Finding G-E Interactions in Quantitative Trait Analysis Using Two-Step Methods

FINDING G-E INTERACTIONS IN QUANTITATIVE TRAIT ANALYSIS USING TWO-STEP METHODS

BY

QIANMIN YANG, B.Math.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2015) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	Finding G-E Interactions in Quantitative Trait Analysis Using Two-Step Methods
AUTHOR:	Qianmin Yang B.Math, (Mathematics/Business Administration) University of Waterloo, Waterloo, Canada
SUPERVISOR:	Dr. Angelo Canty
NUMBER OF PAGES:	xi, 113

Abstract

In recent years, screening approaches known as two-step methods have been proposed to detect gene-environment interactions for genome-wide association studies (GWAS). Genetic and environmental factors are believed to affect disease outcome as well as various quantitative traits such as height and blood pressure. The performance of the two-step methods has not been demonstrated in the quantitative trait setting. This thesis examines the method proposed by Wang and Abbott (2008) for generating genotyped markers in linkage disequilibrium (LD) and takes this approach in simulating data pertaining to a quantitative trait. The simulation results demonstrate that the two-step methods maintain type I error and have power to detect the quantitative trait locus. In this setting, the EG method (Murcray et al., 2009) is influenced by the strength and structure of the gene-environment dependency, the sample type, and the disease model. As such, the power of the EG method can fluctuate depending on the type of data while the DG method (Kooperberg and LeBlanc, 2008) remains fairly robust across a wide range of scenarios. The performance of the combined two-step approaches (EDGE (Gauderman et al., 2013) and H2 (Murcray et al., 2011) methods) tends to favour the more powerful underlying method. The power of the EDGE method can be improved if DG and EG demonstrates similar power while the H2 method can be made more powerful by choosing the appropriate parameters.

Acknowledgements

I would like to use this opportunity to express my sincere gratitude to my advisor Dr. Angelo Canty for his support and guidance throughout the course of my research. His enthusiasm and knowledge has been inspirational. The advice and constructive criticism he provided were pertinent to the success of this thesis. I have learned tremendously under his tutelage and I have truly enjoyed working with Dr. Canty.

I would also like to thank Dr. Joseph Beyene and Dr. Jamila Hamid for partaking in the examination committee. I appreciate their time and considerations on my work.

Lastly, I would like to thank my friends and family for their love and support in helping me achieve my academic and professional goals.

Notation and abbreviations

BMA	Bayesian model averaging
CC	Case-control
CO	Case-only
DSL	Disease susceptibility locus
EB	Empirical Bayes
G x E	Gene-environment interaction
G-E	Gene-environment
GWAS	Genome-wide association study
HW	Hardy-Weinberg
LD	Linkage disequilibrium
LE	Linkage equilibrium
MAF	Minor allele frequency
QTL	Quantitative trait locus
SNP(s)	Single-nucleotide polymorphism(s)

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 GWAS Challenges and the Post-GWAS Era

The completion of the Human Genome Project in 2003 (International Human Genome Sequencing Consortium, 2001, 2004) and the International HapMap Project in 2005 (The International HapMap Consortium, 2005) gave scientists a reference of the human genome sequence and a road map of the genetic variations in several ethnic populations. This new knowledge, coupled with the technological advances in high throughput genotyping technology, made identifying and analyzing a vast amount of DNA data possible. As genotyping technology improved, driving down the cost of obtaining DNA data, genome-wide association studies (GWAS) became an integral research area of genomics in the new millennium.

The era of GWAS was kicked off by a landmark study on age-related macular degeneration which identified two single-nucleotide polymorphisms (SNPs) that are associated with the disease (Klein et al., 2005). This prompted the largest GWAS of its time, the Wellcome Trust Case Control Consortium (WTCCC), which examined

14,000 cases of seven common diseases (coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder, and hypertension) with 3,000 shared controls (Wellcome Trust Case Control Consortium, 2007). The WTCCC identified 24 independent association signals and motivated further collaborative GWAS work with large sample sizes as well as large scale sequencing projects such as the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010). As of January 2nd, 2015, according to the NHGRI GWAS Catalog (Hindorff et al., 2015), the scientific community has conducted 2,021 genome-wide association studies and identified more than 4,400 SNPs that show a significant[1] association with 633 diseases and traits.

GWAS is founded on the common disease-common variant hypothesis, where a common variant is typically defined as having minor allele frequency (MAF) greater than 1%. This hypothesis predicts that common diseases, such as cancer or diabetes, are caused by commonly occurring genetic variants in the human population (Lander, 1996; Reich and Lander, 2001). The basic idea of a GWAS is to identify commonly occurring SNPs that are associated with a phenotype of interest. In a typical case-control study, hundreds of thousands of SNPs are genotyped for a sample of population with the disease and for a similar sample without the disease. A logistic model is fitted for every SNP with the disease status as the response and a test is conducted to determine the significance of the SNP (Balding, 2006; Pearson and Manolio, 2008). For quantitative traits such as height or blood pressure, a similar approach using linear regression can be used to detect significant associations.

While GWAS has been largely successful (Visscher et al., 2012), it does face many challenges. One of which is the small effect sizes of the identified SNPs, often with

[1]Based on the genome-wide significance threshold of 5×10^{-8}

odds ratios less than 1.5 (Hindorff et al., 2009; Manolio, 2013; Manolio et al., 2009). As such, findings from GWAS can only explain a small proportion of the genetic contribution to a disease or a trait. This leads to limited predictive power and raises the issue of the missing heritability (Manolio et al., 2009; Visscher et al., 2008). Another limitation of GWAS is the lack of biological significance of the identified SNPs. A notable proportion of the discoveries reside in the non-coding regions of DNA (Hindorff et al., 2009). These regions do not explicitly contribute to gene function and cannot be assigned to specific biological pathways. Due to these challenges, the clinical applications of GWAS results have been limited despite the volume of discoveries (Manolio, 2013).

Researchers have faced methodological challenges as well, such as small sample sizes, lack of power of tests, multiple testing burden, model misspecification, and computational time of proposed methods (Balding, 2006; Pearson and Manolio, 2008). It should be noted that a typical GWAS utilizes a simplistic model of disease association, e.g. logistic regression with one main effect. As such, complex genetic phenomena such as pleiotropy, epistasis, or gene-environment interactions (G x E) cannot be readily measured using standard GWAS analysis techniques. These biological and methodological challenges have prompted further research beyond the scope of the basic GWAS.

Moving to the post-GWAS era of research, the scientific community has shifted attention away from simple SNP analysis to modelling more complex disease-gene relationships. One promising area of research is the identification of gene-environment interactions. It has been hypothesized that genetic effects can be altered by environmental conditions and certain subpopulations experiencing exposure may exhibit

phenotypic traits that are different from unexposed populations (Aschard et al., 2012; Dempfle et al., 2008; Eichler et al., 2010). If these subpopulations could be identified by their environmental exposure, which elevates their genetic risk to disease, then targeted treatment or personalized medicine can be applied to those who are classified as high risk due to their genetic and environmental dispositions (Thomas, 2010).

G x E analysis can also aid in the identification of new disease associated SNPs. It has been hypothesized that some SNPs may show no marginal association with disease, instead demonstrating an effect only through its interaction with the environment (Kraft et al., 2007). Finding these G x E effects can help to identify such SNPs and broaden the understanding of disease etiology. As such, G x E analysis has been identified as one possible explanation to the problem of the missing heritability (Aschard et al., 2012; Eichler et al., 2010; Manolio et al., 2009).

1.2 G x E Analysis Literature Review

In cross-sectional, case-control studies of a disease status, the most common analysis of gene-environment interaction is to use logistic regression to model the genetic and environmental effects on the disease. Let D denote the disease status ($D = \{0, 1\}$), let G denote the genetic marker coded under some genetic model (i.e. additive or dominant), and let E denote the environmental factor. The traditional case-control (CC) model used in G x E analysis is:

$$\text{logit}(P(D = 1|G, E)) = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} G \times E \tag{1.1}$$

where $\text{logit}(p) = \log(\frac{p}{1-p})$ is the logit function. To assess interaction, the parameter β_{ge} is tested under the null hypothesis of no interaction using an one degree of freedom Wald test. The maximum likelihood estimators of the parameters are asymptotically consistent (Gourieroux and Monfort, 1981) using the traditional case-control approach. This typically results in well maintained type I error rates. However, the method quickly runs into the multiple testing burden as the number of SNPs tested increases. As such, the case-control analysis suffers from limited power (Gauderman et al., 2013; Kooperberg and LeBlanc, 2008; Murcray et al., 2009). For the evaluation of quantitative or categorical traits, linear regression or multinomial regression of a similar form can be used. Note that the environment factor is often coded as a binary variable to represent exposed and unexposed individuals.

An alternative to the case-control analysis is the case-only (CO) analysis proposed by Piegorsch et al. (1994). The authors demonstrated that, for a rare disease under the assumption of gene-environment (G-E) independence, it is possible to obtain an efficient estimate of the G x E effect without studying the controls. The case-only analysis uses logistic regression to model the effect of a SNP, G, on the environment as follows:

$$\text{logit}(P(E = 1|G, D = 1)) = \theta_0 + \theta_{ge}G \tag{1.2}$$

It has been shown that testing the parameter θ_{ge} is equivalent to testing the parameter β_{ge} in Model (1.1) (Murcray et al., 2009; Piegorsch et al., 1994). The case-only approach has also been shown to be the most efficient analysis under the model assumptions (Albert et al., 2001; Mukherjee et al., 2012*b*; Murcray et al., 2009). However, it is extremely sensitive to the gene-environment independence assumption. Even a small degree of G-E association can produce inflated type I errors (Albert et al.,

2001; Li and Conti, 2009). Attempts to assess the G-E independence assumption before applying a case-only approach has been shown to increase the type I error rate as well (Albert et al., 2001). Lastly, it should also be noted that due to the construction of Model (1.2), the marginal genetic effect on disease cannot be readily measured in the case-only analysis of G x E effects.

To strike a balance between the case-control approach and the case-only approach, two notable Bayesian methods have been proposed. Mukherjee and Chatterjee (2008) introduced an empirical Bayes (EB) method which corresponds to a weighted average of the case-only and the case-control estimates of G x E effect. The proposed EB estimate of the interaction parameter is defined as:

$$\hat{\beta}_{EB} = \frac{\hat{\sigma}^2_{CC}}{\hat{\tau}^2 + \hat{\sigma}^2_{CC}}\hat{\theta}_{ge} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2_{CC}}\hat{\beta}_{ge} \tag{1.3}$$

where $\hat{\sigma}^2_{CC}$ is the estimated asymptotic variance of the case-control estimate $\hat{\beta}_{ge}$, and $\hat{\tau}^2$ is a conservative estimate of the uncertainty in the G-E independence assumption. The parameter $\hat{\tau}^2$ is calculated as the maximum likelihood estimate of the odds ratio of the G-E association among the controls (Mukherjee and Chatterjee, 2008):

$$\hat{\tau}^2 = \log\left(\frac{r_{000}r_{011}}{r_{001}r_{010}}\right)$$

where r_{dge} represents the sample data tabulated in the form of Table 1.1.

The EB approach uses the information obtained from the sample data to balance the efficiency of the case-only estimator and the consistency of the case-control estimator. The EB method is robust against G-E dependence and demonstrates better power than the case-control analysis alone (Gauderman et al., 2013; Mukherjee and

	$G=0$		$G=1$		
	$E=0$	$E=1$	$E=0$	$E=1$	Total
$D=0$	r_{000}	r_{001}	r_{010}	r_{011}	n_0
$D=1$	r_{100}	r_{101}	r_{110}	r_{111}	n_1

Table 1.1: Sample data for case-control study of a disease status with binary genetic and environmental factors where r_{dge} is the number of observations for $D=d$, $G=g$, and $E=e$.

Chatterjee, 2008).

The second Bayesian approach for G x E analysis, proposed by Li and Conti (2009), is the Bayesian model averaging (BMA) method. This approach uses a weighting scheme between the CC and the CO models. The weighting scheme is a function of the data and the prior beliefs on the G-E independence assumption. Given the observed data, the posterior probability of the interaction effect is defined as:

$$P(\phi|D) = P(\phi_1|D, M_1)P(M_1|D) + P(\phi_2|D, M_2)P(M_2|D)$$

Where $\phi_1 = \beta_{ge}$ from the CC model, $\phi_2 = \theta_{ge}$ from the CO model, and ϕ denotes the interaction parameter from the Bayesian model averaging approach. $P(M_k|D)$ for $k = 1, 2$ is the posterior probability of each model, with $P(M_k|D) \propto P(D|M_k)P(M_k)$. The prior probability, $P(M_k)$ for $k = 1, 2$, is assigned to each model ($k = 1$ for CC model and $k = 2$ for CO model). A relative weight function, or the prior odds, is defined as: $W = P(M_1)/P(M_2)$. This quantity is chosen based on the prior beliefs of G-E independence. For example, if there is a strong prior belief that gene and environment factors are independent then the weight function W would be small (< 1) to favour the case-only model.

To test the null hypothesis of no interaction, a Wald test can be used by assuming that the interaction estimate is normally distributed. Let $\hat{\sigma}_k$ denote the estimated interaction effect from model k, then the expected interaction effect in this Bayesian setting is defined as:

$$\begin{aligned} E(\phi|D) &= \hat{\phi}_1 P(M_1|D) + \hat{\phi}_2 P(M_2|D) \\ &= \hat{\beta}_{ge} P(M_1|D) + \hat{\theta}_{ge} P(M_2|D) \end{aligned}$$

This quantity, along with the variance of the interaction effect, is used to create the test statistic. Similar to the EB method, the expected interaction effect obtained from the BMA method also corresponds to a weighted average of the interaction estimates from the CC and the CO model. It has been demonstrated that this approach is robust to deviations away from the G-E independence assumption and is typically more powerful than the case-control analysis alone. However, if the prior belief is weighted highly in favour of the CO model (i.e. $W = 0.001$), then type I errors can be inflated if gene and environment factors are not independent (Li and Conti, 2009).

While both Bayesian methods cannot achieve the power of the case-only approach under G-E independence, they strike a balance between bias and efficiency by combining the CC and the CO approaches. The performance of the BMA method is comparable to the EB method (Li and Conti, 2009). Due to the construction of the two Bayesian methods, the resulting estimates can show modest bias when the G-E independence assumption is violated. This means type I error rates can be inflated and the coverage of the associated confidence intervals may be less than nominal (Mukherjee and Chatterjee, 2008; Mukherjee et al., 2012*b*). It should also be noted that the implementation of the Bayesian methods is not as straightforward as the

case-control or case-only analysis.

The previously mentioned methods look for G-E interactions by testing every SNP in the sample. As such, these approaches are exhaustive searches for G x E effects. Due to the multiple testing burden, exhaustive searches have low power to detect small to moderate effects. To address this issue, a class of screening methods have been proposed in recent years. In the literature, this class of G x E analysis is known as the two-step methods. The basic idea is to first evaluate all the SNPs using a screening test and some defined threshold, then formally test the SNPs that pass the first stage for presence of G x E effects. The multiple testing burden is reduced in the second step by considering a much smaller set of markers, thereby increasing the power to detect interactions. However, this method is only valid if the test statistics of the first and second stages are independent. The independence condition is met for quantitative traits (i.e. using linear regression models) (Kooperberg and LeBlanc, 2008). For logistic regression, Dai et al. (2012) demonstrated the general conditions that achieves asymptotic independence of the two test statistics. Simulations under finite samples have shown that the correlation between the two test statistics are small enough to be ignored (Kooperberg and LeBlanc, 2008; Murcray et al., 2009).

The first screening procedure was proposed by Kooperberg and LeBlanc (2008). In their method, SNPs that demonstrate a certain level of marginal genetic effect on disease status are passed onto the second step for formal G x E testing. The method assumes SNPs that interact with the environment will also demonstrate some marginal effect on the disease. The screening model:

$$\text{logit}(P(D = 1|G)) = \beta_0 + \beta_g G \tag{1.4}$$

is used to filter out the irrelevant SNPs by testing the null hypothesis $H_0 : \beta_g = 0$. The SNPs that pass, based on some step 1 threshold α_1, are then tested for G x E effects using Model (1.1) or a Bayesian approach. This is known in the literature as the Disease-Gene (DG) method. The number of SNPs passed in the first step is controlled by the threshold α_1. As α_1 decreases the number of SNPs passed onto the second step also decreases. This can increase the power as the multiple testing burden is reduced. However, stringent values of α_1 could lead to false negative results as relevant SNPs are screened out at the first step. Kooperberg and LeBlanc (2008) demonstrated that their proposed method maintained type I error rates and achieved higher power than the traditional case-control approach under most simulation settings. It should be noted that the CC model showed higher power for large interaction effects in the opposite direction to the main effect. It has also been shown that if there are zero or small marginal genetics effect on disease, this method lacks power in comparison to other two-step methods (Gauderman et al., 2013).

Borrowing ideas from the case-only analysis, Murcray et al. (2009) proposed a screening step which evaluates SNPs based on their G-E association. The screening step uses a G-E association model similar to the case-only Model (1.2), but considers the whole sample of cases and controls. It has been shown that in the presence of a G x E effect, a correlation between the causal SNP and environment factor is induced by the ascertainment of cases at a higher rate than the disease prevalence level (Murcray et al., 2009). Hence the induced correlation can be leveraged as a way to filter out the irrelevant SNPs. The whole sample is used to ensure that the first and second step test statistics remain independent. Similar to the DG method, the screened SNPs are then formally tested for G x E effects in the second step using the case-control

Model (1.1). Note that using a Bayesian approach in the second step would violate the independence requirement between the first and second stage test statistics. This is because the CO test statistic is correlated with the step 1 test statistics of the EG method. As such, the Bayesian test statistic, which is derived from both CC and CO methods, will also be correlated with the step 1 test statistic of the EG method. This approach is known in the literature as the Environment-Gene (EG) method. The EG method maintains type I error rates and has been shown to be more powerful than the case-control analysis. The method is also robust against G-E dependence (Murcray et al., 2009). It has been shown that the EG method performs better than the DG method when there are zero to small marginal genetic effects. However, as the marginal genetic effect of the causal SNP increases, the power of DG method can overtake the power of the EG method (Gauderman et al., 2013).

Building on the DG and EG screening tests, a number of combined two-step approaches utilizing various configurations of the DG and the EG methods have also been proposed. The hybrid (H2) method, proposed by Murcray et al. (2011), uses both screening tests on all the genotyped markers. Any SNP that demonstrates a G-E association or a marginal genetic effect will be formally tested for G x E effects in the second phase using Model (1.1). The procedure of the H2 method is as follows:

1. Test all SNPs for gene-environment association using the EG screening model and threshold α_{1a}

    (a) SNPs that pass are formally tested for G x E effects using Model (1.1) and significance level $\alpha^* = p\alpha/s_a$

2. Test all SNPs for disease-gene association using the DG screening model and threshold α_{1m}

(a) SNPs that pass are formally tested for G x E effects using Model (1.1) and significance level $\alpha^* = (1-p)\alpha/s_m$

3. SNPs that pass both DG and EG tests are formally tested for G x E effects using Model (1.1) and the more liberal significance level $\alpha^* = \max(p\alpha/s_a, (1-p)\alpha/s_m)$

Where s_m is the number of SNPs that passed the DG screening test, s_a is the number of SNPs that passed the EG screening test, and p is a value chosen to be between 0 and 1 for the allocation of type I error rate. It has been shown that the H2 method maintains type I error and is robust against G-E dependence. The H2 method has also been shown to be more powerful than the DG and the EG method alone for some choices of p (Murcray et al., 2011). The H2 method tries to balance the performance of the DG method and the EG method. If there are small marginal genetic effects, the H2 method is more powerful than the DG method, but less powerful than the EG method. Alternatively, if there are large marginal genetic effects, the H2 method becomes more powerful than the EG method, but less powerful than the DG method (Gauderman et al., 2013).

Lastly, Gauderman et al. (2013) proposed the EDGE method by combining the DG and the EG step 1 test statistic into one screening statistic $S_{EDGE} = S_{EG} + S_{DG}$, where S_{EG} is the test statistic of the EG screening step and S_{DG} is the test statistic of the DG screening step. The two test statistics, S_{EG} and S_{DG} are independent and each follows a χ^2-distribution with one degree of freedom under the null hypothesis (Dai et al., 2012). Therefore, the resulting sum, S_{EDGE}, follows a χ^2-distribution with two degrees of freedom under the null hypothesis. Since both S_{DG} and S_{EG} are independent of the CC model test statistic, it follows that S_{EDGE} is also independent of the CC model test statistic. In simulations, the EDGE method maintained

type I error and demonstrated greater power than the previous two-step methods for moderate genetic effects (i.e. odds ratio between 1.1 and 1.25) (Gauderman et al., 2013).

It should be noted that other two-step approaches have been proposed by Hsu et al. (2012) named the Cocktail methods. These approaches utilize both the DG and the EG screening tests in the first step. Based on the results of the first step, all of the SNPs are tested for G x E effects in the second step using either the CC approach or a Bayesian approach. This method is not examined by this thesis in the quantitative trait setting.

Outside of exhaustive searches and two-step methods, there have been a number of other notable G x E analyses proposed. Kraft et al. (2007) introduced the use of a joint test of both marginal and interaction effect to detect genetic associations. The authors used Model (1.1) to test the hypothesis: $H_0 : \beta_g = \beta_{ge} = 0$. Under the null hypothesis, the resulting test statistic follows a χ^2-distribution with two degrees of freedom. Simulations have shown that this joint test of marginal genetic and G x E effects attains good power in a wide range of scenarios. It is generally more powerful than a simple marginal test in detecting associated SNPs and more powerful than the standard case-control approach in detecting interactions (Kraft et al., 2007). Other approaches to finding G x E effects have been proposed by Aschard et al. (2013) and Paré et al. (2010) that utilize nonparametric methods.

1.3 Challenges of G x E Analysis

G x E analysis is not without challenges and limitations. It faces the same methodological issues as GWAS in terms of sample sizes and power to detect. These problems

are exacerbated in G x E analysis since larger sample sizes are needed to reliably detect a modest interaction (Aschard et al., 2012; Thomas, 2010). Outside of the methodological problems, one particular criticism is the lack of biology explained by the identified interactions. This goes back to a hotly debated subject matter on the philosophical differences between statistical interactions and biological interactions (Dempfle et al., 2008; Greenland, 2009; Rothman et al., 1980). Statistical interaction, often defined as a multiplicative departure from an additive model, does not always imply a biological phenomenon.

The biggest challenge is the effect of the environment is often time dependent and measuring it at a single point in time can result in a substantial loss of information (Aschard et al., 2012; Khoury and Wacholder, 2009). It has been hypothesized by scientists that the environment can play a crucial role in the etiology of disease during key development phases such as gestation, infancy, or puberty (NIH G x E Interplay Workshop, 2011). The majority of current G x E research has been focused on retrospective, cross-sectional data such as case-control studies. It is easy to see that time dependent information is lost in this approach such as length of exposure and time of initial exposure. G x E analysis can be more powerful and provide a better understanding of disease etiology in a longitudinal setting. Currently, there is a lack of research extending G x E analysis to longitudinal, cohort data. The current proposed methods have been criticized for using simple approaches in averaging the responses across time periods where the collapsing of information can decrease power (Fan et al., 2012). The applications of the current longitudinal G x E analysis methods are also tailored for genes identified a priori, thus it is unclear how these approaches fare in a genome-wide setting with a large volume SNPs.

For the two-step methods specifically, the application of these approaches in detecting G x E effects has not been applied to other response types such as quantitative traits. It is unclear how the two-step methods will perform in the quantitative trait setting. This is especially true for the EG method since the basis of this approach relies on the oversampling of cases in examining a disease status. The performance of the EG method in the quantitative trait setting subsequently impacts the performance of the combined two-step approaches. It should also be noted that simulation studies used to demonstrate the performance of the two-step methods often assume that the genotyped markers are independent. However, in practice, these markers can exhibit some degree of linkage disequilibrium (LD). The impact of correlation among the SNPs on the performance of two-step methods is also unclear.

1.4 Organization of Thesis

This thesis focuses on the application of the two-step methods in quantitative trait analysis. Two types of sample data are examined. The first type is a random sample where the quantitative trait of interest is measured on the selected individuals. The second type is a case-control sample, such as those used in GWAS, where a quantitative trait is also measured along with the genotyped markers. This thesis also examines the generation of SNPs in linkage disequilibrium. Simulation studies examining the performance of two-step methods for a quantitative trait will consider some degree of correlation among the SNPs.

The second chapter is a detailed analysis on a method of generating SNPs in LD first proposed by Wang and Abbott (2008). This method uses correlated multivariate normal variables to generate correlated binomial variables. Simulation studies are

used to assess the LD measures of the generated SNPs in terms of r^2 and D' based on the input correlation of the normal variables and MAF. Regression analysis is performed to quantify the relationship between the input correlation of the normals and the output correlation of the binomials. Results from the regression analysis provide a general guideline on the selection of input correlations of the multivariate normal variables.

The third chapter utilizes the method of generating SNPs in LD described in Chapter 2 to simulate datasets containing correlated markers. The two types of sample data for a quantitative traits are considered under various settings of marginal genetic and G x E effects. The two-step methods are applied to the simulated datasets to examine their performance in the quantitative trait setting.

A detailed look at the EG method and the various factors that can impact its performance is described in Chapter 4. Lastly, Chapter 5 presents a sensitivity analysis on the step 1 parameters of the two-step methods in the quantitative trait setting.

Chapter 2

Generating SNPs in Linkage Disequilibrium

Simulation studies examining methods used to analyze GWAS data often involve generating markers representing bi-allelic SNPs. A marker, denoted as G, is typically coded numerically to represent the number of copies of the minor allele at a locus. If Hardy-Weinberg (HW) equilibrium is assumed, then G can be generated as a binomial random variable. For large datasets such as the genetic data captured in a GWAS, the SNPs can be separately generated as binomial variables. By generating SNPs in this fashion, the simulation study assumes that the markers are in linkage equilibrium (LE). While the genotyped markers are typically selected to reduce redundancies caused by LD, the resulting SNPs can still exhibit some degree of correlation with one another. To represent the LD seen in GWAS markers, simulation studies should consider generating correlated binomial variables.

Wang and Abbott (2008) introduced a fast and simple method to generate correlated binomial variables using correlated multivariate normal variables which can be

generated very easily. This method has also been used by Guo et al. (2013) in their analysis of multiple traits. The basic premise of this method converts the marginal normal variables into binomial variables based on cutoff values calculated from the MAF using the Hardy-Weinberg equilibrium principle. It has been noted that the resulting binomial variables do not have the same correlation as the original normal variables. However, the change in correlation has not been quantified in the current literature. This chapter examines the effect of this method on the sample Pearson's correlation coefficient and the LD measures of the generated binomial variables. The LD measures considered are r^2 (Hill and Robertson, 1968) and D' (Lewontin, 1964).

2.1 Methods

Let X be a normally distributed random variable. For a given MAF, p, a binomial random variable representing the genotypes can be obtained by converting X using cutoff(s) chosen based on the pre-specified probabilities of the homozygous and heterozygous genotypes under the HW equilibrium assumption and the genetic coding. For example, using the coding $G = \{0, 1, 2\}$, where G represents the number of copies of the minor allele, the cutoffs c_1 and c_2 are chosen such that:

$$P(G = 0) = (1 - p)^2 = P(X \leq c_1)$$
$$P(G = 1) = 2p(1 - p) = P(c_1 \leq X \leq c_2)$$
$$P(G = 2) = p^2 = P(X \geq c_2)$$

For n observations of the normal variable X, a binomial variable can be obtained by comparing each component of X to the cutoff values. If $x_i \leq c_1$, a value of 0 will be

assigned; if $c_1 \leq x_i \leq c_2$, a value of 1 will be assigned; and if $x_i \geq c_2$, a value of 2 will be assigned, for $i = 1, ..., n$.

To generate a set of correlated binomial variables, let $\mathbf{X}$ have a multivariate normal distribution with mean vector $\mathbf{0}$ and some variance-covariance matrix, $\mathbf{\Sigma}$. The marginal distributions of $\mathbf{X}$ are normal, thus binomial variables can be generated using the above procedure with the appropriate cutoff values. If $\mathbf{\Sigma}$ is not a diagonal matrix, then the resulting set of binomial variables will be correlated. In the software program `R` (R Core Team, 2014), this procedure can be easily implemented using the `mvrnorm` function from the `MASS` library (Venables and Ripley, 2002) to generate the multivariate normal variables and using the `qnorm` function to obtain the appropriate cutoff values for converting the normal variables to binomial variables.

2.2 Simulation Study and Results

To examine the correlation of the generated binomial variables using the procedure proposed by Wang and Abbott (2008) under various scenarios, two simulation studies are conducted. For simplicity, the simulation studies consider generating a pair of correlated binomial variables with genotypes $G = \{0, 1, 2\}$ from a bivariate normal variable. The mean vector is set to zero and the two variances are set to one. As such, the resulting variance-covariance matrix represents the correlation between the marginal normal variables. Both simulation studies varied the correlations of the bivariate normal variables from 0 to 1 by increments of 0.01. The number of observations and the simulation replicates are each set to 1,000 in all of these simulations.

The simulation studies consider the effect of the value of MAF if both SNPs have the same MAF and the effect of differing values of the MAFs. At each replicate of

the simulation, the sample Pearson's correlation coefficient, LD measure r^2, and LD measure D' are calculated for the generated binomial variables. The estimates of r^2 and D' are obtained using the `LD` function from the `genetics` library (Warnes et al., 2013) in `R` which uses maximum likelihood to estimate the haplotypes from the given genotypes. The specific simulation settings for each study are detailed next. This thesis uses the terminology input correlation to indicate the Pearson's correlation of the bivariate normal variables and output correlation to indicate the correlation measures (sample Pearson's correlation coefficient, r^2, or D') of the generated binomial variables. From here on, the sample Pearson's correlation coefficient pertaining to the generated binomials will be denoted as Pearson's r and the input correlation used to generate the normal variables will be denoted as ρ.

2.2.1 Effect of the Value of MAFs

In this simulation, the effect of the MAF value on the output correlations of the binomial variables is examined. For simplicity, the MAFs of the two SNPs are assumed to be the same. The MAF values considered are: {0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50}. The associated cutoff values given an additive genetic coding are shown in Table A.1 in Appendix A.

The output correlations (Pearson's r, observed LD measure r, or observed LD measure D') obtained at each simulation replicate are averaged to produce an empirical estimate. A subset of the results is shown in Figure 2.1. The left panel shows the Pearson's r of the generated binomials across the range of input correlations by MAF values: {0.01, 0.1, 0.2, 0.3, 0.4, 0.5}. For very small MAF values, such as 0.01, the curve indicates that larger values of the input correlation are needed to achieve the

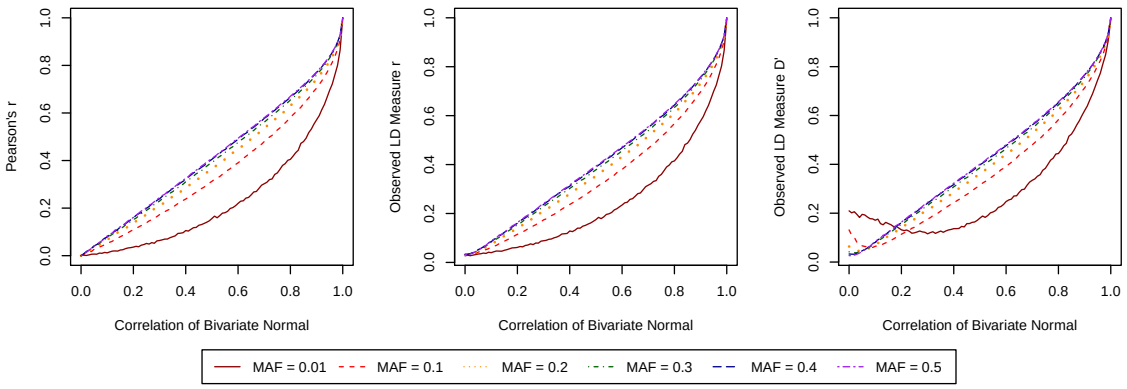

Figure 2.1: Effect of MAF values on the output correlation of generated binomials variables for a pair of SNPs with same MAF value. Left panel shows Pearson's r, middle panel shows observed LD measure r, and right panel shows observed LD measure D'.

desired output correlation. The relationship between input and output correlation is not linear for small values of MAF. However, as the MAF increases, the relationship becomes more linear where roughly the same input correlation can be used to produce the desired output correlation. In general, the nonlinear relationship between input and output correlation is exaggerated for small values of MAF and diminishes for MAF values greater than 0.2.

The middle and right panels of Figure 2.1 show the observed LD measures r and D' values of the generated binomial variables. The results demonstrate similar effects of the MAF value on the LD measures as seen in the right panel for Pearson's r. It should be noted that the observed D' values are inflated for small input correlation values given small MAF values. Table A.2 in Appendix A lists the Pearson's r, LD measure r, and D' of the generated binomial variables using input correlation of 0.50 for the bivariate normal variable. It should be noted that the Pearson's r is a close approximation to the LD measure r.

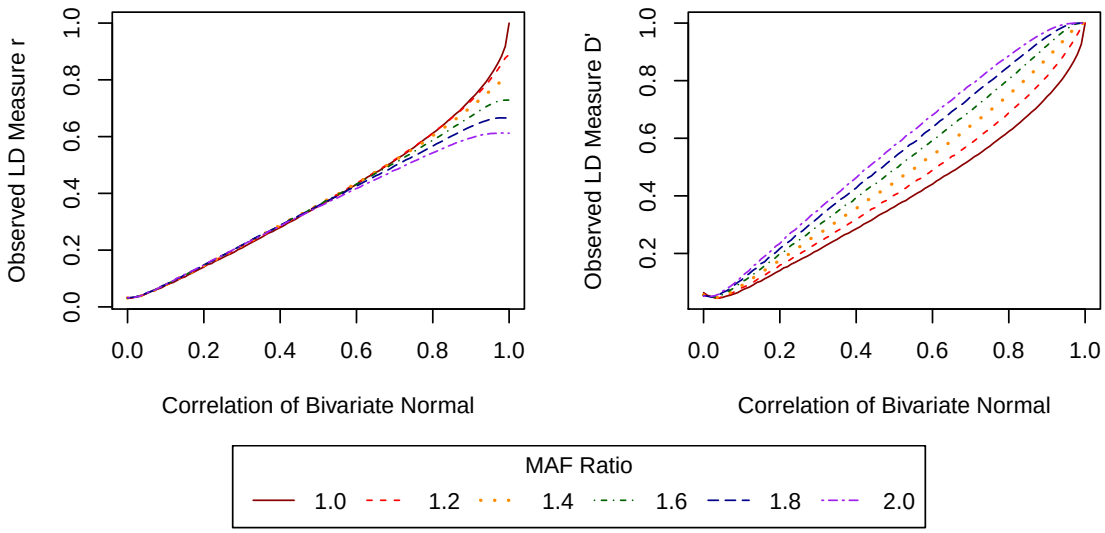

Figure 2.2: Effect of differing MAF values on the output correlation of generated binomial variables for a pair of SNPs by MAF ratio. Left panel shows the observed LD measure r and right panel shows the observed LD measure D'.

2.2.2 Effect of Differing MAF Values

In this simulation, the effect of non-homogeneous MAF values on the output correlation measures is examined. The MAF of the first SNP is held constant at 0.2 while the MAF of the second SNP is varied between 0.2 to 0.4 by increments of 0.02. As such, the MAF of the second SNP is a multiple of the MAF of the first SNP. The multiplicative factor ranges from 1 to 2 by increments of 0.1. This thesis uses the term MAF ratio to represent the multiplicative factors. The observed LD measures from each replicate are averaged to produce the empirical estimates of r and D' in the same fashion as in the previous simulation study.

A subset of the results is presented in Figure 2.2. The left panel shows the effect of differing MAF values on the observed LD measure r across the range of input correlations. Different MAF values do not have a large impact on the output correlation

for input correlation $\rho \leq 0.6$. The impact of differing MAF values are more apparent as the differences between the two MAFs increase for larger input correlation values. As expected, the maximal r achieved by using perfectly correlated normal variables decreases as the difference between the MAF values increases. The right panel shows the effect of differing MAF values on the observed LD measure D'. The impact of differing MAF values on D' is less significant for small input correlation measures, i.e. $\rho \leq 0.2$. As the input correlation increases, the spread of the D' curves increase. The D' values for SNPs with the largest difference in MAF values are always higher than those for SNPs with smaller differences in MAF values. However, using perfectly correlated normal variables results in all the D' values converging at 1.

2.3 Regression Analysis

Given the above simulation results, it is clear that the MAF value impacts the relationship between the input and the output correlations. It would be ideal to quantify this relationship for a given MAF value and provide a general guideline for researchers utilizing the method proposed by Wang and Abbott (2008) in generating SNPs in LD. Having a guideline can help researchers choose the appropriate input correlation, ρ, to achieve the desired output correlation based on the MAF value.

One approach to quantify the relationship between the input and the output correlations is to use linear regression models. In this case, the previous simulation results can be used where ρ is designated as the response and the observed LD measure r is designated as the covariate. Then to obtain a certain level of correlation between SNPs, i.e. desired r^2, the estimated coefficients based on the simulation results can be used to calculate the required input correlation. In this section, various linear

regression models are considered using ρ and the observed LD measure r from the simulation results of Section 2.2.1. Note that in these results, it is assumed that both SNPs have the same MAF.

2.3.1 Model Development

From Section 2.2.1, it is concluded that the relationship between input correlation and the observed LD measure r is not linear. The nonlinearity is exaggerated for small MAF values. As such, it may be naive to directly use the terms ρ and r in a linear model. Alternatively, the Fisher Transformation[1] can be applied to overcome this nonlinear relationship. Figure 2.3 shows a plotted comparison of the relationship between ρ and r, with and without the Fisher Transformation, for a subset of MAF values. Note the data used for Figure 2.3 are the simulated values from Section 2.2.1.

By applying the Fisher transformation to the input and the output correlations, their relationship appears more linear. However, the relationship remains nonlinear for small MAF values, especially for small input and output correlations. As such, a separate linear model is fitted for each MAF value. Three general classes of linear models are considered: the Naive Model, Fisher Model 1, and Fisher Model 2. The naive model regresses ρ on r and r^2 for each MAF value. Fisher Model 1 regresses the Fisher transformed ρ on the Fisher transformed r for each MAF value. Fisher Model 2 is the same as Fisher Model 1 with an additional squared Fisher transformed r term as a covariate. See Table 2.1 for the classes of linear models considered. Each class of linear models thus consists of 11 separate models for the set of MAF values considered. The MAF values are the same as those considered in the previous simulations.

[1]For a given correlation coefficient ρ, the Fisher transformation is defined as $z = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right) = \text{arctanh}(\rho)$

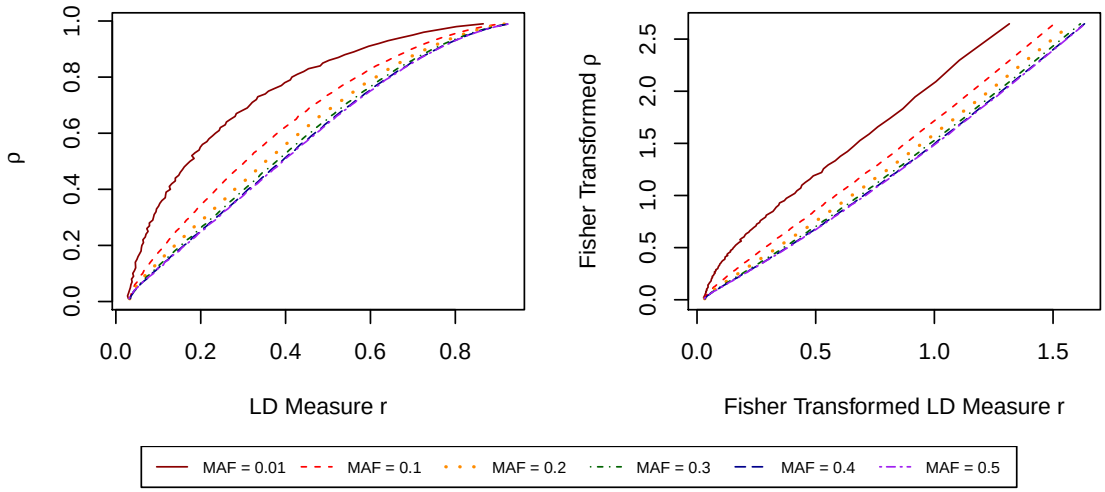

Figure 2.3: Relationship between input correlation, ρ, and output correlation, r of generated binomial variables, with and without Fisher Transformation

Model Name	Linear Models
Naive Model:	$\rho_i = \beta_{0i} + \beta_{1i} r_i + \beta_{2i} r_i^2 + \epsilon_i$
Fisher Model 1:	$fz(\rho_i) = \beta_{0i} + \beta_{1i} fz(r_i) + \epsilon_i$
Fisher Model 2:	$fz(\rho_i) = \beta_{0i} + \beta_{1i} fz(r_i) + \beta_{2i} fz(r_i)^2 + \epsilon_i$

Table 2.1: Linear regression models for input and output correlations of generated binomial variables for a pair of SNPs with same MAF. Note that $i = 1, ..., 11$ indicates the MAF values considered: $\{0.01, 0.05, 0.10, ..., 0.50\}$, fz indicates the Fisher transformation $fz(x) = \text{arctanh}(x)$, and ϵ_i is some normally distributed random noise.

MAF	$\hat{\beta}_0$ (Std. Error)	$\hat{\beta}_1$ (Std. Error)	$\hat{\beta}_2$ (Std. Error)
0.01	0.1789 (0.0014)	2.2313 (0.0072)	-0.4666 (0.0064)
0.05	0.0587 (0.0009)	1.8746 (0.0040)	-0.1196 (0.0034)
0.10	0.0263 (0.0007)	1.6414 (0.0030)	0.0334 (0.0024)
0.15	0.0146 (0.0006)	1.4974 (0.0025)	0.1170 (0.0020)
0.20	0.0064 (0.0006)	1.4187 (0.0023)	0.1541 (0.0018)
0.25	0.0023 (0.0006)	1.3578 (0.0021)	0.1821 (0.0016)
0.30	-0.0015 (0.0005)	1.3194 (0.0020)	0.1967 (0.0015)
0.35	-0.0035 (0.0005)	1.2905 (0.0019)	0.2101 (0.0014)
0.40	-0.0040 (0.0005)	1.2685 (0.0019)	0.2178 (0.0014)
0.45	-0.0052 (0.0005)	1.2606 (0.0019)	0.2207 (0.0014)
0.50	-0.0034 (0.0005)	1.2480 (0.0019)	0.2286 (0.0014)

Table 2.2: Estimated Coefficients of Fisher Model 2 by MAF Value

The simulation results from Section 2.2.1 are used as the data to fit the linear models shown in Table 2.1. The estimated coefficients for Fisher Model 2 are shown in Table 2.2. For the estimated coefficients of the Naive Model and Fisher Model 1, see Table A.3 and Table A.4, respectively, in Appendix A.

Note that the standard errors of the estimated coefficients of Fisher model 2 are fairly constant for MAF ≥ 0.25. This suggests that a linear model including the MAF as a covariate term (as a main effect or as an interaction effect) may be sufficient in capturing the differences attributed to the MAF value. This corresponds to earlier simulation results which indicated a somewhat linear relationship between ρ and LD measure r for MAF > 0.2. This also results in almost parallel lines seen in the Fisher transformed relationships between ρ and r for MAF > 0.2 (Figure 2.3). For small MAF values (MAF < 0.2), the standard errors of the estimated coefficients are no longer constant, which reinforces that separate linear models based on MAF are more appropriate.

2.3.2 Simulation Study and Results

A simulation study is conducted to evaluate the accuracy of the input correlation, estimated from each class of models, in generating the desired output correlation r^2 for a pair of SNPs with genotypes $G = \{0, 1, 2\}$. The range of desired r^2 values considered are 0.1 to 0.9 by intervals of 0.1 and MAF values considered are {0.01, 0.05, 0.01, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50}.

For a desired r^2 value, the estimated input correlation, $\hat{\rho}$, is calculated based the estimated coefficients from each class of linear models for all MAF values considered. The set of $\hat{\rho}$ obtained is then used as the input correlations to generate the binomial variables following the procedure by Wang and Abbott (2008). This process is repeated 1,000 times. At each replicate, an r^2 value for each MAF is estimated based on the generate binomial variables. A small set of the simulation results is displayed in Figure 2.4. The plots show the observed r^2 values by MAF obtained from each model's estimated $\hat{\rho}$ for desired r^2 values: 0.2, 0.5, and 0.8. Table 2.3 show the $\hat{\rho}$ values obtained from each model by MAF value for the desired r^2 of 0.5. For the simulation results across the range of desired r^2 values considered, see Tables A.6, A.7, and A.8 in Appendix A for the observed median values of r^2 produced from the $\hat{\rho}$ values of the Naive Model, the Fisher Model 1, and the Fisher Model 2, respectively.

From the simulation results, it is evident that while the Naive model is reliable for smaller desired r^2 values, it fails as the desired r^2 increases past 0.7. For the desired r^2 value of 0.8, the Naive Model indicates that an input correlation $\hat{\rho} = 1$ should be used for MAF values 0.15 or higher. This results in an overestimate of r^2 for large MAF values and an underestimate for small MAF values. The Fisher Models, on the other hand, are more consistent over the entire range of desired r^2 considered. Fisher

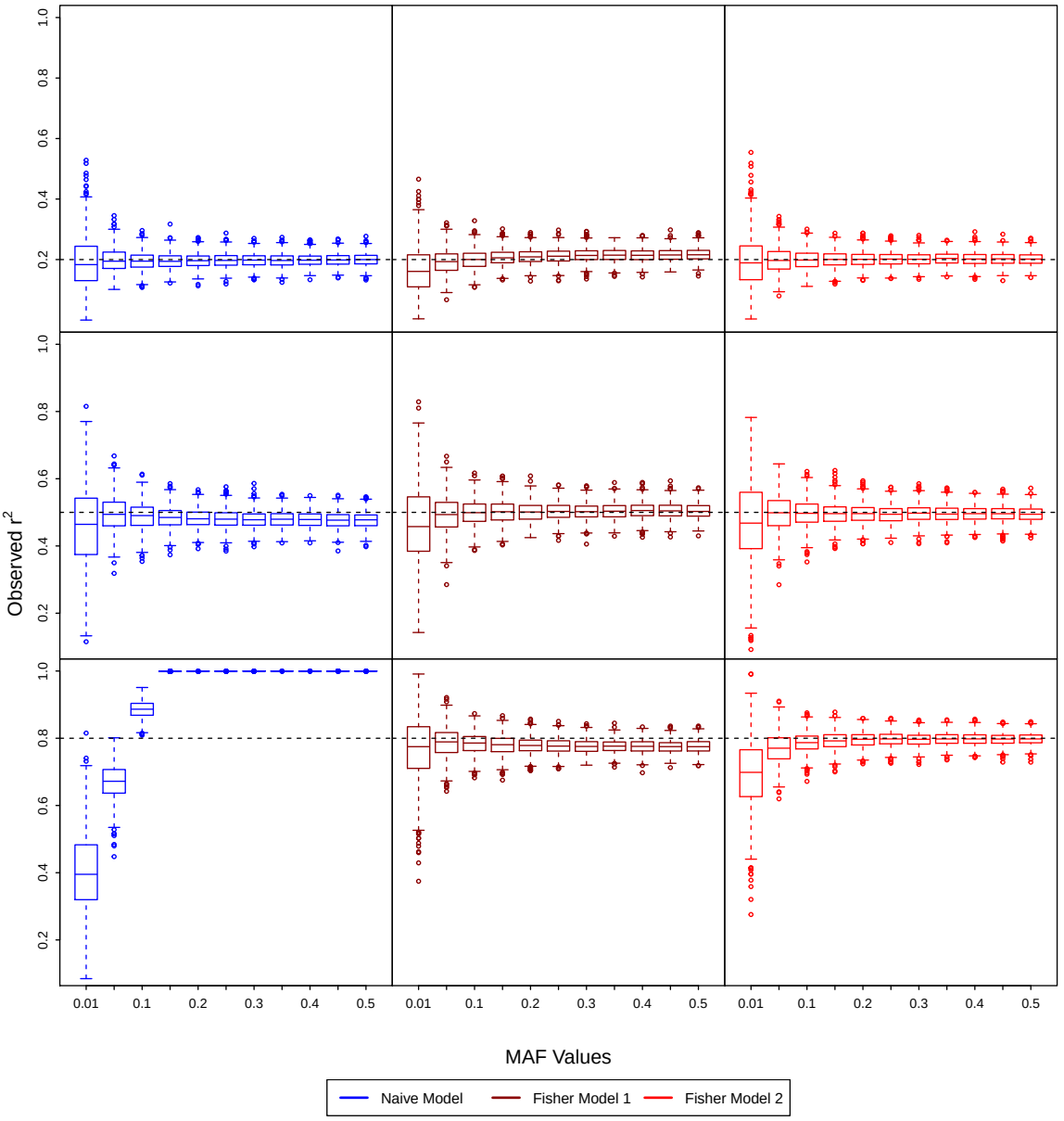

Figure 2.4: Observed r^2 values by linear models and MAF for the desired r^2 values of 0.2, 0.5, and 0.8. The observed r^2 values are estimated from a pair of generated binomial variables using each model's estimate $\hat{\rho}_i$ for some MAF value i. Top panels are results for desired $r^2 = 0.2$. Middle panels are results for desired $r^2 = 0.5$. Bottom panels are results for desired $r^2 = 0.8$.

MAF	Naive Model $\hat{\rho}$	Fisher Model 1 $\hat{\rho}$	Fisher Model 2 $\hat{\rho}$
0.01	0.9442	0.9444	0.9450
0.05	0.9237	0.9239	0.9243
0.10	0.9012	0.9053	0.9050
0.15	0.8841	0.8922	0.8907
0.20	0.8718	0.8827	0.8802
0.25	0.8620	0.8753	0.8718
0.30	0.8547	0.8696	0.8653
0.35	0.8497	0.8658	0.8610
0.40	0.8456	0.8627	0.8573
0.45	0.8438	0.8613	0.8558
0.50	0.8429	0.8607	0.8549

Table 2.3: Estimated input correlation, $\hat{\rho}$, for desired $r^2 = 0.5$ ($r = 0.7071$), by class of linear models and MAF values.

Model 1 shows a slightly inflated median observed r^2 than Fisher Model 2 for desired $r^2 = 0.2$, but it is comparable to the performance of Fisher Model 2 for desired $r^2 = 0.5$. For desired $r^2 = 0.8$, Fisher Model 1 shows a slightly deflated median observed r^2 than Fisher Model 2. Overall, the observed r^2 derived from Fisher Model 2 show the most consistent median value over the range of desired r^2 and MAF values considered.

It should be noted that none of the models demonstrate a consistent ability to produce the desired r^2 when the MAF is at 0.01 or 0.05. All three classes of linear models tend to underestimate the required input correlation value needed to achieve the desired r^2 in these two cases. This is due to small MAFs exaggerating the non-linear relationship between input and output correlations. This nonlinearity is not captured accurately in any of the models considered.

Let $\tilde{r}^2$ represent the estimator of the desired output r^2 value. Then $\tilde{r}^2_m$, for $m = 1, 2, 3$, represents the estimator of r^2 for the three classes of linear models considered. Each class of linear models produces an estimated input correlation, $\hat{\rho}_i$, for a given r^2 and MAF value i. This is then used to generated a pair of correlated binomial

	$\text{var}(\tilde{r}^2)$	$\text{bias}(\tilde{r}^2)$	$\text{MSE}(\tilde{r}^2)$
Naive Model	0.0123	0.0055	0.0124
Fisher Model 1	0.0018	-0.0063	0.0018
Fisher Model 2	0.0021	-0.0076	0.0021

Table 2.4: Variance, Bias, and MSE of Estimator $\tilde{r}^2$ by Linear Model

variables with the same MAF as described in the above simulation. Let $\hat{r}^2_m$ denote the observed r^2 values of the generated binomials by each class of model for all MAFs considered. Then, $\hat{r}^2_m$ can be viewed as observations of the random variable $\tilde{r}^2_m$ for $m = 1, 2, 3$, the three classes of linear models. Thus, the performance of the three classes of models can also be evaluated using the variance, bias, and mean squared error (MSE) of the estimator $\tilde{r}^2_m$. The statistics are estimated as follows:

$$\text{var}(\tilde{r}^2) \approx \frac{1}{NQR} \sum_{i=1}^{N} \sum_{j=1}^{Q} \sum_{k=1}^{R} \left(\hat{r}^2_{ijk} - \bar{r}^2_{ijk}\right)^2$$

$$\text{bias}(\tilde{r}^2) \approx \frac{1}{NQR} \sum_{i=1}^{N} \sum_{j=1}^{Q} \sum_{k=1}^{R} \left(\hat{r}^2_{ijk} - r^2_i\right)$$

$$\text{MSE}(\tilde{r}^2) \approx \text{var}(\tilde{r}^2) + \text{bias}^2(\tilde{r}^2)$$

where $N = 9$, is the number of desired r^2 values considered[2]; $Q = 11$, is the number of MAF values considered[3]; R = 1,000 is the number of replicates in the simulation study; $\hat{r}^2_{ijk}$ is the observed r^2 value at the k-th replicate derived from $\hat{\rho}_{ij}$ given the i-th desired r^2 and j-th MAF value; and r^2_i is the i-th desired r^2 value.

Table 2.4 shows the estimated variance, bias, and mean squared error of each model's estimator $\tilde{r}^2_m$. The variance of Fisher Model 1 and Fisher Model 2 are both lower than the Naive Model as expected. Overall, Fisher Model 1 and Fisher Model 2

[2]Desired r^2 values considered are: {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}.

[3]MAF values considered are: {0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50}

have lower MSE than the Naive model. Based on the MSE and the above simulation results, it is recommended that Fisher Model 2 be used to estimate the required input correlation given a desired output correlation r^2 for a certain MAF value.

It should be noted that although the Fisher transformed models appear to satisfy the assumptions of linear regression better than the Naive model, the residuals of the Fisher transformed models are not normally distributed. This is a result of the nonlinear relationship between input and output correlations that is still significant near the boundaries of no correlation and perfect correlation, especially in the case of small MAF values. As such, this nonlinearity is not accurately captured by a linear nor a quadratic term in any of the models considered. However, the Fisher Models are still better overall fits based on the simulation results. This is also reflected in the R^2 value of the Fisher Models, see Table A.5 in Appendix A.

2.4 Discussion

A further step to consider in the regression analysis is to examine the behaviour of the linear models given different MAF values for the two SNPs. Based on the simulation results on the effect of the differences in MAF value, it is predicted that for small desired r^2 values, the differences in MAF will not result in significant changes in the required input correlation value. However, as the desired r^2 value increases, the differences in the MAF become more important in determining the required input correlation as large differences set a limit on the maximum r^2 achievable. As such, the models considered in this chapter require modifications in order to produce reliable estimates of the input correlation, ρ, when the MAF values differ for large values of desired r^2.

It should be noted that the simulation studies of this chapter only consider the special case of generating a pair of correlated SNPs. In this simple setting, the LD structure of the two SNPs is represented by a single parameter. However, one challenge in implementing this procedure is constructing LD patterns for multiple SNPs. In practice, a group of SNPs will likely be in LD with each other and possibly demonstrate cross correlation with other groups of SNPs. The number of SNPs in LD and their strength of LD will depend on the type of data to be analyzed. Another challenge when considering multiple SNPs in LD is ensuring that the input variance-covariance matrix for the multivariate normal variables is positive definite in order to utilize the method proposed by Wang and Abbott (2008).

Chapter 3

Two-Step G x E Methods for a Quantitative Trait

The current two-step methods for G x E analysis were developed and illustrated in the GWAS setting which typically considers case-control data pertaining to a disease status. This chapter examines the performance of the two-step methods in detecting G x E effects given a continuous quantitative trait and the presence of LD in the genotyped markers. It is assumed that there exist genetic effects (i.e. marginal genetic association or G-E interactions) that impact quantitative phenotypes. Some common examples of quantitative traits that can be affected by genetics as well as the environment are blood pressure, height, and body mass index.

3.1 Methods

The two-step methods examined in the quantitative trait setting are the Disease-Gene (DG) method proposed by Kooperberg and LeBlanc (2008), the Environment-Gene

(EG) method proposed by Murcray et al. (2009), the Hybrid (H2) method proposed by Murcray et al. (2011), and the EDGE method proposed by Gauderman et al. (2013). The application of the two-step methods as well as the exhaustive search method in the quantitative trait setting is described next.

3.1.1 Exhaustive Search for G x E

Let y denote a continuous quantitative trait, G denote the quantitative trait locus (QTL), E denote an environment factor, and D denote the disease status. It is assumed that the disease is associated with the quantitative trait. Then the effects of gene, environment, disease, and G-E interaction on the quantitative trait can be modelled as:

$$y = \beta_0 + \beta_g G + \beta_e E + \beta_d D + \beta_{ge} G \times E + \epsilon \tag{3.1}$$

where $\epsilon \sim N(0, \sigma)$ is some random noise independent of the covariates. To detect a G x E effect, the null hypothesis $H_0 : \beta_{ge} = 0$ is tested. Model (3.1) is the linear equivalent of the CC Model (1.1) described in Chapter 1. It should be noted that there does not exist a linear model that is equivalent to the CO Model (1.1) described in Chapter 1. As such, the Bayesian approaches also not do apply in the quantitative trait setting.

The exhaustive search method fits Model (3.1) for every SNP in the sample. Then presence of the G x E effects are determined based on some significance level α corrected for multiple testing. The ordinary least squares estimates from Model (3.1) are unbiased and consistent. As such, the exhaustive search method is typically expected to maintain the type I error rate. However, the power of this approach decreases as the number of tests increase. Consequently, the exhaustive search method

has poor power in detecting small to moderate G x E effects given a large number of SNPs.

3.1.2 Disease-Gene Two-Step Method

The DG method is based on the hypothesis that a SNP with a G x E effect will likely also demonstrate a marginal genetic effect on the response. As such, filtering based on marginal genetic effect is a justified approach to eliminate irrelevant SNPs. For the DG method in the quantitative trait setting, the screening step fits the model:

$$y = \beta_0^* + \beta_g^* G_i + \epsilon_i^* \tag{3.2}$$

for $i = 1, ..., M$, where M is the total number of SNPs genotyped in the sample and ϵ_i^* is some normally distributed random noise. Note that $\epsilon_i^* \sim N(0, \sigma_i^*)$ and the parameters β_0^*, β_g^*, and σ_i^* are different than those in Model (3.1). Model (3.2) is the linear equivalent of Model (1.4), which is the logistic regression model used in the screening step for the DG method in the GWAS setting as described in Chapter 1.

In the quantitative trait setting, the null hypothesis: $H_0 : \beta_g^* = 0$ is tested for each SNP. A SNP is passed onto the second step if the p-value of its test statistic is less than some step 1 threshold α_1. The full linear model, Model (3.1), is then used to test for the presence of G x E effects at the second step.

3.1.3 Environment-Gene Two-Step Method

In a case-control panel, the oversampling of cases induces a correlation between the environment factor and the disease susceptibility locus (DSL) if a G x E effect exists.

If the QTL and the DSL is the same SNP, then screening SNPs based on any gene-environment associations is justified in the quantitative trait setting given a case-control sample. For binary E, the screening step of the EG method uses the following model:

$$\text{logit}(P(E = 1|G)) = \delta_0 + \delta_g G_i \tag{3.3}$$

for $i = 1, ..., M$, where M is the total number of SNPs genotyped in the study. For each SNP, the null hypothesis: $H_0 : \delta_g = 0$ is tested. The SNP is passed onto the second step if the p-value of its test statistic is less than some step 1 threshold α_1. At second step, Model (3.1) is used to detect the presence of any G x E effects.

3.1.4 EDGE Two-Step Method

For the EDGE method, the test statistic is defined as $S_{EDGE} = S_{EG} + S_{DG}$, where S_{EG} is the test statistic from step 1 of the EG method for null hypothesis: $H_0 : \delta_g = 0$ and S_{DG} is the test statistic from step 1 of the DG method for null hypothesis: $H_0 : \beta_g^* = 0$. Thus the null hypothesis for the EDGE method is: $H_0 : \beta_g^* = \delta_g = 0$. The test statistics, S_{EG} and S_{DG}, are asymptotically χ^2-distributed with one degree of freedom[1]. As such S_{EDGE} is asymptotically χ^2-distributed with two degrees of freedom.

For each SNP in the sample, the test statistic S_{EDGE} is calculated and if its p-value is less than some step 1 threshold α_1 the SNP is passed onto the second step for formal G x E testing using Model (3.1). The EDGE method is a combined two-step

[1]Note that the usual test statistic for the DG screening test has a student's t-distribution. However, in the case of large samples, the distribution is approximately standard normal and thus can be squared to obtain a χ_1^2-distribution. This also applies for the EG screening test statistic if the environment is a continuous variable.

approach using both DG and EG screening criteria to filter out irrelevant SNPs.

3.1.5 Hybrid Two-Step Method

The original H2 method proposed by (Murcray et al., 2011) runs both the DG screening step and the EG screening step to screen for relevant SNPs. In the quantitative trait setting, the H2 algorithm applies without any additional changes than those already specified under the DG and EG methods. For full description of the H2 procedure, see Chapter 1. It should be noted that the H2 method is also a combined two-step approach utilizing both the DG and the EG methods.

3.1.6 Additional Notes

It should be noted that all of the two-step methods considered in this chapter uses Model (3.1) in the second step to test for G x E effects. As such, the power of the two-step methods considered will rely on their ability pass the QTL onto the second step. The validity of the two-step methods relies on the independence of the step 1 test statistic and the step 2 test statistic. In the linear regression setting, the parameter estimate $\hat{\beta}_g^*$ of Model (3.2) is independent of the parameter estimate $\hat{\beta}_{ge}$ of Model (3.1), thus their test statistics are also independent (Kooperberg and LeBlanc, 2008) and the independence condition is satisfied for the DG method. For the EG method, simulation results from this thesis show that the sample covariances of the step 1 and step 2 test statistics for binary environment are negligible and independence of the two statistics can be assumed. See Figure B.1 in Appendix B for a comparison of the step 1 test statistics of the DG and the EG method against the test statistic used in the second step. For the EDGE method, since the DG screening test statistic and

the EG screening test statistic are both independent of the step 2 test statistic, it follows that the sum of the DG and EG screening test statistic is also independent of the step 2 test statistic (Gauderman et al., 2013).

It should also be noted that the Cocktail methods (Hsu et al., 2012) are not considered by this thesis. This is because these methods do not exclude any SNPs from formal G x E testing. Instead, the Cocktail methods utilize the empirical Bayes method[2] (Mukherjee and Chatterjee, 2008) and the CC method[3] in the second step along with weighted hypothesis testing to gain power. Since only the full linear model, Model (3.1), is used in the second step for a quantitative trait, the Cocktail method becomes an exhaustive search as every SNP is tested at the second stage for the presence of G x E effects.

3.2 Simulation Study

A simulation study is used to evaluate the performance of the two-step methods in detecting G x E effects in the quantitative trait setting. This study considers two ways a sample with a continuous quantitative trait can be obtained. The first is by taking a random sample of the population and measuring the quantitative trait. The second is by using a continuous covariate captured by a case-control study as the quantitative trait. This thesis will use the terms random sample and case-control sample, respectively, to denote the type of data obtained.

[2]See Equation (1.3) from Chapter 1
[3]See Model (1.1) from Chapter 1

3.2.1 Generation of SNPs

For the simulation study, a total of 1,000 SNPs are generated to represent the genetic data. In a typical GWAS, the number of markers genotyped can range from 10,000 to 1 million SNPS. However, in the interest of time, this thesis has only considered a small subset of the typical volume of markers seen in a GWAS setting.

Out of the 1,000 SNPs considered for the simulation, 25 are generated in linkage disequilibrium based on the procedure by Wang and Abbott (2008) as described in Chapter 2. The 25 SNPs are divided into five equally sized LD blocks with the QTL located in the first block. Each of the five SNPs within a LD block share a high degree of correlation with each other. The r^2 of the SNPs within each LD block is approximately 0.66. Each of the LD blocks also have some degree of cross correlation. SNPs in adjacent blocks share a moderate level of cross correlation with an approximate r^2 of 0.38. The SNPs that are two blocks apart share a moderately weak level of LD with an approximate r^2 of 0.18. The SNPS that are three blocks apart share a weak level of LD with an approximate r^2 of 0.08. Lastly, SNPs in block 1 and block 5 are in very weak LD with an approximate r^2 of 0.02. Figure 3.1 illustrates the block structure and the LD between the SNPs by block.

The MAF of the QTL is set at 0.134, which approximately gives $P(G = 1) = 0.25$ under a dominant genetic model. For the SNPs in LD with the QTL, their MAFs are selected in a manner that produces the desired LD structure as shown in Figure 3.1. See Appendix D for the details on generating the MAFs of the SNPs in LD with the QTL. The input correlations for the multivariate normal variables are: 0.96 for within block correlation; and 0.85, 0.65, 0.45, 0.25 for the respective cross block correlations. The remaining SNPs in this simulation study are generated as independent binomial

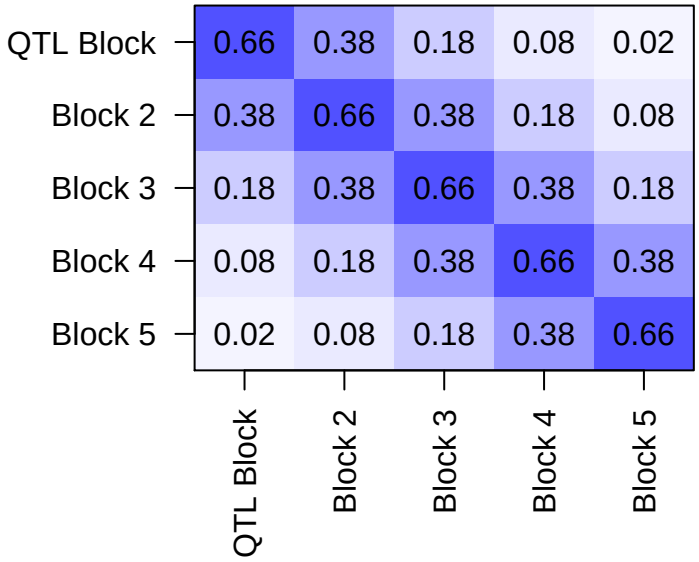

Figure 3.1: Approximate r^2 of SNPs in LD by block. Each block contains 5 SNPs that are in strong LD with each other. The SNPs within each block are also in LD with the SNPs from other blocks with r^2 as specified.

variables with MAFs randomly selected from a uniform distribution between 0.05 and 0.4. 1,000 SNPs are generated for a population of 500,000 individuals based on dominant genetic coding. This means a single cutoff value is used to generated the SNPs in LD. For the QTL, the cutoff $c_1 = 0.6745$ is used to convert the normal variables into binomial variables.

3.2.2 Data Generating Models

To allow for the comparison between random and case-control samples, population data for 500,000 individuals is generated using a G-E association model, a disease model, and a quantitative trait model. The disease and quantitative trait models both assume there exists a single G x E effect. It is assumed that the disease status affects the quantitative trait and not vice versa. As such, the disease status for the population is generated before the quantitative trait.

In this simulation, the environment factor, E, is coded as a binary variable. This is a common practice described in the current literature to represent exposed and unexposed individuals. To simulate a small degree of G-E dependency, 10 SNPs independent of the QTL are randomly selected to have a G-E association. The following model is used to generate the binary environmental variable:

$$\text{logit}(P(E = 1|\mathbf{G})) = \theta_0 + \sum_{i \in A} \theta_{ge} G_i$$

where A is the set of SNPs with a G-E association. The parameter $\theta_0 = \text{logit}(0.2)$, where 0.2 is the underlying exposure rate, the parameter $\theta_{ge} = 0.4$ indicates moderate association between the gene and the environment, and $\mathbf{G}$ are the 10 randomly

selected SNPs in set A. Overall, this model results in a population exposure rate of approximately 59%.

Next, the disease status is generated using the genotypes and the environment variable. The disease model is defined as:

$$\text{logit}(P(D = 1|G, E)) = \gamma_0 + \gamma_g G_D + \gamma_e E + \gamma_{ge} G_D \times E \tag{3.4}$$

where G_D is the DSL, $\gamma_0 = -3$, indicating an underlying disease prevalence rate of approximately 4.7%, and $\gamma_g = \gamma_e = \gamma_{ge} = \log(1.3)$ representing relatively weak associations of the gene, the environment and the G x E effect with the disease. These parameters result in an overall disease prevalence rate of approximately 6.2%. This rate is comparable to the 2011 prevalence of Type 2 diabetes in Canada (Government of Canada, Public Health Agency of Canada, 2011). It should be noted that the third SNP is designated as the DSL which has an $r^2 = 0.66$ with the QTL. The MAF of the DSL is approximately 0.116 which gives $P(G_D = 1) \approx 0.22$. All parameters of the disease model are held fixed throughout the simulation study.

Lastly, the quantitative trait is generated using the genotypes, the environment variable, and the disease status. The quantitative trait model used to generate the data is the same as Model 3.1:

$$y = \beta_0 + \beta_g G + \beta_e E + \beta_d D + \beta_{ge} G \times E + \epsilon$$

where G, the first SNP, is designated as the QTL, $\beta_0 = 0$, $\beta_e = \beta_d = 0.4$ indicating moderate influences of the environment and the disease status on the quantitative trait, and $\epsilon \sim N(0, 1)$. The values for β_g considered for the simulation study are: {0,

0.2, 0.4}. This represents zero, weak, and moderate marginal genetic associations with the quantitative trait. To evaluate the power of the two-step methods in detecting G x E effects, the parameter β_{ge} is varied between -0.4 and 0.4 by increments of 0.1. Varying the β_g and β_{ge} parameters produces a total of 27 scenarios. The quantitative trait variable, y, is generated for each of the 27 scenarios for a population of 500,000. For each of the 27 scenarios, the same set of genotypes, exposure status, and disease status (generated earlier) is used.

It should be noted that the QTL in the quantitative trait model is not the same SNP as the DSL in the disease model. However, the DSL is within the same LD block as the QTL. Since these two causal SNPs are highly correlated, the EG method remains valid for case-control samples. See Appendix D for complete details on generating the population data used for the simulation study.

For each of 27 scenarios considered, a total of 1,000 simulation replicates are performed. At each simulation replicate, a sample of 1,000 individuals is selected from the population data. Random samples are constructed by simple random sampling where each individual has an equal probability of being selected. Case-control samples are constructed by randomly selecting 500 cases and 500 controls based on the disease status. The two-step methods are then applied to the sample data to look for the presence of G x E effects.

The step 1 threshold, α_1, for all two-step methods is set at 0.05. This allows approximately 50 SNPs to be passed onto the second step. The step 1 threshold was chosen based on the literature which suggests an α_1 level that passes between 10 to 100 SNPs (Gauderman et al., 2013). For the H2 method, the step 1 thresholds are set at $\alpha_{1m} = \alpha_{1a} = \alpha_1$ and $p = 0.5$ to favour neither the DG nor the EG method.

3.3 Results

3.3.1 Family-Wise Error Rate

The family-wise type I error rate, $\alpha = 0.05$, is used to control false positives. The Bonferroni correction is used for multiple testing. The type I error rate is calculated as the proportion of null SNPs declared significant at α level 0.05 out of the total number of SNPs tested across all replicates for a given scenario. It should be noted that the type I error rate is calculated based on the 975 independent SNPs when $\beta_{ge} \neq 0$ and all 1,000 SNPs when $\beta_{ge} = 0$.

The type I error rates are shown in Figure 3.2. Panels in left column are the type I error rates of the two-step methods for random samples and panels in the right column are the type I error rates for case-control samples. The top, middle, and bottom panels of Figure 3.2 indicate the cases of zero, weak, and moderate marginal genetic effects, respectively. Overall, the type I error rate is well maintained for all two-step methods for varying degrees of marginal genetic association and G x E effects across sample types. The type I error rate appears to be slightly more varied in case-control samples compared to random samples. The numeric results of the type I error rate are shown in Table A.9 and Table A.10 in Appendix A.

3.3.2 Power

Two types of power are examined by this thesis. The first, is the power to detect a true G x E effect involving the QTL. This is termed power to detect the QTL. The second, is the power to detect G x E effects for any SNPs in LD with the QTL. In this case, it is assumed that the QTL is not in the sample but SNPs in LD with the

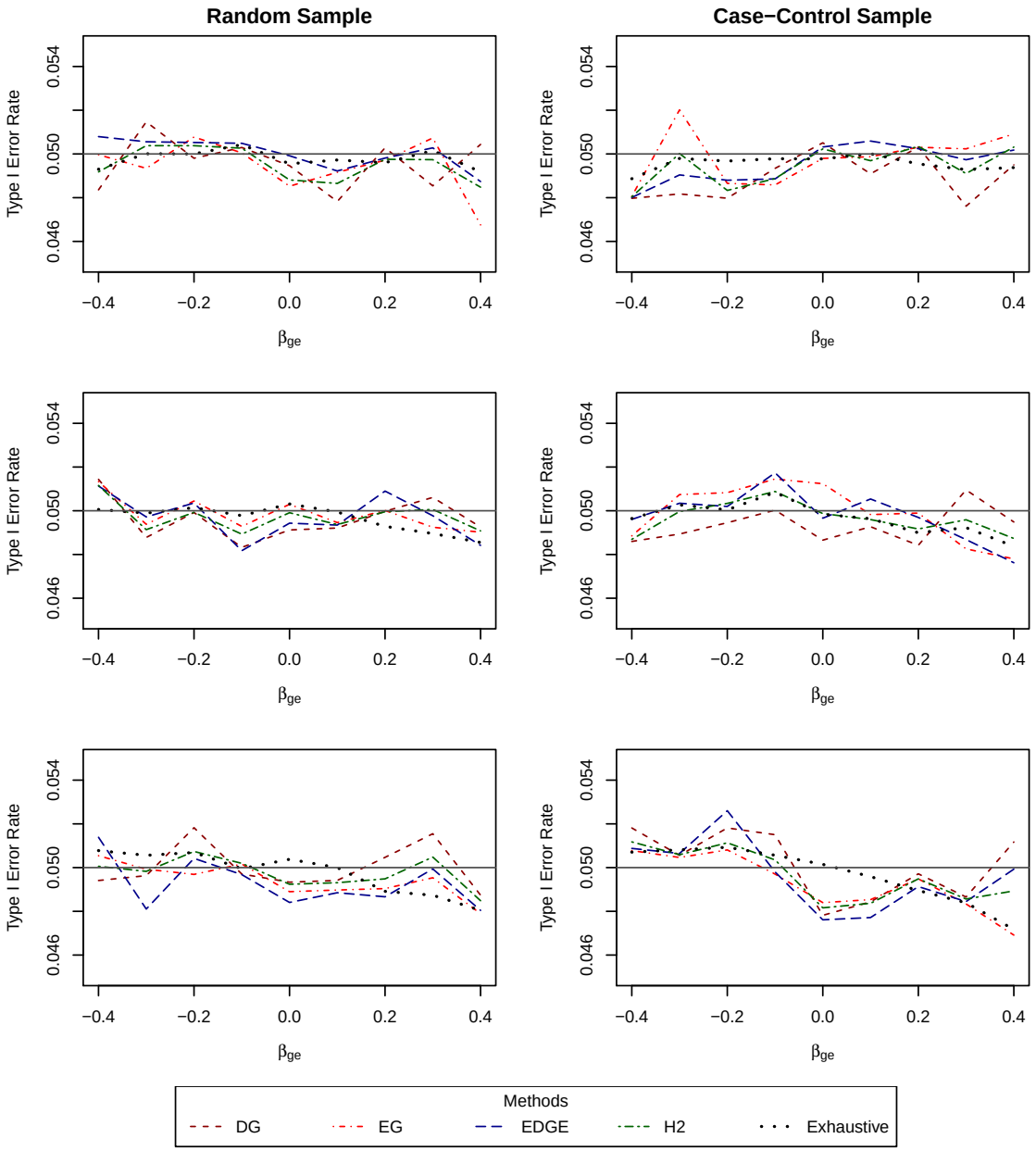

Figure 3.2: Family-wise error rate for all methods in random and case-control samples. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

QTL are genotyped. This is termed power to detect the QTL region. The two types of power are considered since in a typical GWAS, it is unlikely that the true causal SNP is genotyped. However, the basis of GWAS relies on linkage disequilibrium, thus the power to detect SNPs that are in LD with the causal SNP is important.

The power to detect the QTL is defined as the proportion of replicates that declare the QTL significant based on the Bonferroni corrected thresholds for each method. The power to detect the QTL region is defined as the proportion of replicates that declare at least one SNP in LD with the QTL significant based on the Bonferroni corrected thresholds for each method without the QTL in the sample.

The power to detect the QTL is shown in Figure 3.3. The left column show the results pertaining to random samples and the right column show the results pertaining to case-control samples. The top, middle, and bottom panels indicate the cases of zero, weak, and moderate marginal genetic associations, respectively. Overall, the DG and the EDGE methods have comparably the highest power to detect the G x E effect across most scenarios in both sample types. The H2 method demonstrates higher power than both the exhaustive search and the EG method. However, the H2 method typically has lower power than the DG and the EDGE methods. The power of the two-step methods are similar across the two types of samples with case-control samples resulting in a slight increase in power for all methods considered.

It should be noted that in the case when the marginal genetic effect is weak ($\beta_g = 0.2$) and the G x E effect is in an opposing direction, the power of all two-step methods is lower than the exhaustive search method. However, for the case of $\beta_g = 0.4$, the power of the two-step methods, with the exception of the EG method, rebounds and is higher than the exhaustive search method.

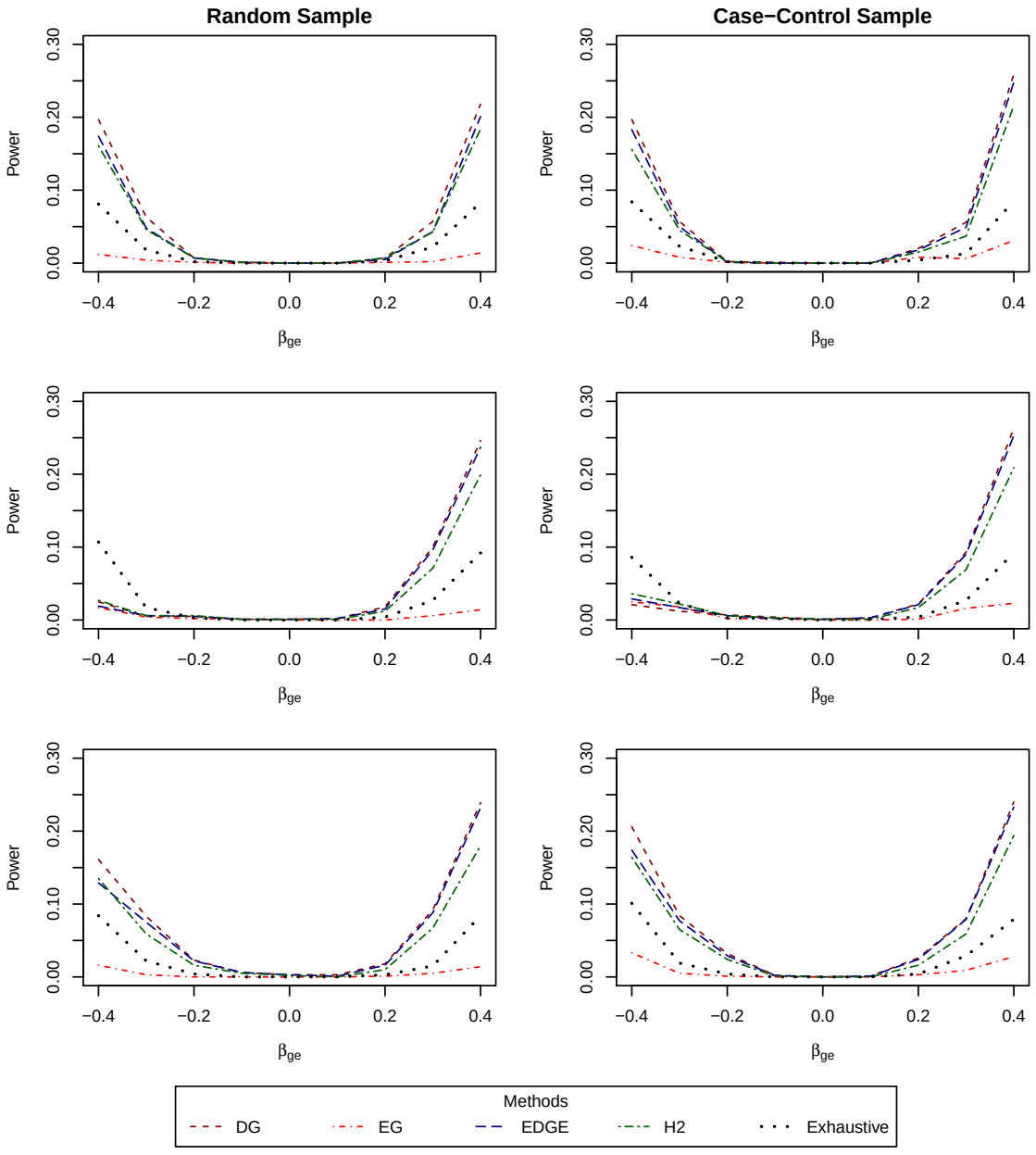

Figure 3.3: Power to detect the QTL for all methods in random and case-control samples. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

The power to detect the QTL region is shown in Figure B.2 in Appendix B. In general, the power of the two-step methods to detect the QTL region is similar to the power to detect the QTL across all scenarios and sample types. In some cases, the power to detect the QTL region is slightly higher than the power to detect the QTL. Again, the opposing effects of marginal genetic association and G x E effect results in low power for $\beta_g = 0.2$ in all samples. Similar to the results for power to detect QTL, case-control samples typically result in higher power to detect, especially in the case where the marginal genetic association and the G x E effect are in the same direction.

3.4 Discussion

The simulation study shows that the family-wise type I error rates are maintained by the two-step methods in the quantitative trait setting. All of the two-step methods, except for the EG method, generally outperformed the exhaustive search. The DG, the EDGE, and the H2 methods all demonstrated good power in detecting the QTL and the QTL region under a wide range of scenarios. The power to detect the QTL and the QTL region is slightly higher in the case-control setting.

It is interesting to note that the EG method performs poorly across all scenarios in both random and case-control samples. It is expected that the EG method should perform poorly in the random samples as cases are not oversampled. As such, the induced correlation between the environment and QTL will not be present in the random sample and the QTL is not frequently passed onto the second step for formal G x E testing. However, the EG method's lack of power in the case-control samples is unexpected. The factors influencing the EG method's power are examined in more detail in Chapter 4.

Chapter 4

Examining the EG Method

From the simulation results of Chapter 3, it is evident the EG method lacks power in detecting G x E effects across most of the simulation scenarios. As a result, the combined two-step approaches, EDGE and H2 methods, fared closer in performance to the DG method but generally exhibited lower power. The expected power gains from utilizing both methods simultaneously are not realized in this case. However, the settings used in chapter 3 are only a few possible ways to generate quantitative trait data. This chapter examines the various factors that impact the power of the EG method and considers alternative G-E association models for a quantitative trait.

4.1 Comparison of Pass Rates

All of the two-step methods considered use the full linear model (see Model 3.1) in the second step. As such, the power of each method is based on its ability to pass the QTL into the second stage for formal G x E testing. The pass rate, defined as the frequency in which a method passes the QTL onto the second step, is thus an indication of the

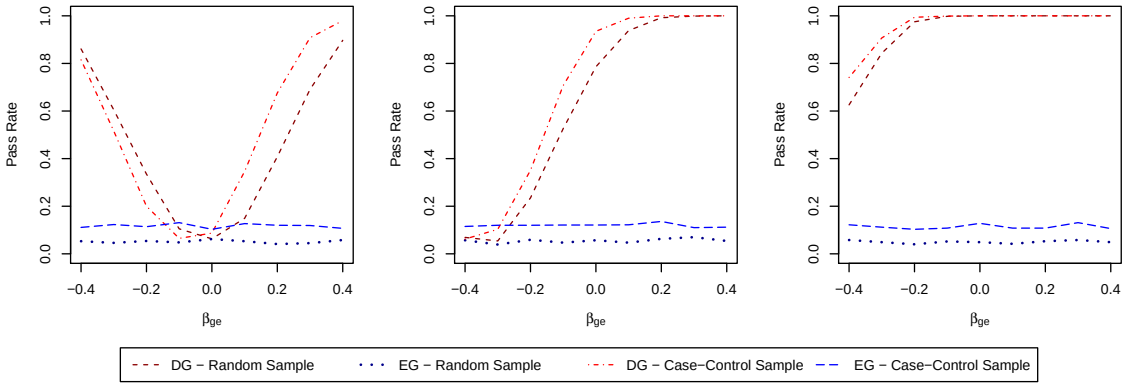

Figure 4.1: Pass rate for DG and EG methods in random and case-control samples from simulation results of Chapter 3. Left panel, zero marginal genetic effect ($\beta_g = 0$). Middle panel, weak marginal genetic effect ($\beta_g = 0.2$). Right panel, moderate marginal genetic effect ($\beta_g = 0.4$).

method's power. The pass rate of the DG and the EG methods from the simulation study in Chapter 3 are shown in Figure 4.1. From this information alone, it is clear the EG method fared poorly in both random and case-control samples across all levels of genetic and G x E effects. Since the performance of the EG screening test depends on correlation between the QTL and the environment, factors affecting this correlation directly impact the pass rate and hence the power. These factors are examined next.

4.2 Factors Affecting Power

4.2.1 Gene-Environment Dependency

A major influence on the EG method is the underlying gene-environment dependency. More specifically, the LD structure of the SNPs associated with the environment and the magnitude of the G-E association both affect the correlation between the QTL

and the environment. As the simulation results from Chapter 3 have shown, even with moderate G-E associations of $\theta_{ge} = 0.4$, the EG method does not perform well. This is because the SNPs that are associated with the environment factor are independent of the QTL. As such, the QTL is not correlated with the environment in random samples and only weakly correlated with the environment in case-control samples. However, if there are some G-E associations with SNPs in LD with the QTL, then it is expected that the QTL will be passed into the second step more frequently by leveraging LD. In this case, as the strength of the G-E association increases, the power of the EG method is expected to increase as well.

4.2.2 Sample Type

The original premise of the EG method is based on the oversampling of cases. If there exists any G x E effects, then the oversampling would induce a correlation between the causal SNPs and the environment factor. However, in the case of quantitative traits, the two ways to obtain a sample as described in Chapter 3 can impact the power of the EG method. The results from Chapter 3 demonstrated that the EG method did poorly in random samples, while its power improved slightly in case-control samples. It should be noted that the sample type would exert less influence on power if there are moderate to strong G-E associations that included SNPs in LD with the QTL.

4.2.3 LD of DSL and QTL

The induced correlation from oversampling cases is associated with the DSL if cases are defined based on the disease status. If the QTL is not the same SNP as the DSL, then the induced correlation from oversampling of cases may not carry over for

the QTL and the environment factor. The LD between the DSL and the QTL is an indication of the strength of the correlation between the QTL and the environment induced from oversampling cases. As LD decreases between the two causal SNPs, the induced correlation between the QTL and the environment also decreases. If the QTL and the DSL are independent SNPs, then the induced correlation from oversampling cases will not be present in the QTL and the environment factor.

4.2.4 G x E Effect in Disease Model

Lastly, the strength of the G x E effect in the disease model (γ_{ge} from Model (3.4)) influences the strength of the induced correlation from oversampling of cases for case-control samples. Stronger G x E effects in the disease model translates to stronger induced correlations between the DSL and the environment. If the QTL is in LD with the DSL, then induced correlation between the QTL and the environment will increase as a result of the stronger G x E effect in the disease model.

4.3 Demonstrating the Effect of the Various Factors

To demonstrate how the various factors affect the performance of the EG method, a small simulation study is conducted to measure the observed correlation between the QTL and the environment by varying the type of G-E association, the strength of LD between DSL and QTL, and the strength of the G x E effect in the disease model, Model (3.4), for random and case-control samples. Three main scenarios of G-E association are considered: a) G-E dependency in independent SNPs only with

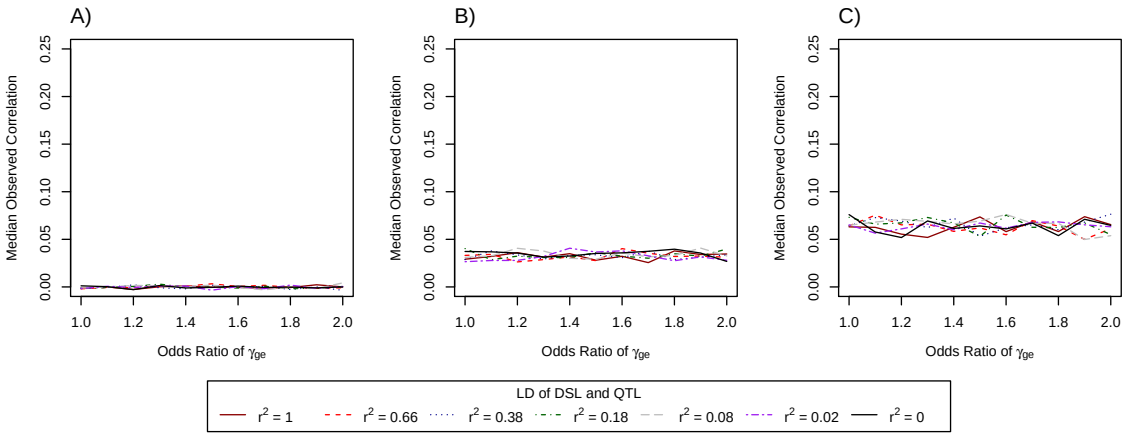

Figure 4.2: Observed median correlation of the QTL and the environment factor by γ_{ge} and strength of LD between DSL and QTL for random samples. Panel A), G-E dependency in independent SNPs only, $\theta_{ge} = 0.4$. Panel B), G-E dependency also in 3 SNPs in LD with QTL, $\theta_{ge} = 0.2$. Panel C), G-E dependency also in 3 SNPs in LD with QTL, $\theta_{ge} = 0.4$.

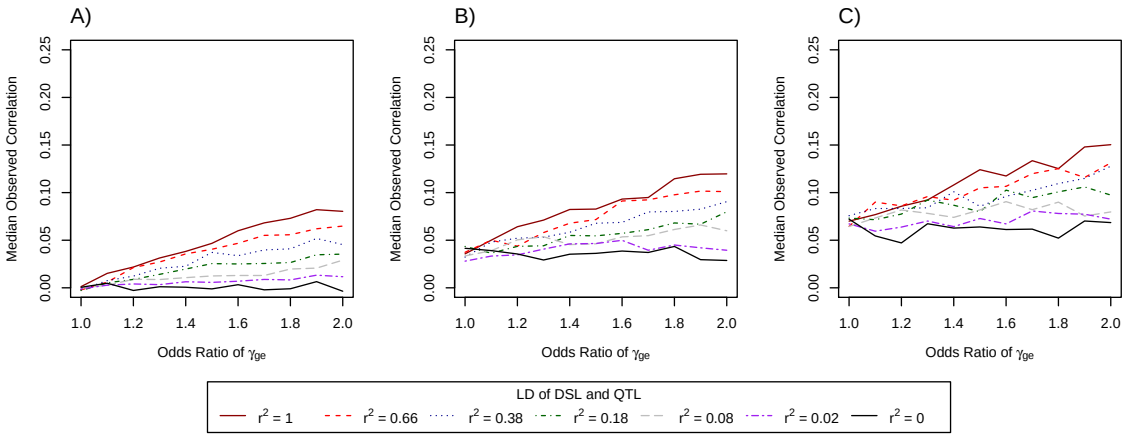

Figure 4.3: Observed median correlation of the QTL and the environment factor by γ_{ge} and strength of LD between DSL and QTL for case-control samples. Panel A), G-E dependency in independent SNPs only, $\theta_{ge} = 0.4$. Panel B), G-E dependency also in 3 SNPs in LD with QTL, $\theta_{ge} = 0.2$. Panel C), G-E dependency also in 3 SNPs in LD with QTL, $\theta_{ge} = 0.4$.

G-E association $\theta_{ge} = 0.4$; b) G-E dependency that include SNPs in LD with the QTL with G-E association $\theta_{ge} = 0.2$; and c) G-E dependency that include SNPs in LD with the QTL with G-E association $\theta_{ge} = 0.4$.

A total of 10 SNPs are selected to have a G-E association. For the cases b) and c) above, three SNPs in LD with the QTL are chosen to have a G-E association. The remaining seven SNPs are independent of the QTL are randomly selected. In the simulation study, SNPs 13, 20, and 23 from blocks 3, 4, and 5 are chosen to have an association with the environment factor. These SNPs have r^2 of 0.18, 0.08, and 0.02, with the QTL respectively. The data generating models from Chapter 3 are used to generate the population data under the three main scenarios considered.

The first SNP in the first LD block structure[1] is set as the QTL. Seven different SNPs are selected to represent the DSL (SNPs 1, 3, 6, 11, 16, 21, and 26). These seven choices represent the varying degrees of LD between the DSL and the QTL from $r^2 = 1$ to $r^2 = 0$. The odds ratio of the G x E effect in the disease model, γ_{ge} from Model (3.4), is varied from 1 to 2 by increments of 0.1. This results in 11 values of the G x E effect, γ_{ge}, considered for the simulation. By varying the γ_{ge} variable and the location of the DSL, a total of 77 cases are created for each of the three main scenarios of G-E association.

For each case, a random sample and a balanced case-control sample of 1,000 observations are selected from the population data and the correlation between the QTL and the environment factor is measured. This was repeated for 5,000 replicates. Figure 4.2 and Figure 4.3 display the median observed correlation of the QTL and the environment factor for random and case-control samples, respectively.

From the results, it is evident that if there exists some G-E association with SNPs

[1]See Section 3.2.1

that are in LD with the QTL, then the baseline correlation between the QTL and the environment increases as θ_{ge} increases for both sample types. The effects of the LD between the DSL and the QTL as well as the G x E effect in the disease model do not impact the observed correlation in random samples. However, these effects on the observed correlation are more prominent in case-control samples.

It should be noted that the observed median correlation between the QTL and the environment factor in G-E association scenarios b) and c) are generally higher that those seen from the simulation results of Chapter 3. Hence, it is expected that the pass rate of the EG method will improve due to increased correlation between the QTL and the environment factor. Figure 4.4 and Figure 4.5 display the pass rate of the EG method for the various cases considered in this simulation for random and case-control samples, respectively.

For random samples, the pass rate only improves by constructing G-E associations in some SNPs that are in LD with the QTL. For case-control samples, when the G-E dependency only exists in the independent SNPs, the pass rate is heavily influenced by the strength of the G x E effect in the disease model and the LD of the DSL and the QTL (γ_{ge} and r^2, respectively). However, when G-E dependency exists in some of the SNPs in LD with the QTL, the effects of γ_{ge} and r^2 are less pronounced on the pass rate of the EG method. In the case of moderate G-E dependency in some of the SNPs in LD with the QTL, $\theta_{ge} = 0.4$, the lowest pass rate hovers around 50% compared to 5% in the case where G-E dependency exist in independent SNPs only. From these results, it is clear that the EG method can perform well in both random and case-control samples depending on the settings of the influential factors.

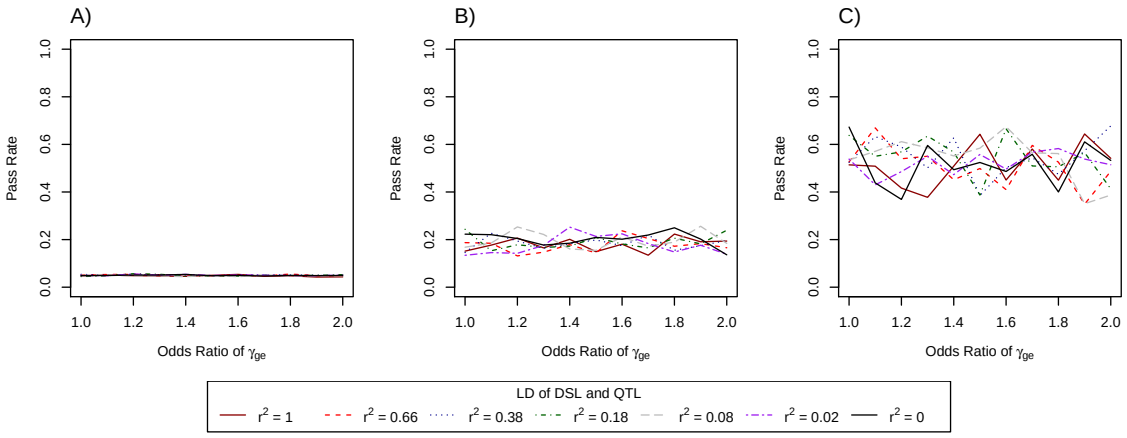

Figure 4.4: Pass rate of the EG method by G x E effect γ_{ge} and strength of LD between DSL and QTL for random samples. Panel A), G-E dependency in independent SNPs only, $\theta_{ge} = 0.2$. Panel B), G-E dependency also in 3 SNPs in LD with QTL, $\theta_{ge} = 0.2$. Panel C), G-E dependency also in 3 SNPs in LD with QTL, $\theta_{ge} = 0.4$.

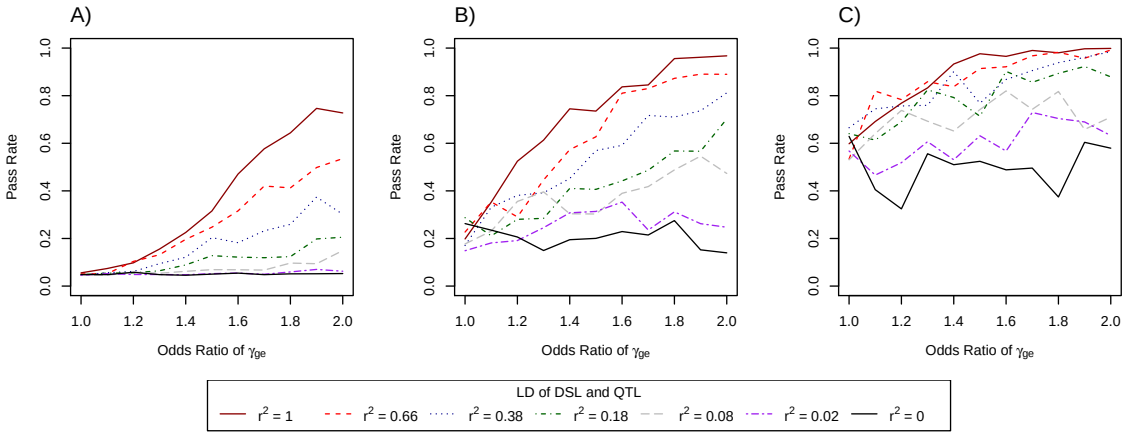

Figure 4.5: Pass rate of the EG method by G x E effect γ_{ge} and strength of LD between DSL and QTL for case-control samples. Panel A), G-E dependency in independent SNPs only, $\theta_{ge} = 0.2$. Panel B), G-E dependency also in 3 SNPs in LD with QTL, $\theta_{ge} = 0.2$. Panel C), G-E dependency also in 3 SNPs in LD with QTL, $\theta_{ge} = 0.4$.

4.4 Simulation Using Alternate Environment Generating Models

A simulation study is conducted to demonstrate that if some SNPs in weak LD with the QTL have a G-E association then the power of the EG method is improved. For this simulation, alternate environment generating models are used to simulate the environment variable. Two levels of G-E association scenarios are considered: weak G-E association $\theta_{ge} = 0.2$ and moderate G-E association $\theta_{ge} = 0.4$. A total of 10 SNPs are selected to have an association with the environment factor including three SNPs in LD with the QTL. The same 10 SNPs in scenarios b) and c) from Section 4.3 are used in this simulation study. All other parameters of the disease model and the quantitative trait model are held the same as the simulation study in Chapter 3. The levels of the marginal genetic effect, β_g, and the G x E effect, β_{ge}, of the quantitative trait model are same as those used in simulation study in Chapter 3. A total of 1,000 replicates are simulated. The family-wise error rate and power are calculated based on the simulation replicates. For description of the calculations, see Section 3.3 from Chapter 3.

4.4.1 Family-Wise Error Rate

Figure 4.6 and Figure 4.7 show the family-wise type I error rates for the two G-E association levels, $\theta_{ge} = 0.2$ and $\theta_{ge} = 0.4$, respectively. The type I error rate is well maintained by all the methods for $\theta_{ge} = 0.4$. However, there appears to be some slight systematic inflation of type I error when $\theta_{ge} = 0.2$ in random samples given some weak marginal genetic effect. This thesis is limited by time to fully explore the underlying

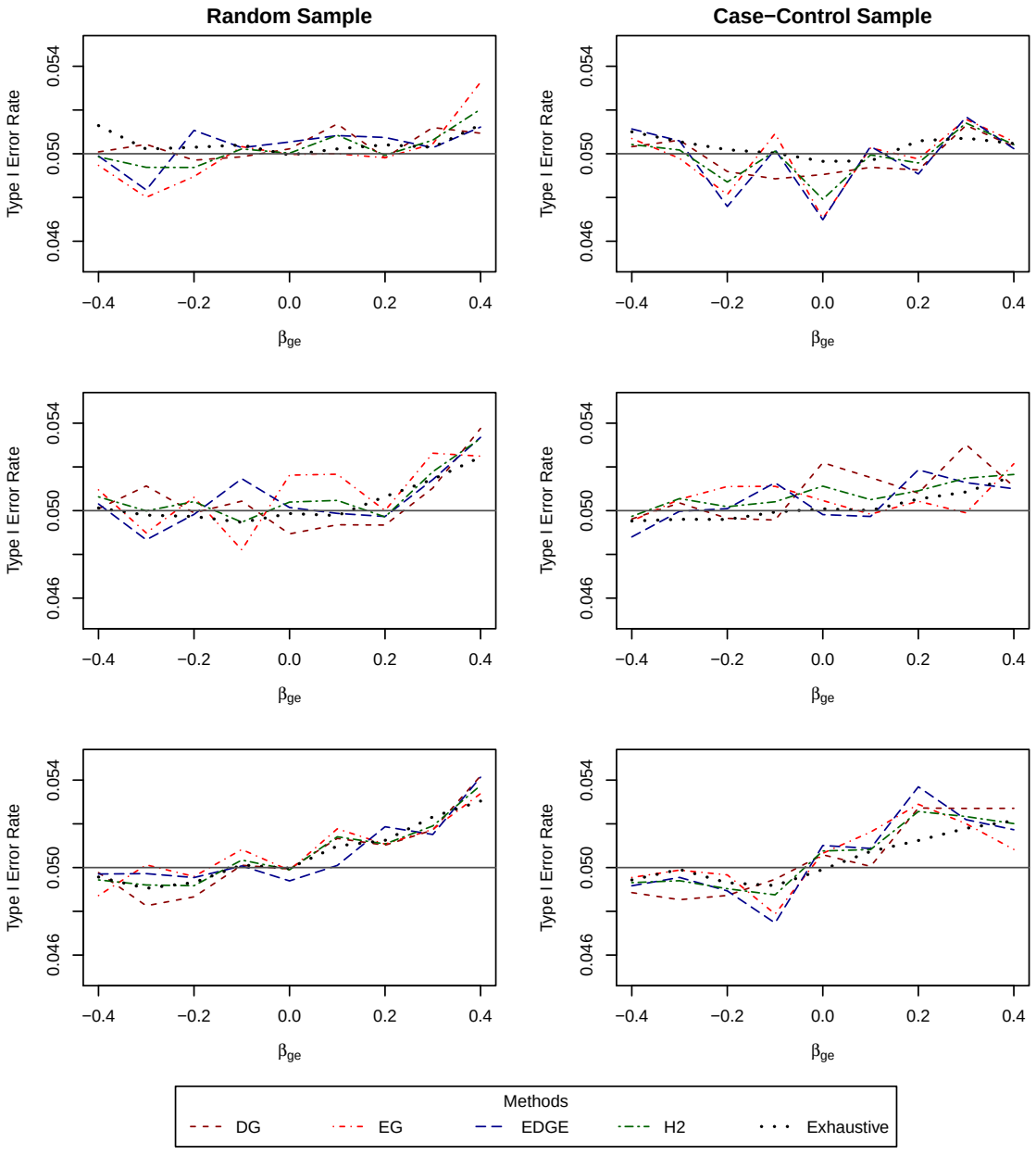

Figure 4.6: Family-wise error rate for all methods in random and case-control samples. G-E association parameter is $\theta_{ge} = 0.2$, including 3 SNPs in LD with QTL. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

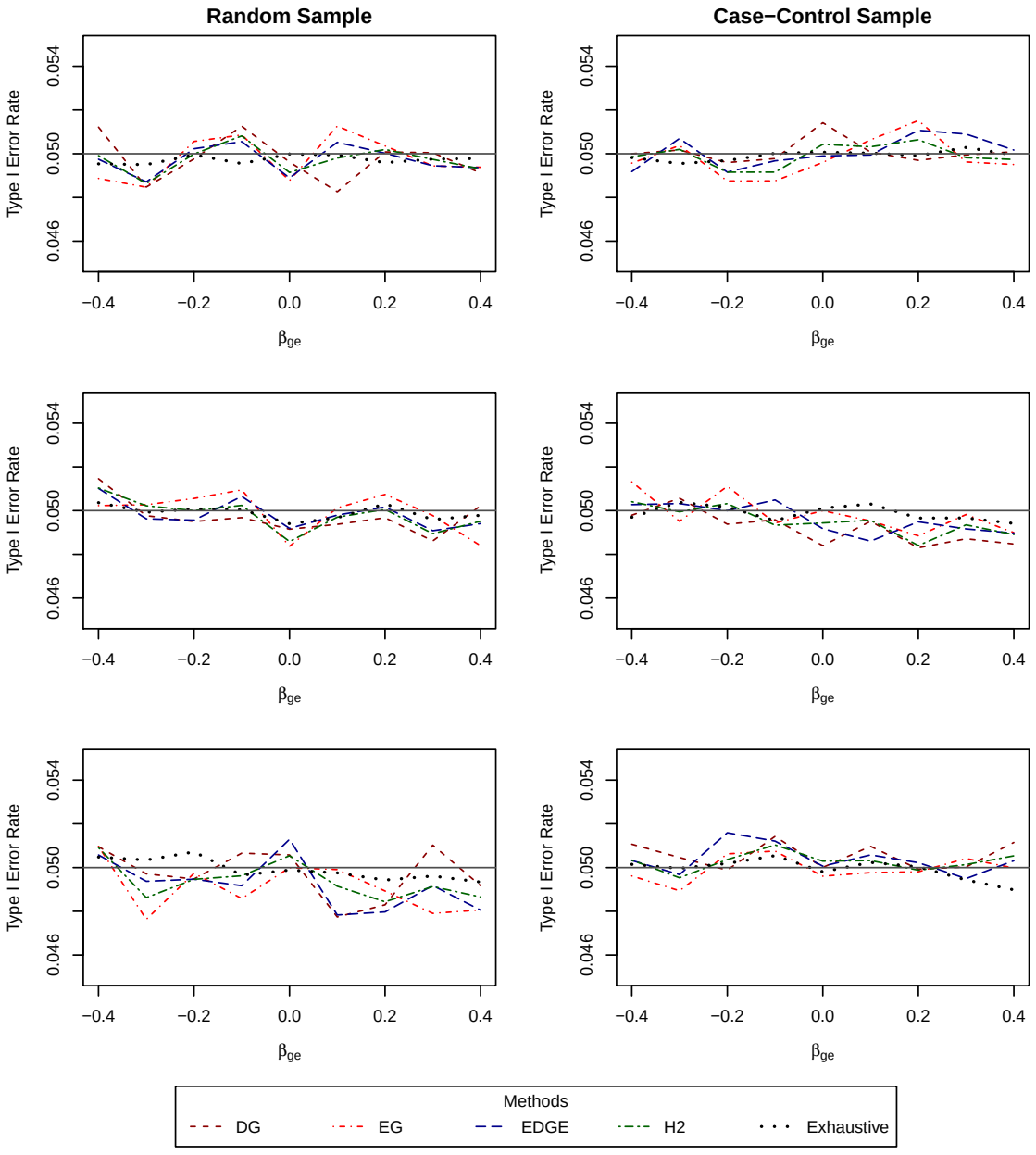

Figure 4.7: Family-wise error rate for all methods in random and case-control samples. G-E association parameter is $\theta_{ge} = 0.4$, including 3 SNPs in LD with QTL. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

causes of this inflation. Some causes such as the random data generation process, and inclusion of SNPs with G-E association in the type I error rate calculations have been ruled out. It should be noted that the slight inflation of the type I error rate does not appear in case-control samples nor in the case of $\theta_{ge} = 0.4$.

4.4.2 Power

Figure 4.8 and Figure 4.9 show the power of the two-step methods for the cases $\theta_{ge} = 0.2$ and $\theta_{ge} = 0.4$, respectively. In the case of $\theta_{ge} = 0.2$, the EG method still performs poorly in random samples, faring about the same or worse than the exhaustive search method. The performance of the EG method improves in case-control samples, doing better than the exhaustive search method. However, the EG method still remains under-powered compared to the other two-step methods in case-controls samples. The notable exceptions is in the case of negative interaction effects given a weak marginal genetic effect of $\beta_g = 0.2$. In this case, the EG method performs well and boosts the power of the EDGE and the H2 methods as a result. It should be noted that in the case of $\theta_{ge} = 0.2$, the DG method is generally the most powerful method across most simulation scenarios.

In the case of $\theta_{ge} = 0.4$, the power of the EG method is improved as a result of the stronger correlation between the QTL and the environment. In both random and case-control samples, the power of the EG method is higher than the exhaustive search method, but is still generally lower than the other two-step methods. The exception again, is in the case of negative interactions with $\beta_g = 0.2$. It should be noted that in some scenarios where both the DG and the EG methods have comparable power, the EDGE method can achieve higher power than the DG method. Overall, the power of

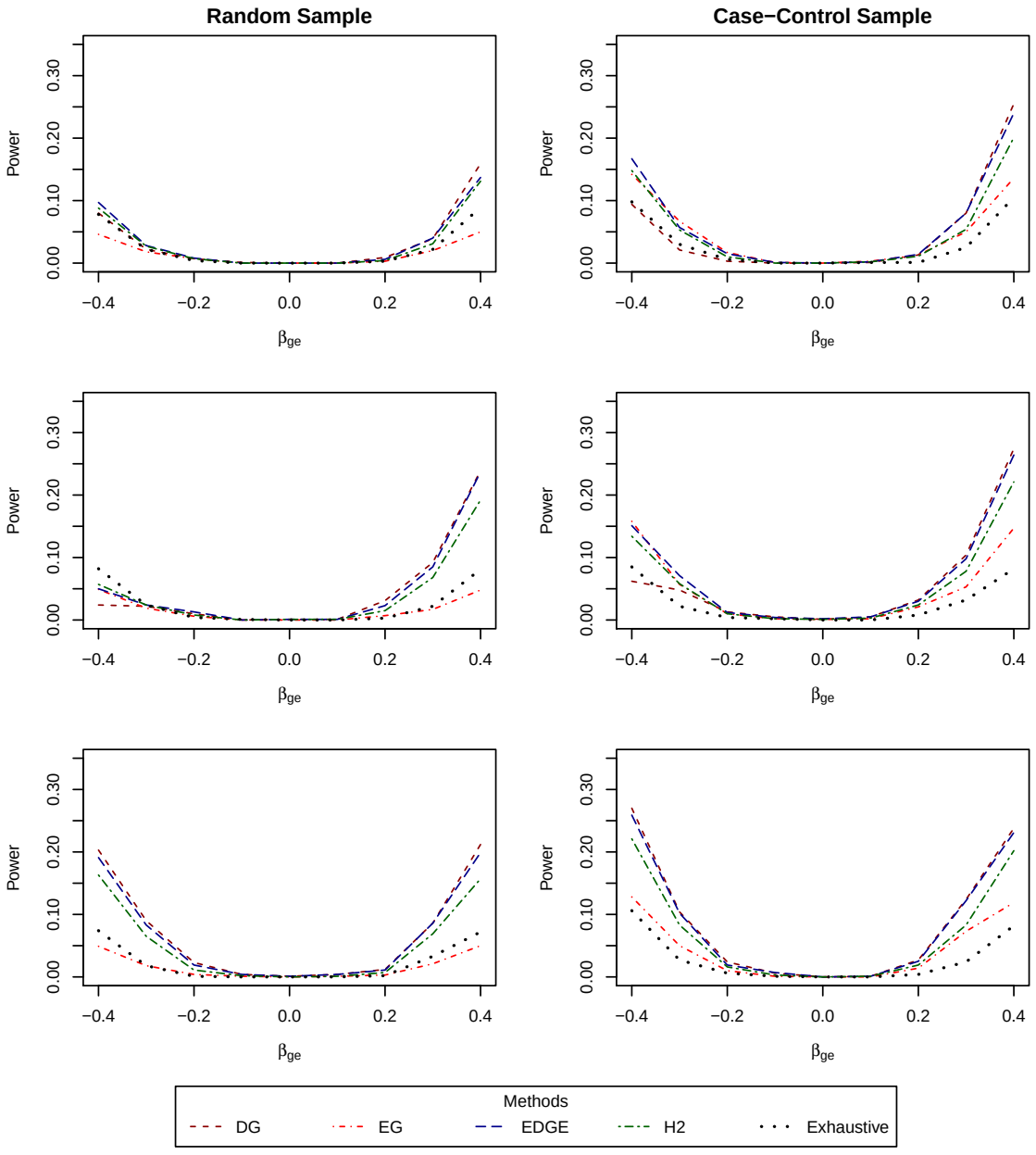

Figure 4.8: Power to detect the QTL for all methods in random and case-control samples. G-E association parameter is $\theta_{ge} = 0.2$, including 3 SNPs in LD with QTL. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

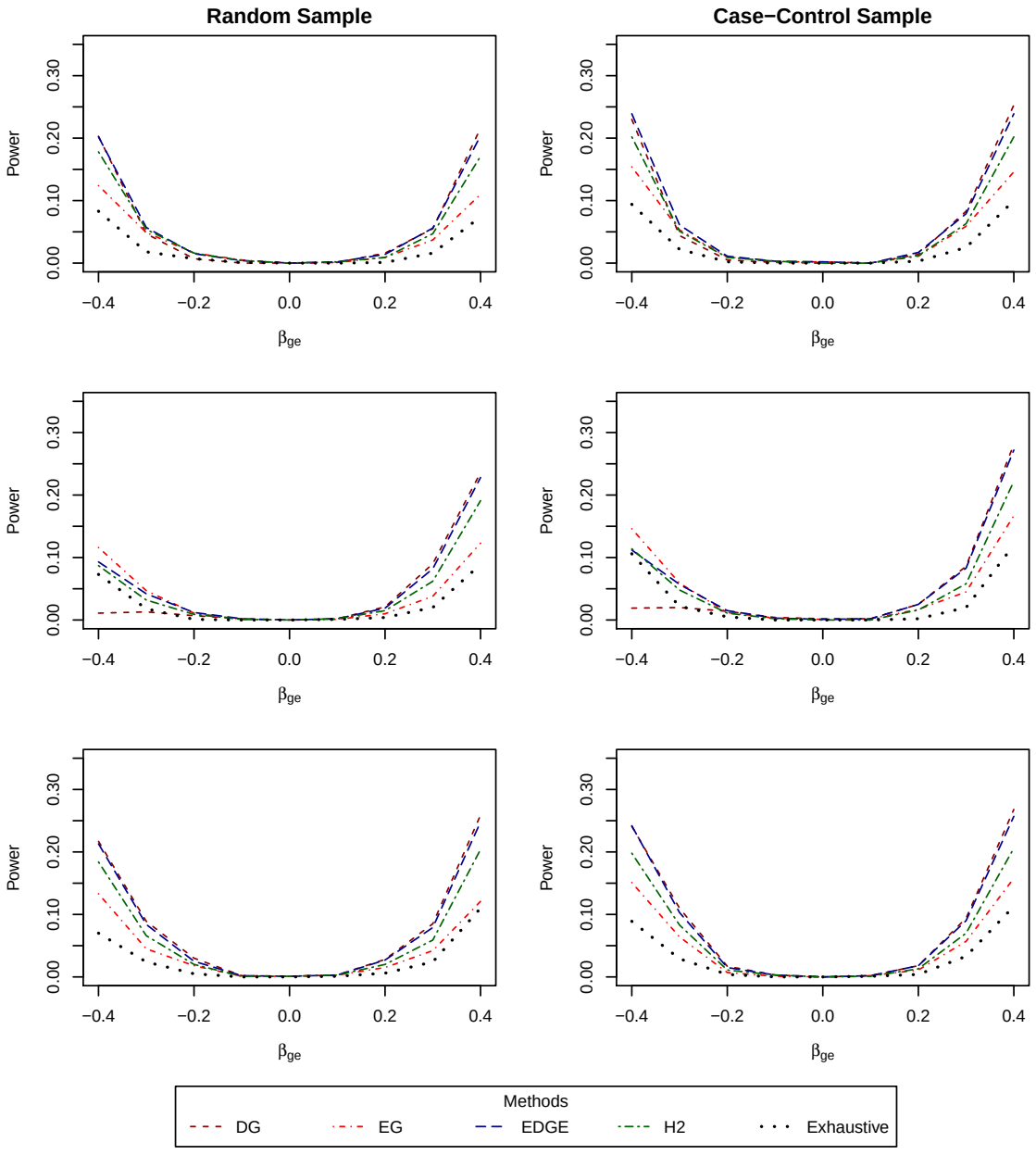

Figure 4.9: Power to detect the QTL for all methods in random and case-control samples. G-E association parameter is $\theta_{ge} = 0.4$, including 3 SNPs in LD with QTL. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

the EG method is improved by considering moderate G-E associations in a few SNPs that are in weak LD with the QTL while all other factors held constant.

It should be noted that the test statistic of the EG screening test is also independent of the test statistic of the null hypothesis: $H_0 : \beta_g = \beta_{ge} = 0$ for the parameters of Model (3.1). As such, a joint χ^2 two degrees of freedom test (Kraft et al., 2007) can be used in the second step for formal G x E testing. Using the joint test in the second step can improve the power of the EG method, see Figure 4.10 and 4.11 for the power of the EG method using the joint test in the cases of G-E association $\theta_{ge} = 0.2$ and $\theta_{ge} = 0.4$, respectively. However, the test statistic of the null hypothesis: $H_0 : \beta_g = \beta_{ge} = 0$ will not be independent of the DG method's screening test statistic. As such, the joint test is not applicable to the DG method or to the combined two-step approaches.

4.5 Discussion

In the GWAS setting, the EG method has been shown to have good power in detecting the DSL under a wide range of scenarios (Gauderman et al., 2013; Murcray et al., 2009). However, the mechanisms that allowed the EG method to perform well may not readily carry over to the quantitative trait setting. Based on the data generating models proposed by this thesis, the EG method can suffer from low power impacted by G-E dependency, sample type, LD of the DSL and the QTL, and the strength of the G x E effect in the disease model. It is easy to construct data that allow the EG method to perform quite well and just as easy to construct data that causes the EG method to fail in the quantitative trait setting. The EG method's performance is heavily dependent on the underlying mechanisms that generate the data.

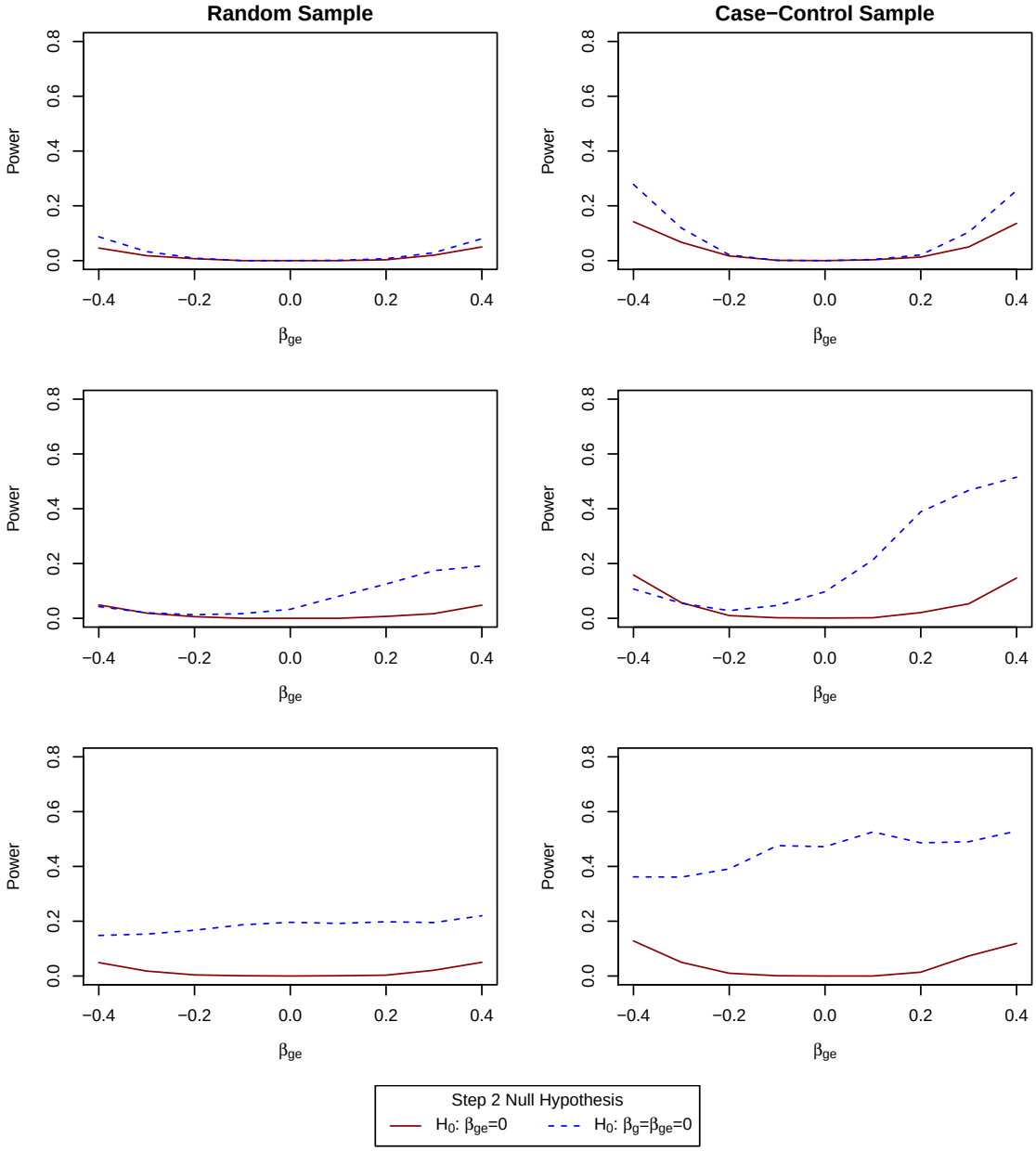

Figure 4.10: Comparison of power to detect the QTL for the EG method in random and case-control samples by two hypothesis tests in the second step. G-E association parameter is $\theta_{ge} = 0.2$, including 3 SNPs in LD with QTL. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

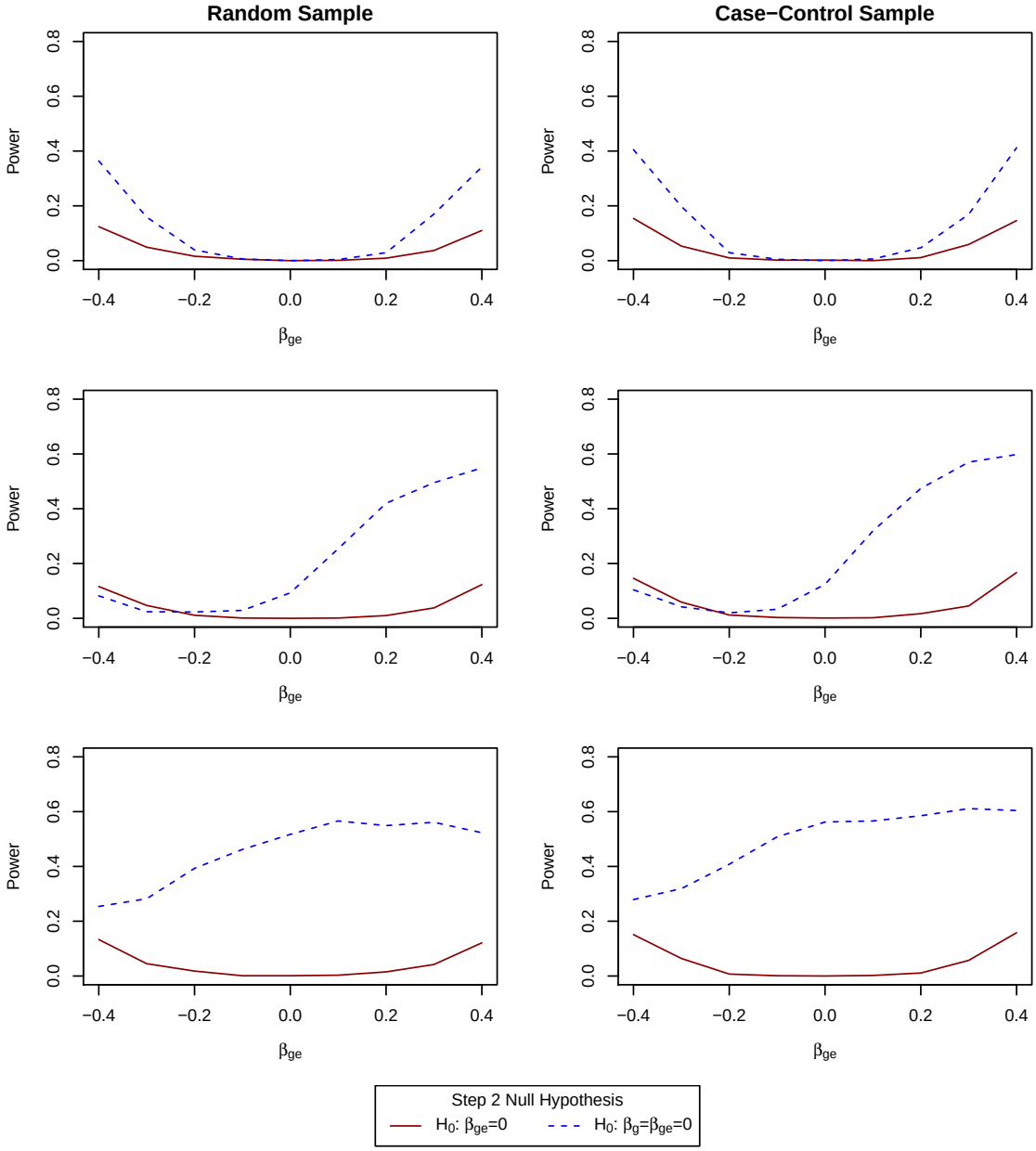

Figure 4.11: Comparison of power to detect the QTL for the EG method in random and case-control samples by two hypothesis tests in the second step. G-E association parameter is $\theta_{ge} = 0.4$, including 3 SNPs in LD with QTL. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

On the other hand, the DG method seems to translate well to the quantitative trait setting. This is because the DG screening test considers the relationship between the quantitative trait and any given SNP. If the QTL also exhibits a marginal genetic effect on the quantitative trait, then the power of the DG method is improved. Other factors such as G-E dependency, sample type, and the G x E effect in the disease model have little to no effect on the DG screening step. As a result, the DG method is more robust against a wide range of scenarios compared to the EG method.

Since the EG screening model utilizes the environment directly, changes to the variable, such as dichotomization, can also affect the method's performance. For continuous or count data environment variables, dichotomization can be done as a way to reduce type I error (Cornelis et al., 2012; Tchetgen Tchetgen and Kraft, 2011) or for simplicity and ease of interpretation (Mukherjee et al., 2012*b*). The dichotomization of the environment variable is common in practice and is expected to affect the performance of the EG method more so than the DG method. Some preliminary work examining normally distributed, Poisson distributed, and zero-inflated Poisson distributed environment variables show that dichotomization can drastically reduce the power of the EG method (see Appendix C). However, these results will likely change based on different cutoff points used for dichotomization. Additional work is needed to fully understand the impact of dichotomizing the environment variable on the EG method.

Chapter 5

Sensitivity Analysis

5.1 Sensitivity to Step 1 Thresholds

The two-step methods considered rely on some chosen step 1 threshold level, α_1, to determine which SNPs will be passed onto the second step for formal G x E testing. Previous literature has shown that the choice of α_1 can impact the power of the method. Typically, more conservative choices of α_1 result in higher power than liberal choices (Gauderman et al., 2013; Kooperberg and LeBlanc, 2008; Murcray et al., 2009).

This section examines the impact of the step 1 thresholds on the power of the two-step methods in the quantitative trait setting. The simulation results from Chapter 3 are examined with various α_1 levels used to control the number of SNPs tested at the second step. The values of α_1 considered are: $\{0.005, 0.01, 0.05, 0.1, 0.5\}$, representing thresholds that range from conservative to moderate to liberal.

The effect of α_1 on the power of the DG and the EG methods are shown in Figure 5.1 and Figure 5.2, respectively. For the DG method, more conservative choices of

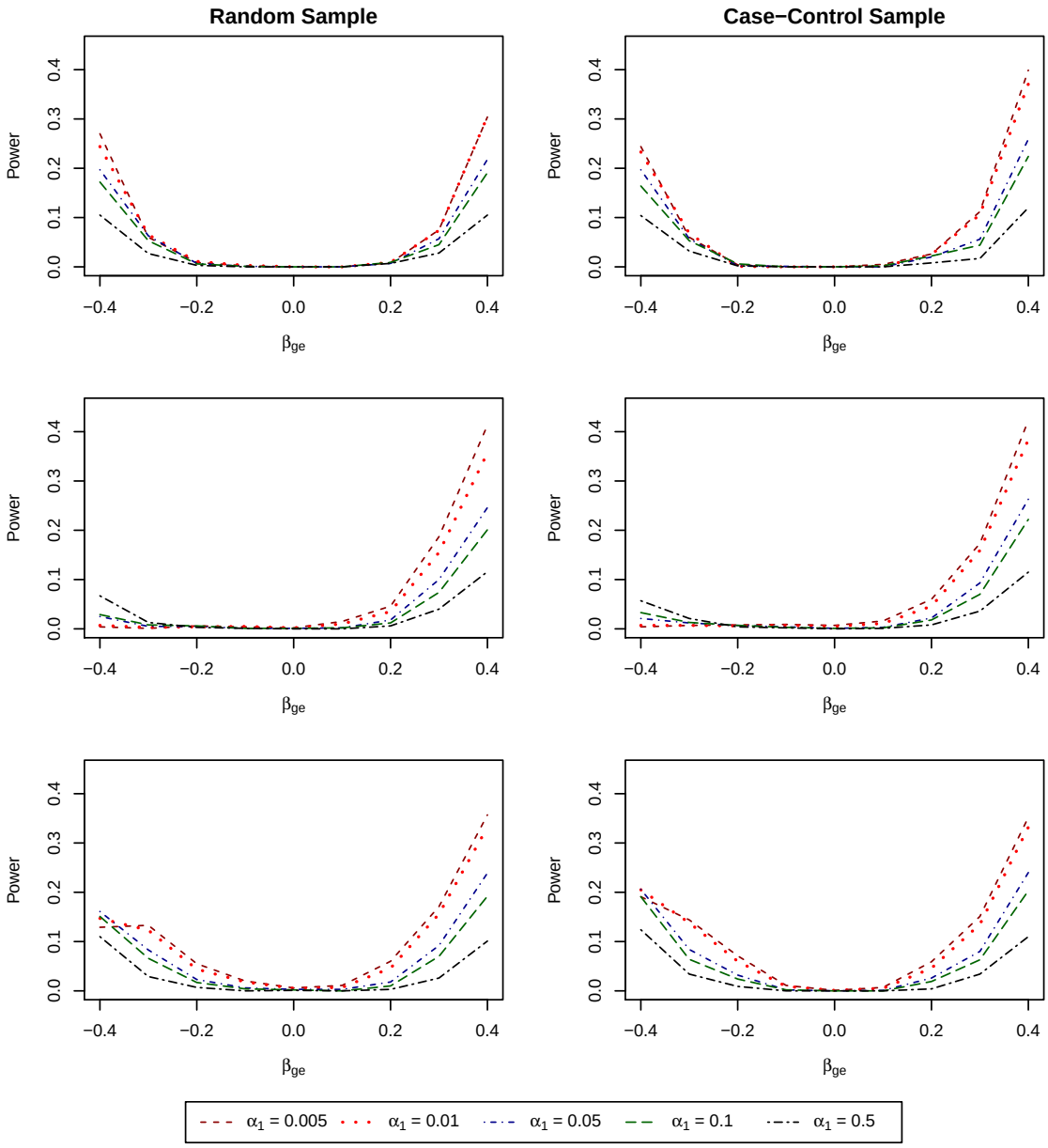

Figure 5.1: Impact of α_1 thresholds on the power of the DG method for random and case-control samples. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

α_1 results in higher power in most scenarios. There are some cases where the step 1 threshold suffers from being too conservative and results in lower power compared to a more liberal choice of α_1. Specifically, conservative α_1 choices in the case of negative interactions given $\beta_g = 0.2$ tend to decrease the power of the DG method compared to using a more liberal threshold.

Contrary to the DG method, the power of the EG method improves with more liberal α_1 choices using the simulation results from Chapter 3. This is due to the poor performance of the EG method under those simulation settings. Given a more liberal step 1 threshold, the pass rate of the EG method improves and the QTL is detected more often in the second step as a result. However, the improvement in power from a more liberal α_1 choice alone does not make the EG method competitive with the other two-step methods. The EG method still suffers from low power despite liberal choices of the step 1 threshold. Using the simulation results from Chapter 4 for the case of moderate G-E association ($\theta_{ge} = 0.4$) including 3 SNPs in weak LD with the QTL, the power of the EG method appears to be less sensitive to the step 1 threshold. See Figure 5.3 for plotted results. In this scenario, moderate choices of α_1 seem to improve the power of the EG method while conservative and liberal choices slightly reduce the power.

The effect of α_1 choices on the power of the H2 method and the EDGE method are shown in Figures B.4 and B.3 in Appendix B, respectively. The power of the H2 method and the EDGE method demonstrate similar responses to the choice of the α_1 as seen in the DG method. Note that for the H2 method, the parameters α_{1a} and α_{1m} are kept equal in this analysis.

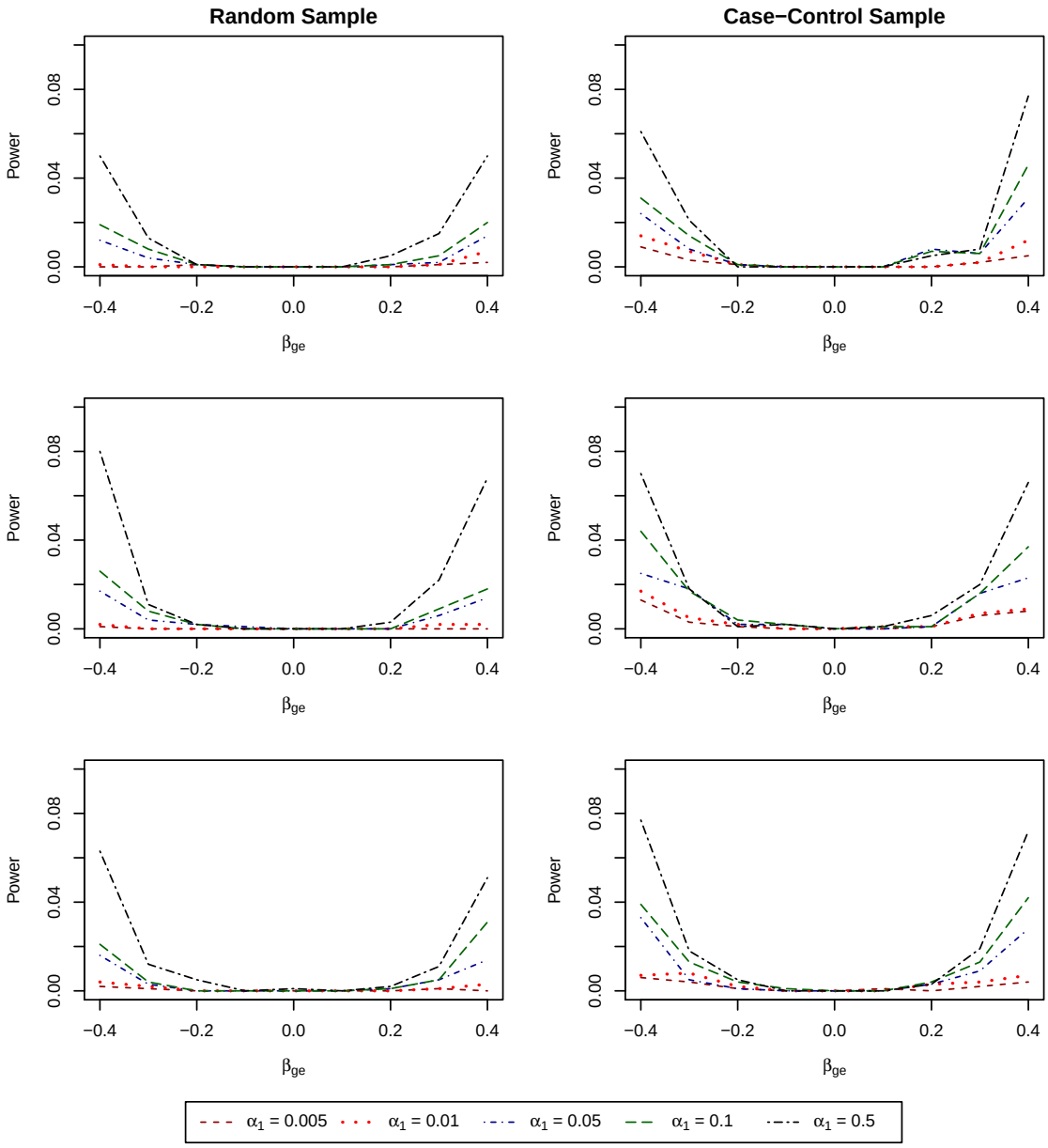

Figure 5.2: Impact of α_1 thresholds on the power of the EG method for random and case-control samples. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

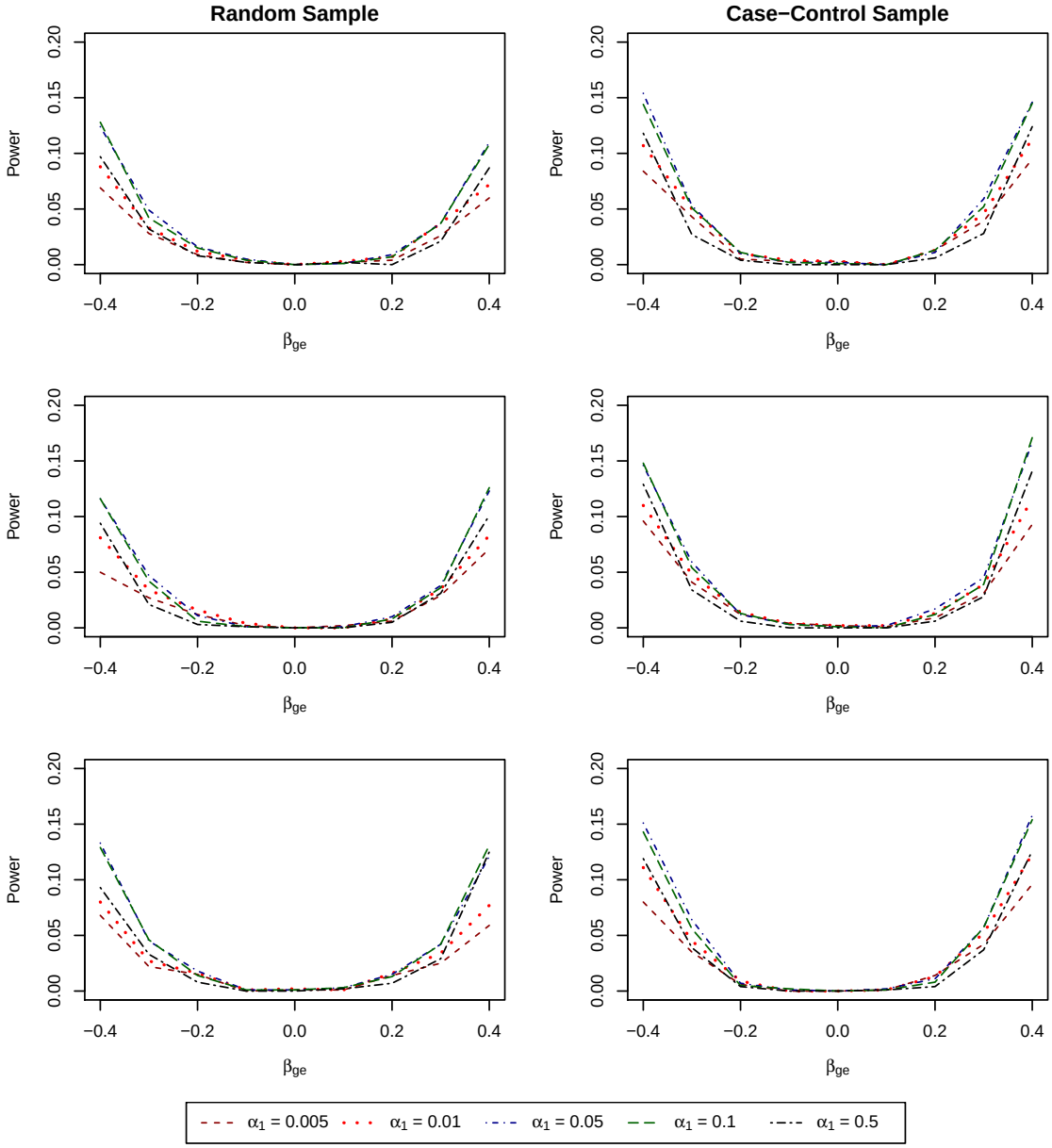

Figure 5.3: Impact of α_1 thresholds on the power of the EG method for random and case-control samples. Population data generated with G-E associations including 3 SNPs in LD with the QTL. Top panels, no marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

5.2 Parameters of the H2 Method

Unlike the other two-step methods, the H2 approach uses two separate choices for the step 1 threshold as it utilizes both the DG and EG screening tests. The method also uses a weighting parameter, p, in the second step to allocate the Bonferroni corrected significance level appropriately based on the number of SNPs passed by each screening test. In this analysis, the choice of the parameter p is varied to examine its effect on the power of the H2 method. Based on the H2 procedure described in Chapter 1, larger values of p favours the EG method and smaller values of p favours the DG method. The values for the parameter p considered are: $\{0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99\}$. The step 1 cutoffs are chosen to maximize the power for each screening test respectively. For the EG screening tests, the liberal choice of $\alpha_{1a} = 0.5$ is used and for the DG screening tests, the conservative choice of $\alpha_{1m} = 0.005$ is used. Note, the simulation results from Chapter 3 are used in this analysis.

The effect of the parameter p on the power of the H2 method is shown in Figure 5.4. Based on the performance of the DG and the EG method from Chapter 3, it is expected that favouring of the DG method will improve the power of the H2 method. This is reflected in choices of $p < 0.5$ with smaller values resulting in slight increases in power. However, in the presence of qualitative interactions, the DG method performs poorly, and larger choices of p increase the power of the H2 method. By fine tuning the parameters of the H2 method, it is possible for the H2 method to achieve higher power than the other two-step methods. However, in practice, the parameters of the H2 method should be chosen ahead of time based on prior beliefs on the performance of the DG and the EG methods.

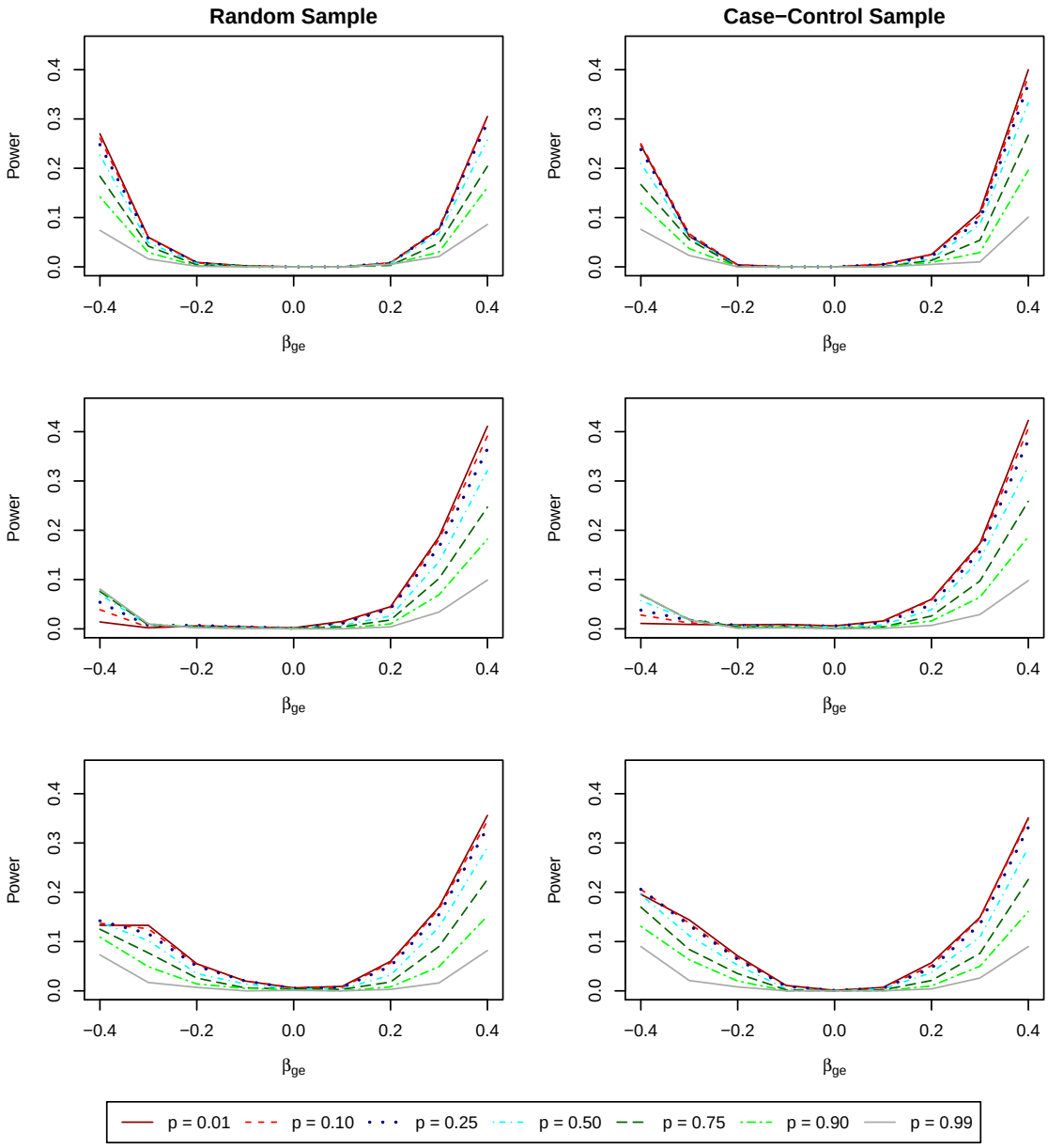

Figure 5.4: Impact of parameter p on the power of the H2 method for random and case-control samples. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

Chapter 6

Discussion and Future Directions

6.1 Discussion

Using simulation studies, this thesis demonstrates that the two-step methods can be successfully applied in the quantitative trait setting for random and case-control samples. The power to detect the QTL is generally higher in case-control samples. The main finding of the simulation studies is that the performance of the EG method is highly reliant on the disease model, the sample type, and the structure of the G-E dependency. The EG method can also be affected by the form of the environment variable and whether it is dichotomized in the sample. In comparison, the DG method's power is consistent across a wide range of scenarios and therefore more robust than the EG method in the quantitative trait setting.

Both the EDGE and the H2 methods utilize a combination of the DG and the EG methods in their screening steps. As such, the characteristics of the EG method are inherited by the combined two-step approaches. Under the quantitative trait setting, they cannot perform well if the EG method lacks power. In many cases, when the EG

method has low power, the combined two-step approaches demonstrate similar power as the DG method. The EDGE method appears to perform best when both the DG and the EG methods have comparable power. In this scenario, the EDGE method can outperform both the DG and the EG methods. However, in cases where one method is more powerful than the other, the power of the EDGE method leans towards the more powerful method but generally fares worse. The same characteristics apply to the H2 method, but this combined approach is generally less powerful than the EDGE method at neutral settings of the H2 parameters.

It should be noted that the H2 method can achieve high power and outperform the other two-step methods if its parameters are chosen to reflect the performances of the DG and the EG methods relative to one another. For example, if it is known a priori that there are marginal genetic effects and G-E independence then it is expected that the DG method would be more powerful than the EG method. Then choosing parameters that favour the DG method can improve the power of the H2 method substantially. However, the extent of a priori knowledge regarding the sample data may be limited in practice.

6.2 Future Directions

Throughout the simulation studies conducted in this thesis, it has been assumed that there is only one environment variable of interest. However, this assumption may not be realistic in practice as there can easily be multiple environmental factors that impact a particular disease or quantitative trait. The treatment of multiple environment factors in two-step methods has not been explored in detail in the current literature. It is easy to see that multiple environment factors can have a big impact on the EG

method and subsequently the combined two-step approaches. Multiple environment factors can also present additional modelling issues such as the presence of correlation between the environment variables. It should also be noted that in recent years there has been more focus on cataloging all the possible exposures associated with a disease or trait (Aschard et al., 2012; Thomas et al., 2012) and it is possible that the volume of environment variables in the sample data can grow quite large. This may require a screening approach to not only filter out the irrelevant SNPs but to also screen for the appropriate exposures before formal G x E testing. It would be interesting to examine the various ways multiple environment factors can affected the screening and testing step of the two-step methods in the GWAS and the quantitative trait setting.

It should be noted that in this thesis, all of the two-step methods utilized the same model in the second step for formal G x E testing. This was the full linear model including all marginal effects and a single two-way interaction term. However, this linear model is often misspecified in practice. In addition with the considerations of multiple environment factors, higher-order interactions may be present and could exacerbate the model misspecification problem. To gain power in the second step, the model used to test for G x E effects should be able to hedge against a certain degree of model misspecification. To account for these nuances, nonparametric data mining methods can be utilized. This area of research has not been explored in detail in the two-step approaches to finding G x E effects. Nonparametric methods may be more applicable in the two-step framework as the volume of data to be explored is pared down by the screening step. As such, some of the computational and dimensionality issues typically experienced by nonparametric methods may be reduced.

Lastly, as many researchers have highlighted (Aschard et al., 2012; Ko et al., 2013;

Mukherjee et al., 2012*a*), longitudinal analysis of G x E effects is the next step to understanding the role of G x E effects on disease etiology. The duration and onset of environment exposures may play a critical role in the development of diseases and affect the associated quantitative traits. While there has been some preliminary work on finding G x E effects in longitudinal analysis (Ko et al., 2013; Mukherjee et al., 2012*a*), it is unclear if the two-step framework can be applied to this type of data. To be successful, the screening step should identify the important genetic and environment factors and the testing step should account for the time dependent structure of the data. Additionally, it would be interesting to explore whether the screening step needs to account for the time structure or can perform well by simply using averages across the time periods.

Appendix A

Supplementary Tables

MAF	Cutoff c_1	Cutoff c_2
0.01	2.0558	3.7190
0.05	1.2959	2.8070
0.10	0.8779	2.3263
0.15	0.5903	2.0047
0.20	0.3585	1.7507
0.25	0.1573	1.5341
0.30	-0.0251	1.3408
0.35	-0.1955	1.1626
0.40	-0.3585	0.9945
0.45	-0.5172	0.8327
0.50	-0.6745	0.6745

Table A.1: Cutoff values for converting normal random variables to binomial random variables by MAF for genotypes $G = \{0, 1, 2\}$.

MAF	Pearson's r	LD Measure r	LD Measure D'
0.01	0.1559	0.1756	0.1814
0.05	0.2575	0.2562	0.2656
0.10	0.3120	0.3057	0.3144
0.15	0.3437	0.3343	0.3426
0.20	0.3661	0.3549	0.3627
0.25	0.3815	0.3687	0.3763
0.30	0.3912	0.3775	0.3847
0.35	0.4007	0.3861	0.3932
0.40	0.4046	0.3896	0.3967
0.45	0.4074	0.3922	0.3991
0.50	0.4083	0.3930	0.3999

Table A.2: Output correlations of binomials variables generate using Wang and Abbott (2008)'s method for a pair of SNPs with same MAF value for input correlation $\rho = 0.5$.

MAF	$\hat{\beta}_0$ (Std. Error)	$\hat{\beta}_1$ (Std. Error)	$\hat{\beta}_2$ (Std. Error)
0.01	0.1746 (0.001)	2.0280 (0.0074)	-1.3288 (0.0101)
0.05	0.0500 (0.006)	2.0106 (0.0040)	-1.0960 (0.0050)
0.10	0.0142 (0.005)	1.8456 (0.0031)	-0.8361 (0.0037)
0.15	0.0013 (0.005)	1.7194 (0.0027)	-0.6660 (0.0031)
0.20	-0.0073 (0.005)	1.6433 (0.0025)	-0.5659 (0.0029)
0.25	-0.0113 (0.004)	1.5799 (0.0024)	-0.4876 (0.0027)
0.30	-0.0143 (0.004)	1.5348 (0.0023)	-0.4326 (0.0026)
0.35	-0.0165 (0.004)	1.5057 (0.0023)	-0.3968 (0.0025)
0.40	-0.0167 (0.004)	1.4798 (0.0022)	-0.3680 (0.0025)
0.45	-0.0181 (0.004)	1.4723 (0.0022)	-0.3583 (0.0025)
0.50	-0.0171 (0.004)	1.4634 (0.0022)	-0.3498 (0.0025)

Table A.3: Estimated Coefficients of Naive Model by MAF Value

MAF	$\hat{\beta}_0$ (Std. Error)	$\hat{\beta}_1$ (Std. Error)
0.01	0.2316 (0.0012)	1.7533 (0.0031)
0.05	0.0783 (0.0007)	1.7436 (0.0014)
0.10	0.0198 (0.0005)	1.6801 (0.0010)
0.15	-0.0108 (0.0005)	1.6376 (0.0009)
0.20	-0.0295 (0.0005)	1.6080 (0.0008)
0.25	-0.0423 (0.0005)	1.5855 (0.0008)
0.30	-0.0514 (0.0005)	1.5689 (0.0008)
0.35	-0.0581 (0.0005)	1.5592 (0.0008)
0.40	-0.0615 (0.0005)	1.5491 (0.0008)
0.45	-0.0640 (0.0005)	1.5458 (0.0008)
0.50	-0.0643 (0.0005)	1.5433 (0.0008)

Table A.4: Estimated Coefficients of Fisher Model 1 by MAF Value

MAF	Naive Model	Fisher Model 1	Fisher Model 2
0.01	0.8203	0.8635	0.8765
0.05	0.9541	0.9670	0.9678
0.10	0.9741	0.9816	0.9816
0.15	0.9804	0.9855	0.9864
0.20	0.9831	0.9869	0.9887
0.25	0.9848	0.9874	0.9899
0.30	0.9859	0.9878	0.9910
0.35	0.9865	0.9879	0.9916
0.40	0.9870	0.9879	0.9920
0.45	0.9871	0.9878	0.9920
0.50	0.9873	0.9876	0.9921

Table A.5: R^2 of Linear Models by MAF Value

	Desired r^2								
MAF	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.01	0.0795	0.1833	0.2984	0.4005	0.4644	0.4836	0.4592	0.3952	0.3326
0.05	0.0954	0.1947	0.3007	0.4028	0.4941	0.5759	0.6373	0.6720	0.6671
0.10	0.0986	0.1948	0.2937	0.3880	0.4904	0.5949	0.7149	0.8864	0.9984
0.15	0.1011	0.1942	0.2899	0.3847	0.4848	0.5963	0.7367	0.9989	0.9989
0.20	0.1015	0.1963	0.2891	0.3812	0.4808	0.5936	0.7402	0.9994	0.9994
0.25	0.1022	0.1971	0.2865	0.3811	0.4798	0.5931	0.7402	0.9993	0.9993
0.30	0.1034	0.1982	0.2877	0.3786	0.4780	0.5905	0.7376	0.9996	0.9996
0.35	0.1052	0.1969	0.2890	0.3795	0.4796	0.5894	0.7379	0.9995	0.9995
0.40	0.1055	0.1980	0.2868	0.3791	0.4787	0.5886	0.7334	0.9994	0.9994
0.45	0.1041	0.1986	0.2897	0.3798	0.4768	0.5886	0.7327	0.9994	0.9994
0.50	0.1068	0.1999	0.2881	0.3805	0.4778	0.5879	0.7317	0.9996	0.9996

Table A.6: Observed median LD measure r^2 by MAF for Naive Model across desired r^2 values, simulation results.

	Desired r^2								
MAF	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.01	0.0725	0.1609	0.2567	0.3517	0.4576	0.5580	0.6708	0.7749	0.8791
0.05	0.0919	0.1932	0.2927	0.4006	0.4934	0.5940	0.6904	0.7892	0.8871
0.10	0.0987	0.2002	0.3013	0.4015	0.4991	0.5923	0.6871	0.7858	0.8805
0.15	0.1025	0.2058	0.3040	0.4054	0.5019	0.5938	0.6854	0.7808	0.8791
0.20	0.1043	0.2089	0.3080	0.4057	0.5011	0.5937	0.6851	0.7784	0.8766
0.25	0.1058	0.2111	0.3120	0.4078	0.5028	0.5951	0.6851	0.7767	0.8740
0.30	0.1067	0.2137	0.3146	0.4099	0.5020	0.5949	0.6841	0.7756	0.8745
0.35	0.1061	0.2146	0.3151	0.4123	0.5034	0.5949	0.6854	0.7765	0.8734
0.40	0.1063	0.2141	0.3159	0.4103	0.5054	0.5934	0.6856	0.7759	0.8720
0.45	0.1075	0.2153	0.3175	0.4120	0.5044	0.5938	0.6859	0.7750	0.8730
0.50	0.1080	0.2159	0.3153	0.4122	0.5028	0.5953	0.6853	0.7749	0.8705

Table A.7: Observed median LD measure r^2 by MAF for Fisher Model 1 across desired r^2 values, simulation results.

	Desired r^2								
MAF	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.01	0.0881	0.1891	0.2823	0.3781	0.4679	0.5478	0.6163	0.6986	0.7551
0.05	0.0949	0.1970	0.3004	0.3997	0.4987	0.5899	0.6848	0.7704	0.8670
0.10	0.0969	0.1984	0.3010	0.3995	0.4972	0.5924	0.6913	0.7870	0.8882
0.15	0.0998	0.2004	0.2978	0.3976	0.4963	0.5938	0.6956	0.7921	0.8962
0.20	0.1003	0.2007	0.2979	0.3983	0.4963	0.5948	0.6945	0.7970	0.8980
0.25	0.1018	0.2016	0.2999	0.3984	0.4944	0.5946	0.6945	0.7980	0.9005
0.30	0.1016	0.2011	0.2989	0.3975	0.4975	0.5940	0.6972	0.7969	0.9009
0.35	0.1024	0.2038	0.2986	0.3998	0.4949	0.5953	0.6960	0.7979	0.9021
0.40	0.1024	0.2018	0.2981	0.3975	0.4961	0.5945	0.6955	0.7980	0.9018
0.45	0.1027	0.2025	0.2995	0.3970	0.4959	0.5951	0.6953	0.7977	0.9022
0.50	0.1014	0.2016	0.3007	0.3986	0.4944	0.5967	0.6937	0.7980	0.9031

Table A.8: Observed median LD measure r^2 by MAF for Fisher Model 2 across desired r^2 values, simulation results.

β_g	β_{ge}	DG	EG	EDGE	H2	Exhaustive
0	-0.4	0.0484	0.0500	0.0508	0.0492	0.0493
	-0.3	0.0515	0.0493	0.0506	0.0504	0.0500
	-0.2	0.0498	0.0508	0.0505	0.0504	0.0500
	-0.1	0.0503	0.0500	0.0505	0.0503	0.0504
	0.0	0.0495	0.0485	0.0499	0.0488	0.0495
	0.1	0.0478	0.0492	0.0492	0.0487	0.0497
	0.2	0.0503	0.0497	0.0498	0.0498	0.0496
	0.3	0.0485	0.0507	0.0503	0.0497	0.0502
	0.4	0.0504	0.0468	0.0487	0.0485	0.0492
0.2	-0.4	0.0514	0.0513	0.0511	0.0512	0.0501
	-0.3	0.0488	0.0494	0.0497	0.0491	0.0499
	-0.2	0.0499	0.0504	0.0504	0.0499	0.0501
	-0.1	0.0483	0.0493	0.0482	0.0489	0.0498
	0.0	0.0491	0.0503	0.0494	0.0499	0.0503
	0.1	0.0492	0.0495	0.0493	0.0494	0.0499
	0.2	0.0500	0.0500	0.0509	0.0500	0.0493
	0.3	0.0506	0.0492	0.0498	0.0501	0.0490
	0.4	0.0493	0.0490	0.0484	0.0491	0.0486
0.4	-0.4	0.0494	0.0505	0.0514	0.0501	0.0508
	-0.3	0.0496	0.0499	0.0481	0.0498	0.0506
	-0.2	0.0518	0.0497	0.0504	0.0507	0.0507
	-0.1	0.0497	0.0502	0.0497	0.0502	0.0500
	0.0	0.0493	0.0489	0.0484	0.0492	0.0504
	0.1	0.0494	0.0490	0.0489	0.0493	0.0500
	0.2	0.0505	0.0490	0.0487	0.0495	0.0489
	0.3	0.0515	0.0495	0.0499	0.0505	0.0487
	0.4	0.0487	0.0479	0.0480	0.0485	0.0480

Table A.9: Type I Error Rate by Two-Step Methods, Random Samples

β_g	β_{ge}	DG	EG	EDGE	H2	Exhaustive
0.0	-0.4	0.0480	0.0481	0.0480	0.0480	0.0489
	-0.3	0.0482	0.0520	0.0490	0.0500	0.0498
	-0.2	0.0480	0.0487	0.0488	0.0483	0.0497
	-0.1	0.0493	0.0486	0.0489	0.0489	0.0498
	0.0	0.0505	0.0498	0.0503	0.0502	0.0498
	0.1	0.0491	0.0499	0.0506	0.0497	0.0500
	0.2	0.0503	0.0503	0.0502	0.0503	0.0495
	0.3	0.0476	0.0502	0.0497	0.0491	0.0493
	0.4	0.0495	0.0509	0.0502	0.0503	0.0494
0.2	-0.4	0.0486	0.0489	0.0496	0.0487	0.0496
	-0.3	0.0489	0.0507	0.0503	0.0500	0.0503
	-0.2	0.0495	0.0508	0.0502	0.0503	0.0500
	-0.1	0.0500	0.0515	0.0517	0.0509	0.0508
	0.0	0.0487	0.0512	0.0497	0.0499	0.0498
	0.1	0.0493	0.0498	0.0505	0.0496	0.0496
	0.2	0.0484	0.0499	0.0497	0.0492	0.0490
	0.3	0.0510	0.0483	0.0487	0.0496	0.0492
	0.4	0.0495	0.0478	0.0476	0.0487	0.0484
0.4	-0.4	0.0518	0.0508	0.0509	0.0512	0.0507
	-0.3	0.0506	0.0505	0.0506	0.0506	0.0508
	-0.2	0.0518	0.0508	0.0526	0.0511	0.0509
	-0.1	0.0515	0.0497	0.0498	0.0504	0.0506
	0.0	0.0478	0.0484	0.0476	0.0482	0.0501
	0.1	0.0484	0.0485	0.0477	0.0484	0.0496
	0.2	0.0497	0.0495	0.0491	0.0495	0.0490
	0.3	0.0487	0.0483	0.0484	0.0486	0.0484
	0.4	0.0512	0.0469	0.0499	0.0489	0.0472

Table A.10: Type I Error Rate by Two-Step Methods, Case-Control Samples

Appendix B

Supplementary Figures

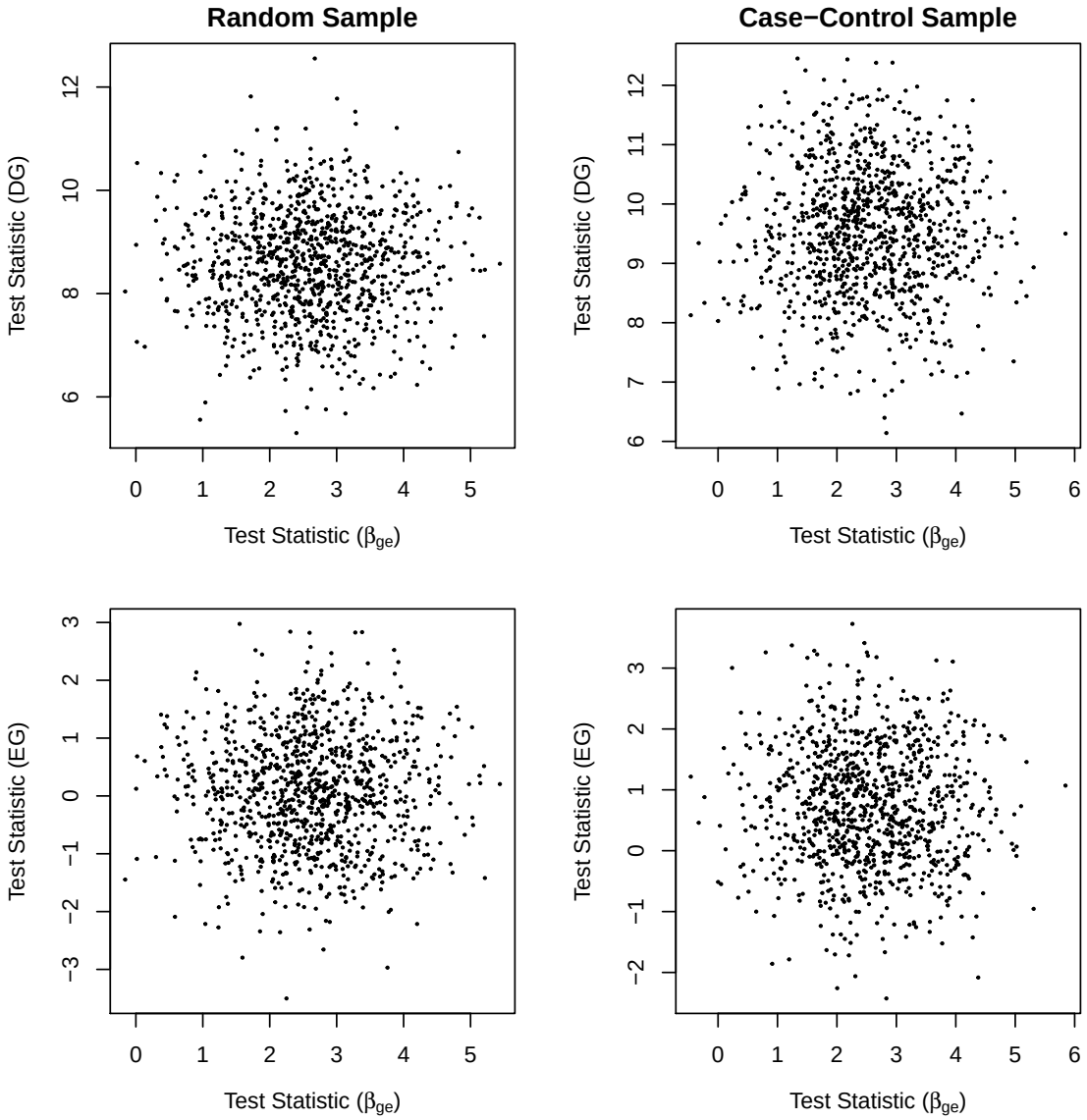

Figure B.1: Comparison of the step 1 test statistics of the DG and the EG methods with the test statistics of $H_0 : \beta_{ge} = 0$ for random and case-control samples. Note that $\beta_g = \beta_{ge} = 0.4$. Top panels show the comparison for the DG step 1 test statistics and bottom panels show the comparison for the EG step 1 test statistics.

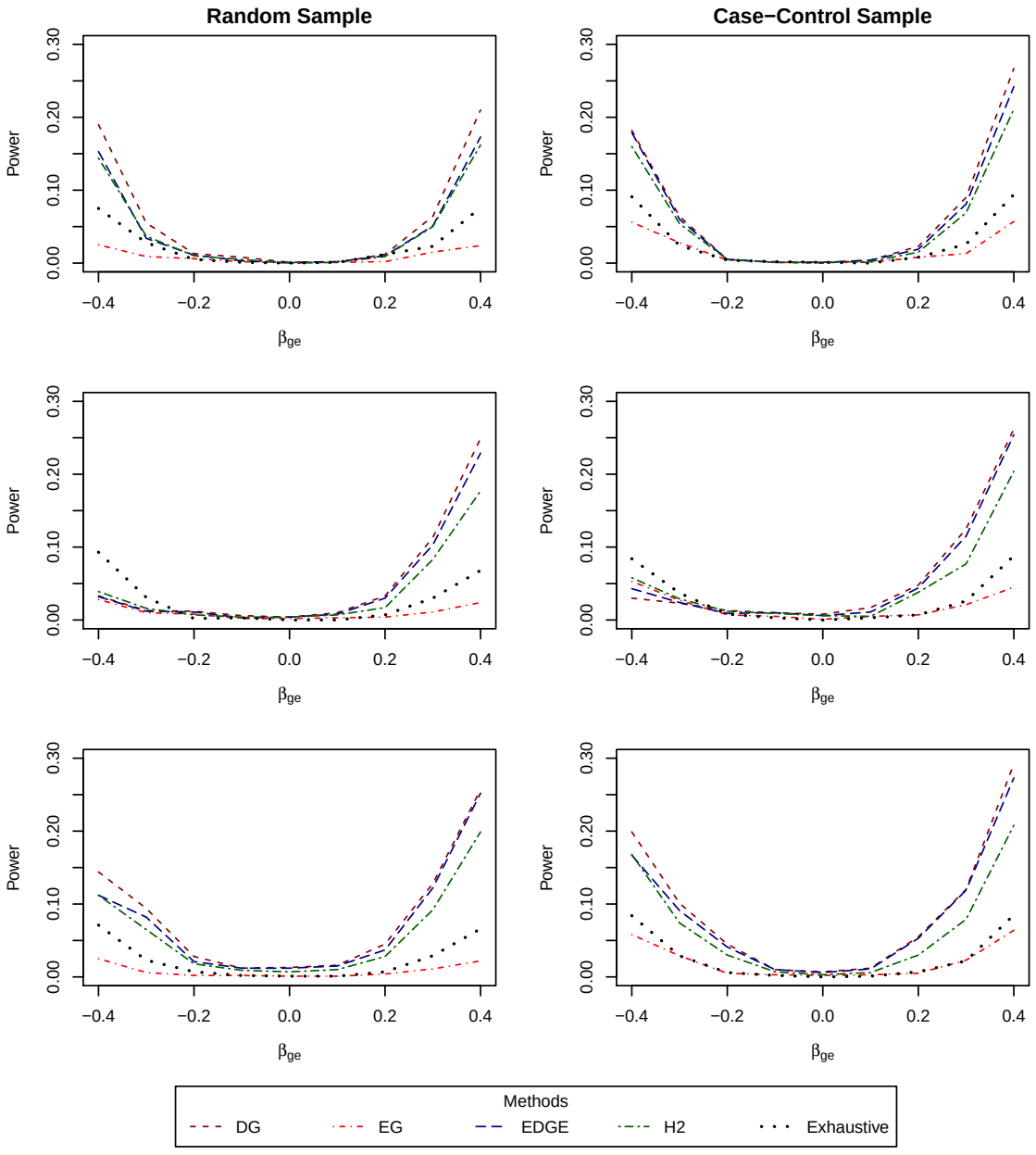

Figure B.2: Power to detect the QTL region for all methods in random and case-control samples given that the QTL is not genotyped. Top panels, zero marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

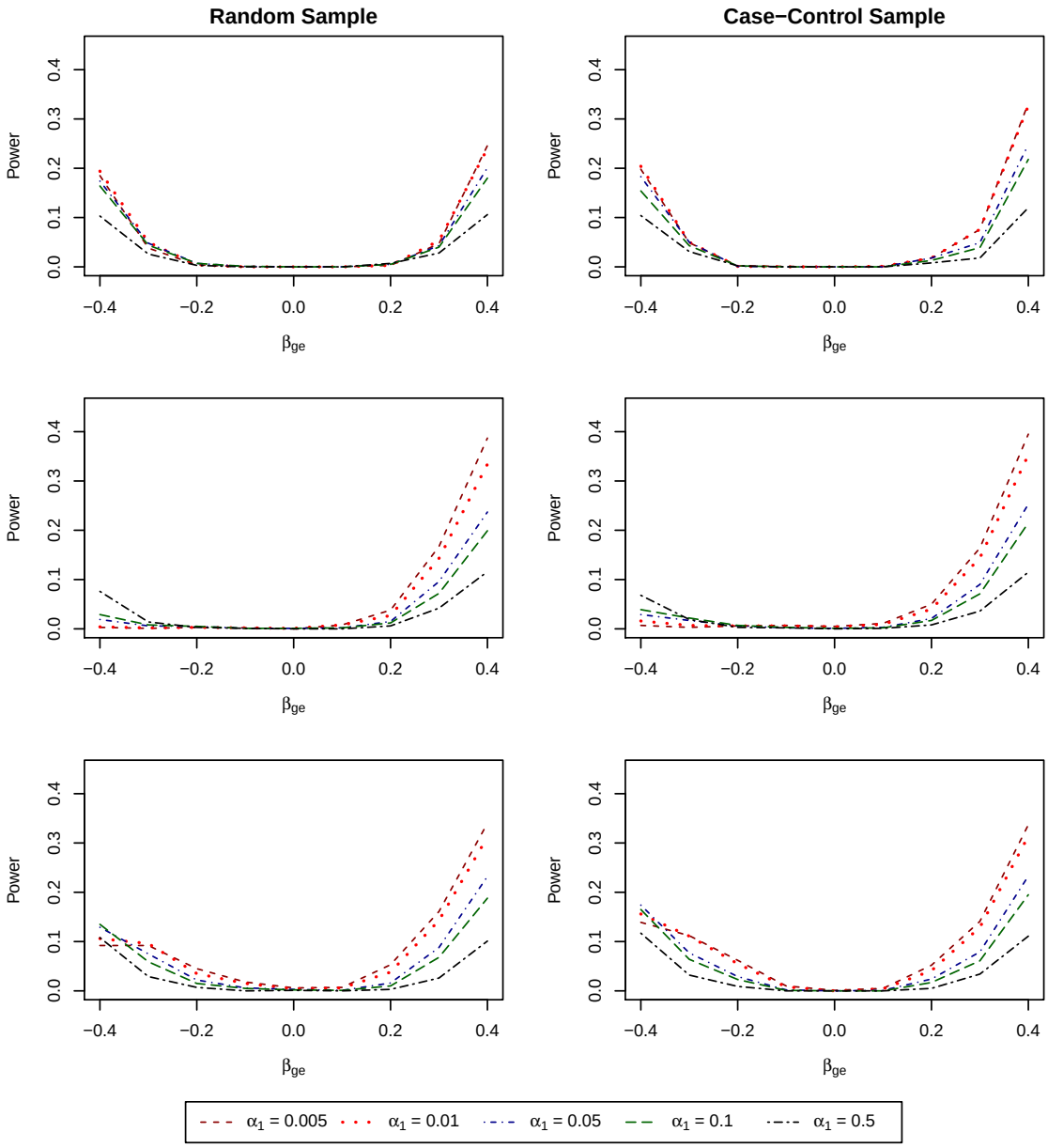

Figure B.3: Impact of α_1 thresholds on the power of the EDGE method for random and case-control samples. Top panels, no marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

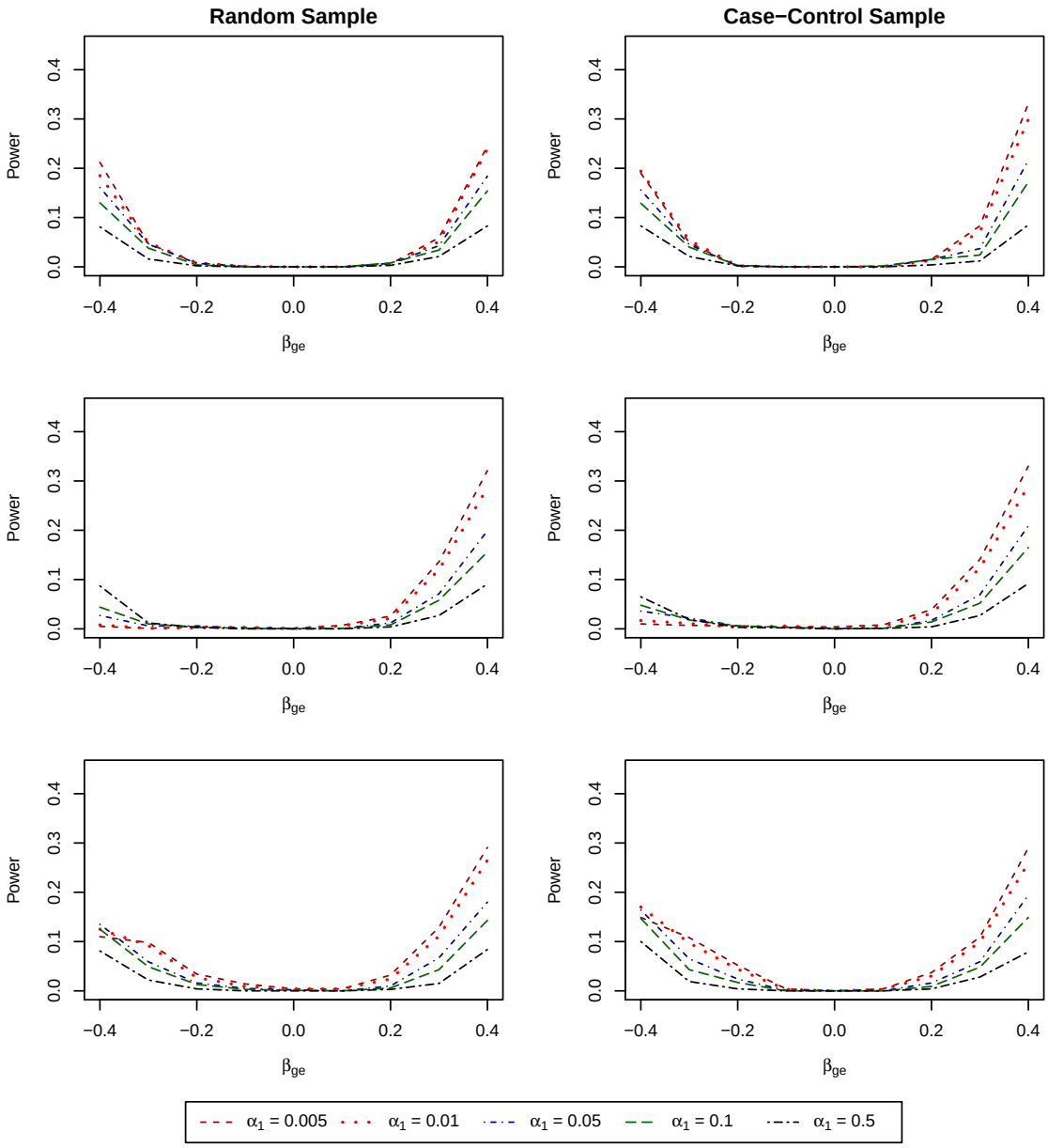

Figure B.4: Impact of α_1 thresholds on the power of the H2 method for random and case-control samples. Top panels, no marginal genetic effect ($\beta_g = 0$). Middle panels, weak marginal genetic effect ($\beta_g = 0.2$). Bottom panels, moderate marginal genetic effect ($\beta_g = 0.4$).

Appendix C

Preliminary Work on Dichotomizing the Environment Variable

A simulation study is conducted to examine the effect of dichotomizing the environment variable on the power of the EG method. Three types of environment variables are considered. The first, is a normally distributed environment variable that is then dichotomized about the median. The second, is a Poisson distributed variable that is then dichotomized by zero and non-zero values. The third, is a zero-inflated Poisson (ZIP) distributed variable that is also dichotomized by zero and non-zero values. The appropriate models are used in the screening step of the EG method based on the form of the environment variable. Linear regression is used for normally distributed environment, log-linear regression is used for Poisson and ZIP distributed environment, and logistic regression is used for the dichotomized environment variables in the screening step.

C.1 Simulation Study

For this simulation, it is assumed that G-E dependency exists among 10 SNPs, 3 of which are in LD with the QTL. The G-E association is set at $\theta_{ge} = 0.4$ to indicate a moderate level of dependency. SNPs 14, 21, and 23 from LD blocks 3 and 5 are selected to have a G-E association. These SNPs have r^2 of 0.18 and 0.02 with the QTL respectively. The three environment variables are generated using the linear predictor of Model (3.2.2) from Chapter 3, where $\theta_0 = \text{logit}(0.2) \approx -1.39$ and $\theta_{ge} = 0.4$. These two parameters are chosen to be the same as those used in the simulation study from Chapter 3. It should be noted that in generating the ZIP environment variables, 50% of the Poisson variables are randomly selected to be coded zero.

Each environment variable is then used to generated the disease data as well as the quantitative trait data. This results in three separate population data frames for a quantitative trait, each obtained from using a different environment variable. The data generating models from Chapter 3 are used in this simulation study with all other parameters held the same. Due to the constraints of time, the marginal genetic effect is fixed at zero and the G x E parameter, β_{ge} is varied from 0 to 0.5 by increments of 0.1. A total of 1,000 replicates are simulated. At each replicate a random sample and a balanced case-control sample of 1,000 observation are selected from the respective populations. The environment variable from the sample is then dichotomized to produce a binary environment variable for each of the three types of environment. For comparison purposes, the DG, the EG, and the exhaustive search methods are performed using the original and dichotomized environment variables.

It should be noted that the dichotomized exposure rates are not the same for the three types of environment variables considered. In random samples, the dichotomized

exposure rates are approximately: 50% for normally distributed environment, 60% for Poisson distributed environment, and 30% for ZIP distributed environment. In case-control samples, the dichotomized exposure rates are approximately: 50% for normally distributed environment, 67% for Poisson distributed environment, and 37% for ZIP distributed environment. Since each of the three environment variables are generated separately and independently, performance of the methods cannot be readily compared across the three types of environment.

C.2 Results

The family-wise type I error rates are displayed in Table C.11 and Table C.12 for random and case-control samples, respectively. The type I error rate is calculated as the proportion of SNPs declared significant given $\alpha = 0.05$ out of all 1,000 SNPs tested at $\beta_{ge} = 0$ for all simulation replicates. The family-wise type I error rate is mostly maintained in the random and case-control samples using the original environment variables. In random samples, using the dichotomized Poisson and zero-inflated Poisson environment variables results in slightly inflated type I error rates. In case-control samples, this inflation in type I error rate for the dichotomized environment variables is higher than seen in random samples.

The power of the EG method for each type of environment variable is shown in Figure C.5. In general, the power to detect the QTL is lower when using the dichotomized environment variable for all three types of environment variables considered. It should be noted that although it appears that the power is higher when using the dichotomized ZIP variables in case-control samples, this is attributed to the inflated type I error rate in this case. The dichotomization of the environment

Environment	DG Method	EG Method	Exhaustive
Normal	0.0489	0.0488	0.0492
Dichotomized Normal	0.0497	0.0510	0.0501
Poisson	0.0503	0.0512	0.0502
Dichotomized Poisson	0.0605	0.0592	0.0465
ZIP	0.0497	0.0500	0.0497
Dichotomized ZIP	0.0739	0.0694	0.0622

Table C.11: Family-wise error rate of DG, EG, and exhaustive search methods for three types of environment variable and their dichotomized counterpart in random samples. Normally distributed environment variables were dichotomized about the median. Poisson and zero-inflated Poisson distributed environment variables were dichotomized by zero and non-zero observations.

Environment	DG Method	EG Method	Exhaustive
Normal	0.0496	0.0498	0.0505
Dichotomized Normal	0.0523	0.0490	0.0502
Poisson	0.0481	0.0474	0.0474
Dichotomized Poisson	0.0796	0.0669	0.0379
ZIP	0.0506	0.0513	0.0510
Dichotomized ZIP	0.1150	0.0962	0.0648

Table C.12: Family-wise error rate of DG, EG, and exhaustive search methods for three types of environment variable and their dichotomized counterpart in case-control samples. Normally distributed environment variables were dichotomized about the median. Poisson and zero-inflated Poisson distributed environment variables were dichotomized by zero and non-zero observations.

variables has a drastic effect on the EG method as expected. The difference in power is especially pronounced for Poisson and zero-inflated Poisson environment variables in random samples. It should be noted that the drop in power associated with dichotomizing these two types of environment variables is affected by the choice of the cutoff. For example, power may be improved if the variables are dichotomized about the median as opposed to by zero and non-zero observations.

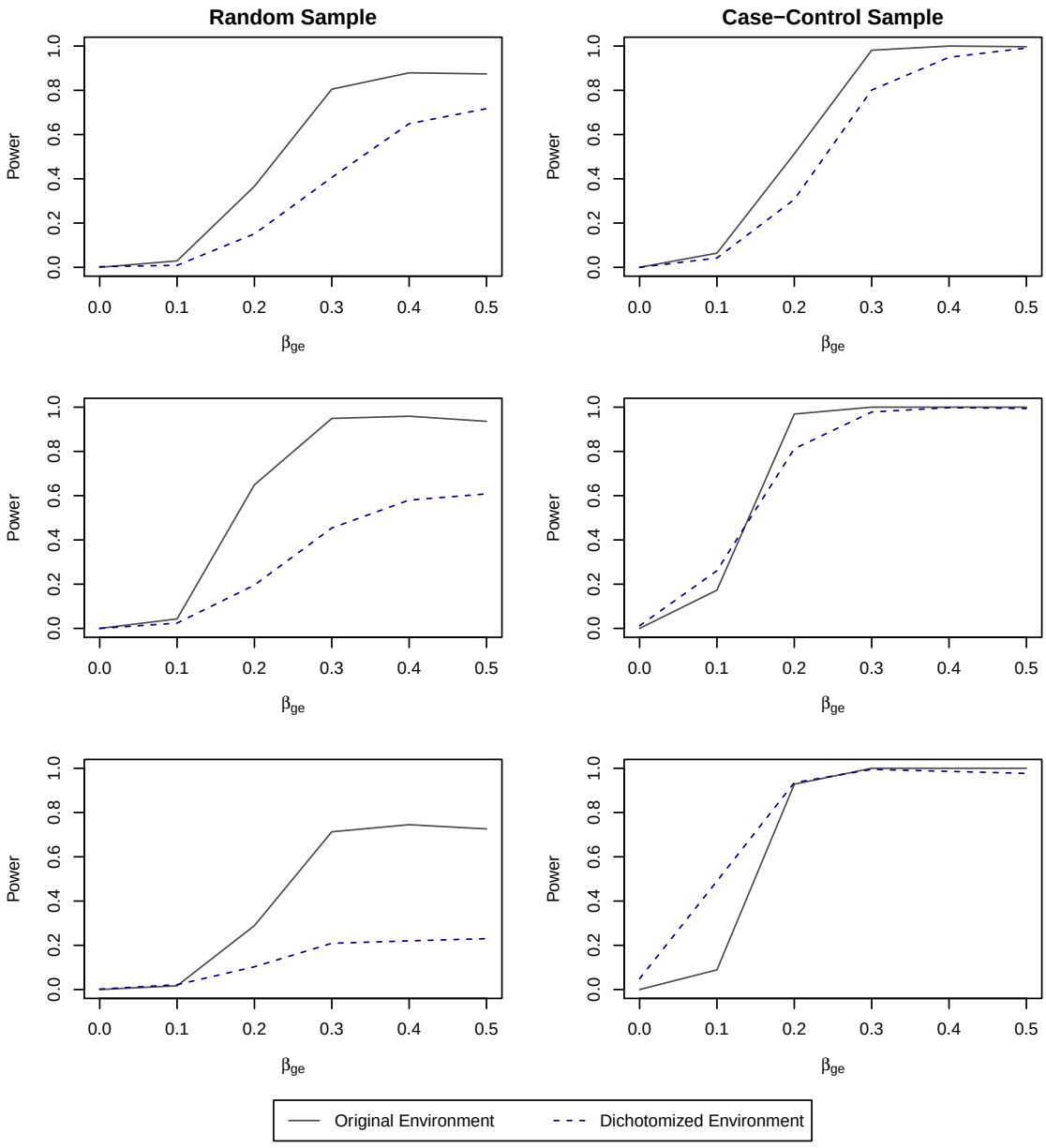

Figure C.5: Impact of dichotomizing environment factor on the power of the EG method to detect the QTL for various environment variables in random and case-control samples. Top panels, environment is normally distributed. Middle panels, environment is Poisson distributed, Bottom panels, environment is zero-inflated Poisson distributed.

Appendix D

Partial R Code

D.1 Functions to Generate SNPs in LD

```
## function to generate variance-covariance matrix for
## multivariate normals, assume all variance = 1 and
## thus correlation = covariance in the off-diagonals
## assume SNPs are in two LD blocks with some degree
## of cross group correlation and assume the given
## correlation vector will specify the LD structure
## as follows: correlation in first group,
## correlation in second group, correlation across groups
generate.sigma.matrix <-
  function(m, m.block1, correlation, variance = 1)
{
  ## create matrices to stop each LD block
  group1 <- matrix( NA, ncol = m.block1, nrow = m.block1 )
  group2 <- matrix( NA, ncol = (m - m.block1),
                    nrow = (m - m.block1) )

  ## input variance into diagonal of matrices
  diag(group1) <- variance
  diag(group2) <- variance

  ## fill off-diagonal elements with given correlations
```

```
  if(length(group1) == 1)
  {
    group1 <- group1
  }else
  {
    group1 <- offdiag(group1, correlation[1])
  }

  if(length(group2) == 1)
  {
    group2 <- group2
  }else
  {
    group2 <- offdiag(group1, correlation[2])
  }
  cross.group <- matrix(correlation[3],
                        ncol = (m - m.block1),
                        nrow = m.block1)

  ## build sigma matrix
  sigma <- rbind(cbind( group1, cross.group ),
                 cbind( t( cross.group ), group2 ))
  return(sigma)
}


## function to fill the off diagonal elements in a matrix
offdiag <- function(mat, value)
{
  ## fill up triangle
  mat[lower.tri(mat)] <- value

  ## fill lower triangle
  mat[upper.tri(mat)] <- value

  return(mat)
}

## function to generate binomial(2) r.v.s from normal r.v.s
## based on the given MAF
```

```
generate.binomial <- function(normal.rvs, maf, coding)
{
  ## if dominant coding
  if(coding == 1)
  {
    q <- maf^2 + 2*maf*(1-maf)
    c <- qnorm ( (1-q) ) ## P(x > c) = q^2+2*q*(1-q)

    ## create vector to store binomial variables
    binomial.rvs <- numeric( length( normal.rvs ) )

    ## parse through the normal r.v.s and assign values 0, 1, 2
    ## based on the cut offs c1 and c2
    for( i in 1 : length( normal.rvs ))
    {
      if( normal.rvs[i] > c )
      {
        binomial.rvs[i] <- 1
      }
    } ## end of for loop for parsing through normal r.v.s
  }

  ## if additive coding
  if(coding == 2)
  {
    ## obtain cut off values
    c1 <- qnorm( ( 1 - maf )^2 ) ## P(x < c1) = p^2
    c2 <- qnorm( 1 - maf^2 )   ## P(x > c2) = q^2

    ## create vector to store binomial variables
    binomial.rvs <- numeric( length( normal.rvs ) )

    ## parse through the normal r.v.s and assign values 0, 1, 2
    ## based on the cut offs c1 and c2
    for( i in 1 : length( normal.rvs ))
    {
      if( normal.rvs[i] >= c2 )
      {
        binomial.rvs[i] <- 2
      }
```

```
      if( c1 < normal.rvs[i] && normal.rvs[i] < c2 )
      {
        binomial.rvs[i] <- 1
      }
    } ## end of for loop for parsing through normal r.v.s
  }
  return(binomial.rvs)
}
```

D.2 Functions to Generate Quantitative Trait Data

```
## function to generated SNPs
## input requires the number of observations, MAF of the DSL,
## and genetic model (additive (2) or dominant (1))
generate.snps <- function(n, dsl.maf, coding, returnmaf = TRUE)
{
  ## generate mvn variables
  ## create variance covariance matrix
  sigma <- generate.sigma.matrix()
  mu <- numeric( 25 )
  ## generate mvn variables using mvrnorm function
  x <- MASS::mvrnorm( n, mu, sigma )

  ## convert mvns to binomials
  ## create blank matrix to store binomials G
  G <- matrix(NA, nrow = n, ncol = 25)

  ## set starting value of MAFs based on DSL and coding
  signs <- sample(c(1,2), 4, replace = TRUE)
  init.mafs <- c(dsl.maf,
                 if(signs[1] == 1){dsl.maf*1.3}else{dsl.maf/1.3},
                 if(signs[2] == 1){dsl.maf*1.5}else{dsl.maf/1.5},
                 if(signs[3] == 1){dsl.maf*1.7}else{dsl.maf/1.7},
                 if(signs[4] == 1){dsl.maf*1.9}else{dsl.maf/1.9})

  maf <- c(runif(4, init.mafs[1]/1.2, init.mafs[1]*1.2),
           runif(5, init.mafs[2]/1.2, init.mafs[2]*1.2),
           runif(5, init.mafs[3]/1.2, init.mafs[3]*1.2),
           runif(5, init.mafs[4]/1.2, init.mafs[4]*1.2),
```

```
          runif(5, init.mafs[5]/1.2, init.mafs[5]*1.2))
  maf <- c(dsl.maf, maf)

  ## use marginal normals to generate binomials
  ## store results into matrix G
  for(i in 1:ncol(x))
  {
    G[, i] <- generate.binomial(x[,i], maf[i], coding)
  } ## end of for loop

  if(returnmaf == FALSE)
  {
    return(G)
  }else
  {
    return(list(G = G, maf = maf))
  }
}

## function to create the variance covariance matrix of the
## multivariate normals, note that this function is built
## specific to the simulation settings considered
## correlation values, number of marginals, and block size
## are set as defaults
generate.sigma.matrix  <- function(M = 25, block.size = 5,
                                   variance = 1,
                                   corr = c(0.96, 0.85,
                                    0.65, 0.45, 0.25))
{
  ## create main diagonal blocks
  g1 <- matrix(NA, nrow = block.size, ncol = block.size)
  diag(g1) <- variance

  ## fill in the off-diagonals of the main diagonal blocks
  g1[lower.tri(g1)] <- corr[1]
  g1[upper.tri(g1)] <- corr[1]

  ## generate the remaining blocks
  g2 <- matrix(corr[2], nrow = block.size, ncol = block.size)
  g3 <- matrix(corr[3], nrow = block.size, ncol = block.size)
```

```
  g4 <- matrix(corr[4], nrow = block.size, ncol = block.size)
  g5 <- matrix(corr[5], nrow = block.size, ncol = block.size)

  ## build sigma matrix
  sigma <- rbind( cbind(g1, g2, g3, g4, g5),
                  cbind(g2, g1, g2, g3, g4),
                  cbind(g3, g2, g1, g2, g3),
                  cbind(g4, g3, g2, g1, g2),
                  cbind(g5, g4, g3, g2, g1))

  return(sigma)
}


## function to generate binomial r.v.s from normal r.v.s
## based on the given MAF and genetic model (additive or dominant)
generate.binomial <- function(normal.rvs, maf, coding)
{
  ## if dominant coding
  if(coding == 1)
  {
    ## calculate P(G=1) and cutoff value for normals
    g.freq <- maf^2 + 2*maf*(1-maf)  ## P(G = 1)
    c <- qnorm ( (1-g.freq) )        ## P(x > c) = q^2 + 2*q*(1-q)

    ## create vector to store binomial variables
    binomial.rvs <- numeric( length( normal.rvs ) )

    ## parse through the normal r.v.s and assign appropriate values
    ## based on the genetic model and cutoffs
    for( i in 1 : length( normal.rvs ))
    {
      if( normal.rvs[i] > c )
      {
        binomial.rvs[i] <- 1
      }
    } ## end of for loop for parsing through normal r.v.s
  }

  ## if additive coding
```

```
  if(coding == 2)
  {
    ## obtain cut off values
    c1 <- qnorm( ( 1 - maf )^2 ) ## P(x < c1) = p^2 = P( G = 0)
    c2 <- qnorm( 1 - maf^2 )   ## P(x > c2) = q^2 = P( G = 2)

    ## create vector to store binomial variables
    binomial.rvs <- numeric( length( normal.rvs ) )

    ## parse through the normal r.v.s and assign values 0, 1, 2
    ## based on the cut offs c1 and c2
    for( i in 1 : length( normal.rvs ))
    {
      if( normal.rvs[i] >= c2 )
      {
        binomial.rvs[i] <- 2
      }
      if( c1 < normal.rvs[i] && normal.rvs[i] < c2 )
      {
        binomial.rvs[i] <- 1
      }
    } ## end of for loop for parsing through normal r.v.s
  }
  return(binomial.rvs)
}

## function to generate linear predictor based on parameters,
## variables, and model type
## model.type = 1: single DSL with GxE interaction
## model.type = 2: two associated Gs, single DSL with GxE interaction,
##                 G1 indep of G2 (LE)
## model.type = 3: two associated G, single DSL with GxE interaction,
##                 G1 dep of G2 (LD, moderate)
generate.linear.predictor <-
   function(G, E, dsl.g, beta.0, beta.g, beta.e, beta.ge, model.type)
{
  ## G x E DSL is also stored in the first column of G matrix
  DSL <- G[, dsl.g]

  ## check model types
```

```
  ## model 1: single DSL with non-zero GxE interaction
  if(model.type == 1)
  {
    LP <- beta.0 + beta.g*DSL + beta.e*E + beta.ge*DSL*E
    g <- NA
  }else if(model.type == 2)
  ## model 2: five associated Gs, single DSL, all Gs in LE
  {
    g <- sample(26:100, 4)
    b <- matrix(beta.g, nrow = 4, ncol = 1)
    LP <- beta.0 + beta.g*DSL + G[,g]%*%b + beta.e*E + beta.ge*DSL*E
  }else if(model.type == 3)
  ## model 3: 5 associated Gs, single DSL, two Gs in moderate/weak LD
  {
    g <- c(sample(16:25, 2), sample(26:100, 2))
    b <- matrix(beta.g, nrow = 4, ncol = 1)
    LP <- beta.0 + beta.g*DSL + G[,g]%*%b + beta.e*E + beta.ge*DSL*E
  }else
  {
    warning("Invalid Model Type")
    LP <- NA
    g <- NA
  }

  ## returns the linear predictor and
  ## the second associated G (if there is one)
  return(list(LP=LP, g=g))
}

## function to generate the case control population data based
## on disease model parameters, this function does not generate
## the quantitative trait population data
generate.cc.pop <-
    function(N, gamma.0, gamma.g, gamma.e, gamma.ge,
             dsl.d, model.type.d,
             dsl.maf, p.e, theta.ge, coding)
{
  library(boot)

  ##### generate G at population size N #####
```

```
snps <- generate.snps(N, dsl.maf, coding)
maf1 <- snps$maf
G1 <- snps$G ## generate the SNPs in LD


total.g <- 100
length.ld <- 25
length.le <- total.g - length.ld

## generate second set of SNPs
G2 <- matrix(NA, ncol = length.le, nrow = N)  in LE

## fill matrix of the second set of SNPs based
## on random MAF and coding
maf2 <- runif(length.le, 0.05, 0.4)

if(coding == 1)
{
  for(i in 1:length.le)
  {
    g.freq <- maf2[i]^2 + 2*maf2[i]*(1-maf2[i])
    G2[, i] <- rbinom(N, coding, g.freq)
  }## end for loop
}else
{
  for(i in 1:length.le)
  {
    G2[, i] <- rbinom(N, coding, maf2[i]^2)
  }## end for loop
}

maf <- c(maf1, maf2)

## combine first and second set of SNPs = 50 SNPs in total
G <- cbind(G1, G2)
##### end of generating SNPs #####

##### generate E (based on multiple Gs) #####
## baseline probability of environment exposure
theta.0 <- logit(p.e)
```

```
  ## randomly sample 10 SNPs from the
  ## set of 75 LE SNPs to have an association with E
  n <- 0.01*1000 ## 1% of SNPs have true G-E association
  g <- sample(26:total.g, n)

  ## generate linear predictor based on sampled Gs
  theta.ge.vec <- rep(theta.ge, n)

  lp <- theta.0 + G[, g]%*%theta.ge.vec

  ## calculate probability of E based on linear predictor
  prob.e <- inv.logit(lp)

  ## generate E vector
  E <- rbinom(N, 1, prob.e)
  #### end of generating exposure status #####

  ##### generate probability of disease #####
  d.linear.pred <-
       generate.linear.predictor(G, E, dsl.d,
                                 gamma.0, gamma.g, gamma.e, gamma.ge,
                                 model.type.d)
  prob.disease <- inv.logit(d.linear.pred$LP)
  D <- rbinom(N, 1, prob.disease)
  d.g <- d.linear.pred$g
  #### end of generating disease status #####

  ## create data frame to store cc population
  cc.population <- data.frame(G, E, D)

  ## create cases and controls indices
  case.index <- which(cc.population$D == 1)
  control.index <- which(cc.population$D == 0)


  return(list(population = cc.population,
              cases = case.index, controls = control.index,
              maf = maf, GE.g = g, D.g = d.g))
}
```

```
## function to generate the case-control population data
## based on disease model parameters
## this function does not generate the quantitative trait
## population data
generate.quant.pop <-
    function(cc.population, beta.0, beta.g, beta.e,
             beta.ge, beta.d, model.type.y, N)
{
  ## extract G, E, and D from case control population
  G <- cc.population[, 1:100]
  E <- cc.population$E
  D <- cc.population$D

  ##### generate quantitative trait y #####
  y.linear.pred <-
       generate.linear.predictor(G, E, dsl.g = 1,
                                 beta.0, beta.g, beta.e, beta.ge,
                                 model.type.y)

  y <- y.linear.pred$LP + beta.d * D  + rnorm(N)

  y.g <- y.linear.pred$g
  #### end of generating quantitative trait y #####


  ## population data frame
  quant.population <- data.frame(G, E, D, y)

  ## generate case index vector and control index vector
  case.index <- which(quant.population$D == 1)
  control.index <- which(quant.population$D == 0)

  ## return results pertaining to the population sample
  ## which G is also associated with y along with DSL
  return(list(population = quant.population,
              cases = case.index,
              controls = control.index, Y.g = y.g))
}
```

```
## end of functions
#########################################################################
#########################################################################
```

Bibliography

Albert, P. S., Ratnasinghe, D., Tangrea, J. and Wacholder, S. (2001), Limitations of the case-only design for identifying gene-environment interactions, *American Journal of Epidemiology* **154**(8), 687–693.

Aschard, H., Zaitlen, N., Tamimi, R. M., Lindström, S. and Kraft, P. (2013), A non-parametric test to detect quantitative trait loci where the phenotypic distribution differs by genotypes, *Genetic Epidemiology* **37**(4), 323–333.

Aschard, H. et al. (2012), Challenges and opportunities in genome-wide environmental interaction (GWEI) studies, *Human Genetics* **131**(10), 1591–1613.

Balding, D. J. (2006), A tutorial on statistical methods for population association studies, *Nature Reviews Genetics* **7**(10), 781–791.

Cornelis, M. C. et al. (2012), Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes, *American Journal of Epidemiology* **175**(3), 191–202.

Dai, J. Y., Kooperberg, C., LeBlanc, M. and Prentice, R. L. (2012), Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction, *Biometrika* **99**(4), 929–944.

Dempfle, A. et al. (2008), Gene-environment interactions for complex traits: definitions, methodological requirements and challenges, *European Journal of Human Genetics* **16**(10), 1164–1172.

Eichler, E. E. et al. (2010), Missing heritability and strategies for finding the underlying causes of complex disease, *Nature Reviews Genetics* **11**(6), 446–450.

Fan, R., Albert, P. S. and Schisterman, E. F. (2012), A discussion of gene-gene and gene-environment interactions and longitudinal genetic analysis of complex traits, *Statistics in Medicine* **31**(22), 2565–2568.

Gauderman, W. J., Zhang, P., Morrison, J. L. and Lewinger, J. P. (2013), Finding novel genes by testing G x E interactions in a genome-wide association study, *Genetic Epidemiology* **37**(6), 603–613.

Gourieroux, C. and Monfort, A. (1981), Asymptotic properties of the maximum likelihood estimator in dichotomous logit models, *Journal of Econometrics* **17**(1), 83–97.

Government of Canada, Public Health Agency of Canada (2011), Diabetes in Canada: Facts and figures from a public health perspective - Public Health Agency of Canada. http://www.phac-aspc.gc.ca/cd-mc/publications/diabetes-diabete/facts-figures-faits-chiffres-2011/chap1-eng.php, Accessed: January 2, 2015.

Greenland, S. (2009), Interactions in epidemiology: Relevance, identification, and estimation:, *Epidemiology* **20**(1), 14–17.

Guo, X., Liu, Z., Wang, X. and Zhang, H. (2013), Genetic association test for multiple traits at gene level, *Genetic Epidemiology* **37**(1), 122–129.

Hill, W. G. and Robertson, A. (1968), Linkage disequilibrium in finite populations, *Theoretical and Applied Genetics* **38**(6), 226–231.

Hindorff, L. A. et al. (2009), Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *Proceedings of the National Academy of Sciences of the United States of America* **106**(23), 9362–9367.

Hindorff, L. A. et al. (2015), A catalog of published genome-wide association studies. http://www.genome.gov/gwastudies/, Accessed: January 2, 2015.

Hsu, L. et al. (2012), Powerful cocktail methods for detecting genome-wide gene-environment interaction, *Genetic Epidemiology* **36**(3), 183–194.

International Human Genome Sequencing Consortium (2001), Initial sequencing and analysis of the human genome, *Nature* **409**(6822), 860–921.

International Human Genome Sequencing Consortium (2004), Finishing the euchromatic sequence of the human genome, *Nature* **431**(7011), 931–945.

Khoury, M. J. and Wacholder, S. (2009), Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies–challenges and opportunities, *American Journal of Epidemiology* **169**(2), 227–230.

Klein, R. J. et al. (2005), Complement factor H polymorphism in age-related macular degeneration, *Science* **308**(5720), 385–389.

Ko, Y.-A., Saha-Chaudhuri, P., Park, S. K., Vokonas, P. S. and Mukherjee, B. (2013), Novel likelihood ratio tests for screening gene-gene and gene-environment interactions with unbalanced repeated-measures data, *Genetic Epidemiology* **37**(6), 581–591.

Kooperberg, C. and LeBlanc, M. (2008), Increasing the power of identifying gene x gene interactions in genome-wide association studies, *Genetic Epidemiology* **32**(3), 255–263.

Kraft, P., Yen, Y.-C., Stram, D. O., Morrison, J. and Gauderman, W. J. (2007), Exploiting gene-environment interaction to detect genetic associations, *Human Heredity* **63**(2), 111–119.

Lander, E. S. (1996), The new genomics: global views of biology, *Science* **274**(5287), 536–539.

Lewontin, R. C. (1964), The interaction of selection and linkage. i. general considerations; heterotic models, *Genetics* **49**(1), 49–67.

Li, D. and Conti, D. V. (2009), Detecting gene-environment interactions using a combined case-only and case-control approach, *American Journal of Epidemiology* **169**(4), 497–504.

Manolio, T. A. (2013), Bringing genome-wide association findings into clinical use, *Nature Reviews Genetics* **14**(8), 549–558.

Manolio, T. A. et al. (2009), Finding the missing heritability of complex diseases, *Nature* **461**(7265), 747–753.

Mukherjee, B. and Chatterjee, N. (2008), Exploiting gene-environment independence for analysis of case-control studies: an empirical bayes-type shrinkage estimator to trade-off between bias and efficiency, *Biometrics* **64**(3), 685–694.

Mukherjee, B. et al. (2012*a*), Principal interactions analysis for repeated measures

data: application to gene-gene and gene-environment interactions, *Statistics in Medicine* **31**(22), 2531–2551.

Mukherjee, B. et al. (2012*b*), Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons, *American Journal of Epidemiology* **175**(3), 177–190.

Murcray, C. E., Lewinger, J. P., Conti, D. V., Thomas, D. C. and Gauderman, W. J. (2011), Sample size requirements to detect gene-environment interactions in genome-wide association studies, *Genetic Epidemiology* **35**(3), 201–210.

Murcray, C. E., Lewinger, J. P. and Gauderman, W. J. (2009), Gene-environment interaction in genome-wide association studies, *American Journal of Epidemiology* **169**(2), 219–226.

NIH G x E Interplay Workshop (2011), Gene-environment interplay in common complex diseases: forging an integrative model - recommendations from an NIH workshop, *Genetic Epidemiology* **35**(4), 217–225.

Paré, G., Cook, N. R., Ridker, P. M. and Chasman, D. I. (2010), On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study, *PLoS Genetics* **6**(6), e1000981.

Pearson, T. A. and Manolio, T. A. (2008), How to interpret a genome-wide association study, *Journal of the American Medical Association* **299**(11), 1335–1344.

Piegorsch, W. W., Weinberg, C. R. and Taylor, J. A. (1994), Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies, *Statistics in Medicine* **13**(2), 153–162.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reich, D. E. and Lander, E. S. (2001), On the allelic spectrum of human disease, *Trends in Genetics* **17**(9), 502–510.

Rothman, K. J., Greenland, S. and Walker, A. M. (1980), Concepts of interaction, *American Journal of Epidemiology* **112**(4), 467–470.

Tchetgen Tchetgen, E. J. and Kraft, P. (2011), On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified, *Epidemiology* **22**(2), 257–261.

The 1000 Genomes Project Consortium (2010), A map of human genome variation from population-scale sequencing, *Nature* **467**(7319), 1061–1073.

The International HapMap Consortium (2005), A haplotype map of the human genome, *Nature* **437**(7063), 1299–1320.

Thomas, D. (2010), Gene-environment-wide association studies: emerging approaches, *Nature Reviews Genetics* **11**(4), 259–272.

Thomas, D. C., Lewinger, J. P., Murcray, C. E. and Gauderman, W. J. (2012), Invited commentary: GE-whiz! ratcheting gene-environment studies up to the whole genome and the whole exposome, *American Journal of Epidemiology* **175**(3), 203–207.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York.

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

Visscher, P. M., Hill, W. G. and Wray, N. R. (2008), Heritability in the genomics era - concepts and misconceptions, *Nature Reviews Genetics* **9**(4), 255–266.

Wang, K. and Abbott, D. (2008), A principal components regression approach to multilocus genetic association studies, *Genetic Epidemiology* **32**(2), 108–118.

Warnes, G., with contributions from Gorjanc G., Leisch, F. and Man, M. (2013), genetics: Population genetics. http://CRAN.R-project.org/package=genetics.

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