000.0;

  // INITIALIZE ROW MEMBERSHIP
  for(i = 0; i < n_row1; i++) row1_ind_evermember[i] = 0;
  for(i = 0; i < n_row2; i++) row2_ind_evermember[i] = 0;

  start = clock();
  // GIVEN DELTA, LOCATE PROBABLE ROWS FOR ROW1
  probmargins( (unsigned int) n1, (unsigned int *) row1, (double) delta,
               (double *) h_lfact, (double *) &egamthetak1, (unsigned long *) &pos1,
               (unsigned long *) &pos1_new, (char *) row1_ind_evermember,
               (unsigned long *) row1_ind_update, (unsigned long *) row1_ind_old,
               (double) pi1, (double) pi2, (double*) row1_prob );

  // GIVEN DELTA, LOCATE PROBABLE ROWS FOR ROW2
  probmargins( (unsigned int) n2, (unsigned int *) row2, (double) delta,
               (double *) h_lfact, (double *) &egamthetak2, (unsigned long *) &pos2,
               (unsigned long *) &pos2_new, (char *) row2_ind_evermember,
               (unsigned long *) row2_ind_update, (unsigned long *) row2_ind_old,
               (double) pi1, (double) pi2, (double*) row2_prob);
```

```
probtruncuncondrefset( (unsigned int) n1, (unsigned int) n2, (unsigned int*) row1,
                       (unsigned int*) row2, (unsigned long) pos1,
                       (unsigned long) pos2, (unsigned long) pos1_new,
                       (unsigned long) pos2_new, (unsigned long *) row1_ind_update,
                       (unsigned long *) row2_ind_update,
                       (unsigned long *) row1_ind_old,
                       (double) t_iota, (double) var_const, (double *) &prob,
                       (double *) row1_prob, (double *) row2_prob, 0.0 );

finish = clock();
dura  = (double)(finish - start) / CLOCKS_PER_SEC;
dura /= 60;
fopen_s(&wstream1, "results.txt", "a");
fprintf_s( wstream1, "%d %d %0.10f %0.10f %0.10e %0.10e %d %d %0.10e %5.3f %3.3f\n",
                   (unsigned int) n1, (unsigned int) n2, (float) pi1, (float) pi2,
                   (double) egamthetak1, (double) egamthetak2, pos1, pos2,
                   (double) prob, (float) delta, (float) dura );
fclose(wstream1);

if( prob >= (epsilon * (egamthetak1 + egamthetak2) / (1.0 - epsilon)) ) ind = 0;
else ind = 1; i = 2;

while(ind)
{
  prob_old = prob;
  curr = time(0);
  printf("Iteration %d: %s", (unsigned int) i, ctime(&curr));
  start = clock();
  multiplier = (double) 0.1;
  delta += (double) log( (double) multiplier);

  // GIVEN DELTA, LOCATE PROBABLE ROWS FOR ROW1
  probmargins( (unsigned int) n1, (unsigned int *) row1, (double) delta,
               (double *) h_lfact, (double *) &egamthetak1, (unsigned long *) &pos1,
               (unsigned long *) &pos1_new, (char *) row1_ind_evermember,
               (unsigned long *) row1_ind_update,
               (unsigned long *) row1_ind_old,
               (double) pi1, (double) pi2, (double*) row1_prob);

  // GIVEN DELTA, LOCATE PROBABLE ROWS FOR ROW2
  probmargins( (unsigned int) n2, (unsigned int *) row2, (double) delta,
               (double *) h_lfact, (double *) &egamthetak2, (unsigned long *) &pos2,
               (unsigned long *) &pos2_new, (char *) row2_ind_evermember,
               (unsigned long *) row2_ind_update,
               (unsigned long *) row2_ind_old,
               (double) pi1, (double) pi2, (double*) row2_prob);

  // UPDATE THE P-VALUE
  probtruncuncondrefset( (unsigned int) n1, (unsigned int) n2, (unsigned int*) row1,
                         (unsigned int*) row2, (unsigned long) pos1,
                         (unsigned long) pos2, (unsigned long) pos1_new,
                         (unsigned long) pos2_new, (unsigned long *) row1_ind_update,
                         (unsigned long *) row2_ind_update,
                         (unsigned long *) row1_ind_old, (double) t_iota,
                         (double) var_const, (double *) &prob, (double *) row1_prob,
                         (double *) row2_prob, (double) prob_old);
```

```
      finish = clock();
      dura  = (double)(finish - start) / CLOCKS_PER_SEC;
      dura /= 60;
      fopen_s(&wstream1, "results.txt", "a");
      fprintf_s(wstream1, "%d %d %0.10f %0.10f %0.10e %0.10e %d %d %0.10e %5.3f %3.3f\n",
                          (unsigned int) n1, (unsigned int) n2, (float) pi1, (float) pi2,
                          (double) egamthetak1, (double) egamthetak2, pos1, pos2,
                          (double) prob, (float) delta, (float) dura );
      fclose(wstream1);

      if( prob >= (epsilon * (egamthetak1 + egamthetak2) / (1.0 - epsilon)) ) ind = 0;
      i++;
    }
  }
}
```

D.2 R Code for GEM Implementation

Contents of the `DLL_test_main.R` file:

```
### LOAD THE .DLL FOR GEM
#
dyn.load('GEM_DLL.dll')
#
### C FUNCTION TO CARRY OUT THE GEM METHODOLOGY -- BINARY ENVIRONMENTAL FACTOR;
### DEFAULT CANDIDATE PATTERNS; I = USER INPUT, R = FUNCTION RETURN
#
GEM_2e = function(n.controls, n.sample, n.snps, dat, n.perms, nsnps.per.geneset)
{
  y = .C( "GEM_2e",               ### C FUNCTION NAME (I)
          as.integer(n.controls), ### NUMBER OF CONTROLS (I)
          as.integer(n.sample),   ### SAMPLE SIZE (I)
          as.integer(n.snps),     ### HOW MANY SNPs (I)
          as.integer(dat),        ### THE DATA -- n.snps x n.sample MATRIX (I)
          as.integer(n.perms),    ### HOW MANY COLUMN PERMUTATIONS (I)
          as.integer(nsnps.per.geneset),  ### HOW MANY SNPs PER GENE (VECTOR) (I)
          maxT.n   = double(n.snps * 11), ### maxT P-VALUES (ADJ AT SNP LEVEL) (R)
          maxT.a   = double(n.snps * 11), ### maxT P-VALUES (ADJ AT GENE LEVEL) (R)
          margins  = integer(10 * n.snps), ### COLUMN/ROW MARGINS BY SNP (R)
          raw_wald = double(11 * n.snps),  ### WALD TEST STAT BY PATTERN/SNP (R)
          log_OR   = double(11 * n.snps),  ### ODDS-RATIO (LOG) BY PATTERN/SNP (R)
          selog_OR = double(11 * n.snps)   ### se(LOG-OR) (R)
        )

  table.margins = as.data.frame(matrix(y[[9]], nrow = n.snps, ncol = 10, byrow = TRUE))
  for(i in 1:6) names(table.margins)[i] = paste('nGE', i, sep = '')
  names(table.margins)[7:10] = c('nonmisscase', 'nonmisscontrol', 'nonmisstot',
                                 'misscontrol')
  wald      = as.data.frame(matrix(y[[10]], nrow = n.snps, ncol = 11, byrow = TRUE))
  maxT.nom  = as.data.frame(matrix(y[[7]], nrow = n.snps, ncol = 11, byrow = TRUE))
  maxT.adj  = as.data.frame(matrix(y[[8]], nrow = n.snps, ncol = 11, byrow = TRUE))
  log.OR    = as.data.frame(matrix(y[[11]], nrow = n.snps, ncol = 11, byrow = TRUE))
  se.log.OR = as.data.frame(matrix(y[[12]], nrow = n.snps, ncol = 11, byrow = TRUE))
```

```
  list(tm = table.margins, ts = wald, maxT_nom = maxT.nom, maxT_adj = maxT.adj,
       logOR = log.OR, selogOR = se.log.OR)
}
#
### C FUNCTION TO CARRY OUT THE GEM METHODOLOGY -- THREE-LEVEL ENVIRONMENTAL FACTOR;
### USER-SPECIFIED CANDIDATE PATTERNS; I = USER INPUT, R = FUNCTION RETURN
#
GEM_3e_flex = function(n.controls, n.sample, n.snps, dat, n.perms,
                       nsnps.per.geneset, ind.mat, num.tests.per.snp)
{
  y = .C( "GEM_3e_flex",          ### C FUNCTION NAME (I)
          as.integer(n.controls), ### NUMBER OF CONTROLS (I)
          as.integer(n.sample),   ### SAMPLE SIZE (I)
          as.integer(n.snps),     ### NUMBER OF SNPs (I)
          as.integer(dat),        ### DATA (MATRIX) (I)
          as.integer(n.perms),    ### NUMBER OF COLUMN PERMUTATIONS (I)
          as.integer(nsnps.per.geneset), ### NUMBER OF SNPs PER GENE (I)
          maxT.n   = double(n.snps * num.tests.per.snp), ### maxT P-VAL (ADJ SNP LVL) (R)
          maxT.a   = double(n.snps * num.tests.per.snp), ### maxT P-VAL (ADJ GENE LVL) (R)
          margins  = integer(13 * n.snps), ### COLUMN/ROW TABLE MARGINS BY SNP (R)
          raw_wald = double(num.tests.per.snp * n.snps), ### WALD TEST STAT BY PAT/SNP (R)
          as.integer(ind.mat), ### CANDIDATE PATTERN INDICATOR MATRIX (I)
          as.integer(num.tests.per.snp) ### NUMBER OF CANDIDATE PATTERNS (I)
        )

  table.margins = as.data.frame(matrix(y[[9]], nrow = n.snps, ncol = 13, byrow = TRUE))
  for(i in 1:9) names(table.margins)[i] = paste('nGE', i, sep = '')
  names(table.margins)[10:13] = c('nonmisscase', 'nonmisscontrol', 'nonmisstot',
                                  'misscontrol')
  wald = as.data.frame(matrix(y[[10]], nrow = n.snps, ncol = num.tests.per.snp,
                       byrow = TRUE))
  maxT.nom = as.data.frame(matrix(y[[7]], nrow = n.snps, ncol = num.tests.per.snp,
                           byrow = TRUE))
  maxT.adj = as.data.frame(matrix(y[[8]], nrow = n.snps, ncol = num.tests.per.snp,
                           byrow = TRUE))
  ind      = as.data.frame(matrix(y[[11]], nrow = num.tests.per.snp, ncol = 9,
                           byrow = TRUE))
  n.tests  = y[[12]]

  list(tm = table.margins, ts = wald, maxT_nom = maxT.nom, maxT_adj = maxT.adj,
       ind.mat = ind, num.tests = n.tests)
}
```

CURRICULUM VITAE

William L. Welbourn, Jr.

EDUCATION

Ph.D. Mathematical Sciences, May 2012, Utah State University, Logan, UT.

Dissertation title: Robust computational tools for multiple testing with genetic association studies.

Dissertation committee: Christopher Corcoran, Chair; Adele Cutler; Kady Schneiter; John Stevens; and, Ronald Munger.

M.Sc. Biostatistics, May 2006, University of Southern California, Los Angeles, CA.

Thesis title: An investigation of the performance for ordered subset analysis to a gene-environment interaction model.

Thesis committee: Kimberly Siegmund, Chair; W. James Gauderman; and, Bryan Langholz.

B.A. Mathematics – Probability and Statistics Concentration, January 2001, California State University at Fullerton, Fullerton, CA (Advisor: James Friel).

AWARDS AND HONORS

Technology Expert Award. Department of Mathematics and Statistics, Utah State University (2010).

Research Writing Award. Department of Mathematics and Statistics, Utah State University (2009).

Inducted into the Golden Key International Honour Society. Utah State University Chapter (2009).

Inducted into The Honor Society of Phi Kappa Phi. University of Southern California Chapter (2006).

GRADUATE ASSISTANTSHIPS

Graduate Research Assistant. Fall 2007 – Fall 2011, Department of Mathematics and Statistics, Utah State University.

Graduate Teaching Instructor. Fall 2007 – Spring 2012, Department of Mathematics and Statistics, Utah State University.

- Fall 2007, Spring 2008: MATH 1050 – *College Algebra.*
- Fall 2008, Spring 2009: STAT 2000 – *Statistical Methods.*
- Summer 2009: STAT 2300 – *Business Statistics.*
- Fall 2009, Fall 2010: SAS/R Computer Lab Instructor for STAT 5100/5200 – *Linear Regression and Time Series/Experimental Designs.*
- Spring 2010: Recitation Leader (three sections), STAT 3000 – *Statistics for Scientists.*
- Summer 2010, Summer 2011, Spring 2012: STAT 3000 – *Statistics for Scientists.*

Graduate Teaching Assistant. Summer 2005 – Fall 2005, Division of Biostatistics, Keck School of Medicine, University of Southern California.

- Summer 2005: SPSS Computer Lab Instructor, PM 510 – *Principles of Biostatistics.*
- Fall 2005: Teaching Assistant (Stanley Azen, Instructor), PM 510– *Principles of Biostatistics.*

PUBLICATIONS

M. Slattery, A. Lundgreen, **B. Welbourn**, B. Caan, and C. Corcoran, "Oxidative balance and colon and rectal cancer: interaction of lifestyle factors and genes," *Carcinogenesis, under review*, 2011.

M. Slattery, A. Lundgreen, **B. Welbourn**, C. Corcoran, R. Wollff, "Genetic variation in selenoprotein genes, lifestyle, and risk of colon and rectal cancer," *manuscript in preparation.*

W. Welbourn Jr, C. Corcoran, M. Slatery, and A. Lundgreen, "A permutation approach to detect gene-environment interaction in genetic association studies," *manuscript in preparation.*

W. Welbourn Jr and C. Corcoran, "The GEM package," *R package in preparation.*

RESEARCH INTERESTS

Statistical Computing. Developing C and CUDA-C statistical applications to harness the power of intra-PC parallel (i.e., multi-threaded applications) computing; Resampling theory; Database management within the Structured Query Language (SQL), interfacing with each of the R and C languages. Particularly, management of extravagant genetic databases; Graphical user interfacing (GUI) programming within the R software environment.

Categorical Data Analysis. Applications involving the multinomial and binomial distributions; logistic regression modeling.

Survival Analysis Methods. Parametric regression models and non-parametric methods (e.g., Kaplan-Meier survival curves, log-rank test); cox-proportional hazard models.

COMPUTER LANGUAGES

Proficiency: R, SAS, Mathematica, SQL (via interfacing with R by way of the RSQLite package).

Skilled: C, CUDA for C, LaTeX(including Beamer class and interfacing with R via Sweave), HTML, SQL (via interfacing with C).