COMPARISON OF METHODS INCORPORATING COVARIATES INTO AFFECTED SIB PAIR LINKAGE ANALYSIS

by

Hui-Ju Tsai

BS, National Yang-Ming University, 1990

MS, National Yang-Ming University, 1994

MPH, Yale University, 1997

Submitted to the Graduate Faculty of

Department of Human Genetics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2004

UNIVERSITY OF PITTSBURGH

FACULTY OF GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Hui-Ju Tsai

It was defended on

March 2nd, 2004

and approved by

M. Michael Barmada, Ph.D., Assistant Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Bernie Devlin, Ph.D., Associate Professor, Department of Psychiatry, School of Medicine, University of Pittsburgh

Eleanor Feingold, Ph.D., Associate Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Robert E. Ferrell, Ph.D., Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Daniel E. Weeks, Ph.D., Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh
Dissertation Director

COMPARISON OF METHODS INCORPORATING COVARIATES INTO AFFECTED SIB PAIR LINKAGE ANALYSIS

Hui-Ju Tsai, Ph.D.

University of Pittsburgh, 2004

Complex diseases such as type 2 diabetes, hypertension and psychiatric disorders have been major public health problems in US. In order to increase the power in the linkage analysis of complex traits, genetic heterogeneity has to be taken into account. During the past few years, several methods have been proposed for dealing with this issue by incorporating covariate information into the affected sib pair (ASP) analysis. However, it is still not clear how these approaches perform under different gene-environment (G x E) interactions. The covariate statistics evaluated in this study are: (1) mixture model; (2) general conditional-logistic model (LODPAL); (3) multinomial logistic regression models (MLRM under no dominance, no additive and min-max restriction); (4) extension of the maximum-likelihood-binomial approach (MLB); (5) ordered-subset analysis (OSA with three different rank orders: high-to-low, low-to-high and optimal-slice); (6) logistic regression modeling (COVLINK). Based on the chromosome-based approach, we have written simulation programs to generate data under various G x E models and disease models. We first define the empirical statistical significance thresholds using C2, the environmental risk factor, under the null hypothesis. We then evaluate the power of the covariate statistics when different covariates are used. We also compare the performance of the covariate statistics with the model-free methods (S_{all} and S_{pair}). In all three G x E interaction models, most covariate methods perform better when using C1, the covariate with

G x E interaction effect, than when using C2 or the random noise covariate C3, except for MLB and the low-to-high OSA method. Comparing with the model-free methods (using S_{all} as the baseline), mixture model and the high-to-low OSA method perform the best of the covariate statistics when using C1. However, when using C2 or C3, most covariate statistics provide less power than S_{all}. Only MLB has comparable power to S_{all} across all genetic models. According to our results, in different G x E interactions, one should apply the appropriate covariate statistic and include the suitable type of covariates carefully.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

To my parents and teachers

ACKNOWLEDGEMENTS

During the past several years, I have exposed my life to various fields and have met many wonderful people. My life would be utterly different if I have not been through all these experiences.

"Why don't you consider working in a lab?" – Sherry Chen

Sherry was one of my best friends in college. At that time, we just graduated from college. None of us would like to be a nurse, our major in college. She decided to continue her study on Molecular Biology. It sounded not a bad idea. I therefore had spent four years in the same field and received my first master degree.

"Are you sure you want to continue your study on Molecular Biology?" – Fung-Fang Wang

Fung-Fang was my advisor in the Department of Biochemistry at National Yang-Ming University. I had worked on the project regarding expression and regulation of TGF-β in her lab. She always reminded her students to think about what was your interest. You just wanted the degree or really had interest in this field. After working 12 hours per day for two years, I was exhausted when I completed my study at that time. If this were not the right field for me, what else could I do? I decided to give myself another try. I ended up continuing my graduate study in the Department of Epidemiology and Public Health at Yale University.

During the two years at Yale, I had to adjust the differences in several aspects: culture, language and study. Yale has an elegant campus and lots of intelligent people. I am glad that I had stayed there for two years. In these two years, I had received solid training on Epidemiology and Biostatistics. I also realized that in order to develop a successful career, I had to be independent, tough and strong-minded.

"Have you decided what you would do next?" – Brenda Cartmel

Brenda was my second advisor at Yale. She supervised my summer internship and monitored my work on a cancer prevention trial. She originally came from England and understood that it was not easy for foreign students to adjust their life and study at the same time in US. She did not only give me great support on the project, but also concerned about my professional plan. In the second year, I applied Ph.D. programs in US. Unfortunately, I only received several admission, but did not obtain scholarship from any Ph.D. program. I therefore decided to move back to Taiwan after receiving my second master degree. Later, I got a job offer in the Division of Biostatistics and Bioinformatics at National Health Research Institutes.

"Do you want to work on a genetic project?" – Chao A. Hsiung

Chao is the director of the division. I worked as a data coordinator and statistician there. Like most of my colleagues, I was in charge of some clinical trials in the beginning. When she decided to participate in an international hypertension genetic project, she encouraged me to join. This project was a window that let me have the opportunity to touch the fields of Statistical Genetics and Genetic Epidemiology. Because of this project, I decided to come back US again to continue my Ph.D. study in the Department of Human Genetics at University of Pittsburgh.

"Yep, I will be happy to be your advisor." – Daniel E. Weeks

At Pitt, I have worked with Dan, my advisor. He is renowned for his research in Statistical Genetics and Genetic Epidemiology. He has dedicated his effort to my study in various ways. I have learned the background of genetic studies from his courses. I have learned how to manipulate genetic programs and write computer programs from him. I have participated in two cooperative genetic projects under his supervision. He has provided me lots of excellent ideas in genetic research. He genuinely cares about my professional career and gives me many valuable suggestions. I am truly lucky to have him as my mentor.

Dr. Robert Ferrell, the chairman of the Department of Human Genetics and my first-year academic advisor, had provided me advices regarding course work and looking for the research advisor. Dr. Eleanor Feingold helped me to settle down in our department when I was still in Taiwan and just arrived at Pitt. She has provided me great suggestions when I have statistical problems. Dr. Bernie Devlin gave me some useful materials regarding my dissertation and shared with us his comments for our preliminary results. I have taken the courses of linkage analysis and quantitative genetics instructed by Drs. Michael Barmada and Daniel E. Weeks. He has supervised our department Linux machines, which make me finish my simulation work smoothly.

Many people in the Department of Human Genetics have offered me an enjoyable, friendly and supportive environment. Nandita Mukhopadhyay, Haydar Sengul, Guy Brock and Brian Reck have answered my technical questions. Jeanette Norbut, Michele Lavalley and Gloria Duval have taken care my registration, stipend, facility support, and administrative-related stuff. Sarita Alvarez Sanguedolce and Xiaojing Ye, my classmates, had worked with me on our course work during the first two years. We always give each other a hand when one of us needs it.

"Would you like to join our bonfire party?" – Barb Lanser

Barb Lanser and Rod Lanser are my Canadian friends whom I have met in Pittsburgh. They have been offering their hospitality. Whenever I go over their house, it makes me feel like I am back to my home in Taiwan again. I have been meeting many nice people during the past few years: Jung-Ying Tzeng, Chi-Yi Yu, Beatriz Rico-Verdin, Martha Victoria, Yah-Tyng Sheu, EunRyoung Sa, Giovanna Vignolo and Mirian Anzoise. I always have fun whenever hanging out with them. Their friendship enriches my life here.

"You want to go back to school again?!" – Wei-Hsiung Tsai

Wei-Hsiung Tsai and Ling-Tzee Chang, my parents and my most loyal supporters, have always concerned about my well-being. Whenever I switch my interest to another field, they try to understand my motivation and provide me their full support. We have had one "conference call" weekly, at least. They do their best to monitor my progress and push me working hard. They are my lifelong mentors.

Raising a Ph.D. student is like planting a tree. It takes time and effort to make sure it has grown well. I am very grateful to meet so many magnificent gardeners in my life. I hope I will become a good gardener and be able to plant not only a tree, but also a forest!

INTRODUCTION

Genetic heterogeneity is an important issue that should be taken into account while performing affected sib pair (ASP) linkage analysis in complex diseases. Various methods have been proposed for dealing with genetic heterogeneity by incorporating covariate information into the models. But the power of these approaches requires investigation.

In this study, we generate data under three types of gene-environment (G x E) interaction models and eight different disease models, respectively. We then evaluate the performance of the following covariate statistics: the mixture model, the multinomial logistic regression model approaches, the general conditional-logistic model, the extension of the Maximum-Likelihood-Binomial approach, ordered-subset analysis, and logistic regression modeling. We also compare the power of the covariate statistics with the model-free methods (allele sharing statistics: S_{all} and S_{pairs}) and the quantitative-trait linkage (QTL) approaches (variance-component linkage analysis and regression-based quantitative-trait linkage analysis).

The specific aims are as follows:

1. Write an efficient data simulation program based on the chromosome-based approach.
2. Compute the empirical significance thresholds of each covariate method, based on data simulated under the null.

3. Estimate the power of covariate statistics using different types of covariates, based on simulation under various G x E interaction models.

4. Compare the power of covariate statistics with the power of model-free methods and QTL approaches.

1. GENETIC HETEROGENEITY AND GENE-ENVIRONMENT INTERACTION

1.1. BACKGROUND AND SIGNIFICANCE

The twentieth-first century has been described as the post-genome era. The completion of the first draft of Human Genome Sequence provided comprehensive genomic data to the scientific community. Mendel's laws were proposed in 1900's, since then many genetic studies have been conducted and achieved significant results for monogenic diseases. Particularly, since the early 1980s, the development of new laboratory technologies has offered scientists golden opportunities to explore the molecular level of DNA sequences. Since then, thousands of molecular markers have been genotyped and applied in gene mapping. The various statistical methods applied to such marker data in linkage analyses have successfully localized susceptibility genes determining simple Mendelian traits. During the past two decades, the emphasis of gene mapping has shifted to complex traits such as type 2 diabetes, hypertension and psychiatric disorders. However, varied etiologies and genetic heterogeneity in the complex traits can cause difficulties in detecting true peaks representing evidence for linkage and localizing disease genes within those peaks (Figure 1.1). To date, even with high-density genome-wide marker data, searching for susceptibility genes for complex traits is still an ongoing and challenging issue to geneticists.

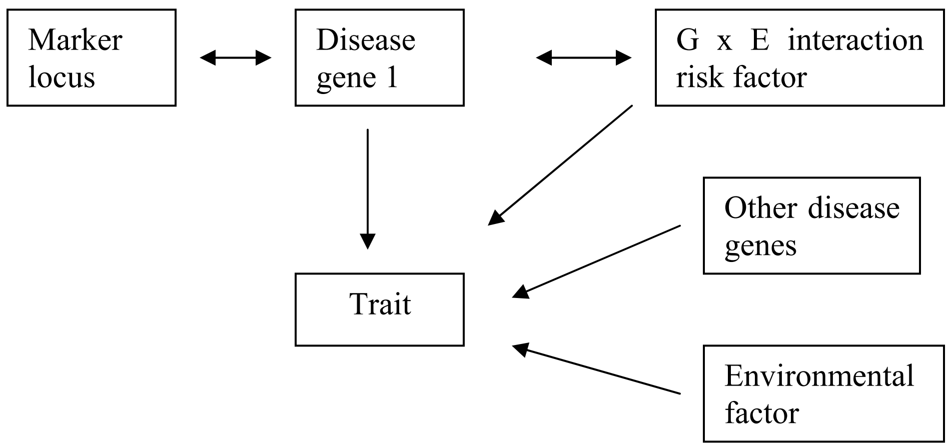


Figure 1.1 A simplified model of the etiological factors predisposing to a complex trait (modified from Terwilliger and Weiss 1998)

1.2. GENETIC HETEROGENEITY

1.2.1. Definition

Genetic heterogeneity is often observed in complex diseases. Two types of genetic heterogeneity should be distinguished: locus heterogeneity and allelic heterogeneity. Allelic heterogeneity can be defined as: (1) individuals having different alleles at the same locus leading to the same phenotype; (2) individuals having different alleles at the same locus leading to different phenotypes. We here only consider the first definition. When allelic heterogeneity exits, it may cause small effective sample size in linkage studies and case-control association studies.

Locus heterogeneity occurs when the trait in some pedigrees is due to a susceptibility gene located in one region, while the same trait in other pedigrees is caused either by another gene located elsewhere or by an environmental factor. Locus heterogeneity impacts on association studies, as well as on linkage analysis by reducing of the effective sample size (the details about the impact on linkage analysis will be provided in Section 1.2.5). Power will increase if we correctly model the effects of genetic heterogeneity (Hanson and Knowler 1998).

1.2.2. Examples

Heterogeneity is expected in conditions where a general biochemical pathway has failed. Generally, the end products of long biochemical pathways are almost always heterogeneous. If susceptibility loci regulating different steps in a pathway are mutated, they may lead to the same disease outcome. Mental retardation, blindness, deafness and various types of cancer are typical examples of disease exhibiting locus heterogeneity. A striking example is Usher syndrome, an autosomal recessive trait. Phenotypes of Usher syndrome are hearing loss and retinitis pigmentosa. Mutations in at least six unlinked loci have been identified: *USH1A* (14q31-qter), *USH1B* (11q13-q14), *USH1C* (11p13-p14), *USH2A* (1q32-q34), *USH2B* (3p24.2-p23) and *USH3* (3q) (Strachan and Read 1996).

1.2.3. Methods for detecting genetic heterogeneity

Several statistics have been developed for detecting genetic heterogeneity. Using a parametric approach, Schaid et al. (2001) developed a regression-based extension of the mixture likelihood. The mixture likelihood for a pedigree was denoted as: $L(\alpha, \theta) = \alpha * L(\theta) + (1- \alpha)*L(0.5)$, where $L(\theta)$ was defined as the likelihood for a pedigree at recombination fraction θ, and α was defined as the probability that a pedigree was linked to the susceptibility locus of interest. They modeled pedigree features using logistic regression to determine the probability that a pedigree was linked or unlinked and estimated the mixture likelihood by putting the recombination fraction and pedigree features in the model. Then they tested whether the pedigree features differentiated the linked and unlinked groups given linkage. This regression-based method was applied to the analysis of a familial prostate cancer study.

In addition, tests for genetic heterogeneity in identical-by-descent (IBD) allele sharing of affected relatives have been suggested (Mirea et al. 2004). These authors added a family-level covariate into the linear and exponential likelihood models (Whittemore 1996; Kong and Cox 1997). Under linkage, they evaluated S_{all} and S_{pairs} allele-sharing scoring functions between two covariate-defined family subgroups. Then they tested the null hypothesis H_0: no covariate-associated heterogeneity vs. the alternative hypothesis H_A: covariate-associated differences in IBD allele sharing.

1.2.4. Stratification vs. heterogeneity

Accounting for gene-gene interaction using stratification has also been proposed to deal with genetic heterogeneity (Leal and Ott 2000). In other words, one would stratify pedigrees by the IBD sharing proportions at one locus, then perform linkage analyses at a second locus. When there is genetic heterogeneity of IBD allele sharing between groups, stratification may increase the power. The approach was applied to an affected-sib-pair study of type I diabetes (Cordell et al. 1995). Cordell et al. (1995) stratified the data on the basis of IBD sharing patterns at one locus, then tested for excess IBD sharing at the other locus. However, the stratification approach may encounter several problems such as small sample sizes in each subset and multiple testing. Also, when there is no difference in IBD allele sharing between groups (homogeneity of IBD sharing), stratification can lead to power loss.

1.2.5. Impact on linkage analysis

Both parametric methods and model-free approaches have been widely applied to linkage analysis (detailed frameworks for both approaches are provided in Appendix A). However, when dealing with complex diseases, both approaches have their limitations. Generally, the development of complex diseases not only depends on genetic factors, but also depends on environmental factors, gene-gene interaction and gene-environment interaction. Disease allele frequencies can vary across different ethnicities. Also, genetic and clinical heterogeneity often are present. In parametric linkage analysis, in order to provide adequate power, modes of inheritance of the disease loci have to be accurately specified. However, the mode of inheritance is often unclear and may not follow any known simple Mendelian pattern. Penetrances therefore may be hard to estimate precisely. Model-free methods may not perform well either because of genetic heterogeneity. The effective sample size becomes small when genetic heterogeneity exists. Previous work showed that as the proportion of unlinked families increases, the power decreases significantly (Figure 1.2) (Weeks and Harby 1995). If the effect of genetic heterogeneity is not considered, both approaches may only provide limited power.

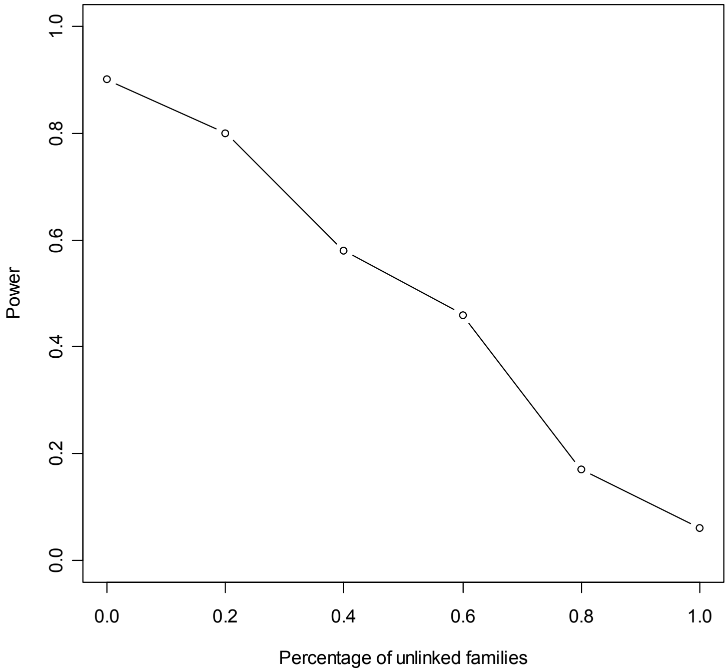


Figure 1.2 Effect of heterogeneity: recessive model with disease allele frequency = 0.27; a 2-allele marker linked at θ = 0.00; 300 affecteds/replicate. (Modified from Weeks and Harby 1995)

Huang and Vieland (2001) compared the parametric HLOD approach with model-free methods, the mean sharing test and the maximum likelihood score (MLS), in the presence of genetic heterogeneity (details about HLOD will be provided in Chapter 2, and about model-free methods in Appendix A). They simulated affected sib pair (ASP) data sets under different disease models with two different disease allele frequencies separately. The data sets had 200 ASPs, but only contained a certain proportion of linked ASPs. The pooled set consisted of two subsets with 70% and 20% proportion of linked ASPs individually. The results showed that the HLOD, mean sharing test and MLS using the pooled set had less power than when only using the set with 70% linked families. They concluded that both parametric and model-free statistics could suffer a loss of power when there is a heterogeneity effect.

Detecting linkage signals in familial breast cancer is a successful example of dealing with heterogeneity in linkage analysis. Genetic heterogeneity of breast cancer in the recruited families appears to be stratified by age of onset. Previous work showed that linkage of breast cancer yields a lod score of 5.98 to *D17S74* residing at chromosome 17q21 in the early age-of-onset families (Hall et al. 1990). But LOD scores for late age-of-onset families were negative. In this study, they first ranked families according to the average age-of-onset within each family. Then they calculated each family's lod score and recorded the accumulated lod scores based on the low-to-high rank order. The maximum lod score of 5.98 occurred in the subset with mean age-of-onset less than 47.4. Later the *BRCA1* gene associated with early-onset familial breast cancer was discovered in the region of chromosome 17q21. The approach here was the same idea as the later developed ordered-subsets analysis (Hauser et al. 1998).

1.3. GENE-ENVIRONMENT INTERACTION

1.3.1. Background

As we mentioned in Section 1.1, the etiologies of complex diseases are usually a mixture of genetic effects, environmental risk factors, gene-gene interaction, and gene-environment (G x E) interaction. Disease in different individuals may be caused by different susceptibility genes or environmental risk factors. The heterogeneity issue is discussed in Section 1.2. In addition to locus heterogeneity and allelic heterogeneity, environmental risk factors and G x E interactions (covariate-related genetic heterogeneity) cannot be ignored when mapping susceptibility genes for complex diseases.

1.3.2. Examples

Various types of G x E interactions can be observed in complex traits. Ottman (1990; 1996) suggested five different models of G x E interaction and provided examples under these models. These biologically plausible models illustrate different possible relationships between genetic effect, environmental effect, and their interaction effect on disease outcome.

In Model A, the susceptibility gene does not cause the disease risk directly, but it increases the level of the risk factor (Figure 1.3A). In this model, the risk factor can be a quantitative trait due to the susceptibility gene. The relationship between the phenylketonuria (PKU) gene, blood levels of phenylalanine, and mental retardation is an example of this model. In this example, blood levels of phenylalanine are not only an intermediate product in the PKU pathway, but also can be an environmental risk factor.

In Model B, the susceptibility gene exacerbates the effect of the risk factor (Figure 1.3B). When the risk factor is absent, the gene has no effect. The risk factor has a direct effect on the

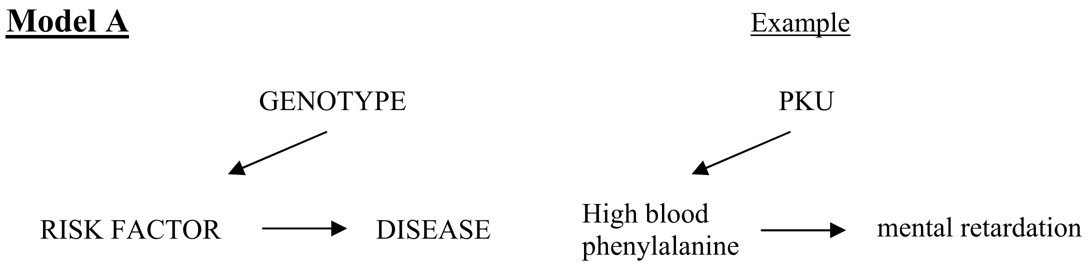


Figure 1.3A The relationship between a high-risk genotype and an environmental exposure in model A (Modified from Ottman 1990)

disease risk. An example is the relationship between xeroderma pigmentosum, ultraviolet radiation, and skin cancer.

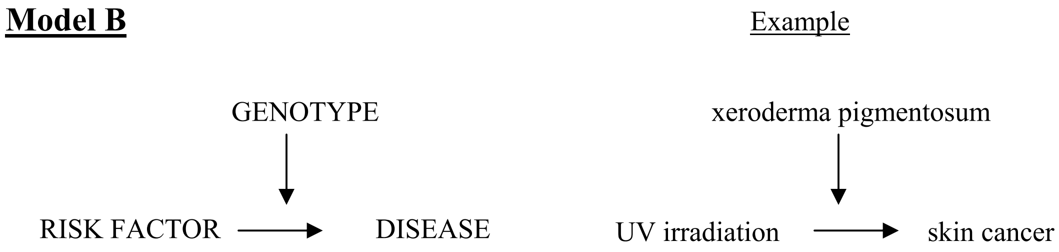


Figure 1.3B The relationship between a high-risk genotype and an environmental exposure in model B (Modified from Ottman 1990)

In Model C, the susceptibility gene directly affects the disease risk. The risk factor exacerbates the genetic effect (Figure 1.3C). The risk factor has no effect on subjects with a low-risk genotype. An example of this model is porphyria variegata. Individuals with porphyria variegata, an autosomal dominant disease, have skin problems with various severity levels. If the subjects with the susceptibility genotype are exposed to barbiturates, a non-harmful exposure for the general public, they may be paralyzed and/or die because of the acute effect of this exposure.

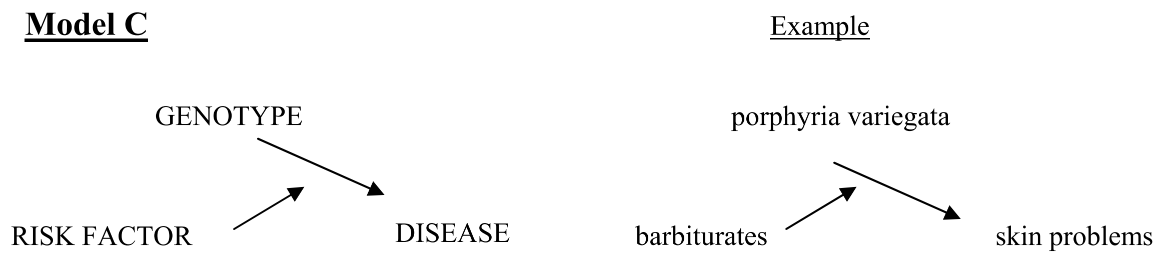


Figure 1.3C The relationship between a high-risk genotype and an environmental exposure in model C (Modified from Ottman 1990)

In Model D, neither the susceptibility gene nor the risk factor affects disease risk by itself, but when both are present, they increase the disease risk (Figure 1.3D). For example, individuals with glucose-6-phosphate dehydrogenase (G6PD) deficiency, an X-linked recessive disease, will develop severe anemia if they consume fava beans. Individuals without G6PD deficiency can eat fava beans without causing anemia.

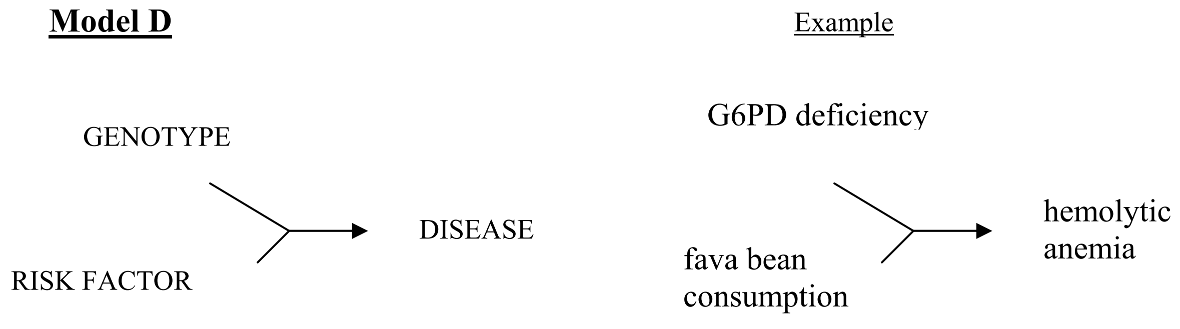


Figure 1.3D The relationship between a high-risk genotype and an environmental exposure in model D (Modified from Ottman 1990)

In Model E, the susceptibility gene and the risk factor each have a direct influence on disease outcome (Figure 1.3E). Chronic obstructive pulmonary disease (COPD) is an example of this model. Non-smokers with α-1-antitrypsin deficiency have an increased risk of developing COPD. The disease risk increases in smokers regardless of α-1-antitrypsin deficiency.

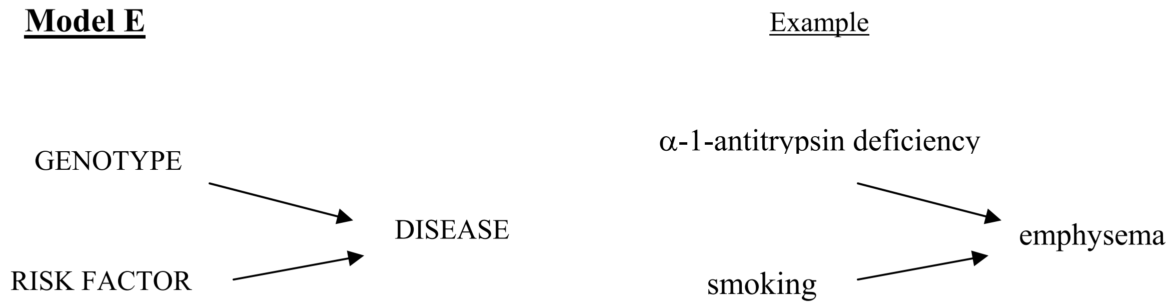


Figure 1.3E The relationship between a high-risk genotype and an environmental exposure in model E (Modified from Ottman 1990)

Similar to the five models described by Ottman (1990; 1996), Khoury et al. (1993) described six biologically plausible patterns of G x E interaction. The magnitude of disease risk varies in different patterns with various directions of the genetic effect and the environmental exposure (Table 1.1 and Figure 1.4). We first define the background risk as I, the genetic risk with exposure absence as IR_g, the environmental risk with susceptibility genotype absence as IR_e, and the G x E interaction risk (when both genetic risk and environmental risk are present) as IR_{ge}. The magnitude of the background risk, I, and the G x E interaction risk, IR_{ge}, are the same in the six patterns. But the genetic risk and the exposure behave differently in the different patterns. In Pattern 1, neither the susceptibility genotype nor the exposure alone increases the risk. Therefore, Pattern 1 corresponds to Ottman's Model D. In Pattern 2, the exposure increases disease risk, but the gene does not have any effect when the environmental factor is absent. This is the same as Ottman's Model B. Complimentary to Pattern 2, in Pattern 3, the susceptibility genotype alone can increase risk, but not the exposure. Ottman's Model C matches Pattern 3. In Pattern 4, which is similar to Ottman's Model E, presence of either the susceptibility genotype or the exposure increases the risk. The susceptibility genotype in Pattern 5 has a protective effect, but the exposure alone does not increase the risk. In contrast to Pattern 5, the susceptibility genotype in Pattern 6 has a protective effect, but the exposure alone increases the risk.

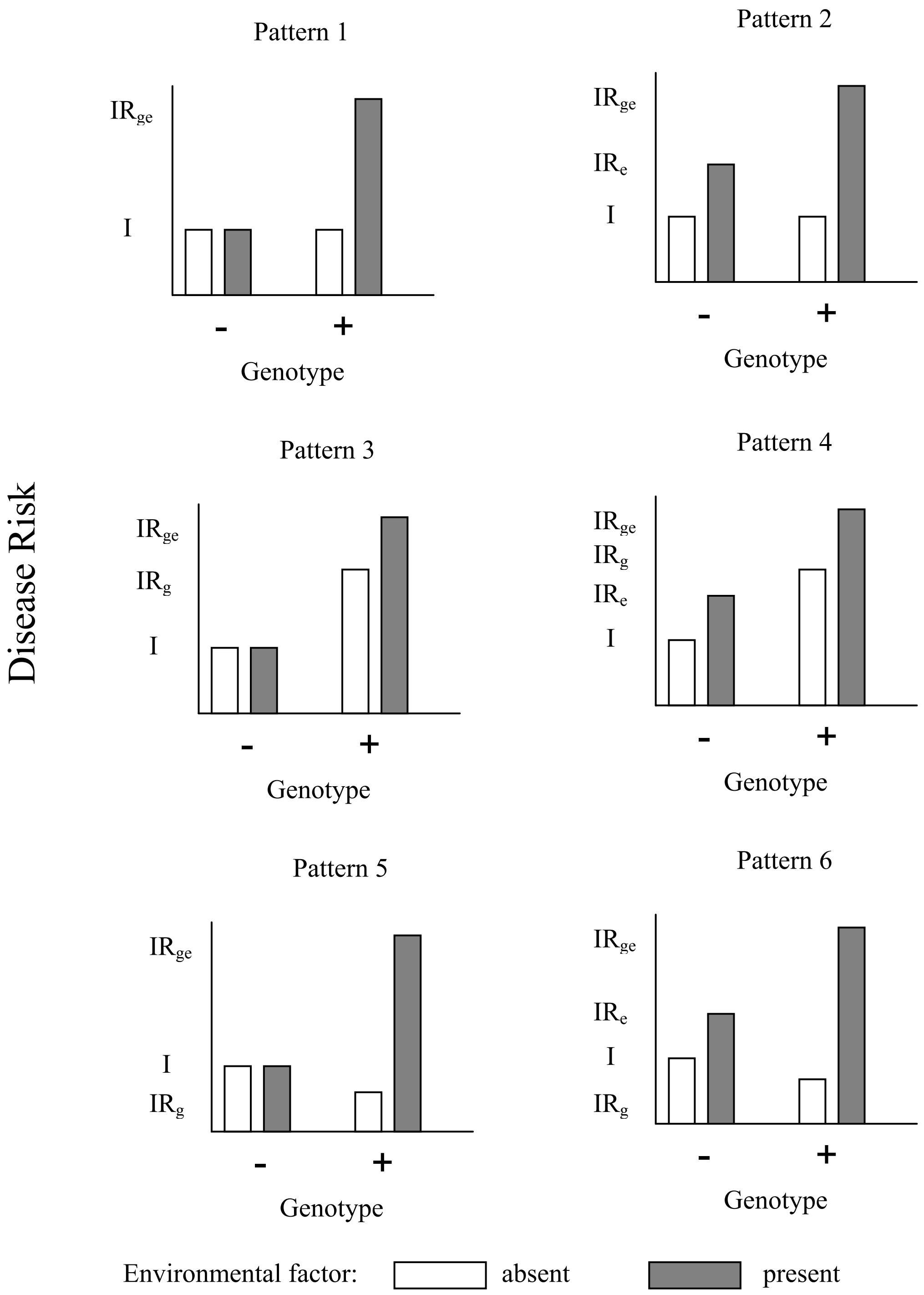
Table 1.1 Six patterns (matching Figure 1.4) of genetic effect, exposure and G x E interaction observed in complex diseases (From Khoury et al. 1993)

	Effect on disease risk of	
Patterns	Genotype in absence of environment	Environment in absence of genotype
1	No effect; $R_g = 1$	No effect; $R_e = 1$
2	No effect; $R_g = 1$	Increases risk; $R_e > 1$
3	Increases risk; $R_g > 1$	No effect; $R_e = 1$
4	Increases risk; $R_g > 1$	Increases risk; $R_e > 1$
5	Decreases risk; $R_g < 1$	No effect; $R_e = 1$
6	Decreases risk; $R_g < 1$	Increases risk; $R_e > 1$

From Khoury et al. (1993)
NOTE: R_g: genetic risk w/ exposure absence; R_e: environmental risk w/ susceptibility gene absence.

1.3.3. Impact on segregation analysis

Several works have surveyed the effects of different G x E interactions on segregation analysis (Eaves, 1984; Tiret et al., 1993). Tiret el al. (1993) applied a regressive model, similar to the class D regressive model proposed by Bonney (1984), specifying equal sib-sib correlations in



I: background disease risk; IR_g: genetic risk w/ exposure absence; IR_e: environmental risk w/ susceptibility genotype absence; IR_{ge}: both genetic risk and environmental risk are present

Figure 1.4 The magnitudes of genetic effect, exposure and G x E interaction in different patterns (From Khoury et al. 1993)

major genotypes. But their segregation analysis model accounted for the G x E interaction effect. Parameters in the regressive model were the gene frequency, the three genotype-specific means, and the three genotype-specific slopes. The results showed that ignoring the G x E interaction effect decreases power to detect a major gene effect and leads to biased parameter estimates.

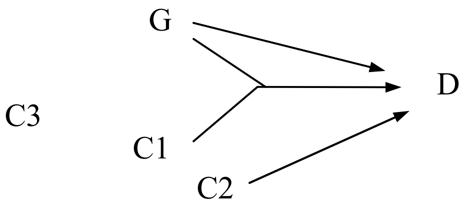
1.3.4. Types of gene-environment interaction models

As we described in Section 2.2, different types and directions of G x E interaction have different effects on disease outcome. Therefore, it is important that our proposed G x E interaction models should be biologically plausible. Herewith, we propose three biologically reasonable G x E interaction scenarios, which may commonly occur in complex diseases. In the later chapters, we will investigate how the statistical methods behave under these different G x E interaction scenarios.

We consider three different types of gene-environment (G x E) interaction models. For simplicity, each of these models has only one disease locus G and one environmental factor C1, which interacts with the disease locus G. Each model also has an environmental factor C2 that influences disease risk, but does not interact with the genetic effect. We also simulate a random covariate C3 that has no effect at all on disease liability.

Overall, the liability model is used to assign individual's affection status according to a threshold in the liability distribution. The liability model is influenced by G, covariate C1, covariate C2, a polygenic effect PG and a random error E. However, whether G or C1 has direct effect on the liability depends on the different types of G x E interaction models. Hence, the components in the liability models vary across different G x E models (details are in the following sections). We then define individual's affection status, giving a desired prevalence, using a liability threshold.

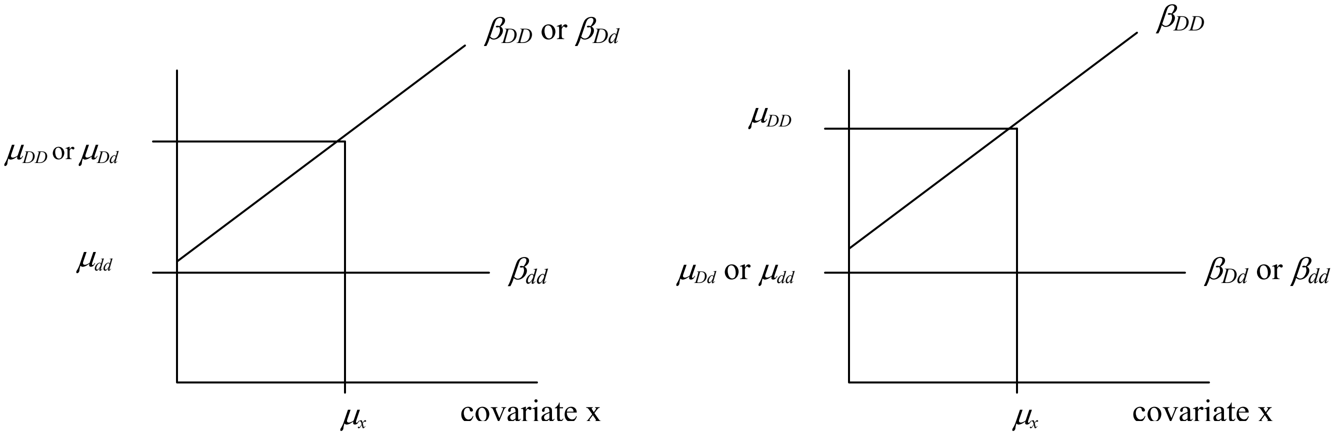
Type I G x E interaction model: in the Type I G x E model, the disease liability is determined by one disease locus G, two covariates C1 and C2, a polygenic effect PG and a random effect E (Figure 1.5). The susceptibility gene has a direct effect on disease outcome. Covariate C1 does not have a direct effect, but it influences the disease risk by interacting with the susceptibility gene. The magnitude of C1's effect depends on the genotype-specific slopes, β_G (Figure 1.6) (Tiret 1993). If an individual carries a high-risk genotype, β_G is equal to one. Otherwise, β_G is equal to zero. In addition, we generate another environmental risk factor (covariate C2). Covariate C2 does not interact with the susceptibility gene G, but directly influences disease outcome. Since the outcome of interest is a complex trait, we also take into account the polygenic effect PG and a random noise effect E as well. The Type I G x E model is similar to a combination of Ottman's Models D and E (1990; 1996).



LI: liability
G: genetic effect
C1: covariate with G x E interaction
C2: environmental covariate
C3: random noise covariate

$$LI = \mu + G + \beta_G * C1 + C2 + PG + E$$

Figure 1.5 Type I model

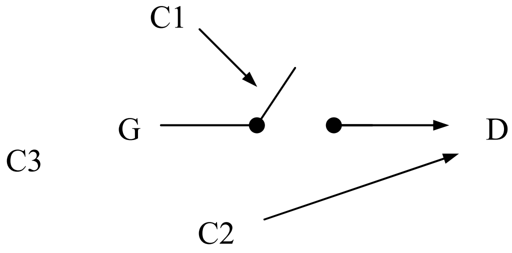


Dominant model

Recessive model

Figure 1.6 Interaction between covariate and susceptibility genotypes under different disease models for the Type I G x E interaction model.

Type II G x E interaction model: the Type II G x E model uses one of two disease liabilities, depending on the value of covariate C1. One liability is determined by one disease locus G, one covariate C2, a polygenic effect PG and a random effect E (Figure 1.7). The other is determined only by one covariate C2, a polygenic effect PG and a random effect E. Whether or not the susceptibility gene G is included in the liability model depends on the covariate C1 level. If C1's level exceeds the threshold (e.g., C1 ≥ 0), the genetic effect is included in the liability. Otherwise, there is no genetic effect in the liability. We also consider an independent exposure: covariate C2. The Type II G x E model is similar to Ottman's Model C, where the risk factor exacerbates the genetic effect, but has no direct effect on disease outcome.

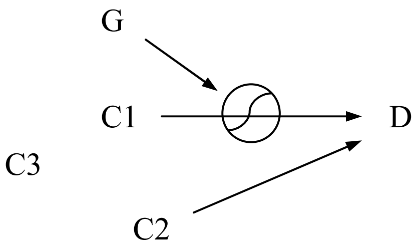


LI: liability
G: genetic effect
C1: covariate with G x E interaction
C2: environmental covariate
C3: random noise covariate

If C1 $\geq$ 0, $LI = \mu + G + C2 + PG + E$
If C1 < 0, $LI = \mu + C2 + PG + E$

Figure 1.7 Type II model

Type III G x E interaction model: in the Type III G x E model, the disease liability is controlled by two covariates C1 and C2, a polygenic effect PG and a random effect E (Figure 1.8). The susceptibility gene does not have a direct effect on disease liability, but it influences the disease outcome via changing the covariate C1 level. Covariate C2 is an independent environmental risk factor, like the covariate C2 in the Type I and Type II models. The susceptibility gene exacerbates the risk factor's effect, but has no direct effect on disease outcome in Ottman's Models A and B, which are close to the Type III model.



LI: liability
G: genetic effect
C1: covariate with G x E interaction
C2: environmental covariate
C3: random noise covariate

$$LI = \mu + C1 + C2 + PG + E$$

Figure 1.8 Type III model

1.4. G x E INTERACTION VS. IBD SHARING PATTERNS

The expected IBD sharing patterns in sib pairs under different types of G x E interaction may deviate from those under the null hypothesis when the covariate has no effect. Greenwood and Bull (1999) presented four G x E models to illustrate the covariate effect on the IBD sharing patterns (Table 1.2). In model A (similar to our Type I model), the covariate increases the disease penetrance so that the chance of developing disease is higher for exposed individuals carrying high-risk genotypes. When both sibs are exposed and the high-risk genotypes have a greater effect, the deviations from the null IBD sharing are more significant. In model B (similar to our Type II model), the covariate is necessary for the susceptibility gene to have an effect. The expected IBD sharing patterns in concordant unaffected pairs and discordant pairs are the same as the null. The power to detect linkage would be low if one recruits all types of sib pairs (both concordant and discordant). In model C, the G x E interaction is not on an additive scale. Overdominance is assumed, so that heterozygotes have higher risk than homozygotes. The deviations of the IBD sharing patterns are observed. In model D, the gene has a protective effect in unexposed subjects, but when the covariate occurs, the gene has a harmful effect in exposed subjects. The IBD sharing patterns in discordant pairs may fall outside the possible-triangle boundaries (Holmans 1993) (the details of the possible-triangle boundaries are provided in Appendix A).

In addition, Guo (2000) also estimated the expected IBD sharing patterns in three examples of G x E interactions: (1) genetic factors interact with exposures in an additive fashion; (2) genetic factors increase disease risk in unexposed individuals, but decrease the risk in exposed individuals; (3) exposures have strong effect, but very mild G x E interaction effect. There is a

Table 1.2 Expected IBD sharing patterns in sib pairs under different G x E models (From Greendwood and Bull 1999)

Model	Exposure	q^{*}	f_0^{**}	f_1	f_2	z_0^{***}	z_1	z_2
A	Neither sib	.05	.05	.10	.30	.24	.50	.27
	Both sib		.05	.20	.60	.19	.50	.31
	One sib					.22	.50	.28
B	Neither sib	.05	.05	.05	.05	.25	.50	.25
	Both sib		.05	.20	.60	.19	.50	.31
	One sib					.25	.50	.25
C	Neither sib	.01	.05	.40	.05	.15	.49	.37
	Both sib		.45	.80	.45	.23	.50	.27
	One sib					.20	.49	.30
D	Neither sib	.20	.424	.0424	.0042	.21	.50	.30
	Both sib		.10	1.00	1.00	.17	.49	.34
	One sib					.37	.50	.12

Note:
* q: disease allele frequency
** f_0, f_1, f_2: exposure-specific penetrances
*** z_0, z_1, z_2: expected IBD sharing patterns

strong genetic effect in unexposed subjects, but a small effect in exposed subjects. The results showed theoretically that the expected IBD sharing patterns and the required sample sizes vary under different G x E interaction behaviors.

1.5. SUMMARY

In order to reveal true signals in linkage analysis of complex traits, the assumptions related to genetic heterogeneity in the statistical models cannot be ignored and must be taken into account very carefully. Otherwise, neither parametric nor nonparametric approaches in linkage analysis can provide adequate power due to the small effective sample size. We discussed the impact of genetic heterogeneity on linkage analysis here. We will describe how to deal with the heterogeneity issue in Chapter 2.

2. STATISTICAL METHODS FOR DEALING WITH HETEROGENEITY

2.1. INTRODUCTION

Since both parametric and nonparametric approaches ignoring heterogeneity cannot provide adequate power to detect signals in linkage analysis for complex diseases, several statistics taking into account the heterogeneity issue have been proposed for the detection of linkage, based on both parametric and nonparametric frameworks. We will introduce parametric approaches in Section 2.2, and various methods of incorporating covariate information into affected-sib-pair (ASP) or affected-relative-pair (ARP) analysis, based on a nonparametric framework in Section 2.3.

2.2. PARAMETRIC APPROACHES VS. GENETIC HETEROGENEITY

Based on the parametric framework, the M-test, the B-test and the admixture model have been proposed to deal with genetic heterogeneity.

2.2.1. M-test

The method proposed by Morton (1956), called the M-test, divides families into pre-defined subsets based on clinical features or ethnic background. To test whether the recombination fraction (θ) varies between subsets, three hypotheses are considered. The first hypothesis H_0 is $\theta = ½$ for all subsets. “The relative support for this hypothesis is measured by the value of the log-likelihood function of the entire dataset evaluated at $\theta = ½$, which we denote as ln L_0" (Sham 1998). The second hypothesis H_1 is that all subsets have the same θ, but $\theta < ½$. " The relative support for this hypothesis is the maximum log-likelihood of the entire data set over $0 \leq \theta \leq ½$, which we denote as ln L_1" (Sham 1998). The third hypothesis H_2 is that θ varies between subsets. "The relative support for this hypothesis is obtained by maximizing the likelihood function over $0 \leq \theta \leq ½$, for each subset of families separately, and then summing the subset-specific maximum log-likelihood. We denote the value of the log-likelihood maximized over subset-specific recombination fractions as ln L_2" (Sham 1998).

For *n* subsets, one can compare H_0 with H_1 using the likelihood ratio test statistic 2* (ln L_1 - ln L_0), which tests for linkage. This is a one-sided χ^2 test with 1 degree of freedom (df). One can also compare H_1 with H_2 using 2* (ln L_2 - ln L_1), which tests for heterogeneity. It is an asymptotically one-sided χ^2 distribution with (*n*-1) df.

2.2.2. B-test

Similar to the M-test, Risch (1988) suggested a Bayesian approach, which is often called the B-test. The θ values in different subsets are assumed to follow a beta distribution with two parameters. One can estimate these two parameters from the posterior distribution of θ. The test statistic 2* (ln L_2 - ln L_1) here is also used to test for heterogeneity (L_1 and L_2 are the same notations as the ones in the M-test). It is a one-sided χ^2 distribution with 1 df. Ott (1999) concluded that "the B-test is generally conservative and often more than the M-test (Risch 1988)".

2.2.3. Admixture model

To take into account genetic heterogeneity, the method introduced by Smith (1959) used a likelihood composed of a mixture of two types of families, one with linkage and one without linkage. This method is known as the admixture model. Let the proportion of families with linkage be α, and (1-α) for families without linkage. Then the likelihood of the ith family can be written as: $L_i(\alpha, \theta) = \alpha^*L_i(\theta) + (1-\alpha)^*L_i(½)$, where $L_i(\theta)$ is the likelihood, evaluated at $\theta = \theta_l$ ($0 \leq \theta_l \leq ½$). The null hypothesis H_0 is either $\theta = ½$ or $\alpha = 0$. The second hypothesis H_1 is $0 \leq \theta \leq ½$, assuming $\alpha = 1$. The third hypothesis H_2 is $0 \leq \theta \leq ½$ and $0 \leq \alpha \leq 1$. The relative support for the null hypothesis is calculated by the value of the log-likelihood function evaluated at $\theta = ½$ and $\alpha = 0$, which we denote as ln L_0. The relative support for the second hypothesis is the maximum value of the log-likelihood function on the line $\alpha = 1$, which we denote as ln L_1. The relative support for the third hypothesis is the maximum value of the log-likelihood function in the entire rectangle space defined by $0 \leq \theta \leq ½$ and $0 \leq \alpha \leq 1$, which we denote as ln L_2.

Hence, the likelihood ratio test statistic 2* (ln L_1 - ln L_0) is used to test for linkage. It is a one-sided χ^2 distribution with 1df. The second likelihood ratio test statistic 2* (ln L_2 - ln L_1) is applied to test for locus heterogeneity. It is a one-sided χ^2 distribution with 1 df. The third likelihood ratio test statistic 2* (ln L_2 - ln L_0) is used to test for linkage, but assuming heterogeneity ($0 < \alpha < 1$) (often called the A-test). Because the two parameters θ and α are separate under the alternative hypothesis but are confounded under the null hypothesis, the asymptotic distribution of this statistic is not a χ^2 distribution. An approximation is that it is a 50:50 mixture of a probability mass at zero and the larger of two independent χ^2 random variables each with one degree of freedom (Faraway 1993).

Based on the admixture model, the lod score corresponding to the likelihood ratio for linkage assuming heterogeneity can be defined as: $\log_{10}[L(\hat{\theta},\hat{\alpha}) / L(½, 0)]$, which is often called the heterogeneity lod score (HLOD score). The admixture model has been implemented in the HOMOG program (Ott 1999). Although use of a heterogeneity parameter, α, can make parametric linkage analysis more robust to model misspecification, the limitation of this approach is that family features, such as body mass index, that may be used to differentiate linked and unlinked families are not used.

Whittemore and Halpern (2001) reported that the estimation of the proportion, α, of pedigrees linked to the disease gene based on the admixture likelihood approach is sometimes problematic. For instance, the estimation of α is valid only when all the following assumptions are met: the disease mutations have no effect on family size, the mutations are rare, and the mutations of all susceptibility genes have equal penetrances. Even when all assumptions were met, their work showed that the HLOD score is estimated on the basis of an incorrect likelihood function. Additionally, Hodge et al. (2001) cited several publications investigating the violations of the HLOD assumption: each family must either be linked or unlinked (Goldin 1992, Vieland et al. 2001). Hodge et al. (2001) pointed out these publications supporting the application of HLOD score to detect a linkage signal, even under the wrong assumed heterogeneity model.

2.3. METHODS FOR INCORPORATING COVARIATES INTO AFFECTED-SIB-PAIR ANALYSIS

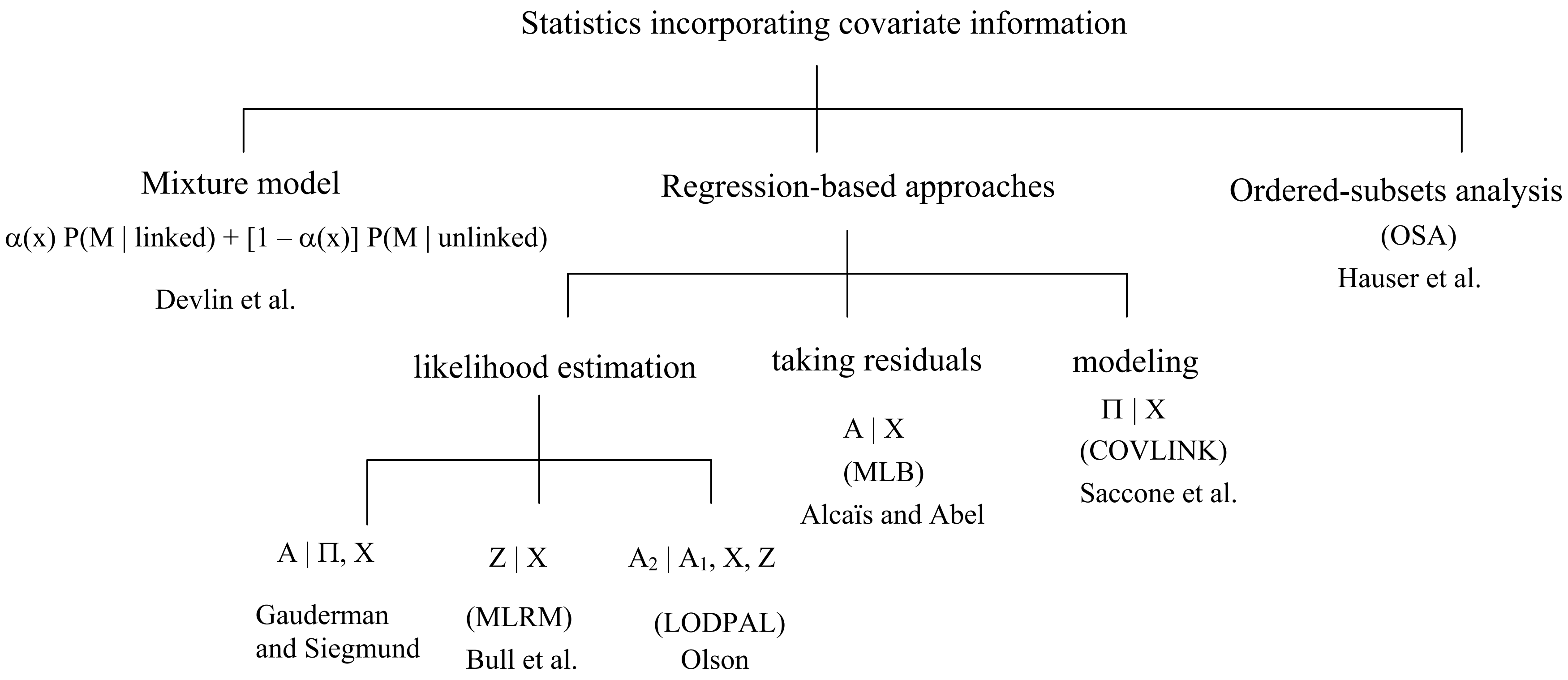
2.3.1. Background

We described several statistics for taking into account genetic heterogeneity in a parametric framework in Section 2.2. Based on nonparametric approaches, several methods have been suggested to incorporate covariate information into affected-sib-pair (ASP) or affected-relative-pair (ARP) analysis. An overview of the different covariate statistics is presented in Figure 2.1. Overall, there are three different approaches using covariate information in ASP or ARP analysis: mixture model (Devlin et al. 2002b), regression-based statistics (Greenwood and Bull 1997; Olson 1999; Rice et al. 1999; Gauderman and Siegmund 2001; Alcaïs and Abel 2001; Saccone et al., 2001) and ordered-subsets analysis (Hauser et al. 1998). The details of these methods will be described in the following sections.

2.3.2. Mixture models (Devlin et al. 2002b)

The idea of mixture models is that under genetic heterogeneity, the collection of family data is a mixture of two groups, where one group is linked to the suspected gene and the other is not. Devlin et al. (2002) suggested two mixture models, the "pre-clustering" model and the "Cov-IBD" model, for clustering linked and unlinked groups using covariate information. The pre-clustering model first uses covariate information to cluster families and then tests for excess IBD sharing independent of the covariate information. The Cov-IBD model jointly uses the covariate information and IBD sharing to simultaneously cluster linked and unlinked groups while maximizing the likelihood.

Devlin et al. (2002) evaluated the performance of their mixture models by generating simulated data that mimicked breast cancer families and their pedigree features. They also applied the



X: covariate; M: marker data; Z: inheritance vector; A: affection status;
A_1, A_2: affected sib pair; Π: probability of IBD sharing

Figure 2.1 Overview of covariate statistics

mixture model approach to an anorexia nervosa study. By incorporating several behavioral covariates into the model, they detected linkage signals at several regions on different chromosomes (Devlin et al., 2002a).

The mixture model approach may be summarized as follow. For each family, we observe marker data M_i, and family-level covariates X_j. We measure the family-level covariate information by taking the average of all family members, both affecteds and unaffecteds. The full likelihood for a sib-pair is specified as a mixture model:

$$\alpha(X_j)P(M_i \mid \text{linked}) + [1 - \alpha(X_j)]P(M_i \mid \text{unlinked}),$$

where $\alpha(X_j)$ is an estimate of the probability that the subject belongs to the cluster of interest; $P(M_i \mid \text{linked})$ is computed as a function of λ_s, which is the recurrence risk ratio for a sibling of an affected individual; $P(M_i \mid \text{unlinked})$ is computed assuming $\lambda_s = 1$.

In the pre-clustering model, $\alpha(X_j)$ is first estimated by using covariate information only, and so the probabilistic clusters are determined without using the IBD sharing information. Then the likelihood is maximized as a function of λ_s. An alternative approach is to maximize the likelihood jointly with regards to covariates ($\alpha(X_j)$) and IBD sharing patterns (λ_s). This is called the Cov-IBD model. The likelihood-ratio test is asymptotically distributed as a 50%:50% mixture of zero and one-sided χ^2 with one degree of freedom (df). In this study, we implement the pre-clustering model in our R code and analyze data with this approach.

2.3.3. General conditional-logistic model (Olson 1999): LODPAL

Olson (1999) proposed a general conditional-logistic model for ARP analysis. This conditional-logistic method has been extended to incorporate covariate effects.

The relevant part of the likelihood can be formulated as:

$$\sum_{Z}[P(A_2 \mid A_1, X, Z)]P(Z \mid M),$$

where A_1 and A_2 are affected sibs (sib1 and sib2), X is pair-level covariate information, M is marker data, Z is the IBD sharing vector (Z = 0, 1, or 2), $P(A_2 \mid A_1, X, Z)$ is a function of two sets of parameters β_j and δ_j: $P(A_2 \mid A_1, X, Z) = \sum_{j=0,1,2} e^{\beta_j + \delta_j X}$ with $\delta_0 = 0$, $\beta_j = \log_e \lambda_j$, λ_j is the relative risk of the relative pair sharing j IBD and δ_j are the estimated coefficient parameters of covariate X.

Instead of this original general conditional-logistic model, we evaluate the performance of their modified version as proposed in their prostate cancer study (Goddard et al., 2001). The modified version is implemented in the LODPAL program in S.A.G.E. package (Elston 2001). They applied the min-max restriction on the relative risk, λ, recommended by Whittemore and Tu (1998) to reduce the number of parameters from two to one. Previous work showed that a min-max one-parameter restriction approach is more robust than the traditional two-parameter methods for most genetic models (Whittemore and Tu 1998). However, for some covariate values, the likelihood estimate can be negative. In these cases, the likelihood estimate is set to an extremely small positive value to avoid computational difficulty. LODPAL does not constrain on the covariate information. It is difficult to decide whether one should constrain on the covariate information, because outliers may provide useful information. The question is which is the "right" regression line for the data set as a whole or the bulk of it [Olson, personal communication]. We herein implement this one-parameter restriction modified version in our R code.

Although several options are provided to measure pair-level covariate information in LODPAL, we follow the same procedures suggested by Olson (1999). We center the covariate values: x - $\bar{x}$, where x is the covariate values and $\bar{x}$ is the mean of x, then take the average pair-level covariate value into the model. The test statistic asymptotically follows a 50%:50% mixture of χ^2 distributions with P and P + 1 df where P is the number of covariates.

Olson (1999) applied this model to one simulated data set and to one Type 1 diabetes data set. This covariate-based linkage analysis was also used to identify potential chromosomal regions linked to prostate cancer (Goddard et al., 2001), to locate the second locus for a very-late-onset form of Alzheimer disease (Olson et al., 2002a), and to map possible regions linked to systemic lupus erythematosus (Olson et al., 2002b).

2.3.4. Multinomial logistic regression model (Bull et al. 2002): MLRM

Greenwood and Bull (1997) developed a method for incorporating covariate information into ASP linkage analysis based on a multinomial logistic regression approach, which estimates the proportion of expected IBD sharing conditional on covariates in affected sib pairs.

For each affected sib pair, the likelihood for the marker data M given pair-level covariates X is:

$$P(M \mid X) = \sum_{Z} P(M \mid Z) P(Z \mid X) = \sum_{Z} \frac{P(Z \mid M) P(M)}{P(Z)} P(Z \mid X),$$

where Z is the IBD sharing vector (Z = 0, 1, or 2), and $P(Z \mid X)$ is a function of the parameter vector β: $P(Z \mid X) = \dfrac{e^{\beta_j X}}{1 + e^{\beta_0 X} + e^{\beta_1 X}}$ for j IBD sharing = 0, 1, 2 and $\beta_2 = 0$.

Previous work showed that the likelihood-ratio test with IBD sharing constraints increases power (Holmans 1993) (details are provided in Appendix A). However, they pointed out that IBD sharing patterns can fall outside the possible triangle designated by Holmans (1993) under certain gene-environment interactions. Similarly, Guo (2000) also indicated that the genetic triangle constraints may no longer hold when there is an environmental effect and/or G x E interaction.

In order to improve power in the covariate model, they proposed three alternative restricted models. (1) Average constraints: this method proposes that the expected value of the constrained allele-sharing estimates has to fall within the plausible genetic region, where expectation is taken over the covariate distribution of affected sib pairs' population. (2) Subgroup-triangle constraints: if the covariate is categorical, one can apply the constraints in the usual possible triangle to each subgroup defined by the covariate, then sum the LOD scores across all subgroups. (3) Simultaneous-boundary constraints: for each value of the covariate, this approach can be thought of as constraining the allele sharing to one of the boundaries ($z_1 = .5$, $z_1 = 2*z_0$, and $z_1 = .355 + .58*z_0$) (Greenwood and Bull 1999).

Under the assumption of no linkage, 2ln(10)*LOD score is asymptotically distributed as a χ^2 with 2P + 2 degrees of freedom where P is the number of covariates. For the first two constrained models, they suggested the use of Monte Carlo *p*-values for the significance of tests. For the simultaneous-boundary constraints model, the test statistic has an asymptotic χ^2 distribution with (P +1) degrees of freedom, instead of (2P + 2) degrees of freedom.

Greenwood and Bull (1997) applied their method to a bipolar affective disorder study, but failed to find significant heterogeneity and didn't observe the linkage signal previously found on chromosome 18 (Greenwood and Bull 1997). In a study of Canadian families with inflammatory bowel disease (Rioux et al. 2000), previous work had suggested a susceptibility locus on chromosome 5. Bull et al. (2002) applied their method to a subset of 167 pedigrees from this study and found that the linkage signal increases after incorporating covariate information. Their results showed heterogeneous effects of diagnostic subtypes and age of onset (Bull et al. 2002).

Greenwood and Bull (1999) reported that: (1) in most situations, the simultaneous-boundary constraints approach under no dominance assumption ($z_1 = .5$) has the best power, compared to the unconstrained model and the other constrained models suggested in their work; (2) but this approach under the no additive assumption ($z_1 = 2*z_0$) tends to have very low power, since the probability of sharing one allele IBD was estimated to be 0.497, which is very close to the null ($z_1 = 0.5$), in the unconstrained model with no covariates; (3) the min-max restriction approach has power between those of the two other simultaneous-boundary constraints approaches.

We here take the mean of ASP as the measure of pair-level covariate information, and analyze data using our R code, which implements the simultaneous-boundary constraints approach under three different assumptions: no dominance variance (corresponding to $z_1 = .5$), no additive variance (corresponding to $z_1 = 2*z_0$), and use of the min-max restriction (equal to $z_1 = .355 + .58*z_0$).

2.3.5. Pearson's logistic regression residuals (Alcaïs and Abel 2001): MLB

Alcaïs and Abel (2001) used a logistic regression approach to account for phenotypic and covariate effects. In contrast to the logistic regression approaches applied in LODPAL, MLRM and COVLINK, they regressed out the covariate effects, and then computed Pearson's residuals. The residuals were treated as a quantitative phenotype and analyzed using an extension of the Maximum-Likelihood-Binomial (MLB) linkage approach (Abel et al. 1998; Alcaïs and Abel 1999). The MLB is based on the binomial distribution of the numbers of affected sibs carrying a given parental allele. Under the null hypothesis of no linkage, each affected sib should have a 50% chance of receiving allele A from a parent who has an AB genotype. The test for linkage is whether the probability is higher than 50% in affected sibs. The test statistic is asymptotically distributed as a 50%:50% mixture of zero and one-sided χ^2 distribution with one df. The Maximum-Likelihood-Binomial approach is implemented in the MLB program.

Alcaïs and Abel (2001) accounted for the familial correlation by applying the generalized estimating equations (GEE) approach (Liang and Zeger 1986). The Pearson's residuals *R* here are obtained by regressing out the effect of individual-level covariate *X* on the family members (both affected and unaffected), and taking into account familial correlation. Then the likelihood for the marker data M_i given *R* is computed by introducing a latent binary variable *T* for the sibship that models the linkage information between the quantitative trait and the markers. This allows one to write:

$$P(M \mid R) = \sum_T P(M \mid T, R) P(T \mid R) = \sum_T P(M \mid T) P(T \mid R),$$

where $P(M \mid T)$ is a function of parameter γ, the probability that a sib with $T_i = 1$ received an A allele from an AB parent and a sib with $T_i = 0$ received a B allele from an AB parent. The AB genotypes here represent assumed genotypes at the putative trait.

It is important to elucidate that either the generalized linear model (GLM) (Wedderburn 1974) or GEE would provide the correct Pearson's residuals, even though GEE allows us to deal with familial correlation. The only difference between GEE and GLM is that variances of the estimated coefficients based on GEE are smaller than those based on GLM when familial correlation occurs.

To test the MLB extension, Alcaïs and Abel (2001) conducted a simulation study, which contained a genetic factor, one binary genotype-dependent covariate and a quantitative covariate. The results showed that compared to the MLB method without accounting for covariate information, power increases when the allele frequency is rare, no matter whether the trait is dominant or recessive, and when unaffected siblings are included in the analysis.

2.3.6. Ordered-subsets analysis (Hauser et al. 1998): OSA

Ordered-subsets analysis (OSA) carries out stratified linkage analysis using family-level covariate information. First, the ASM module of GeneHunter-Plus (Kong and Cox 1997) is used to obtain the LOD scores of all families before adding the covariate information. Then the OSA methods are implemented as follows: calculate the covariate mean value for each family, using the covariate values of affected sibs only. Rank the pedigrees based on their family-specific mean values. Starting with the family having the highest mean value, add one family at a time, in rank order. After including each family, re-estimate the single gene effect parameter, δ, and compute the maximum LOD score (MLS) for each subset. Finally report the best MLS over all ordered subsets of the families. This is the high-to-low (H → L) OSA statistic. Repeat the same procedure by starting with the family having the lowest family-specific mean value to perform the low-to-high (L → H) OSA statistic. We also run OSA using the optimal slice option. This option first defines the subset of adjacent families from the covariate distribution that maximizes the LOD score. Then repeat the same procedures starting with consecutively higher ranks (e.g., then starting with the second rank, the third rank, etc.).

Significance can be obtained by performing a Monte Carlo permutation test for the subset with the MLS. However, one should be careful that the p-value reported in the OSA program measures whether use of the covariate information significantly increases the lod score, not the p-value for the likelihood-ratio test statistic. Hence, for the OSA methods, we do not use the p-value reported in OSA program. Instead, we obtain the empirical p-values (at 1%, 5% and 10% levels) estimated based on our 16,000 replicates generated under the null hypothesis of no linkage.

The OSA approach has been applied to the Finland-United States Investigation of Non-Insulin-Dependent Diabetes Mellitus Genetics (FUSION) study (Ghosh et al. 2000) and an autism study (Shao et al., 2003). Based on OSA analyses, both studies found some suggestive linkage regions related to the complex traits of interest.

2.3.7. Logistic regression for predicting the IBD sharing probability (Rice et al. 1999): COVLINK

Rice et al. (1999) introduced a logistic regression approach that includes covariates as independent variables. They predict the probability of IBD sharing in sib pairs using covariate information as predictors in the logistic regression model. Then they test whether or not the probability of IBD sharing is significantly more than 50% and whether or not there are significant covariate effects. They tested their method using simulated sibling data from the Genetics Analysis Workshop 11 (GAW 11). Later, they extended this logistic regression approach to cousin pairs, and applied it to data from the GAW 12 (Rice et al. 1999, Saccone et al., 2001).

They use logistic regression to model IBD sharing as a function of covariates. The logistic model of the IBD sharing probability Π regressed on pair-level covariates X_i can be written as:

$$\mathrm{Log}_e\left(\frac{\Pi}{1-\Pi}\right) = \beta_0 + \beta_\iota X_i,$$

where X_i is the ith covariate in a vector of covariates and β_ι is the ith coefficient in a vector of the corresponding estimated coefficients. Their test statistic follows a two-sided χ^2 distribution with one df. They implemented this method in the COVLINK program.

We measure the pair-level covariate information by taking the mean of sib pairs (both concordant and discordant pairs). We implement their method in our R code. Although they use all sib pairs' information (affected-affected, affected-unaffected and unaffected-unaffected), the major limitation of this method is that it only uses relative pairs who have unambiguously determined IBD sharing probabilities.

2.3.8. Linear regression model approach (Gauderman and Siegmund 2001)

Gauderman and Siegmund (2001) presented theoretical work in taking into account environmental factors. Their approach, the "mean-interaction" test, was based on the framework of the mean test statistic (defined in Appendix A). They first set up three factors: genetic relative risk in unexposed individuals (R_g), exposure relative risk in noncarriers (R_e) and G x E interaction relative risk (R_{ge}). G x E interaction models were specified by varying these factors. Affected sib pairs were divided into three subgroups: both exposed (EE), only one exposed (EU), and both unexposed (UU). Then they computed the expected proportion of IBD sharing in three subgroups under different G x E interaction models.

In the conventional mean test for linkage, one uses a z-statistic: $z = \sqrt{N}(\overline{\pi} - 0.5)/\sigma$ to test the null hypothesis that $\pi = 0.5$, where $\overline{\pi}$ is the mean of π_i, $i = 1, 2, \ldots,$ N sib pair and σ is the standard deviation. Equivalently, one can fit the degenerate regression model: $\pi_i = \pi = \varepsilon_i$, where ε_i is assumed to be independent and have normal distribution with mean zero and variance σ^2. For the mean test based on this regression model, the likelihood ratio test (T_π) can be formed and has a 50:50 mixture of two χ^2 distributions with zero and one df. They extended the mean test by adding environmental factors simultaneously into the linear regression model for IBD sharing estimation. If we denote covariates as X_i, the regression model then is written as:

$$\pi_i = \pi + \beta * X_i + \varepsilon_i.$$

Based on this extended regression model, they proposed a likelihood ratio test, the "mean-interaction" test. The likelihood of their approach can be written as $P(\text{A} \mid \Pi, X)$, where A is affection status, Π is IBD sharing and X is pair-level covariate. And the test can be defined as:

$$\mathrm{T}_{\pi\beta} = -2\{\ln(L[\pi = 0.5, \beta = 0]) - \ln(L[\hat{\pi}, \hat{\beta}])\}.$$

The test statistic has a 50:50 mixture of two χ^2 distributions with P and P+1 df (P is the number of covariates). They compared the power of the "mean-interaction" test with the mean test and concluded that incorporating G x E interaction into the linkage test can increase power. We do not investigate the performance of the "mean-interaction" test here.

2.3.9. Analytical distribution of covariate statistics

Most papers about covariate statistics describe the theoretical asymptotic distribution and degrees of freedom, except for OSA. However, whether the analytical distributions of the covariate statistics are accurate still requires investigation. Especially when covariates are included in the analysis, the theoretical asymptotic distribution may not hold (Olson 2002c, d; Devlin et al. 2002b, c). The most straightforward way to address this issue is to calculate type I error rates based on empirical distributions. We will discuss these issues in Chapter 5.

2.4. SUMMARY

2.4.1. Comparison of the methods

One of the specific aims is to examine whether or not the covariate statistics increase power to detect linkage. Does power increase by adding covariates into the statistical model? Or does it lead to a loss of power because of increased degrees of freedom? Therefore, it is not only important to evaluate the performance of the covariate statistics, but also important to compare their power to that of conventional methods that ignore covariate information. Since conventional model-free methods ignore covariate information, we use the power of the model-free methods (allele-sharing score statistics) as a baseline for our comparisons.

In addition, we can treat our covariates as quantitative traits, testing for linkage of the covariate itself. For this purpose, we use two approaches: variance-component linkage analysis and regression-based quantitative-trait linkage analysis, computing their power.

Allele-sharing statistics: allele-sharing statistics are widely used in model-free approaches which measure the excess of IBD sharing in affected relatives (Whittemore and Halpern 1994; Kruglyak et al 1996; Kong and Cox 1997). Two score functions, S_{pairs} and S_{all}, are commonly applied to capture the IBD sharing probabilities in affected pairs. These two score functions weight the affected individuals in the pedigrees differently. The details of these two score functions and allele-sharing statistics are provided in Appendix A.

The likelihood-ratio test based on these score functions is asymptotically distributed as a 50%:50% mixture of zero and χ^2 with one df. We analyze our simulated data using allele-sharing statistics (ignoring covariate information) and compute S_{pairs} and S_{all} by Merlin (Abecasis et al. 2002). Previous work showed that S_{all} performs consistently well over various genetic models (Sengul et al. 2001). We therefore use S_{all} as the baseline to compare with the covariate methods.

Variance-component linkage analysis: variance-component linkage analysis (VC) estimates the covariance between relatives as a function of the IBD sharing at a QTL (Fulker and Cherny 1996; Almasy and Blangero 1998). For instance, in the simple additive model, the covariance matrix for a family can be written as:

$$\Omega = \sum_{i=1}^{n} \hat{\Pi}_i \sigma_{ai}^2 + 2\Phi \sigma_g^2 + \mathrm{I}\sigma_e^2$$

where $\hat{\Pi}_i$ is the matrix of the estimated proportions of genes that the relative pairs share IBD at the *i*th QTL, Φ is the kinship matrix, I is an identity matrix, σ_{ai}^2 is the additive genetic variance due to the *i*th QTL, σ_g^2 is the overall genetic variance and σ_e^2 is the variance of the random effect. We then test the null hypothesis: the additive genetic variance, σ_{ai}^2, is equal to zero vs. the alternative hypothesis: $\sigma_{ai}^2 > 0$. The test statistic is asymptotically distributed as a 50%:50% mixture of zero and χ^2 with one df. We analyze our data using the Merlin program, which implements the VC method. We input the original covariate values without standardization as the trait values in VC analysis.

Regression-based quantitative–trait linkage analysis: Sham et al. (2002) developed a method that regresses the IBD sharing between relative pairs on the squared sums and squared differences of the relative pairs' trait values (Sham et al 2002). The weighted-least-squares estimators of the regression coefficients can be written as a function of three covariance matrices: (1) the covariance matrices of squared sums and squared differences; (2) the covariance matrix of estimated IBD sharing probabilities; and (3) the covariance matrix between the estimated IBD sharing and the squared sums and squared differences. The elements in the covariance matrix (3) are proportional to the additive variance due to a QTL. A test for linkage is to test whether the additive variance is equal to zero. The test statistic is distributed as a 50%:50% mixture of zero and χ^2 with one df. We also analyze data using this approach implemented in the Merlin program. We use the default settings in Merlin: mean zero, variance 1.5 and heritability 0.8, even though these may differ from the values in the simulated models.

2.4.2. Abbreviations

For simplicity, we name the ‘multinomial logistic regression model’ approaches (Bull et al. 2002): under no dominance assumption as ‘no-dominance MLRM’, under no additive assumption as ‘no-additive MLRM’ and under the min-max restriction as ‘min-max MLRM’, the ‘general conditional-logistic model’ (Olson 1999) as ‘LODPAL’, the ‘Pearson’s logistic regression residuals extension of the MLB approach’ (Alcaïs and Abel 2001) as ‘MLB’ and the ‘logistic regression for predicting the IBD sharing probability’ (Rice et al. 1999) as ‘COVLINK’. We will use these abbreviated forms in the following chapters (Table 5.1).

Table 2.1 Summary of covariate statistics

Statistic	Definition	Reference
Mixture model	$\alpha(X_j)P(M_i \mid \text{linked}) + [1 - \alpha(X_j)]P(M_i \mid \text{unlinked})$, $\alpha(X_j)$: an estimator of the probability that the subject is belonged to the cluster of interest; $P(M_i \mid \text{linked})$: a function of λ_s, which is the recurrence risk ratio for a sibling of an affected individual; $P(M_i \mid \text{unlinked})$ is computed assuming $\lambda_s = 1$.	Devlin et al. 2002
LODPAL	$\sum_Z [P(A_2 \mid A_1, X, Z)]P(Z \mid M)$, Z: IBD sharing vector ($Z = 0, 1, \text{or } 2$); $P(A_2 \mid A_1, X, Z)$: a function of two sets of parameters β_j and δ_j: $P(A_2 \mid A_1, X, Z) = \sum_{j=0,1,2} e^{\beta_j + \delta_j X}$ with $\delta_0 = 0$.	Olson 1999
MLRM	$P(M \mid X) = \sum_Z P(M \mid Z)P(Z \mid X) = \sum_Z \frac{P(Z \mid M)P(M)}{P(Z)} P(Z \mid X)$, Z: IBD sharing vector ($Z = 0, 1, \text{or } 2$); $P(Z \mid X)$: a function of the parameter vector β: $P(Z \mid X) = \frac{e^{\beta_j X}}{1 + e^{\beta_0 X} + e^{\beta_1 X}}$ for $j = 0, 1, 2$ and $\beta_2 = 0$.	Bull et al. 2002
MLB	$P(M \mid R) = \sum_T P(M \mid T, R)P(T \mid R) = \sum_T P(M \mid T)P(T \mid R)$, R is the Pearson's residuals obtained by regressing out the effect of covariate X; $\mathrm{P}(M \mid T)$: a function of parameter γ -- the probability that a sib with $T_i = 1$ received an A allele from an AB parent and a sib with $T_i = 0$ received a B allele from an AB parent. The AB genotypes here represent assumed genotypes at the putative trait location.	Alcaïs and Abel 2001
OSA	$P(M, A)$ by stratified on family-specific mean values of X	Hauser et al. 1998
COVLINK	$\mathrm{Log}_e(\frac{\Pi}{1 - \Pi}) = \beta_0 + \beta_\iota X_i$.	Rice et al. 1999

Notation: "A" denotes an affected sib pair A_1, A_2, "M": marker data, "X": covariates, "Z": inheritance vector, "Π": IBD sharing probability = $P(Z \mid M)$, α: proportion of linked families.

3. DATA FEATURES AND SIMULATION

3.1. DATA FEATURES

3.1.1. Overview

We generate simulated data sets with varied sibship sizes and with different underlying disease models. As described in Section 1.3.4 in Chapter 1, there are several components influencing the underlying liability model: one genetic factor G, one covariate C1 with a G x E interaction effect, one independent environmental covariate C2, a polygenic effect PG, and a random error E. Meanwhile, we also generate one random noise covariate C3, which is not part of the liability. The relative effect of each component is modified through the variance proportions of each component, which are described in Tables 3.1A, B, C. Based on the empirical liability, we set a threshold for defining the affection status, that gives the desired prevalence. We only ascertain pedigrees with two or more affected sibs.

3.1.2. Pedigree structure

We simulate nuclear families with affected and unaffected children. The simulated data sets contain a mixture of various sibship sizes, with sibship size varying from 2 to 5. The sibship size distribution follows a truncated negative binomial distribution. In terms of ascertainment, we only include the pedigrees containing 2 or more affected sibs. After conducting several pilot analyses, we decided to simulate 100 pedigrees with 2 or more affected sibs in each replicate.

3.1.3. Disease models

We consider single-locus two-allele dominant and recessive models with different disease allele frequencies. Data sets generated under different G x E interaction models and different disease models have varied penetrances and phenocopy rates. According to the different G x E interaction models, the disease gene has either direct or indirect effect on the disease outcome (Figures 1.5, 7, 8). We consider four different disease allele frequencies for both dominant and recessive models. The disease allele frequencies used for the different dominant models are 0.01, 0.02, 0.05, and 0.1. The disease allele frequencies used for the different recessive models are 0.1, 0.2, 0.3, and 0.4.

3.1.4. Liability and affection status

To define affection status, we use a liability threshold model described in Section 1.3.4 in Chapter 1 where a quantitative liability value is computed for each individual. We generate 10,000 families to obtain an empirical liability distribution for different G x E interaction and disease models. A threshold is then determined so that 5% of the simulated liability values fall above this threshold, giving a 5% prevalence for the disease. Each individual's affection status then depends on his/her underlying quantitative disease liability value. When the subject’s liability value falls within the top 5%, the subject is affected. The individual whose liability value is below the top 5% threshold is unaffected. The contribution of each component is defined on the basis of its variance proportions.

3.1.5. Marker data

We simulate two sets of marker data evenly spaced every 5 cM along one autosomal chromosome with no missing data. The ‘unlinked’ set contains marker data without any disease locus residing on the chromosome. We then use the unlinked set to estimate empirical false positive rates. The other set, the ‘linked’ set, contains a disease locus in addition to the markers. The linked data set allows us to evaluate power.

The length of the simulated chromosome is 169 cM, which mimics the length of chromosome 10 as taken from the recent deCode map (Kong et al. 2002). Each simulated data set contains 33 markers. The disease locus resides at a position 50 cM along the chromosome. The unlinked dataset doesn’t contain any disease gene. The heterozygosity of all markers is set to 0.8 and the number of alleles at each marker is five. All markers have equally frequent alleles. We will describe how we simulate the maker data in Section 3.2.

3.1.6. Covariate data

Types of covariates: there are various types of covariates such as discrete or continuous variables that one can incorporate into G x E interaction models. For instance, gender is a dichotomous variable, age can be either an ordinal categorical variable or a continuous variable, and blood glucose concentration is a continuous variable. However, not all of the covariate methods can take into account discrete covariates. For example, discrete covariates are not suitable for the OSA statistics and the MLB approach. All the covariate statistics we consider here can handle a continuous covariate. Therefore, we only simulate continuous covariates.

Number of covariates: conceptually, there is no limitation to how many covariates one could include in the model. However, we usually do not know exactly which exposures are associated with disease outcome. Generally, the more covariates in the model, the more complicated the underlying disease liability. It also makes the results more difficult to interpret. Furthermore, when we incorporate more covariates, we also introduce more degrees of freedom. Since the major aim of our study is to evaluate the performance of the covariate methods, we start from the simplest scenario first. Even though we generate three different covariates (C1, C2, C3), when we analyze the simulated data, we only take into account one covariate at a time, instead of simultaneously considering the effects of all three covariates.

3.2. SIMULATIONS

3.2.1. Overview

We simulate data under the three types of G x E interaction models that we described in Chapter 1. For G x E interaction model, we generate data under either dominant or recessive disease models with 4 different disease allele frequencies, separately.

Next, we analyze each simulated replicate using covariate statistics, model-free methods, and QTL approaches. We first use the unlinked marker data and the covariate C2 (the covariate without a G x E interaction effect) and the covariate C3 (the random noise covariate) to obtain the empirical null thresholds (details will be described in Chapter 5). We then analyze the linked marker data and various covariates to determine power of each statistic (details will be described in Chapter 6).

3.2.2. Data simulation of three G x E interaction models

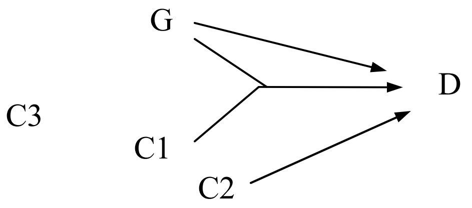
General simulation schema: for all G x E interaction models, we first simulate the parents' disease locus genotypes using the distribution defined by the disease allele frequencies assuming Hardy-Weinburg equilibrium and linkage equilibrium. The children's disease locus genotypes are generated conditional on the parents' genotypes. The covariates C1 and C2 for parents and offspring are simulated from two multivariate normal distributions. The parents and offspring's polygenic effect are generated from a normal distribution N(0, σ_{PG}^2) and N($\frac{P_1+P_2}{2}$, $\frac{\sigma_{PG}^2}{2}$), respectively, where P_1 and P_2 are the polygenic values for the parents. The random error of parents and offspring is obtained from a normal distribution N(0, σ_E^2).

Covariate C3 is simulated from one of three different normal distributions with means (114.46, 0, -64.59) and the same variance (35.75). For each individual, we first pick one normal distribution randomly, and then simulate covariate C3 from the selected normal distribution with the corresponding mean and variance. These three means correspond to three genotype distributions of a bi-allelic locus with allele frequencies 0.24 and 0.76 under Hardy-Weinburg equilibrium. This locus is independent of the disease liability.

The difference in the three types of G x E interaction models is how we use the covariate C1 information to define the liability model, and we transform the covariate C1 values in the Type III model. We will describe the details in the following sections. The parents and offspring's affection status are assigned according to the threshold obtained from the empirical liability distribution, which gives a 5% disease prevalence. We then ascertain families with 2 or more affected sibs. For computational efficiency, we only generate maker data for the ascertained families.

Type I G x E interaction model: the simulation procedures of the Type I model were described above. The simulation schema is shown in Figure 3.1. The liability model is a function of one disease locus G, gene and covariate interaction (G x C1), covariate C2, PG, and E. We assign individual's affection status and ascertain families based on the threshold computed from this liability distribution.

In the Type I model, the disease liability is determined by G, the covariates C1 and C2, PG, and E (Figure 1.5). Covariate C1 is generated from a multivariate normal distribution with mean zero and correlation 0.8; likewise, C2 is generated from a multivariate normal distribution with mean zero and correlation zero. The mean values of PG, and E are set to zero. The variances of C1, C2, PG, and E are assigned so as to generate the desired variance proportions. For both the dominant and recessive models, the variance proportions are as given in Table 3.1A.



LI: liability
G: genetic effect
C1: covariate with G x E interaction
C2: environmental covariate
C3: random noise covariate

$$LI = \mu + G + \beta_G * C1 + C2 + PG + E$$

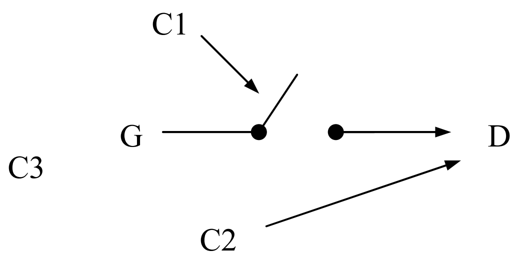
Figure 1.5. Type I model

Table 3.1A Variance proportion of each component in the liability under the Type I model

Disease model	Disease allele frequencies	Proportion of variance due to each component				
		Disease gene	G x E interaction	Environmental factor	Polygenetic effect	Random effect
Dm1 – Dm4	0.01 – 0.1	10 %	10 %	20 %	40 %	20 %
Rm1 – Rm4	0.1 – 0.4	10 %	10 %	20 %	40 %	20 %

Note: LI = μ + G + G * C1 + C2 + PG + E

Type II G x E interaction model: most simulation procedures were described in Section 3.2.2.1. Figure 3.2 presents the simulation schema. In the Type II model, when C1 exceeds the threshold value (e.g. C1 ≥ 0), the liability is computed as a function of G, C2, PG and E. Otherwise, G does not enter in and the liability is a function of covariate C2, PG and E (Figure 1.7). The definition of affection status, and pedigrees' ascertainment are same as in the Type I model.



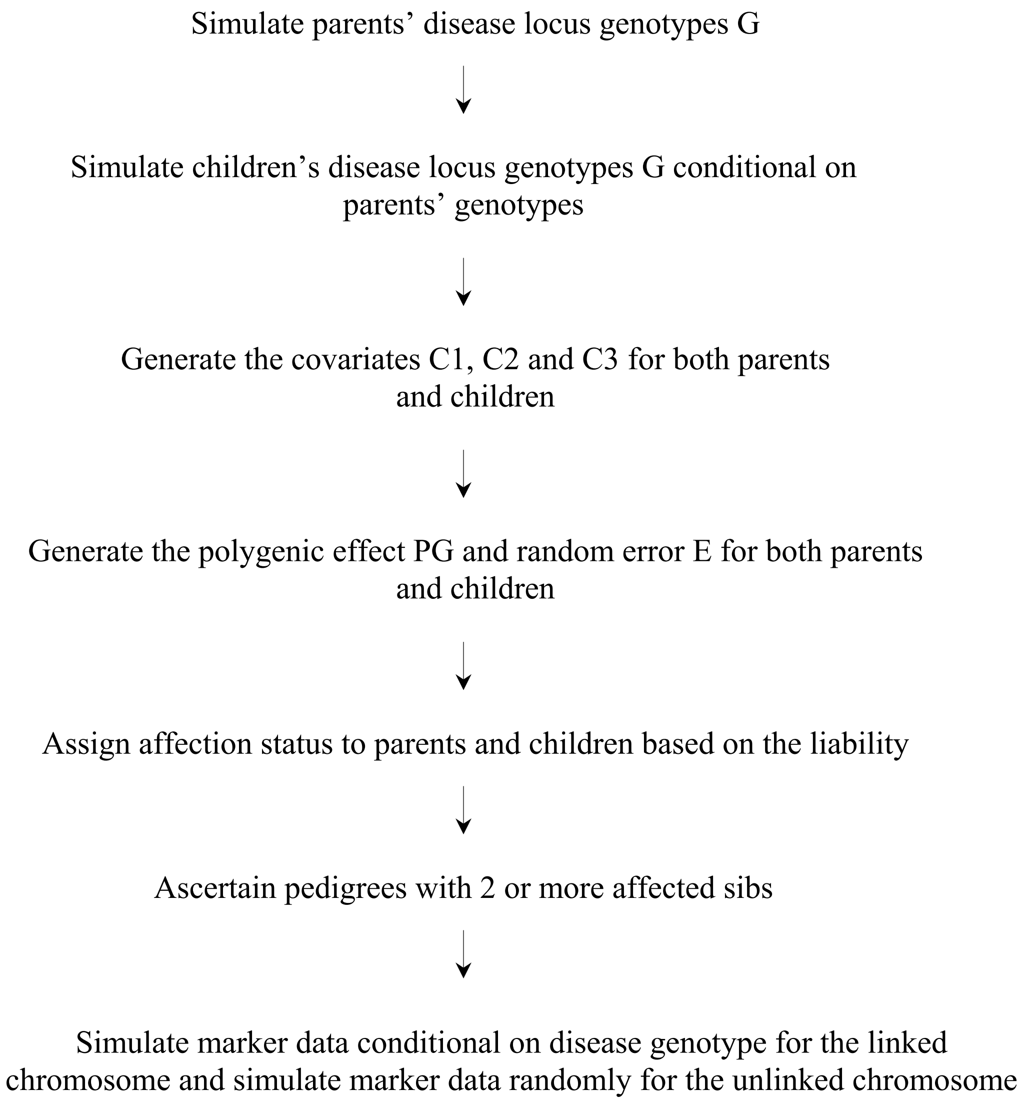
LI: liability
G: genetic effect
C1: covariate with G x E interaction
C2: environmental covariate
C3: random noise covariate

If C1 ≥ 0, $LI = \mu + G + C2 + PG + E$
If C1 < 0, $LI = \mu + C2 + PG + E$

Figure 1.7. Type II model

The liability is constructed from one of two different underlying models, depending on the covariate C1 value, as described in Section 3.2.2.3.1. Covariate C1 is generated from a multivariate normal distribution with mean zero, variance 1, and correlation 0.6; similarly, covariate C2 is generated from a multivariate with mean zero, and correlation 0.6. The mean values of PG, and E are equal to zero. As in the Type I model, variances of C2, PG, and E are assigned so as to generate the desired variance proportions.

In order to differentiate the power of each method, we adjust the genetic effect by using two genetic variances (GV = 20% or 30%). The details of the variance proportion of each component are as given in Table 3.1B.



Simulate parents’ disease locus genotypes G

Simulate children’s disease locus genotypes G conditional on parents’ genotypes

Generate the covariates C1, C2 and C3 for both parents and children

Generate the polygenic effect PG and random error E for both parents and children

Assign affection status to parents and children based on the liability

Ascertain pedigrees with 2 or more affected sibs

Simulate marker data conditional on disease genotype for the linked chromosome and simulate marker data randomly for the unlinked chromosome

Figure 3.1 Data generation schema under the Type I model

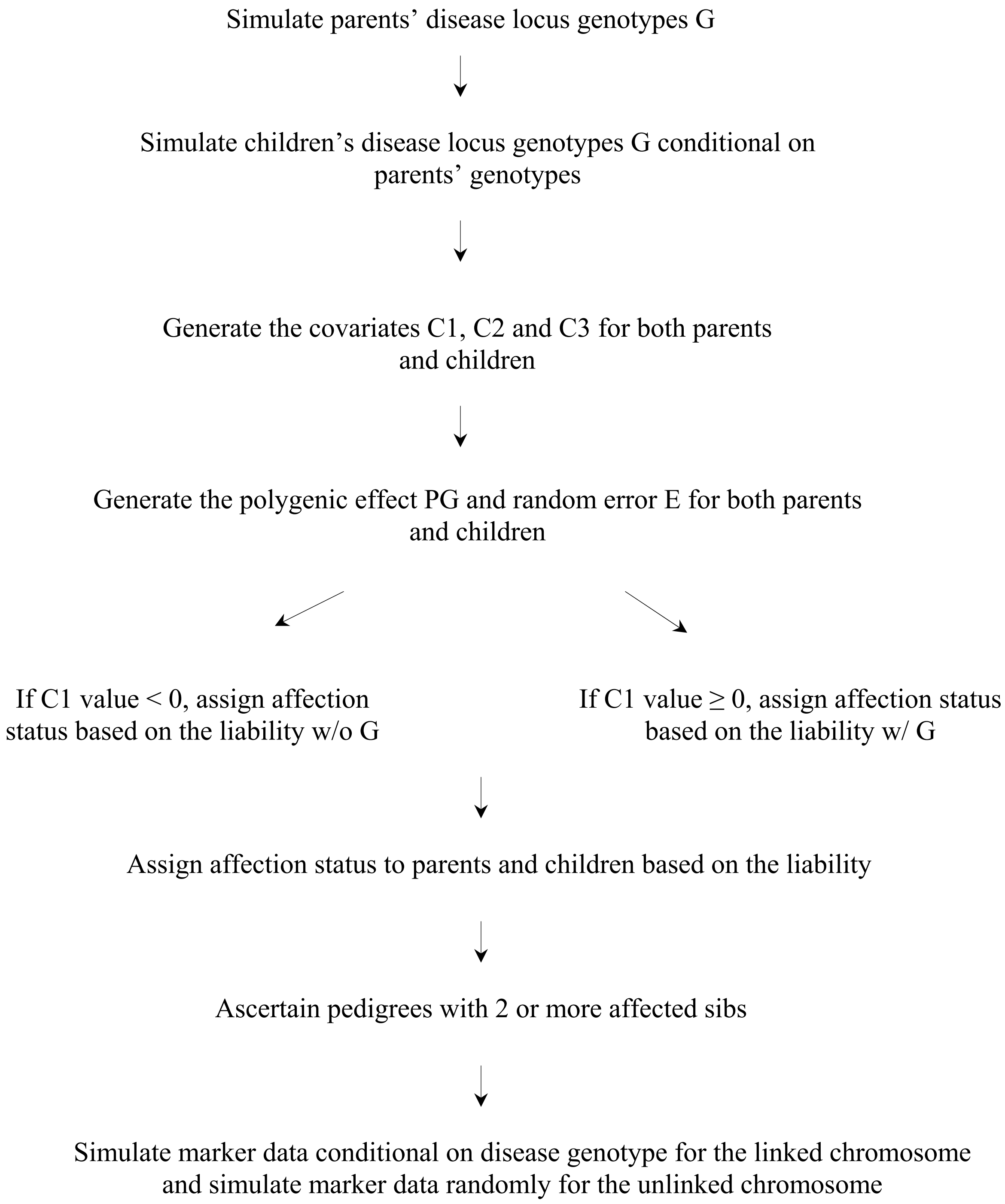
Table 3.1B Variance proportion of each component in the liability under the Type II model

Disease model	Disease allele frequencies	Proportion of variance due to each component			
		Disease gene	Environmental factor	Polygenetic effect	Random effect
20 % genetic variance					
C1 $\geq$ 0					
Dm1 – Dm4	0.01 – 0.1	20 %	20 %	40 %	20 %
Rm1 – Rm4	0.1 – 0.4	20 %	20 %	40 %	20 %
C1 < 0					
Dm1 – Dm4	0.01 – 0.1	--	25 %	50 %	25 %
Rm1 – Rm4	0.1 – 0.4	--	25 %	50 %	25 %
30 % genetic variance					
C1 $\geq$ 0					
Dm1 – Dm4	0.01 – 0.1	30 %	20 %	40 %	10 %
Rm1 – Rm4	0.1 – 0.4	30 %	20 %	40 %	10 %
C1 < 0					
Dm1 – Dm4	0.01 – 0.1	--	29 %	57 %	14 %
Rm1 – Rm4	0.1 – 0.4	--	29 %	57 %	14 %

Note: C1 $\geq$ 0, LI = μ + G + C2 + PG + E; C1 < 0, LI = μ + C2 + PG + E

Type III G x E interaction model: simulation procedures were similar to those used for the other two models (Figure 3.3). We first simulate the disease genotype and covariate C1 for each subject. For individuals with the high-risk genotype, covariate C1 value is transformed using a sigmoidal curve: f(x) = 30 / (5 + 7.5 * exp(-3.5 * x)) (Figure 3.4). For subjects with the low-risk genotype, covariate C1 value is re-generated from a normal distribution N(0, 0.01) (Figure 1.8). The definition of affection status, and ascertainment were described in Section 3.2.2.1.

The components in the liability are C1, C2, PG, and E (Figure 1.8). The susceptibility gene affects the liability by altering the covariate C1 value in a disease genotype-dependent manner. C1 is first generated from a multivariate normal distribution with mean 2, standard deviation 1, and correlation 0.6. Then the C1 value is modified according to individual's disease genotype as



Simulate parents' disease locus genotypes G

Simulate children's disease locus genotypes G conditional on parents' genotypes

Generate the covariates C1, C2 and C3 for both parents and children

Generate the polygenic effect PG and random error E for both parents and children

If C1 value < 0, assign affection status based on the liability w/o G

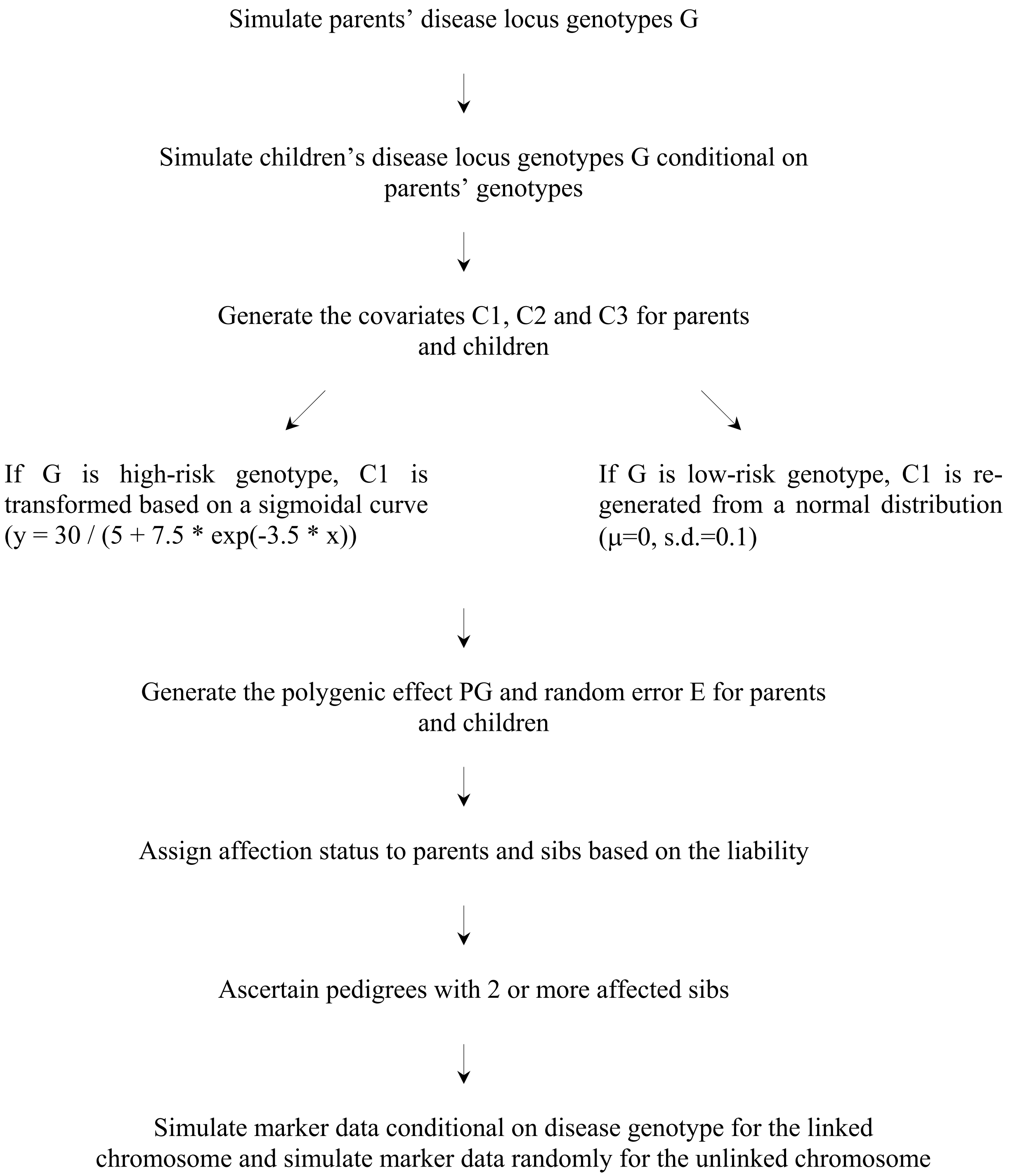
If C1 value ≥ 0, assign affection status based on the liability w/ G

Assign affection status to parents and children based on the liability

Ascertain pedigrees with 2 or more affected sibs

Simulate marker data conditional on disease genotype for the linked chromosome and simulate marker data randomly for the unlinked chromosome

Figure 3.2 Data generation schema under the Type II model



Simulate parents’ disease locus genotypes G

Simulate children’s disease locus genotypes G conditional on parents’ genotypes

Generate the covariates C1, C2 and C3 for parents and children

If G is high-risk genotype, C1 is transformed based on a sigmoidal curve (y = 30 / (5 + 7.5 * exp(-3.5 * x))

If G is low-risk genotype, C1 is re-generated from a normal distribution (μ=0, s.d.=0.1)

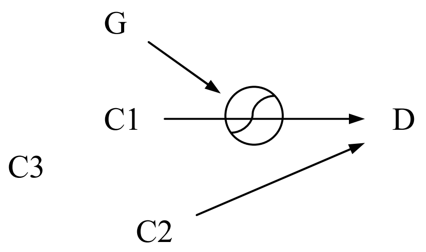
Generate the polygenic effect PG and random error E for parents and children

Assign affection status to parents and sibs based on the liability

Ascertain pedigrees with 2 or more affected sibs

Simulate marker data conditional on disease genotype for the linked chromosome and simulate marker data randomly for the unlinked chromosome

Figure 3.3 Data generation schema under the Type III model



LI: liability
G: genetic effect
C1: covariate with G x E interaction
C2: environmental covariate
C3: random noise covariate

$$LI = \mu + C1 + C2 + PG + E$$

Figure 1.8. Type III model

described in Section 3.2.2.4.1. C2 is generated from a multivariate normal distribution with mean zero and correlation 0.6. Mean values of PG, and E are equal to zero. As in the Type I and II models, we assign different variances for C2, PG, and E so as to generate the matching variance proportions. The corresponding variance proportions are as given in Table 3.1C.

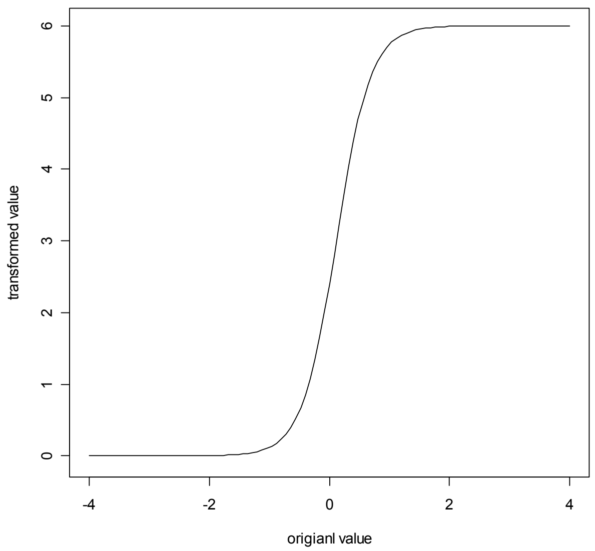


Figure 3.4 Sigmoidal curve for transforming the covariate values for individuals with a high-risk genotype

Table 3.1C Variance proportion of each component in the liability under the Type III model

Disease model	Disease allele frequencies	Proportion of variance due to each component			
		Genotype-dependent covariate	Environmental factor	Polygenetic effect	Random effect
G = high-risk genotype (G_{HR})					
Dm1 – Rm4	0.01 – 0.1	7.4 %	20.6 %	41.1 %	30.9 %
Rm1 – Rm4	0.1 – 0.4	7.4 %	20.6 %	41.1 %	30.9 %
G = low-risk genotype (G_{LR})					
Dm1 – Dm4	0.01 – 0.1	0.11 %	22.2 %	44.39 %	33.3 %
Rm1 – Rm4	0.1 – 0.4	0.11 %	22.2 %	44.39 %	33.3 %

Note:

1. $LI = \mu + C1 + C2 + PG + E$
2. C1 is generated under a multivariate normal distribution with $\mu = 2$, sd = 1.
   If G is G_{HR}, C1 value is transformed to follow one sigmoidal curve:
   $(y = 30 / (5 + 7.5 * \exp(-3.5 * x)))$.
   If G is G_{LR}, C1 value is re-generated from a normal distribution ($\mu = 0$, sd = 0.1).

3.2.3. Data generation

Disease genotypes and covariate data: as described in Section 3.1.3, we consider 8 different disease models for each type of G x E interaction models. How we generate covariate data was provided in Section 3.2.2.

Marker data simulation

Chromosome-based method vs. recombination-fraction approach: many simulation programs (e.g. SIMLINK, SIMULATE and SLINK) are available to generate marker data (Boehnke 1986; Weeks et al. 1990; Ott and Terwilliger 1992; Terwilliger and Ott 1994). One common approach is the recombination-fraction method, as implemented in the SLINK program (Weeks et al. 1990). Based on the recombination-fraction method, for the "unlinked" chromosome, one can first simulate genotypes with phase at each marker independently for all founders under Hardy-Weinberg equilibrium (HWE). Next, one simulates marker segregation from the simulated parents to their offspring based on the law of independent segregation and the recombination fraction between each adjacent pair of markers. For the "linked" chromosome, one can create marker genotypes conditional on the disease phenotype, disease allele frequencies and genetic map distance between the adjacent markers. The conditional probability can be written as:

$$P(\mathbf{g} \mid \mathbf{h}) = P(g_1 \mid \mathbf{h}) * P(g_2 \mid g_1, \mathbf{h}) * P(g_3 \mid g_1, g_2, \mathbf{h})\ldots,$$

where $\mathbf{g} = (g_1, g_2, \ldots, g_n)$, a vector of the n subjects' multilocus genotypes in the same family, and $\mathbf{h} = (h_1, h_2, \ldots, h_n)$, a vector of these n subjects' phenotypes. However, one drawback of this approach is that the simulation process can become extremely slow as the number of markers increases.

Instead of generating the marker genotypes conditional on genetic distance to the neighboring marker, we employ the chromosome-based approach (Terwilliger et al. 1993) to simulate our marker data. In this procedure, the location and number of crossover points along a chromosome decide whether the offspring marker data is derived from the paternal or maternal chromosome. The benefit of this approach is that it is not only computationally efficient, but also allows us to model interference.

The steps of the chromosome-based approach are as follows:

a. We first generate the parents' marker data and disease locus genotypes under Hardy-Weinberg equilibrium and linkage equilibrium. Each parent is a "founder" with two chromosomes.
b. For each offspring, we simulate two vectors storing the location and number of crossover points for both parents according to a gamma distribution with shape equal to 3.9 and rate equal to 7.8 (for details, see Section 3.2.3.2.2) (McPeek and Speed 1995; Broman and Weber

2000). We utilize these vectors to pick the parents' genotype data transmitted from their paternal or maternal chromosome to each offspring.

c. In the "unlinked" case, we first pick the grandpaternal or grandmaternal chromosome, as an initial chromosome, from one of the parents by 50% chance. Once we start with the initial chromosome at position zero, whether we switch to the other chromosome depends on the number (odd or even) and location of crossover points. For example, we pick mother's maternal chromosome at position zero as the initial chromosome by 50% chance. While the first crossover event (odd number) occurs at 30 cM and the second crossover event (even number) occurs at 45 cM, we switch to the mother's paternal chromosome from 30 cM to 45 cM, then switch back to the mother's maternal chromosome from 45 cM till we hit the third crossover event.

d. In the "linked" case, we generate offspring's disease genotype conditional on the parents' disease genotypes. Meanwhile, we index the allele of offspring's disease genotype derived from father's paternal or maternal side; we also similarly index the allele coming from the mother. Therefore, for the offspring's marker data, we start with the same chromosome as the one having disease genotype at position zero if the number of crossover events before the disease locus location is even. If the number of crossover points is odd, we start with the other chromosome. Once we know which chromosome to start with, the rule for switching to the other chromosome is the same as Step c. For example, if the allele of offspring's disease genotype derived from father's paternal side, and the number of crossover events before the disease locus location is odd, we start with father's maternal chromosome.

<u>Interference and map function</u>: interference is generally divided into two aspects: chromatid interference and chiasma (or crossover) interference. Chromatid interference means the choice

of non-sister chromatids involved in a crossover is not independent. Chiasma interference means the number and position of crossover in the given region interferes with the number and position of crossovers in the adjacent region. We only model chiasma interference here. Applying mathematical models for the crossover-point distribution subject to chiasma interference is a substantial topic. The Haldane (1919) map function satisfies the no chiasma interference assumption; in contrast, the Kosambi (1944) map function assumes interference. Several mathematical distribution models have been suggested to model interference (McPeek and Speed 1995). McPeek and Speed compared 6 different mathematical fitted models. Based on their stochastic simulation, they showed that point process models such as gamma model (details provided in the following paragraph) often provides a very good fit to the observed data.

Two types of mathematical models are commonly used for generating crossover distribution: the count-location model and gamma model. In the count-location model (Karlin and Liberman, 1978; Risch and Lange 1979), the number of chiasma (crossover points) on the four-strand bundle is first generated, following a "count" distribution such as a Poisson distribution. Then the locations of crossover points are independently and evenly distributed, following a "location" distribution such as a uniform distribution. On the other hand, in the gamma model, the locations of crossover points on the four-strand bundle are determined based on a stationary renewal process, following a gamma distribution (Fisher et al. 1947; Foss et al. 1993). Broman and Weber (2000) have reported the gamma model fits the genotype data from CEPH families very well. We therefore employ gamma model to generate the number and locations of crossover points. The detailed gamma model is as follows:

Let x_i be the genetic distances between chiasmata on the four-strand bundle. The locations of chiasmata can be generated according to a stationary renewal process with gamma interarrivals.

The x_i are independent and follow a gamma distribution with shape (v) equal to 3.9 and rate (λ) equal to 7.8. The values of shape and rate were obtained from the Table 3 in the Broman and Weber paper (2000). Therefore, x_i has the gamma density, which can be denoted as:

$$f(x_i) = \frac{\lambda^v}{\Gamma(v)} x_i^{v-1} e^{-\lambda x_i}$$

where the gamma function: $\Gamma(v) = \int_0^\infty x^{v-1} e^{-x} dx$ and the mean is equal to v / λ.

The density of the first point x_0 needs to be treated differently from the other points and can be written as:

$$g(x_0) = \frac{\lambda^{v+1}}{\Gamma(v+1)} \int_x^\infty s^{v-1} e^{-\lambda s} ds \, .$$

In order to properly generate the first point, we apply rejection sampling to obtain the first point for the non-integer shape parameter v (Ripley 1987). The steps of rejection sampling are described as follows:

a. First, find a suitable distribution *h(x)*, known as the envelope distribution, which is always above the first point distribution $g(x_0)$ (Figure 3.5). We employ the distribution: $2.8*e^{-1.5*x}$ as the envelope distribution *h(x)* here.

b. Next, choose a point y, randomly from a uniform distribution between zero and A, where A is the total area under the envelope distribution *h(x)*. Then obtain the corresponding point x by inverting the function *H(x)*, which is the integral from zero to x of *h(y)*.

c. Then pick a random uniform point U between zero and one. Accept the point x if $U \leq f(x) / h(x)$. If $U > f(x) / h(x)$, repeat steps b and c until an acceptable point is generated.

Once we obtain the first point position, we then generate the other locations by using the gamma distribution with shape and rate of 3.9 and 7.8. Under no interference assumption, the final locations of crossover points can be obtained by "thinning" the chiasma on the four-strand bundle: chiasmata are retained independently as crossover points with 50% probability.

3.2.4. Simulation programs and procedures

We used the R programming language to write our data simulation program (Appendix B) and the code for several covariate statistics: the mixture model (the "pre-clustering" model), LODPAL, the multinomial logistic regression model approaches (no-dominance MLRM, no-additive MLRM, and min-max MLRM), and COVLINK (Appendix C). We first generate data using the simulation program, and apply several Perl scripts for data editing and filtering. We then analyze the data using the original programs for MLB and the OSA methods, the Merlin program, which implements model-free methods and QTL approaches, and our R code for the

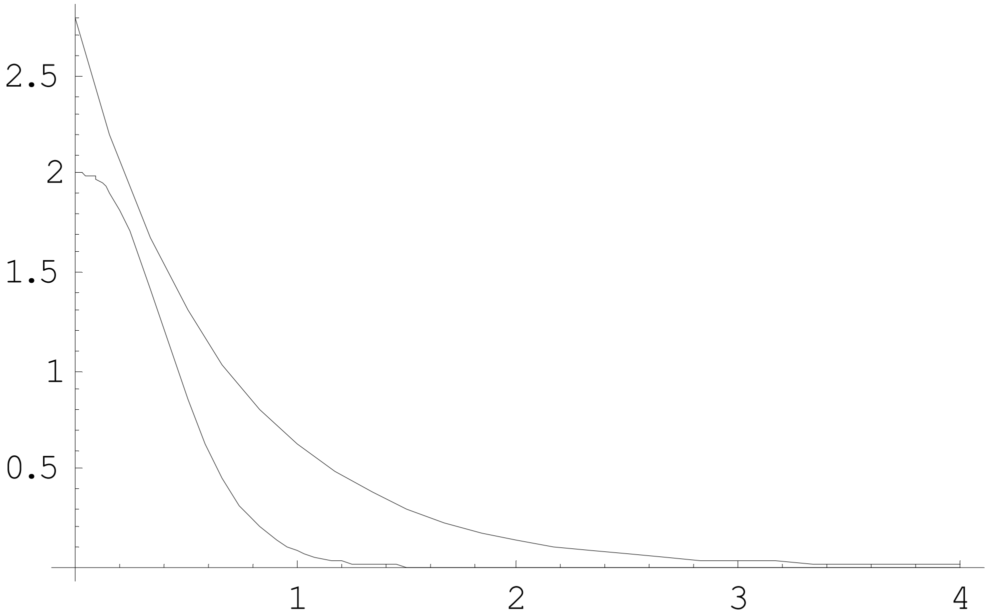


Figure 3.5 Distribution of first point in gamma model and the desired envelope function

other covariate methods. We write shell scripts to connect data generation and data analysis, and record the outputs. All the simulations were done using the Linux operating system. In average, it takes about 20 minutes to complete data simulation and data analysis for one replicate of 100 pedigrees when running on a 500 MHz processor, or 12 minutes when running on a 1.53 GHz processor.

4. VALIDATION AND PROPERTIES OF THE SIMULATED DATA

4.1. VALIDATION OF SIMULATION PROGRAMS

To make sure our simulation programs generate data correctly, we validate the programs in several ways as described in the following sections.

4.1.1. Variance proportion of each liability component

We simulate data under eight different disease models for each of the G x E interaction models. In each model, we generate one genetic factor G, three covariates (C1, C2, C3), a polygenic factor PG and a random error effect E. The variance proportion of each component varies across the liability models (Tables 3.1A, B, C). We first calculate the analytical variance for each component, corresponding to the desired variance proportion. Next, we generate 10,000 families and calculate the mean and variance of each component. We then check whether these values are very close to the analytical means and variances. According to our results, the simulated means and variances are quite close to the analytical values (data not shown).

4.1.2. Liability threshold vs. affection status

We first simulate 10,000 pedigrees to obtain the empirical liability distribution and calculate top 5% distribution as the cut-off point to assign affection status. To check whether the liability threshold gives 5% disease prevalence, we generate another 10,000 families and assign individuals' affection status according to the threshold. We then compute the population prevalence by two ways: (1) count the total number of affecteds, and then divided by the total number of subjects; (2) count the number of pedigrees whose first sib is affected and then divided by total family number. The outputs show that all liability thresholds provide approximate 5% overall prevalence (data not shown).

4.1.3. Recombination fraction between markers

The simulated chromosome length is 169 cM. We generate 33 markers evenly distributed along the chromosome with 5 cM spacing. As described in Chapter 3, the chiasma (crossover points) are generated from a gamma distribution with shape 3.9 and rate 7.8. Because the first point distribution needs to be treated differently, we employ rejection sampling to generate the location of the first point.

We check whether using a gamma distribution for marker generation provides the appropriate recombination fraction between markers. Therefore, we first simulate 500 pedigrees with marker data across 22 chromosomes under gamma distributions with proper shape and rate parameters. The lengths of these 22 chromosomes mimic the lengths of 22 autosomal chromosomes as taken from the recent deCode map (Kong et al. 2002). The shape and rate parameters in gamma distributions are taken from the Table 3 in the Broman and Weber paper (Broman and Weber 2000). We then count the crossovers between each marker pair and calculate the recombination fractions. We find that the recombination fraction between each adjacent marker pair is very close to 0.05, as expected (data not shown).

4.1.4. Statistics code

Except for COVLINK, we have the authors' programs implementing their covariate methods. But for computational efficiency, we wrote R code implementing several covariate statistics, as described in Chapter 3. To check the accuracy of our R code, we first generate various data sets with 100, 200, or 300 pedigrees. We then analyze data using our R code and the authors' original programs, and compare the outputs. The results indicate that the outputs from our R code are close to those from the authors' programs (data not shown). Slight deviations exit in our outputs because the optimization algorithms applied in our code and the available programs are different. We employ our R code to analyze the data for these statistics.

4.2. PROPERTIES OF SIMULATED DATA

4.2.1. Proportion of linked families

We examine the proportion of linked families. The definition of a "linked family" here is a family with two or more affected sibs carrying a high-risk disease genotype. We generate 500 replicates, where each replicate contains 100 ascertained pedigrees. Then we calculate the mean and standard deviation of the proportion of linked families for each model. We denote "the proportion of linked families" as α.

The results are presented in Table 4.1. As the disease allele frequency increases, α increases, for both dominant and recessive traits. When we look at each G x E interaction model separately, we find α in the Type III model is more than 90% in the common disease models ($p = 0.1$ in dominant mode; $p = 0.4$ in recessive mode). In the Type II model, α in the models with 20% genetic variance is always lower than the models with 30% genetic variance. In general, α in the Type II models is lower than the other models.

4.2.2. Recurrence risk ratio for siblings

Table 4.1 Percentage of linked families in different models, based on 500 replicates per model

	Disease allele frequencies			
Dominant trait	0.01	0.02	0.05	0.1
Type I model	36.68% ± 4.80%	61.93% ± 4.89%	74.87% ± 4.30%	79.51% ± 3.88%
Type II model (GV = 20%)	14.39% ± 3.55%	17.95% ± 3.79%	26.11% ± 4.38%	35.05% ± 5.03%
Type II model (GV = 30%)	16.79% ± 3.86%	24.67% ± 4.33%	32.99% ± 4.84%	43.12% ± 4.83%
Type III model	35.55% ± 4.81%	58.68% ± 4.96%	86.10% ± 3.42%	95.71% ± 2.08%
	Disease allele frequencies			
Recessive trait	0.1	0.2	0.3	0.4
Type I model	23.06% ± 4.11%	54.93% ± 5.11%	69.65% ± 4.90%	74.76% ± 4.48%
Type II model (GV = 20%)	7.33% ± 2.64%	14.29% ± 3.51%	20.54% ± 4.25%	28.60% ± 4.55%
Type II model (GV = 30%)	7.68% ± 2.55%	20.39% ± 4.05%	27.15% ± 4.52%	36.07% ± 4.81%
Type III model	14.35% ± 3.61%	50.10% ± 4.57%	77.65% ± 4.19%	92.00% ± 2.69%

Note:

The definition as a “linked family” is the family with two or more affected sibs carrying high-risk disease genotype.

We compute the recurrence risk ratio for siblings under different models. We pick the first two sibs in each family to compute the recurrence risk ratio. Two recurrence risk ratios are defined as: (1) λ_s: the probability that the 2nd sib is affected given the 1st sib is affected divided by the population prevalence *K*, [*P*(the 2nd sib affected | the 1st sib affected) / population prevalence *K*]; (2) $\lambda_{s,\ HR}$ the probability that the 2nd sib is affected and carries the high-risk genotype given the

1st sib is affected and carries the high-risk genotype divided by the population prevalence *K*, [*P*(the 2nd sib affected & high-risk genotype | the 1st sib affected & high-risk genotype) / population prevalence *K*]. For each model, we simulate 10,000 families before ascertainment, assign the affection status (according to the threshold giving 5% disease prevalence) to each individual and compute the recurrence risk ratio.

Generally speaking, the range of λ_s is between 3.4 and 5.8, and the range of $\lambda_{s,\,HR}$ is between 2.4 and 9.3. There is no particular pattern observed across different disease models, no matter which G x E interaction models (Table 4.2). Although α increases as the disease allele frequencies increase (Table 4.1), the recurrence risk ratio does not follow the same trend. One possible explanation may be that the liability model is determined by a combination of several components, not genetic factors only. In other words, the etiologies of the first two affected sibs in the same family do not have to be the same. For example, when the disease allele frequencies are less common, it is possible that one sib is affected due to a high-risk genotype and the other sib is affected due to a high environmental covariate value.

4.2.3. Covariate values in affecteds and unaffecteds

We compute the average covariate values in affecteds and unaffecteds before and after ascertainment across the models, respectively (Tables 4.3 A, B). Under the Type I and III models, the values of C1 and C2 in affecteds are higher than those in unaffecteds. Since C1 and C2 affect the disease liability, affecteds should have higher values than unaffecteds. In addition, since C1 is strongly influenced by disease locus genotypes under the Type III model, we would expect to observe that affecteds after ascertainment have higher values than those before ascertainment.

Under the Type II model, the C2 values in affecteds are higher than those in unaffecteds, but not the C1 values. Since C2 involves in liability, we would expect affecteds to have higher values than unaffecteds. However, C1 in the Type II model is used to determine whether genetic factor has an effect on liability. Since the proportion of the linked families is less than 50% across all the Type II models, it is not surprising that the average C1 values in affecteds are often negative, and are not higher than those in unaffecteds.

The C1 and C2 values in unaffecteds across three types G x E models are close to the mean in the general population. However, in the Type III model, we first generate individuals' C1 values from a multivariate normal distribution with mean 2 and variance 1, and then update the C1 values according to individuals' genotypes. Therefore, the original mean C1 value in the general population is not available here.

The C3 values do not differ between affecteds and unaffecteds, and do not differ before and after ascertainment. Also the C3 values in both groups across three types G x E models are close to the mean in the general population. Since C3 is a random noise covariate, it should not vary between affecteds and unaffecteds. It is sensible to obtain such results.

Table 4.2 Recurrence risk ratio for siblings in different models, based on 10,000 families per model

		Disease allele frequencies			
Dominant trait		0.01	0.02	0.05	0.1
Type I model	λ_s*	5.40	4.50	3.56	3.84
	$\lambda_{s, HR}$**	9.33	7.59	4.74	4.73
Type II model	λ_s	5.28	4.84	4.36	4.52
(GV = 20%)	$\lambda_{s, HR}$	6.42	4.16	3.48	3.82
Type II model	λ_s	5.56	4.88	4.70	4.69
(GV = 30%)	$\lambda_{s, HR}$	6.05	4.89	3.38	3.97
Type III model	λ_s	3.87	4.39	5.62	5.13
	$\lambda_{s, HR}$	5.26	7.27	6.36	5.05
		Disease allele frequencies			
Recessive trait		0.1	0.2	0.3	0.4
Type I model	λ_s	3.40	3.98	3.74	3.58
	$\lambda_{s, HR}$	5.10	5.02	4.04	4.83
Type II model	λ_s	5.42	3.90	3.68	3.99
(GV = 20%)	$\lambda_{s, HR}$	4.58	2.39	2.93	2.54
Type II model	λ_s	5.77	5.00	4.49	3.94
(GV = 30%)	$\lambda_{s, HR}$	6.36	5.66	3.57	2.64
Type III model	λ_s	3.40	4.26	5.01	4.40
	$\lambda_{s, HR}$	3.98	4.75	5.69	4.47

Note:

* $\lambda_s = P$(the 2nd sib affected | the 1st sib affected) / population prevalence K

** $\lambda_{s, HR} = P$(the 2nd sib affected & high-risk genotype | the 1st sib affected & high-risk genotype) / population prevalence K

Table 4.3A Average covariate values in affecteds and unaffecteds in different dominant models, based on 10,000 pedigrees per model

			Disease allele frequencies							
Dominant trait			0.01		0.02		0.05		0.1	
		Mean **	Aff	Un-aff	Aff	Un-aff	Aff	Un-aff	Aff	Un-aff
Type I model	C1 $_{\text{b/f}}$ *	0	0.339	-0.034	0.528	-0.037	0.914	-0.042	1.159	-0.058
	C1 $_{\text{a/f}}$ *		0.582	0.296	1.159	0.585	1.664	0.997	1.803	1.190
	C2 $_{\text{b/f}}$	0	0.327	-0.021	0.428	-0.01	0.628	-0.034	0.842	-0.044
	C2 $_{\text{a/f}}$		0.341	0.271	0.214	-0.074	0.349	-0.137	0.583	-0.192
	C3 $_{\text{b/f}}$	62.39	61.83	63.29	63.56	62.11	61.54	61.81	62.22	62.00
	C3 $_{\text{a/f}}$		62.41	61.30	61.98	61.90	62.07	62.26	62.13	62.65
Type II model (GV = 20%)	C1 $_{\text{b/f}}$	0	-0.418	0.029	-0.374	-0.006	-0.196	0.004	0.035	0.010
	C1 $_{\text{a/f}}$		-0.362	-0.376	-0.298	-0.299	-0.202	-0.206	-0.076	-0.100
	C2 $_{\text{b/f}}$	0	0.240	-0.012	0.342	-0.017	0.549	-0.034	0.754	-0.029
	C2 $_{\text{a/f}}$		0.275	0.136	0.386	0.186	0.647	0.348	0.879	0.479
	C3 $_{\text{b/f}}$	62.39	62.30	63.62	65.24	62.96	61.63	63.31	62.56	63.25
	C3 $_{\text{a/f}}$		63.15	63.23	63.07	62.34	63.72	63.44	63.02	63.28
Type II model (GV = 30%)	C1 $_{\text{b/f}}$	0	-0.357	0.028	-0.316	0.010	-0.110	0.021	0.015	-0.010
	C1 $_{\text{a/f}}$		-0.319	-0.345	-0.236	-0.266	-0.108	-0.144	0.003	-0.046
	C2 $_{\text{b/f}}$	0	0.203	-0.011	0.264	-0.013	0.459	-0.020	0.615	-0.025
	C2 $_{\text{a/f}}$		0.219	0.097	0.316	0.149	0.507	0.256	0.708	0.369
	C3 $_{\text{b/f}}$	62.39	63.42	63.17	62.49	63.76	62.70	63.13	62.13	62.79
	C3 $_{\text{a/f}}$		63.07	62.70	63.39	62.65	63.08	63.07	62.78	63.05
Type III model	C1 $_{\text{b/f}}$	NA	1.268	0.054	2.333	0.101	4.444	0.354	5.510	0.842
	C1 $_{\text{a/f}}$		2.370	0.462	3.712	0.885	5.224	1.711	5.710	2.464
	C2 $_{\text{b/f}}$	0	1.196	-0.071	1.079	-0.022	0.878	-0.065	0.932	-0.057
	C2 $_{\text{a/f}}$		1.284	0.759	1.102	0.614	0.958	0.479	1.067	0.549
	C3 $_{\text{b/f}}$	62.39	63.14	62.99	63.62	63.26	64.88	63.68	62.64	62.99
	C3 $_{\text{a/f}}$		63.36	63.08	63.45	63.89	63.49	63.01	63.25	62.96

Note:

* b/f: before ascertainment; a/f: after ascertainment; ** The mean value in the general population

Table 4.3B Average covariate values in affecteds and unaffecteds in different recessive models, based on 10,000 pedigrees per model

			Disease allele frequencies							
Recessive trait			0.1		0.2		0.3		0.4	
		Mean**	Aff	Un-aff	Aff	Un-aff	Aff	Un-aff	Aff	Un-aff
Type I model	C1 $_{b/f}$*	0	0.035	-0.023	0.505	-0.030	0.948	-0.024	1.088	-0.065
	C1 $_{a/f}$*		0.308	0.215	1.047	0.719	1.569	1.080	1.800	1.244
	C2 $_{b/f}$	0	0.262	-0.012	0.408	-0.022	0.565	-0.024	0.807	-0.031
	C2 $_{a/f}$		0.191	-0.037	0.254	-0.068	0.367	-0.104	0.522	-0.151
	C3 $_{b/f}$	62.39	62.04	62.44	62.54	62.15	63.30	62.63	61.83	62.44
	C3 $_{a/f}$		62.02	62.55	62.80	62.19	62.64	62.51	62.36	61.58
Type II model (GV = 20%)	C1 $_{b/f}$	0	-0.459	0.019	-0.309	0.016	-0.157	-0.002	-0.046	0.005
	C1 $_{a/f}$		-0.468	-0.427	-0.379	-0.336	-0.271	-0.247	-0.116	-0.128
	C2 $_{b/f}$	0	0.167	-0.008	0.348	-0.017	0.517	-0.029	0.700	-0.034
	C2 $_{a/f}$		0.200	0.094	0.406	0.205	0.633	0.337	0.822	0.445
	C3 $_{b/f}$	62.39	63.06	62.72	62.40	62.33	62.41	61.81	61.82	62.38
	C3 $_{a/f}$		62.92	62.73	62.80	62.69	63.11	62.02	62.87	62.77
Type II model (GV = 30%)	C1 $_{b/f}$	0	-0.443	0.026	-0.248	0.012	-0.151	0.001	-0.007	0.006
	C1 $_{a/f}$		-0.455	-0.403	-0.307	-0.256	-0.199	-0.159	-0.073	-0.087
	C2 $_{b/f}$	0	0.146	-0.009	0.283	-0.015	0.435	-0.019	0.581	-0.032
	C2 $_{a/f}$		0.165	0.070	0.330	0.158	0.502	0.251	0.661	0.345
	C3 $_{b/f}$	62.39	62.17	62.72	59.62	62.73	64.28	62.45	61.72	62.55
	C3 $_{a/f}$		62.23	62.16	62.18	62.23	62.06	62.53	61.80	62.57
Type III model	C1 $_{b/f}$	NA	0.725	0.027	2.463	0.114	4.237	0.327	5.279	0.684
	C1 $_{a/f}$		0.913	0.081	3.153	0.457	4.792	1.123	5.545	2.001
	C2 $_{b/f}$	0	1.344	-0.046	1.061	-0.072	0.934	-0.056	0.893	-0.041
	C2 $_{a/f}$		1.544	0.876	1.212	0.691	0.999	0.549	1.037	0.563
	C3 $_{b/f}$	62.39	62.76	63.05	63.17	63.34	62.56	63.24	61.66	63.08
	C3 $_{a/f}$		63.25	63.44	62.94	63.19	63.49	62.72	62.05	62.90

Note:

* b/f: before ascertainment; a/f: after ascertainment; ** The mean value in the general population

4.2.4. Covariate values of sib pairs

We compare the covariate values, after ascertainment, between affected-affected, affected-unaffected and unaffected-unaffected sib pairs across the models, respectively. Under the Type I and II models, the C1 values are highly correlated between all three kinds of sib pairs. However, correlation of the C2 values decreases as the disease allele frequencies increase between all three kinds of sib pairs. In other words, as the disease is getting more common, more linked pedigrees are ascertained (Figures 4.1 – 4.24).

Under the Type III model, when the disease is rare, we observe more sib pairs with low C1 values. When the disease is common, we observe more affected pairs with high C1 values and most unaffected pairs with low C1 values. The C2 values do not change too much between pairs across the models (Figures 4.25 – 4.32).

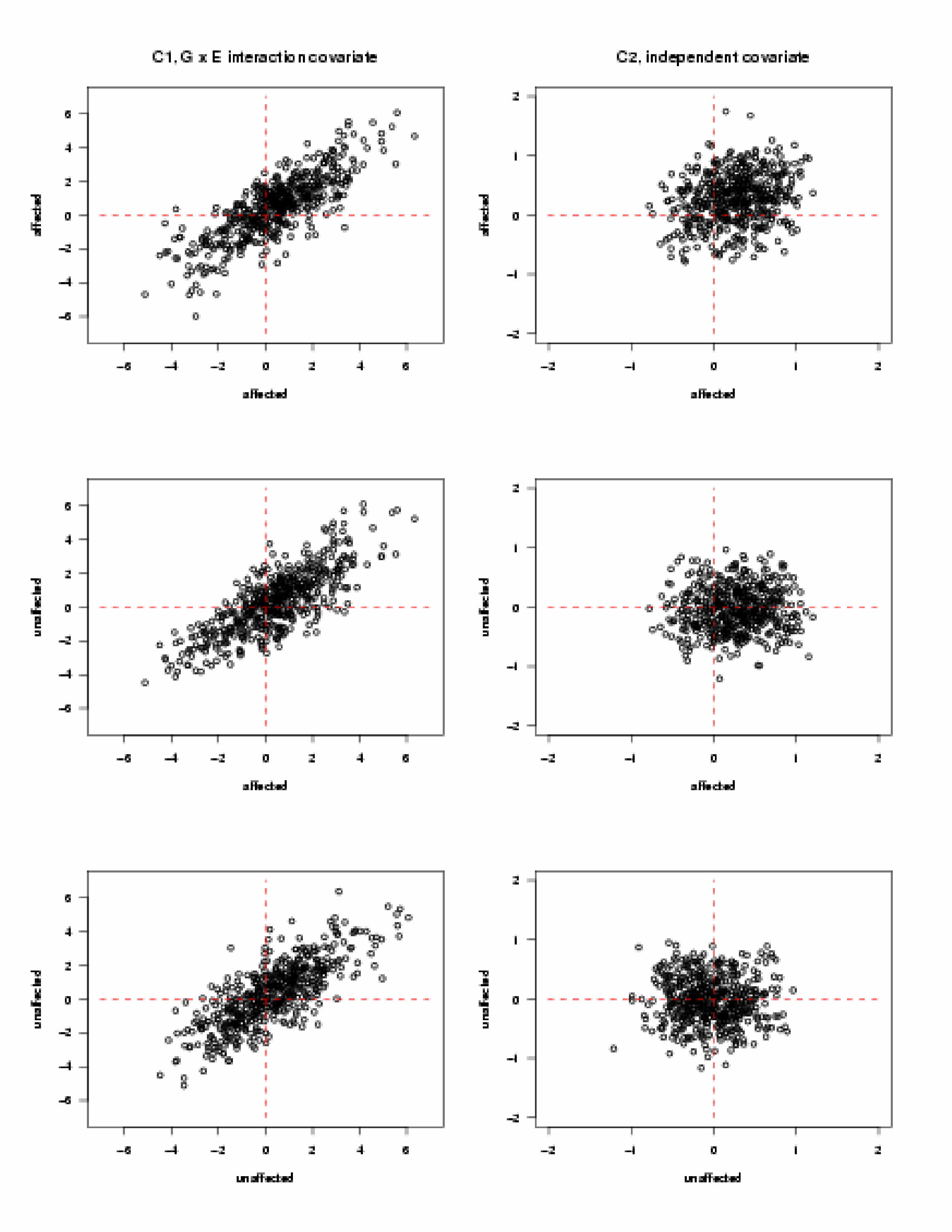


Figure 4.1 Covariate values in different sib pairs under Type I dominant model with disease allele frequencies equal to 0.01, based on 500 pedigrees

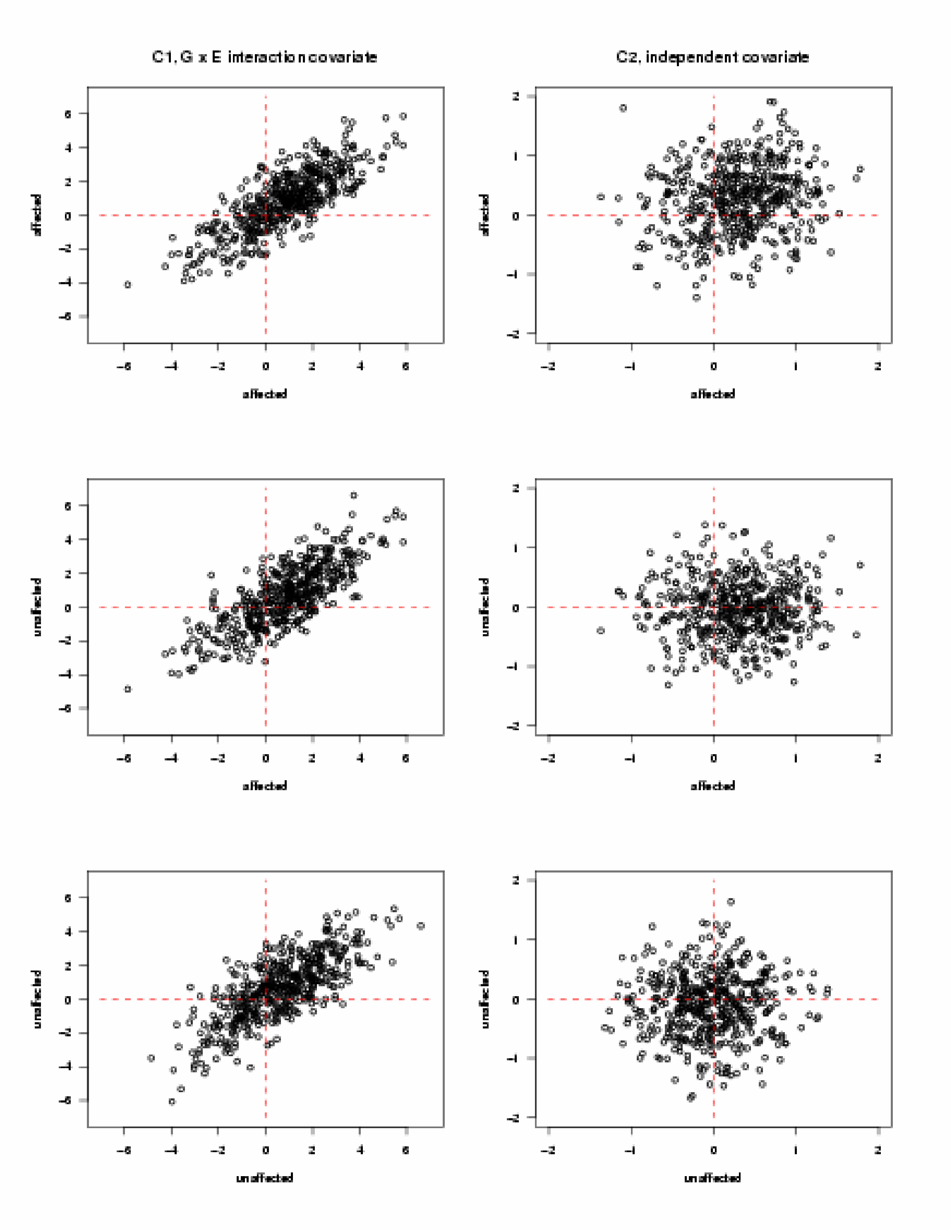


Figure 4.2 Covariate values in different sib pairs under Type I dominant model with disease allele frequencies equal to 0.02, based on 500 pedigrees

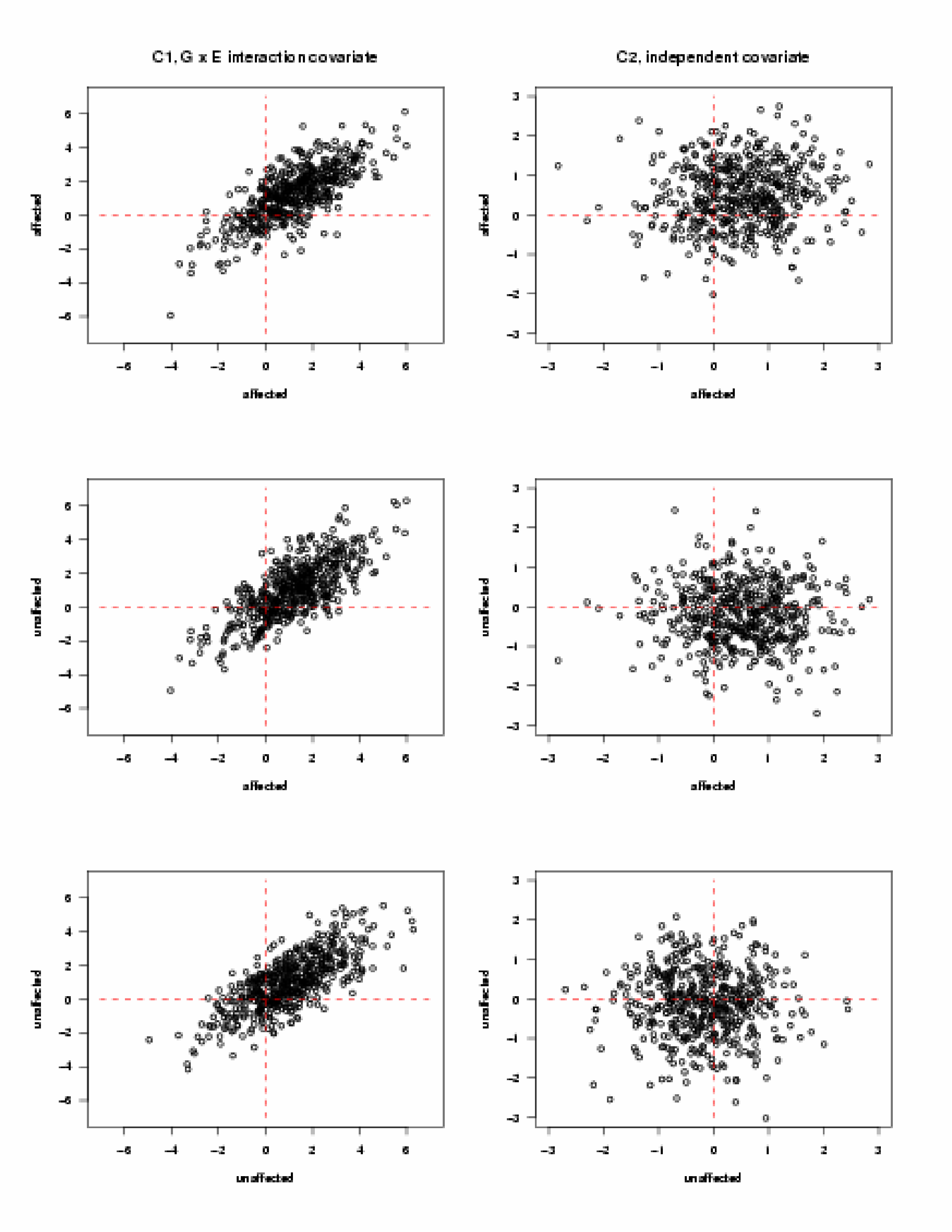


Figure 4.3 Covariate values in different sib pairs under Type I dominant model with disease allele frequencies equal to 0.05, based on 500 pedigrees

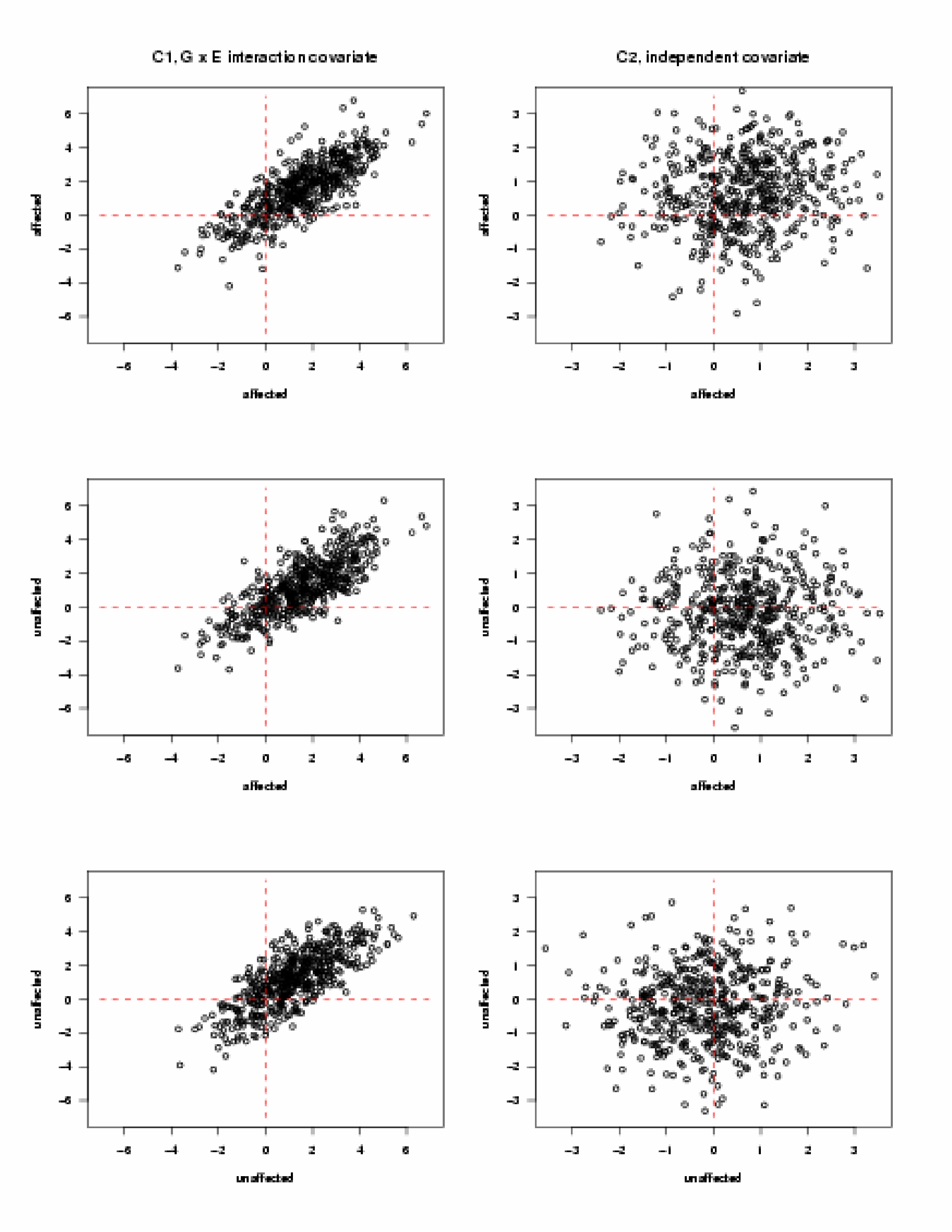


Figure 4.4 Covariate values in different sib pairs under Type I dominant model with disease allele frequencies equal to 0.1, based on 500 pedigrees

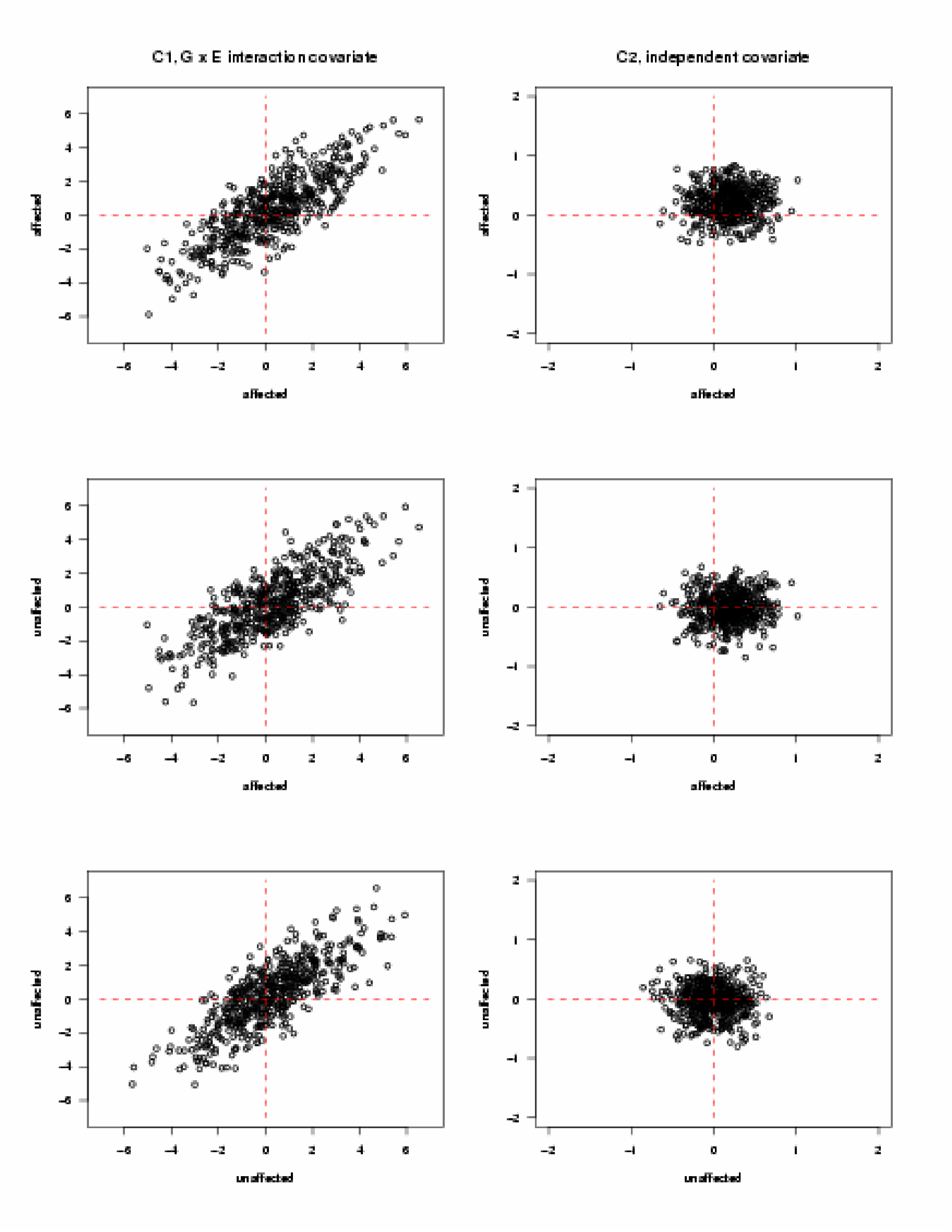


Figure 4.5 Covariate values in different sib pairs under Type I recessive model with disease allele frequencies equal to 0.1, based on 500 pedigrees

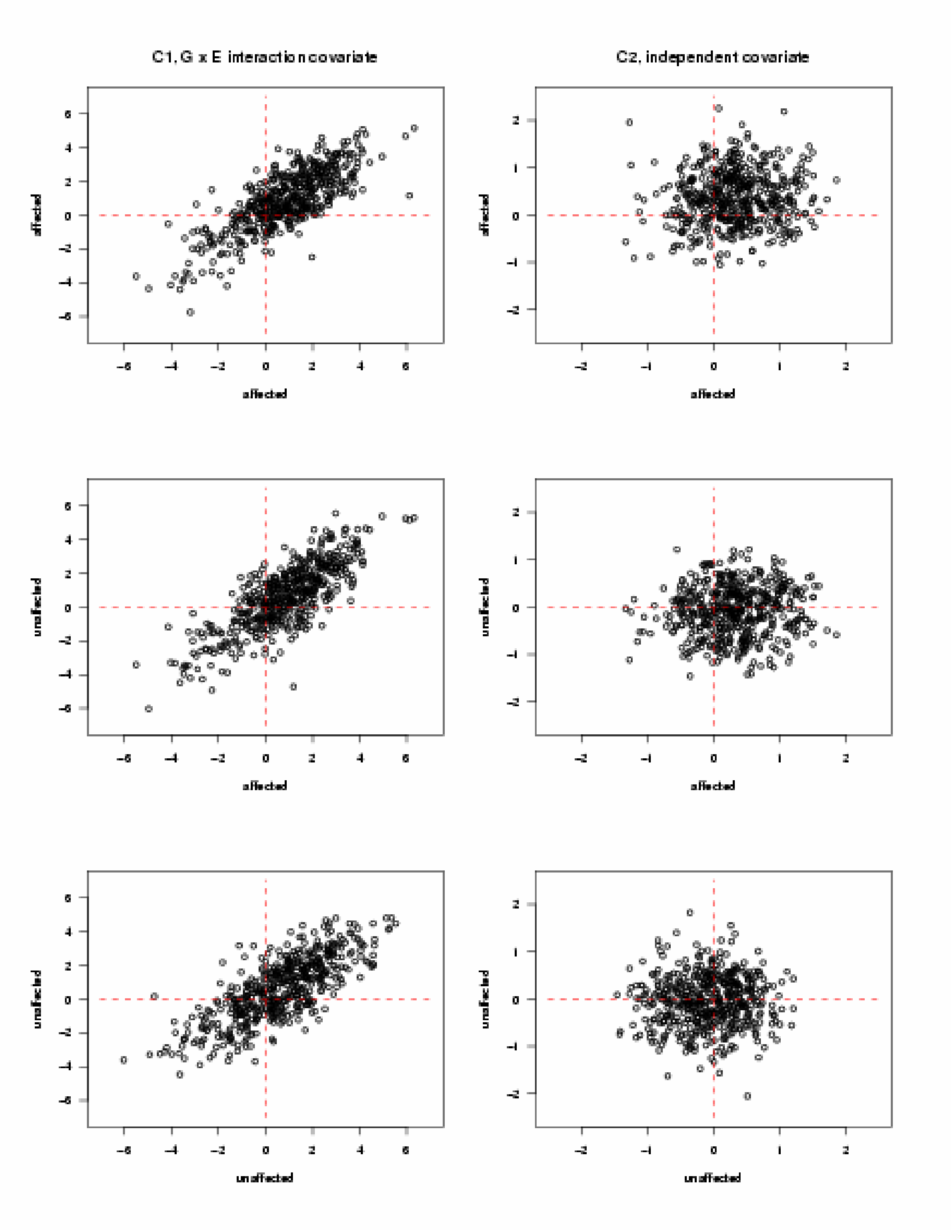


Figure 4.6 Covariate values in different sib pairs under Type I recessive model with disease allele frequencies equal to 0.2, based on 500 pedigrees

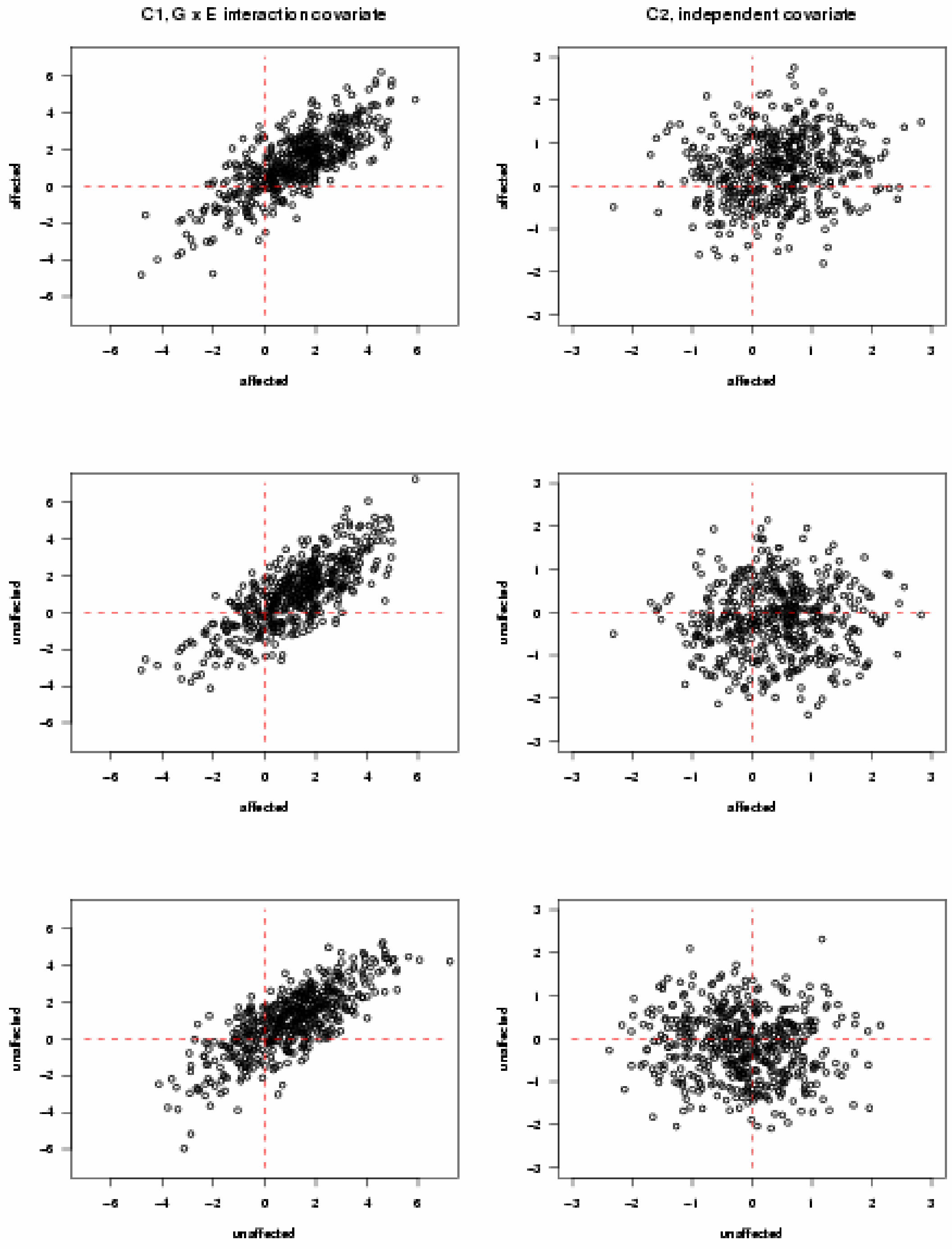


Figure 4.7 Covariate values in different sib pairs under Type I recessive model with disease allele frequencies equal to 0.3, based on 500 pedigrees

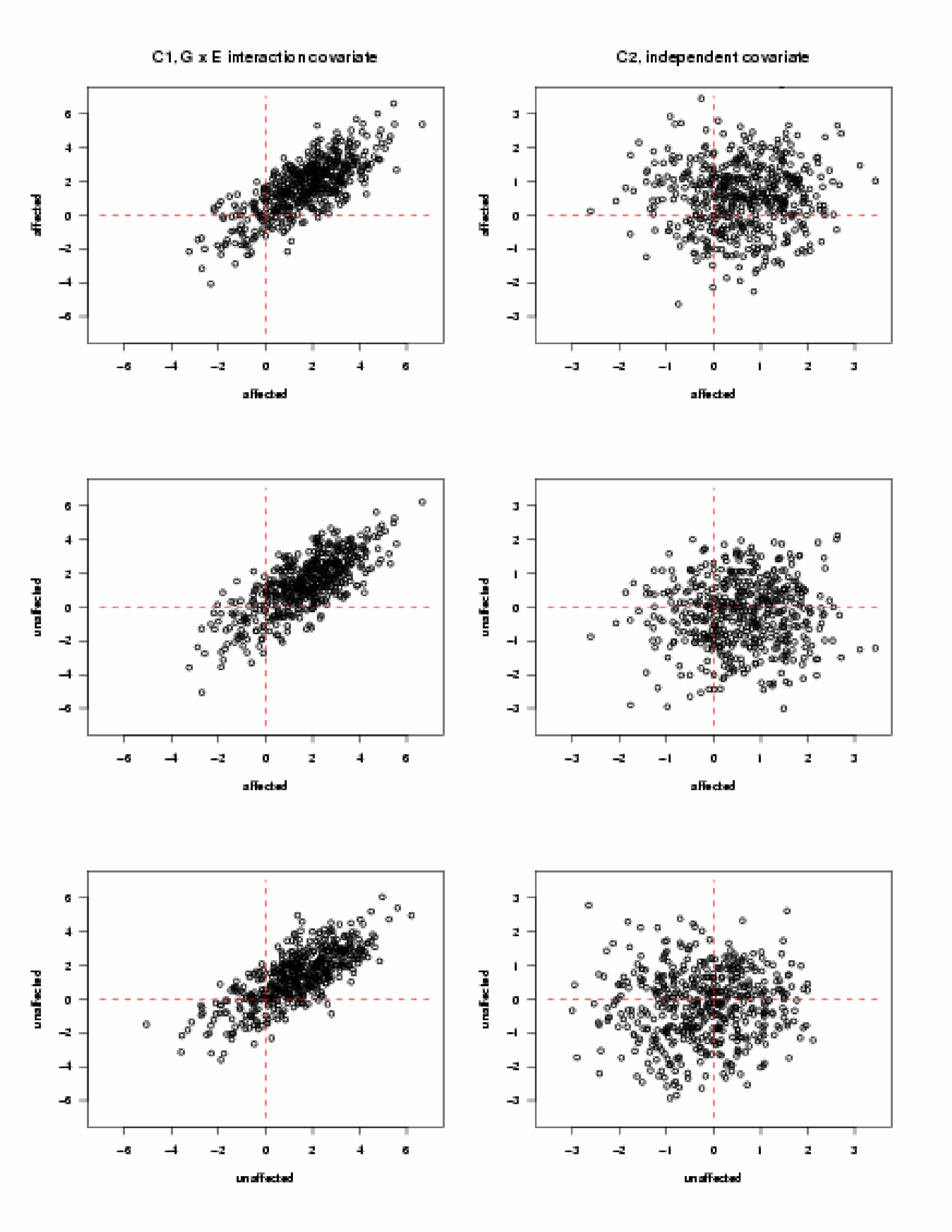


Figure 4.8 Covariate values in different sib pairs under Type I recessive model with disease allele frequencies equal to 0.4, based on 500 pedigrees

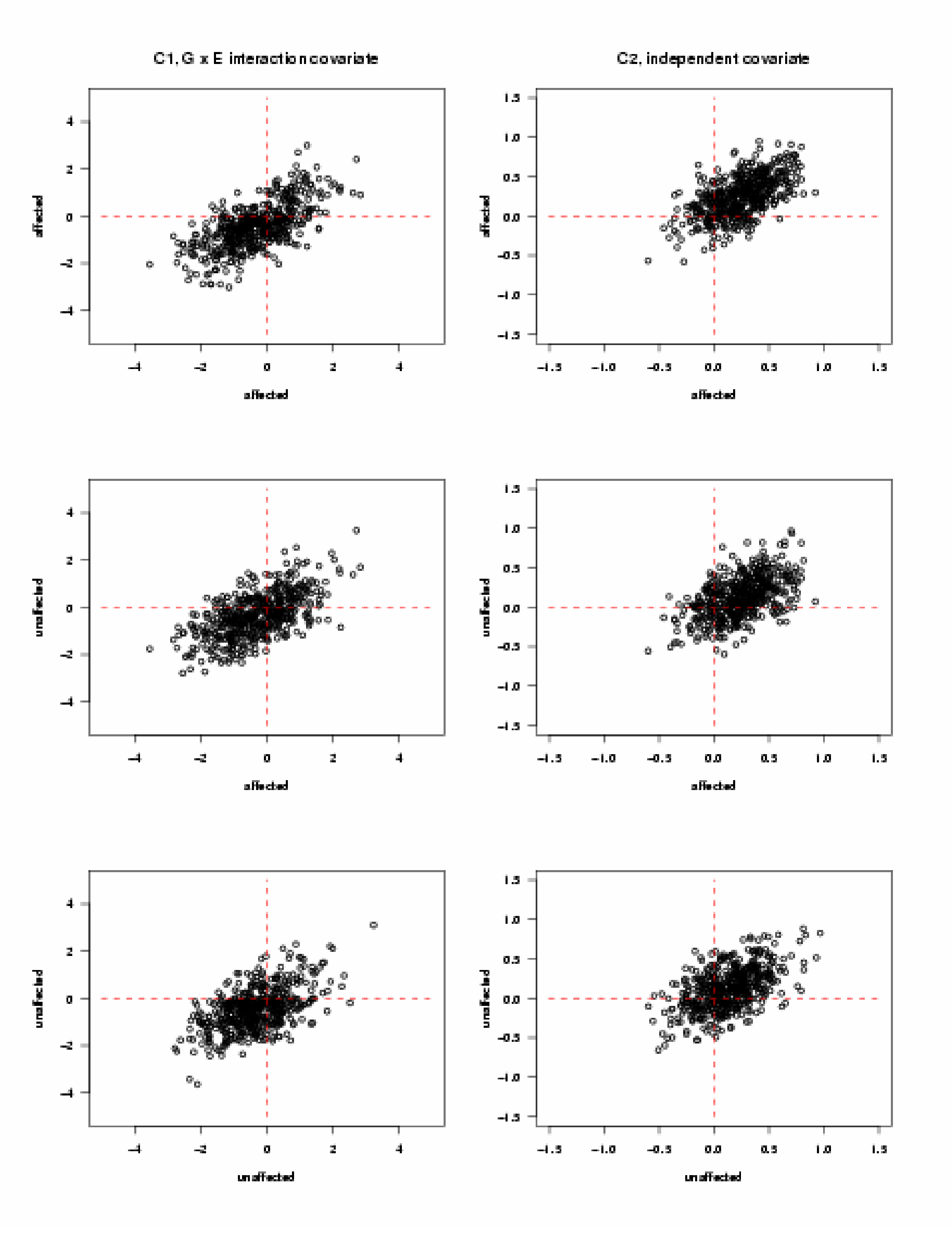


Figure 4.9 Covariate values in different sib pairs under Type II dominant model with disease allele frequencies equal to 0.01 and genetic variance equal to 20%, based on 500 pedigrees

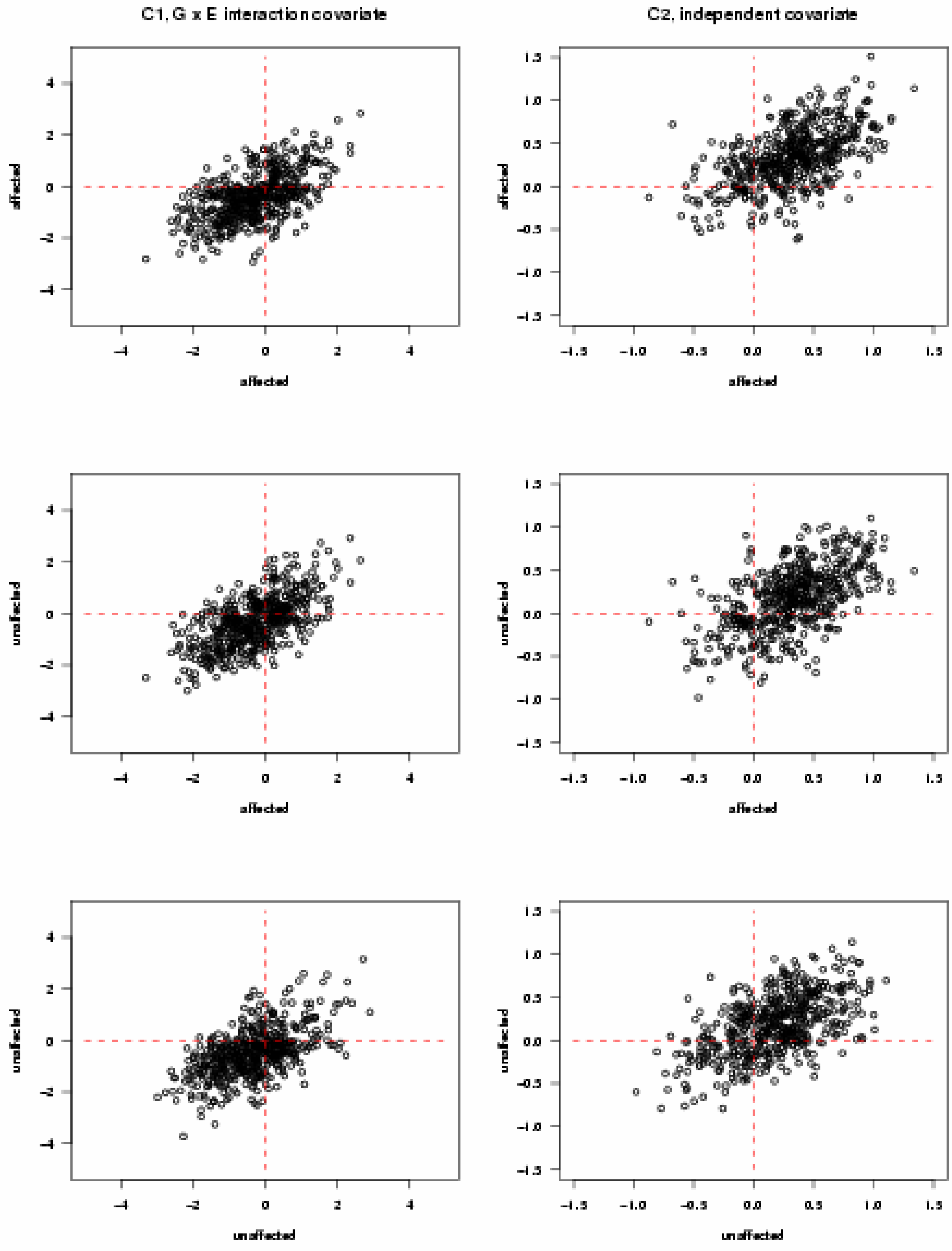


Figure 4.10 Covariate values in different sib pairs under Type II dominant model with disease allele frequencies equal to 0.02 and genetic variance equal to 20%, based on 500 pedigrees

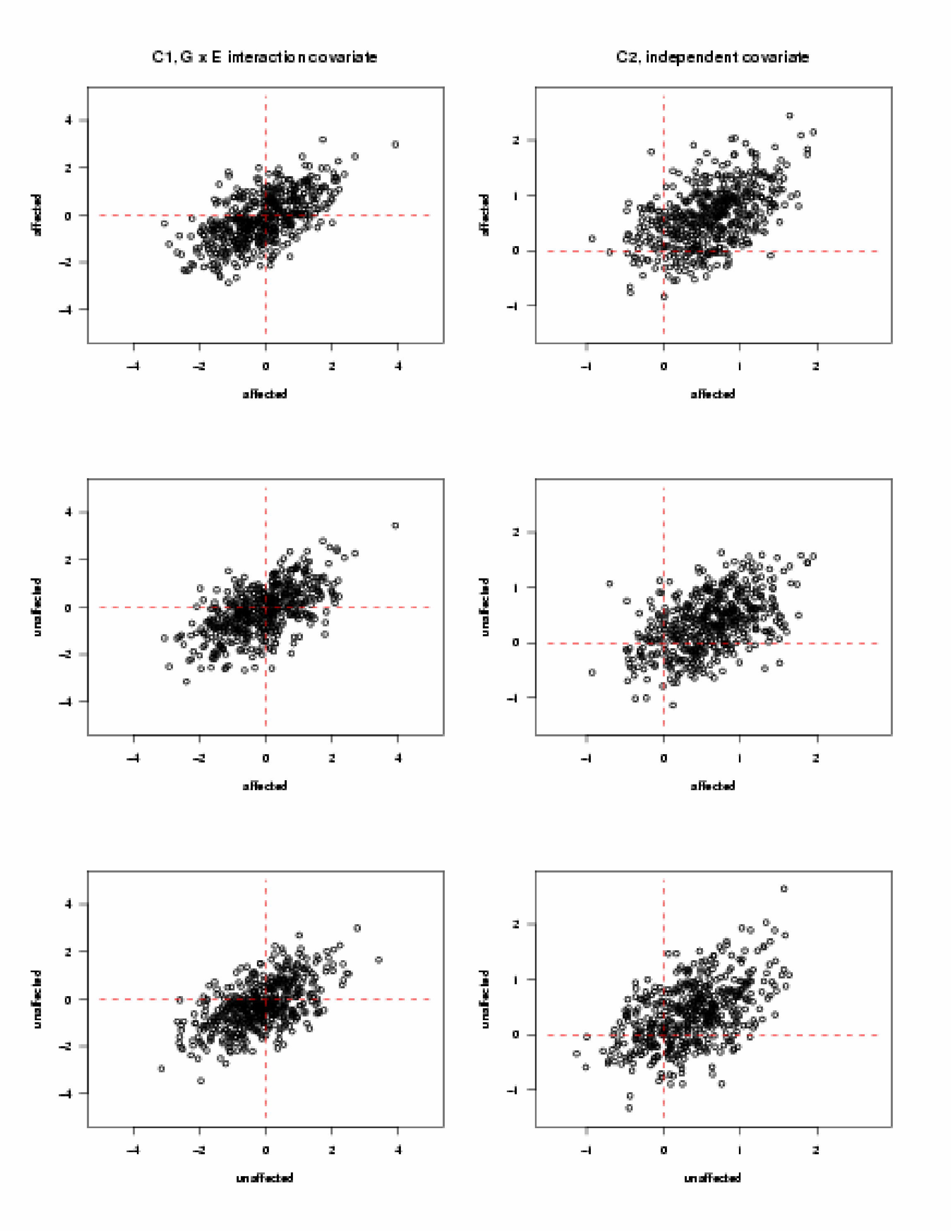


Figure 4.11 Covariate values in different sib pairs under Type II dominant model with disease allele frequencies equal to 0.05 and genetic variance equal to 20%, based on 500 pedigrees

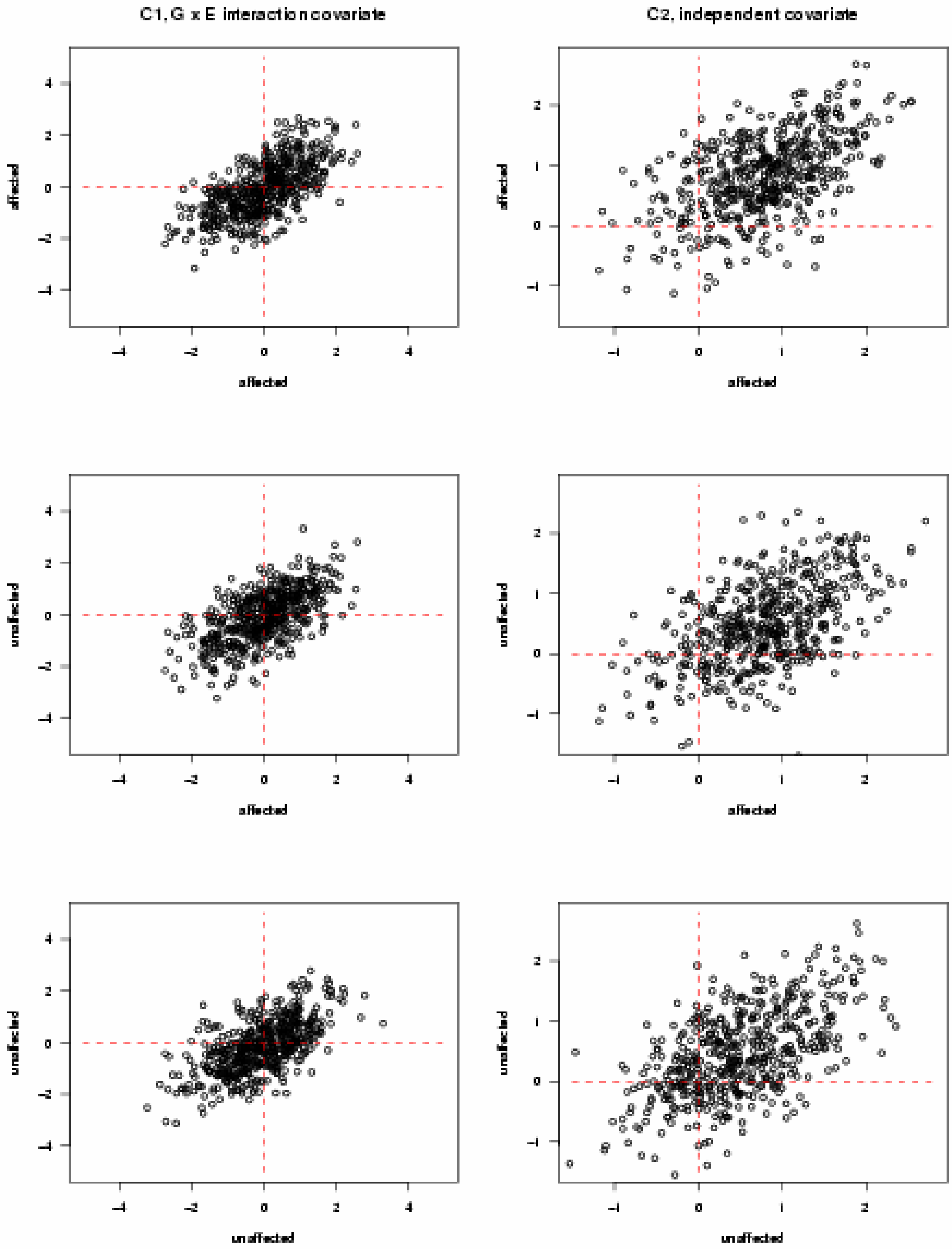


Figure 4.12 Covariate values in different sib pairs under Type II dominant model with disease allele frequencies equal to 0.1 and genetic variance equal to 20%, based on 500 pedigrees

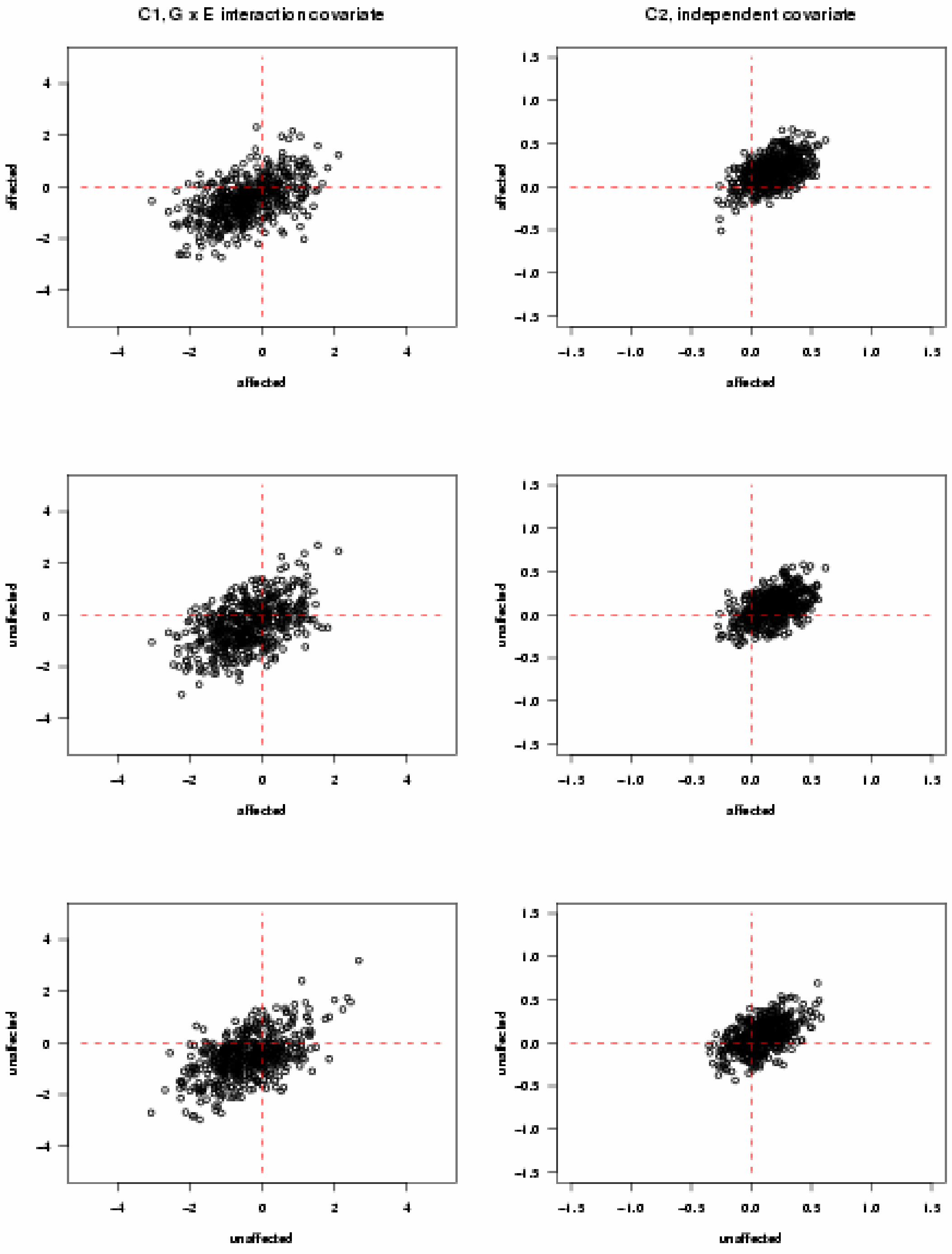


Figure 4.13 Covariate values in different sib pairs under Type II recessive model with disease allele frequencies equal to 0.1 and genetic variance equal to 20%, based on 500 pedigrees

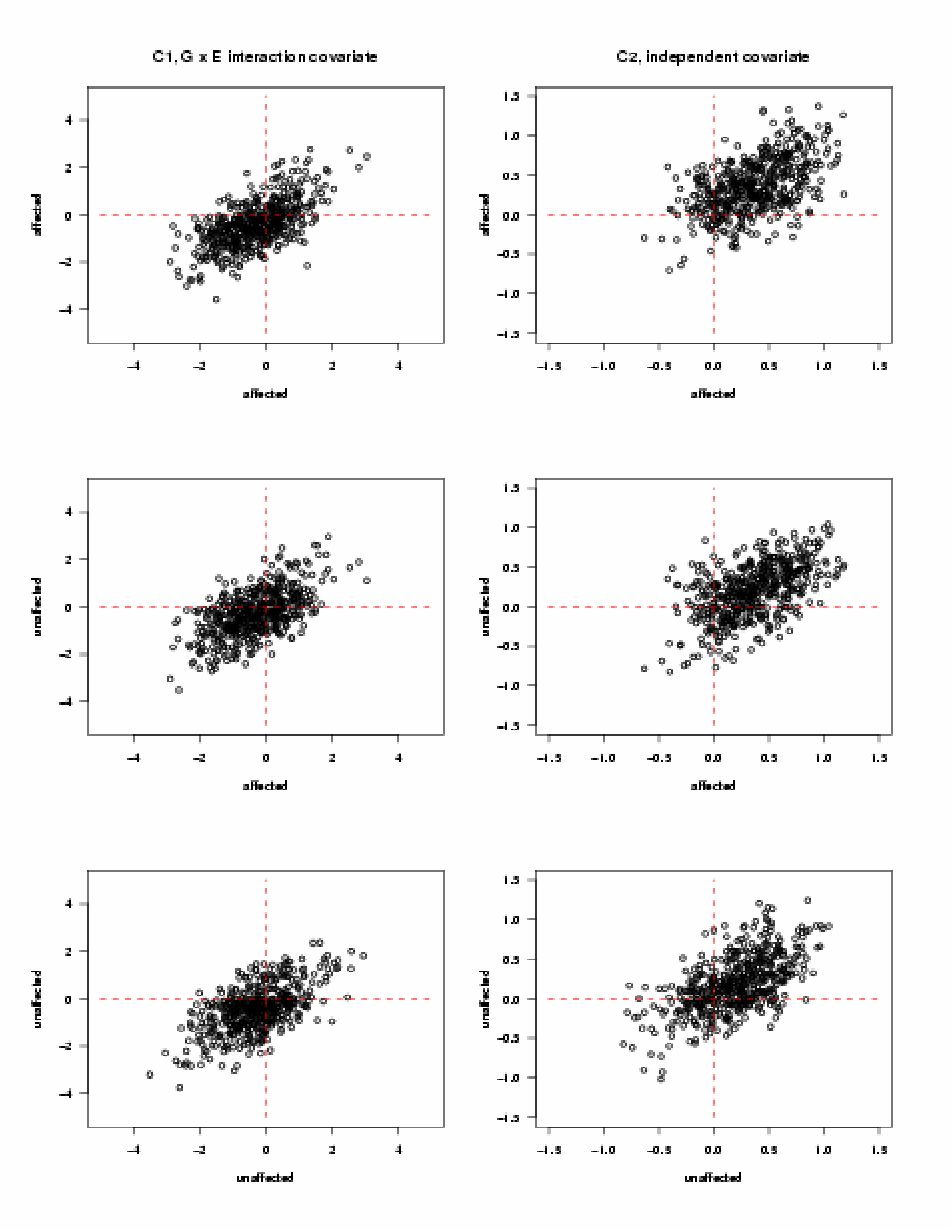


Figure 4.14 Covariate values in different sib pairs under Type II recessive model with disease allele frequencies equal to 0.2 and genetic variance equal to 20%, based on 500 pedigrees

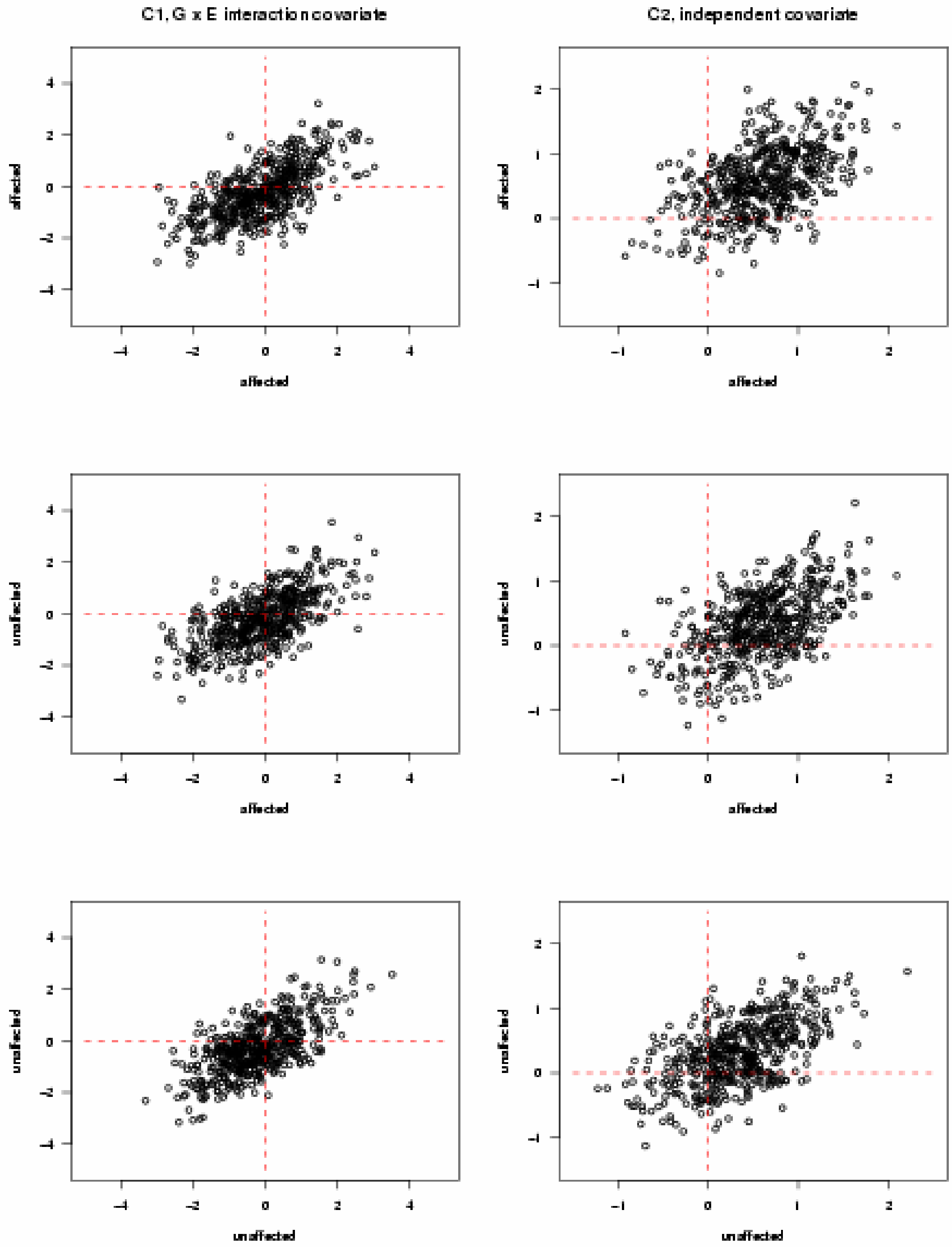


Figure 4.15 Covariate values in different sib pairs under Type II recessive model with disease allele frequencies equal to 0.3 and genetic variance equal to 20%, based on 500 pedigrees

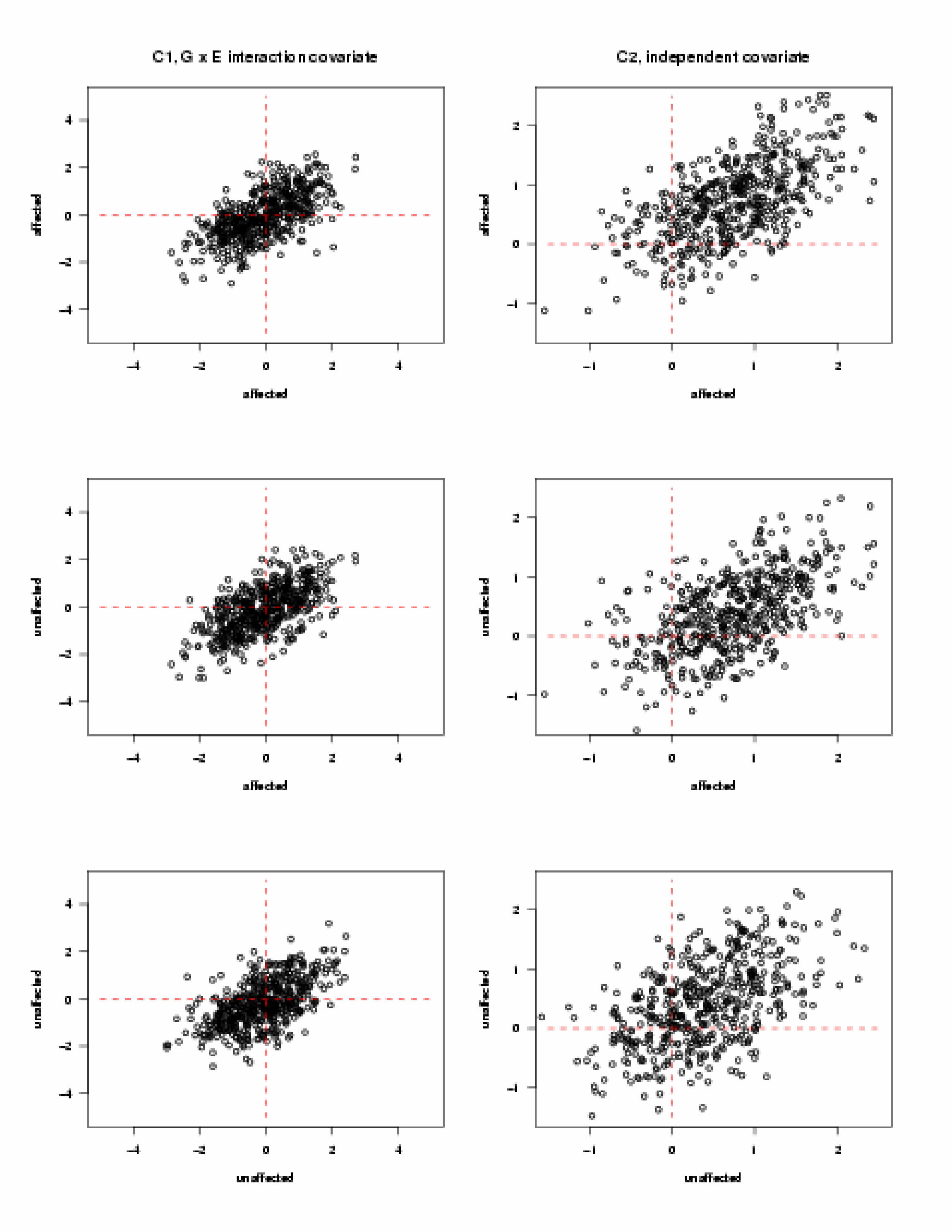


Figure 4.16 Covariate values in different sib pairs under Type II recessive model with disease allele frequencies equal to 0.4 and genetic variance equal to 20%, based on 500 pedigrees

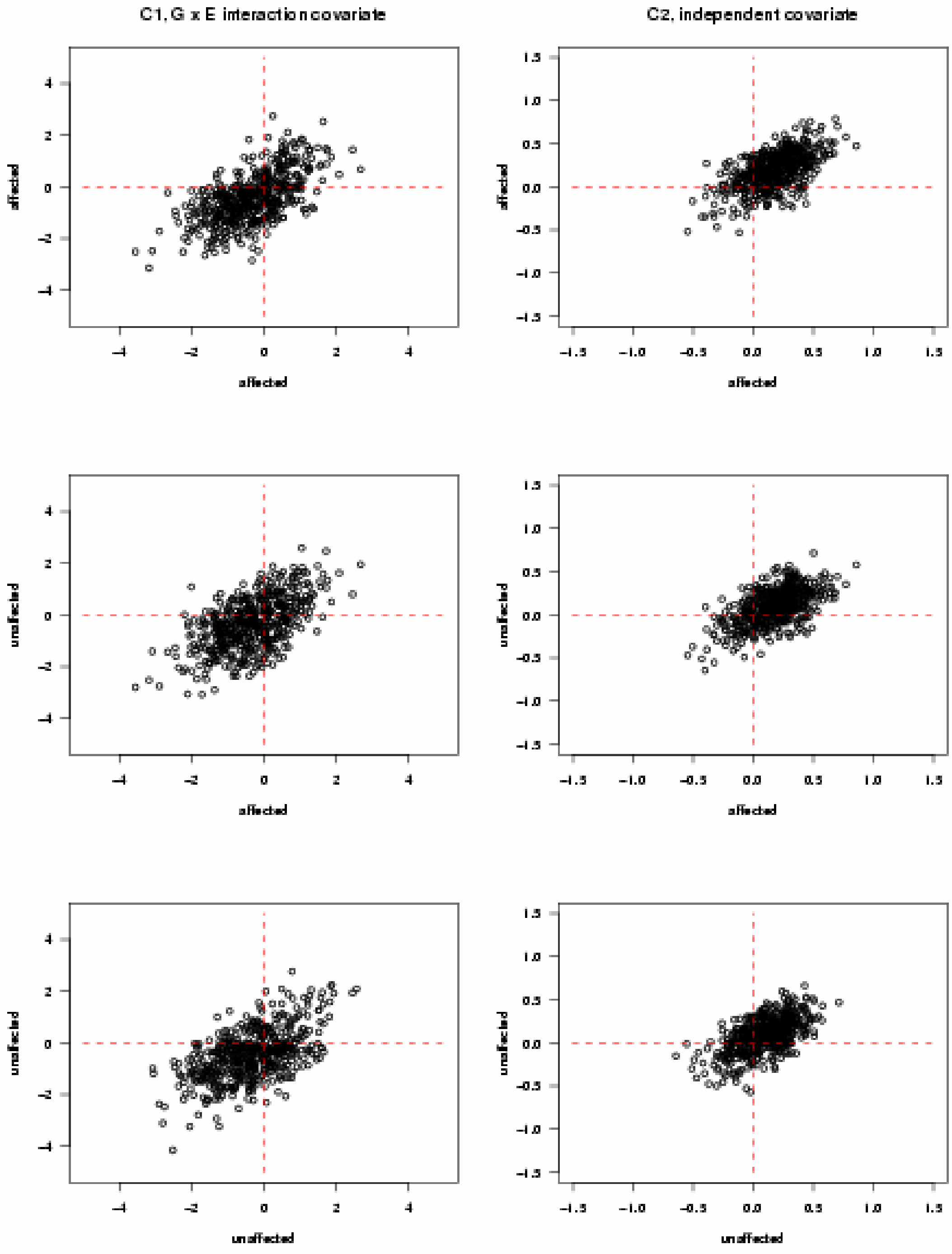


Figure 4.17 Covariate values in different sib pairs under Type II dominant model with disease allele frequencies equal to 0.01 and genetic variance equal to 30%, based on 500 pedigrees

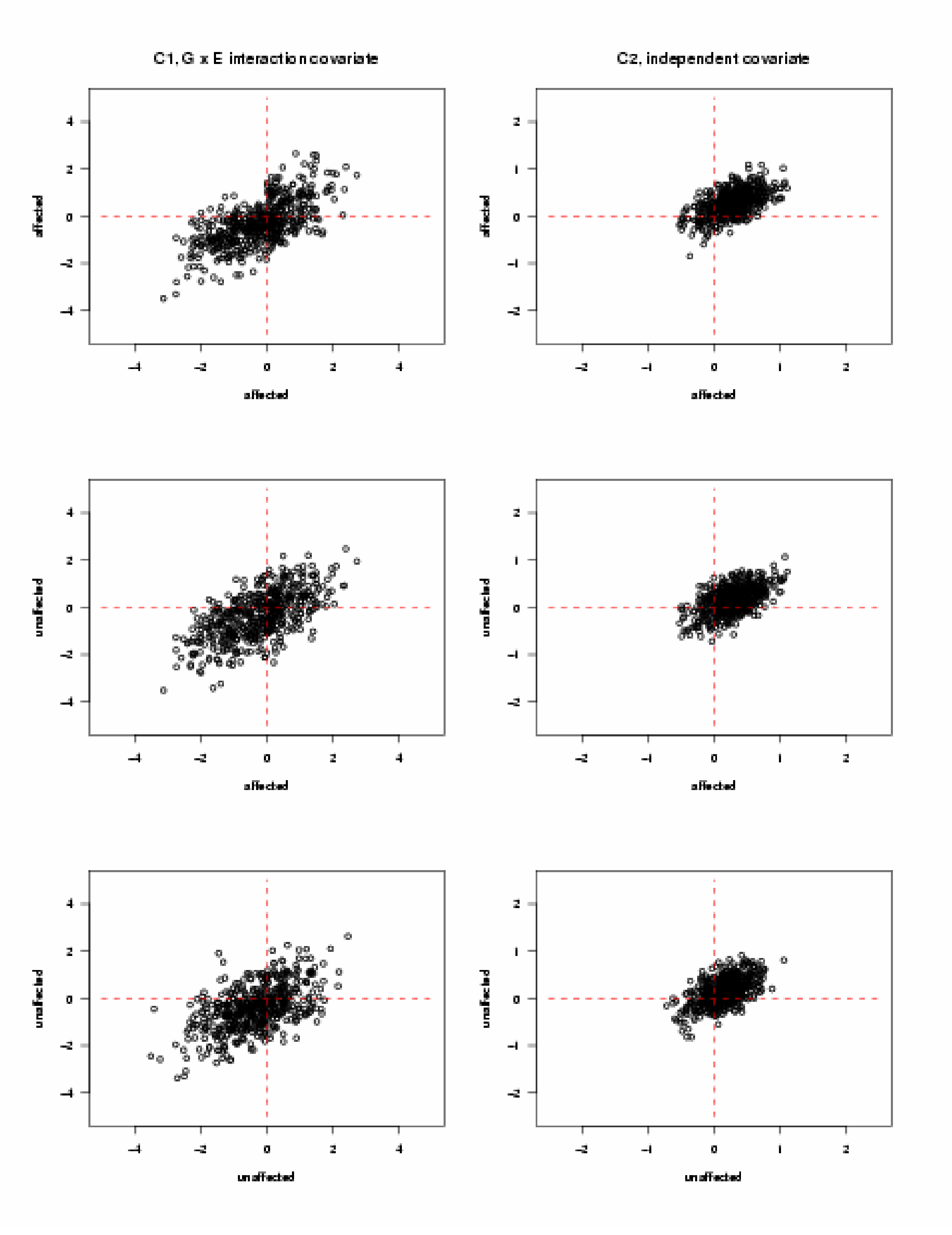


Figure 4.18 Covariate values in different sib pairs under Type II dominant model with disease allele frequencies equal to 0.02 and genetic variance equal to 30%, based on 500 pedigrees

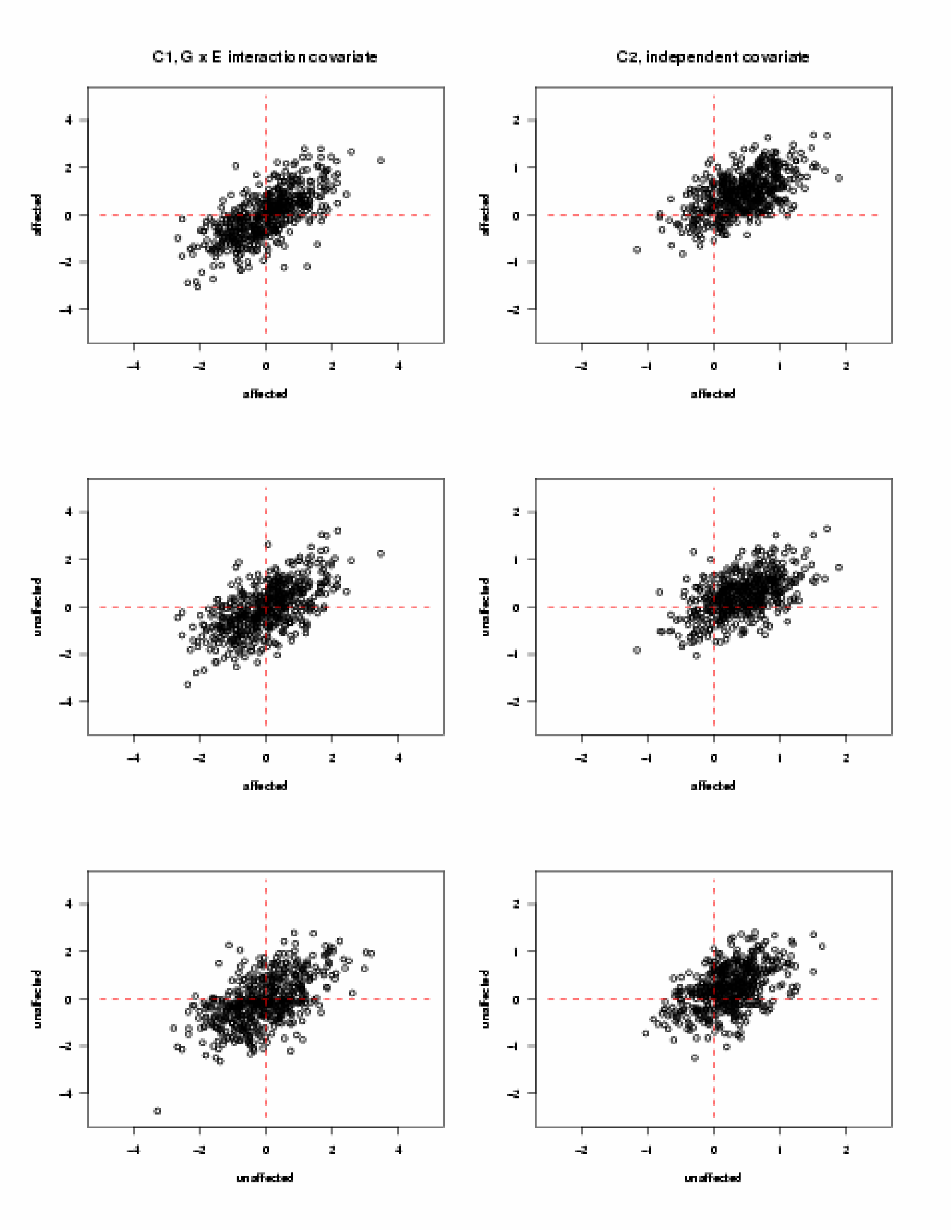


Figure 4.19 Covariate values in different sib pairs under Type II dominant model with disease allele frequencies equal to 0.05 and genetic variance equal to 30%, based on 500 pedigrees

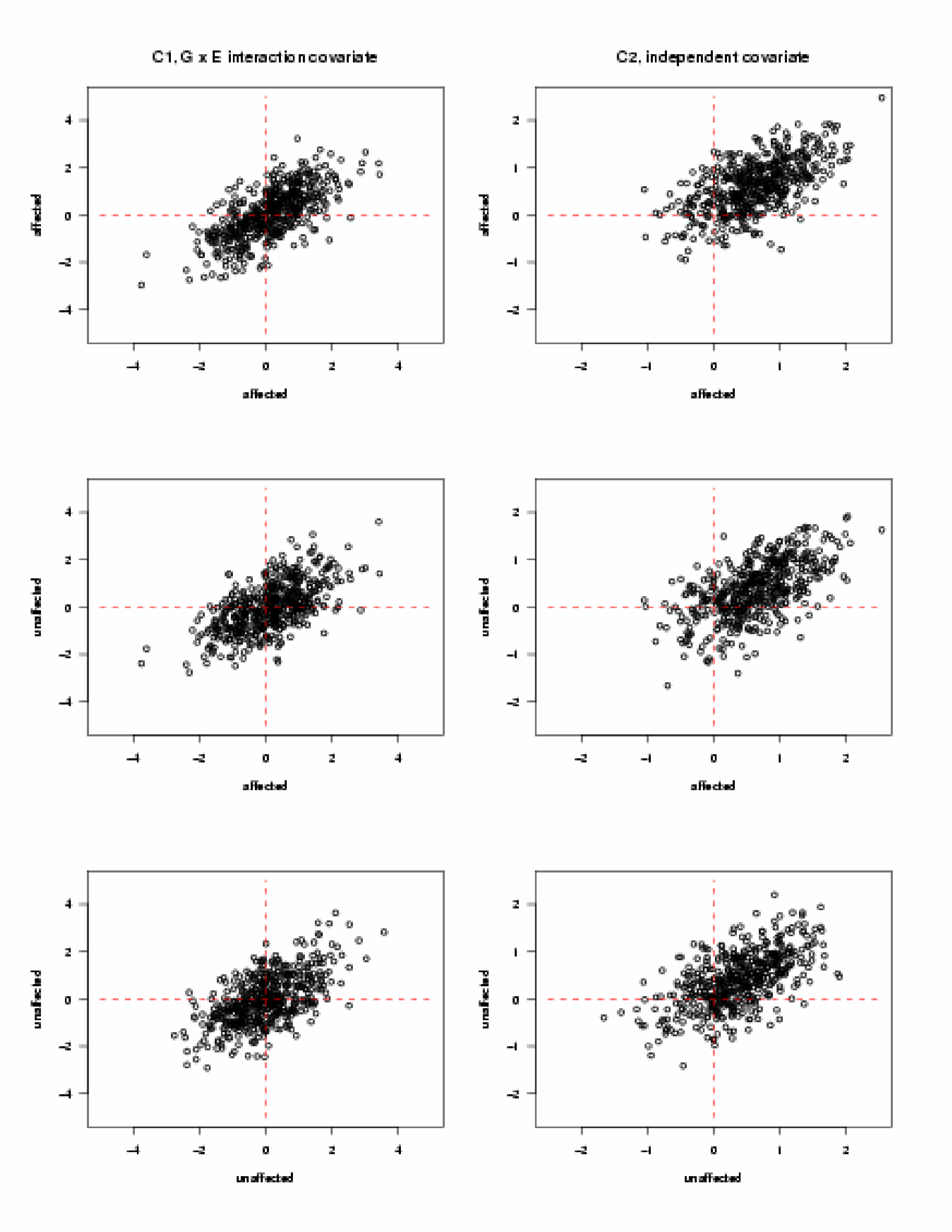


Figure 4.20 Covariate values in different sib pairs under Type II dominant model with disease allele frequencies equal to 0.1 and genetic variance equal to 30%, based on 500 pedigrees

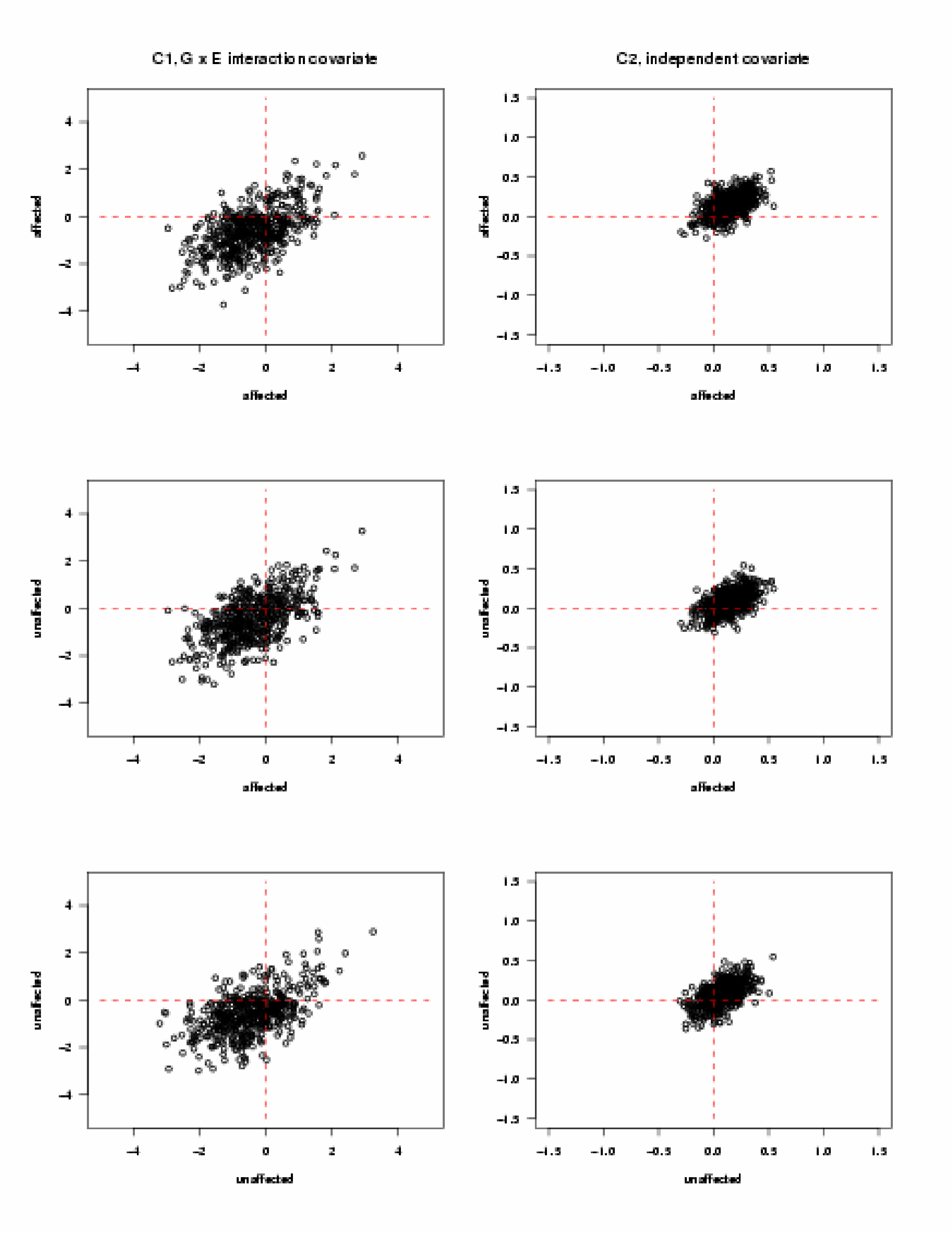


Figure 4.21 Covariate values in different sib pairs under Type II recessive model with disease allele frequencies equal to 0.1 and genetic variance equal to 30%, based on 500 pedigrees

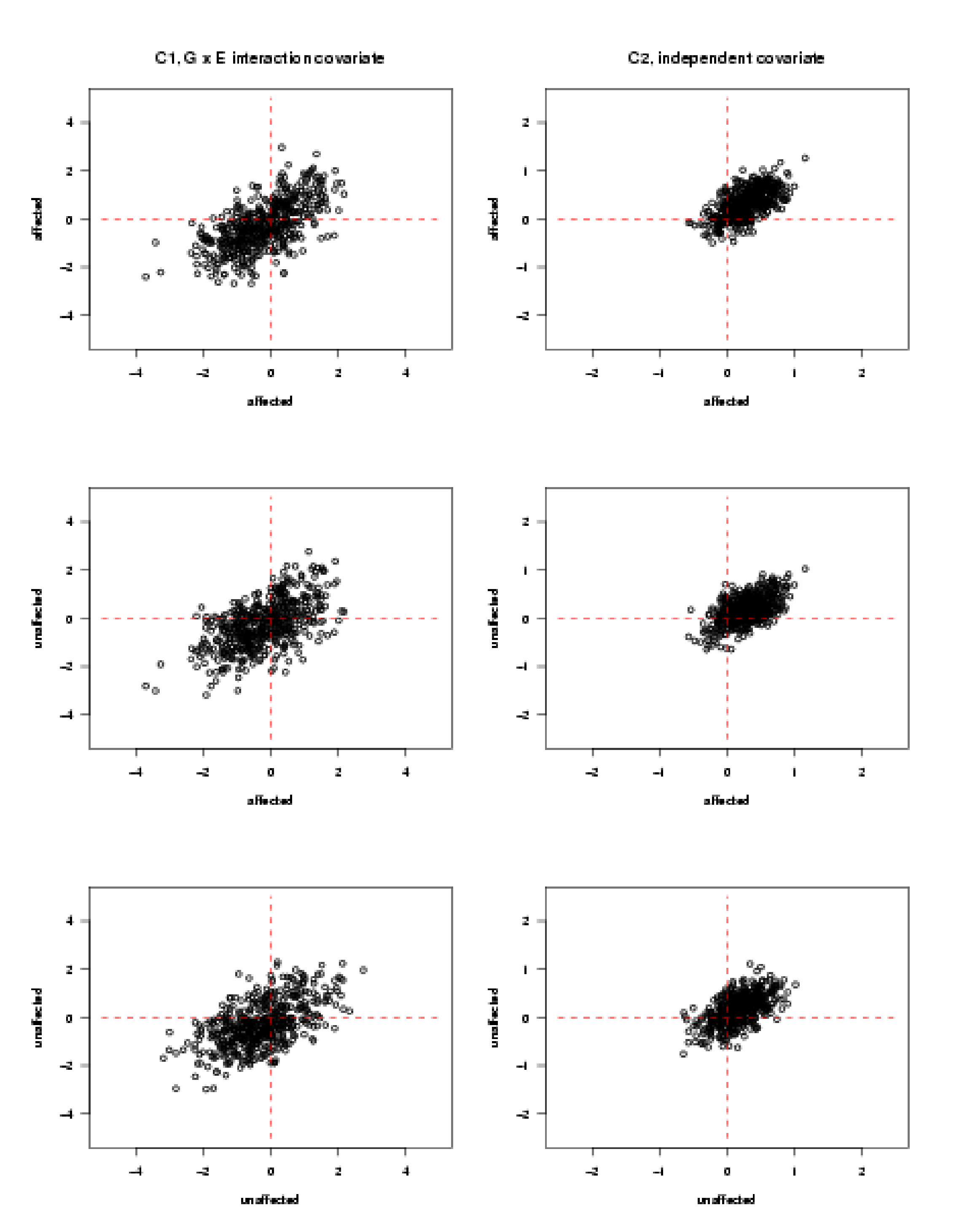


Figure 4.22 Covariate values in different sib pairs under Type II recessive model with disease allele frequencies equal to 0.2 and genetic variance equal to 30%, based on 500 pedigrees

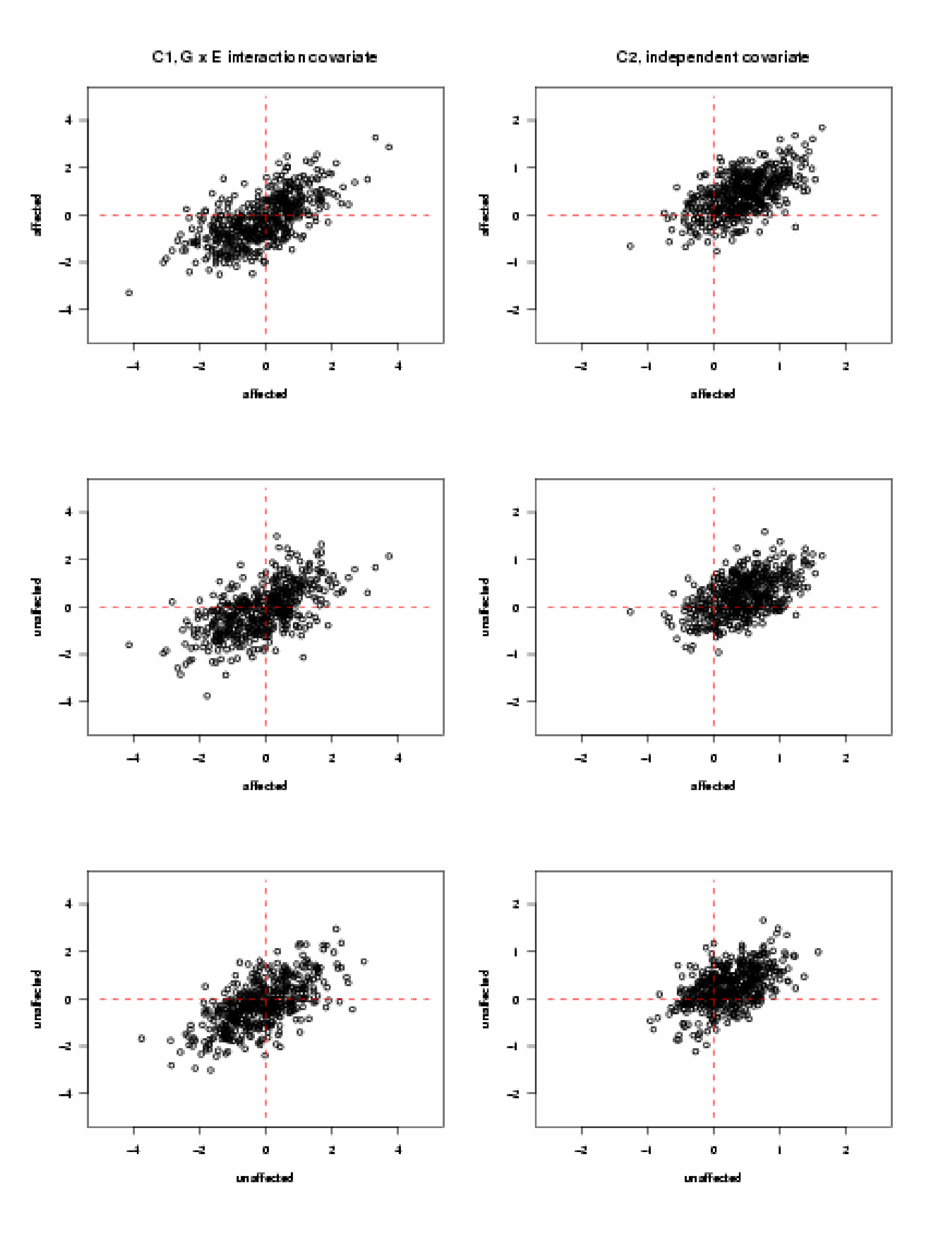


Figure 4.23 Covariate values in different sib pairs under Type II recessive model with disease allele frequencies equal to 0.3 and genetic variance equal to 30%, based on 500 pedigrees

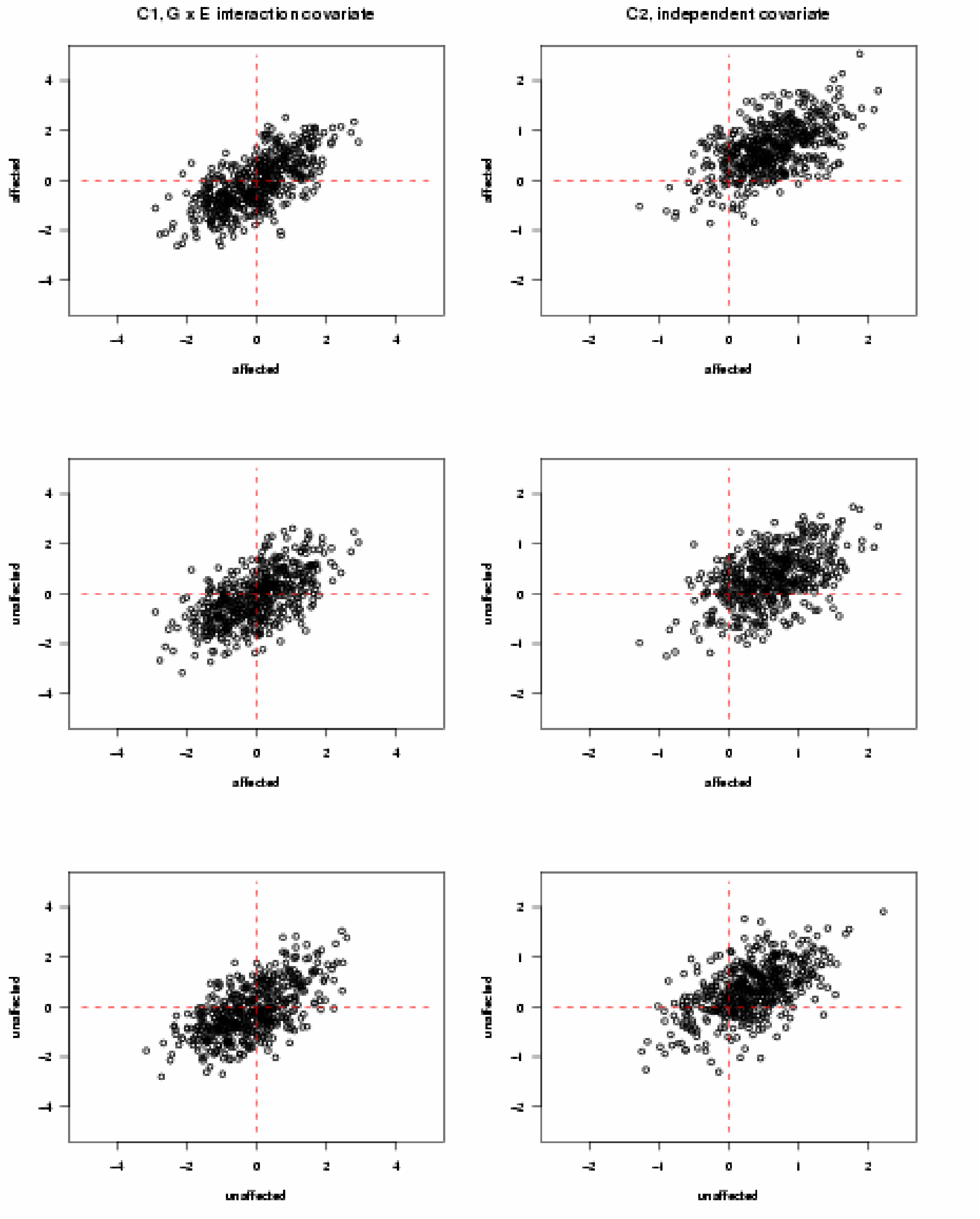


Figure 4.24 Covariate values in different sib pairs under Type II recessive model with disease allele frequencies equal to 0.4 and genetic variance equal to 30%, based on 500 pedigrees

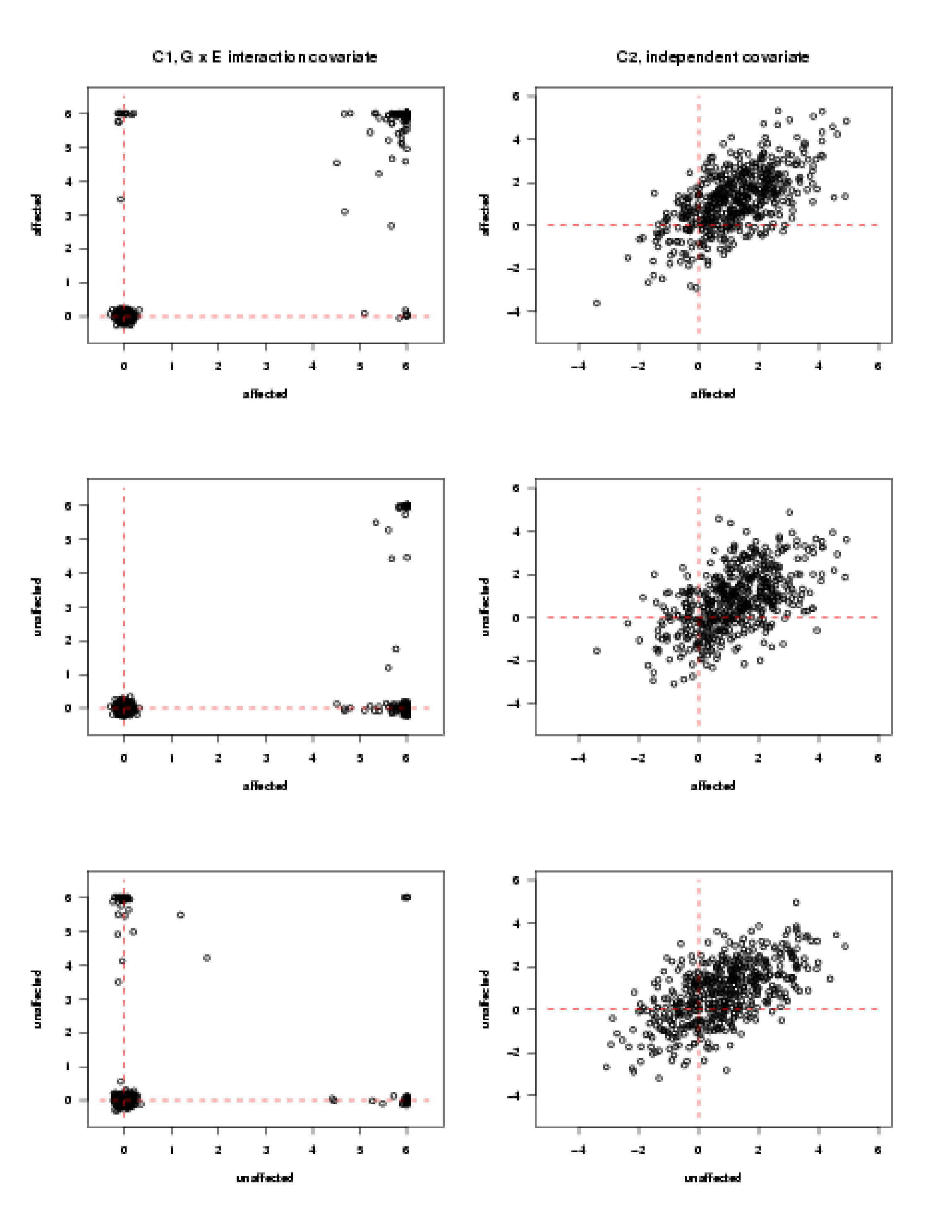


Figure 4.25 Covariate values in different sib pairs under Type III dominant model with disease allele frequencies equal to 0.01, based on 500 pedigrees

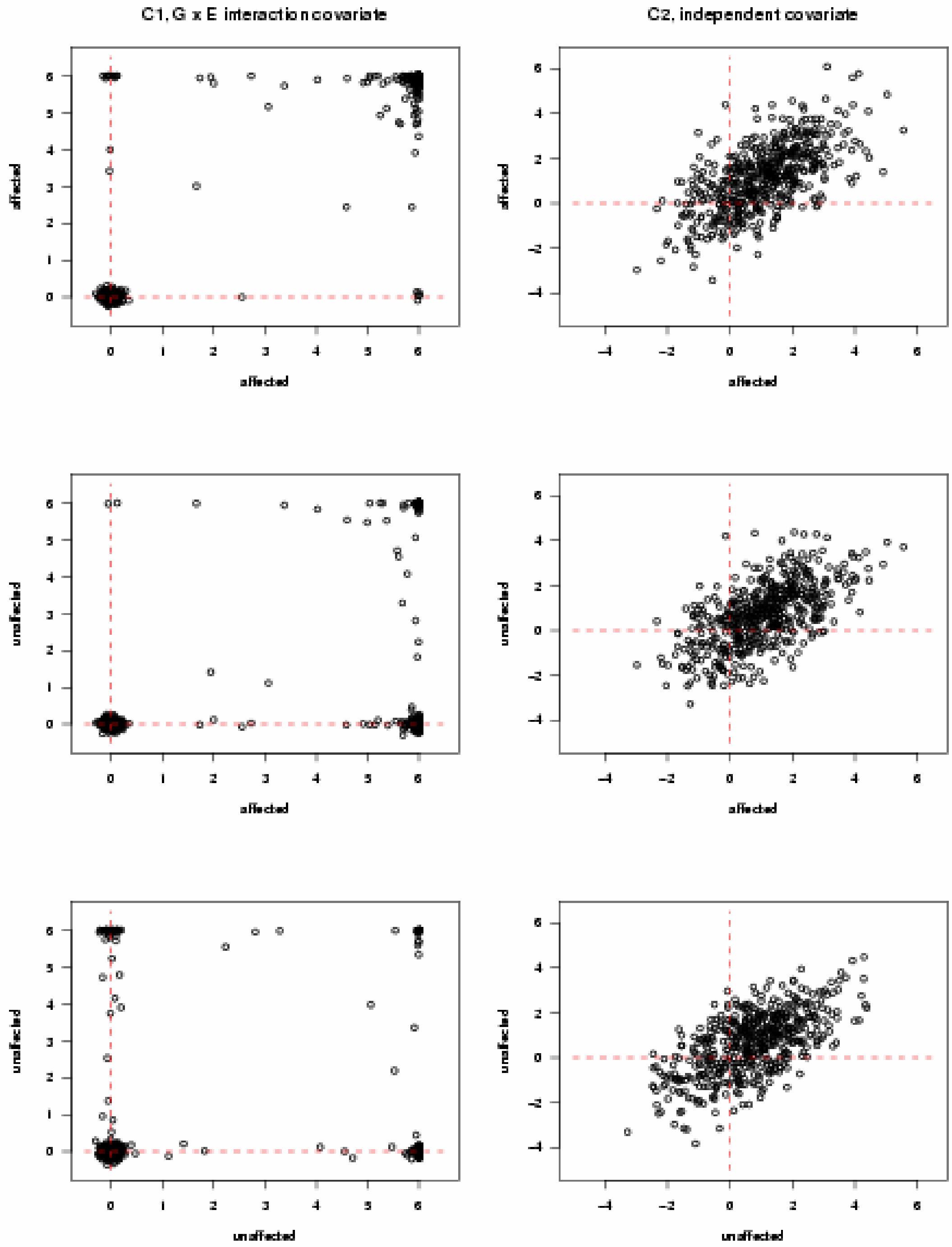


Figure 4.26 Covariate values in different sib pairs under Type III dominant model with disease allele frequencies equal to 0.02, based on 500 pedigrees

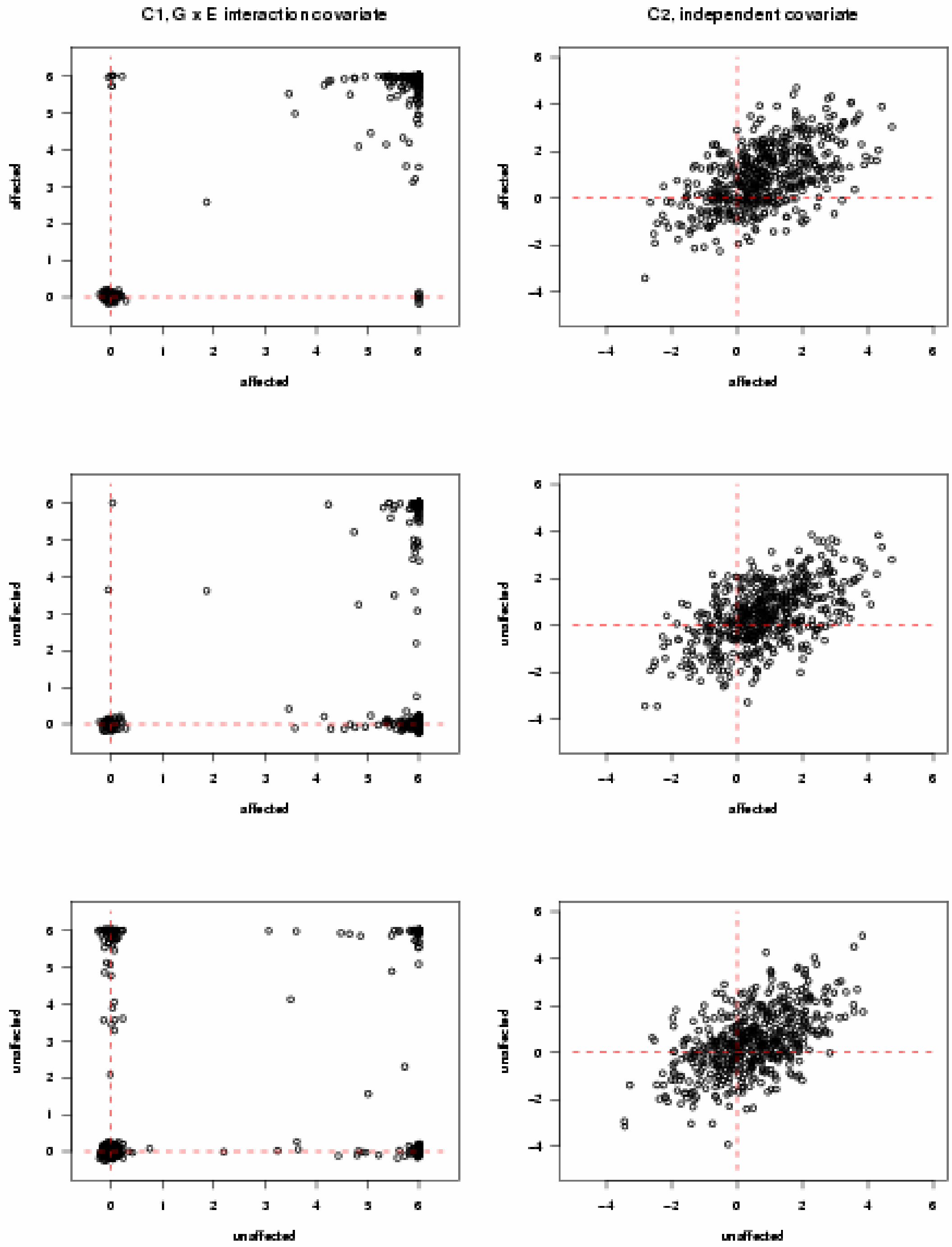


Figure 4.27 Covariate values in different sib pairs under Type III dominant model with disease allele frequencies equal to 0.05, based on 500 pedigrees

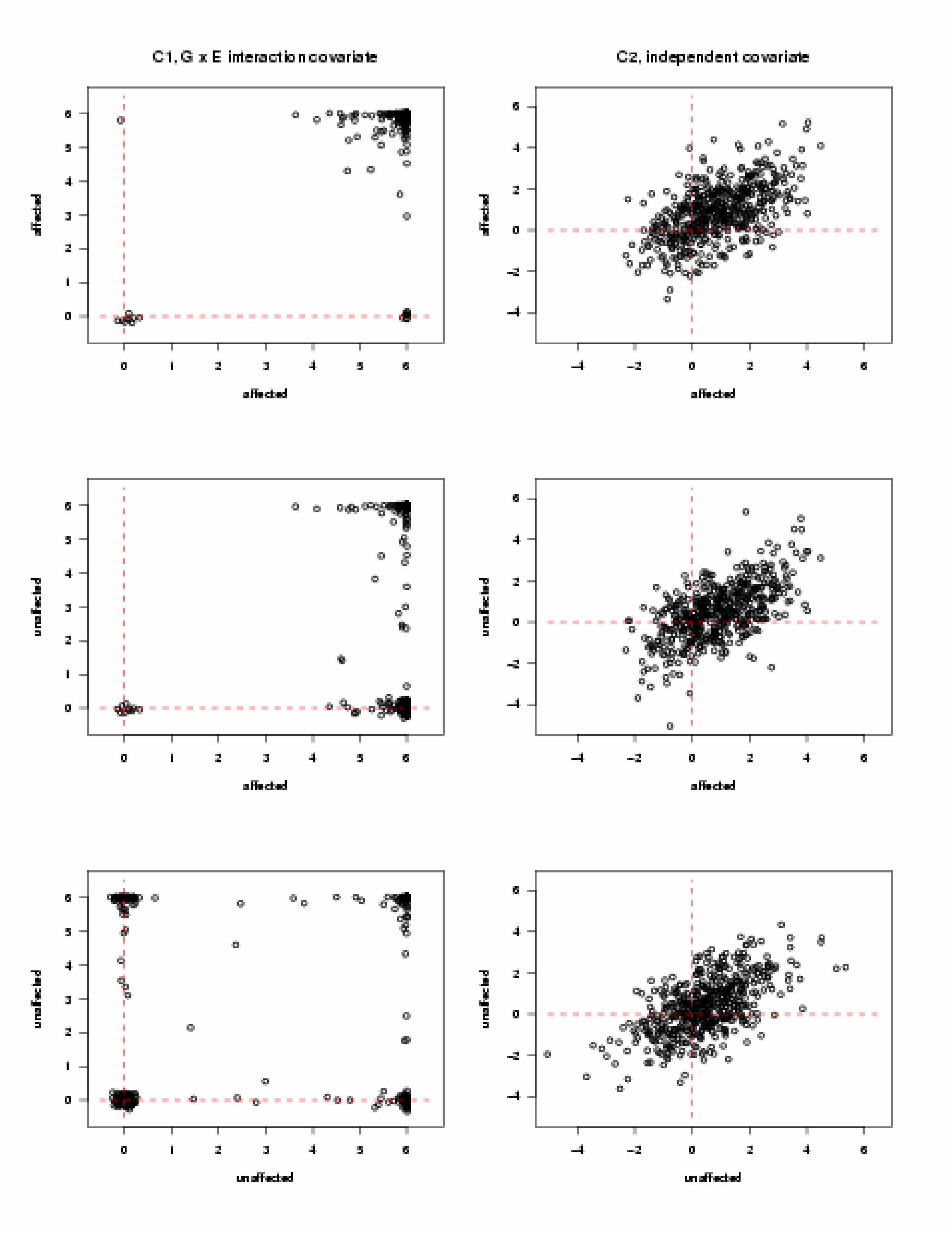


Figure 4.28 Covariate values in different sib pairs under Type III dominant model with disease allele frequencies equal to 0.1, based on 500 pedigrees

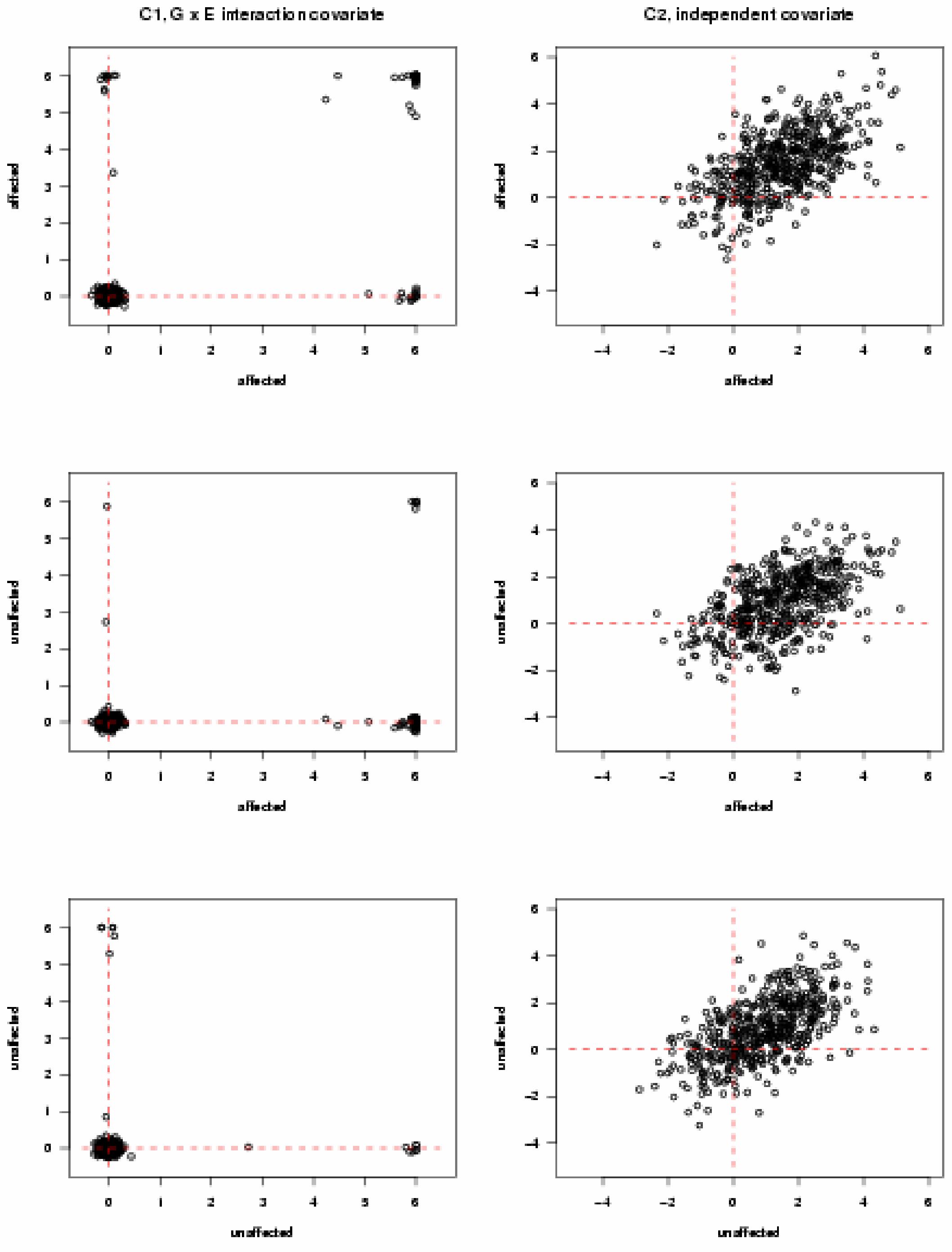


Figure 4.29 Covariate values in different sib pairs under Type III recessive model with disease allele frequencies equal to 0.1, based on 500 pedigrees

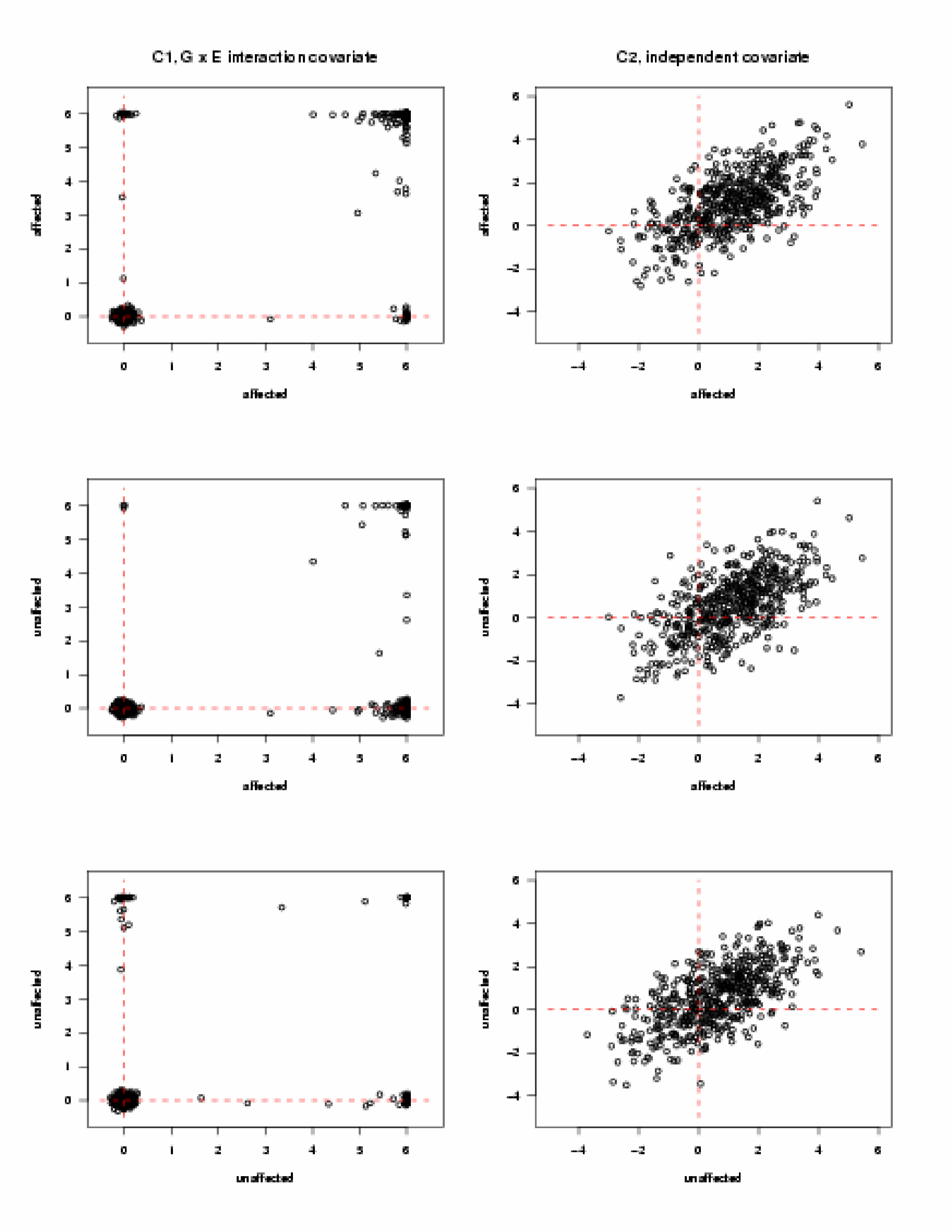


Figure 4.30 Covariate values in different sib pairs under Type III recessive model with disease allele frequencies equal to 0.2, based on 500 pedigrees

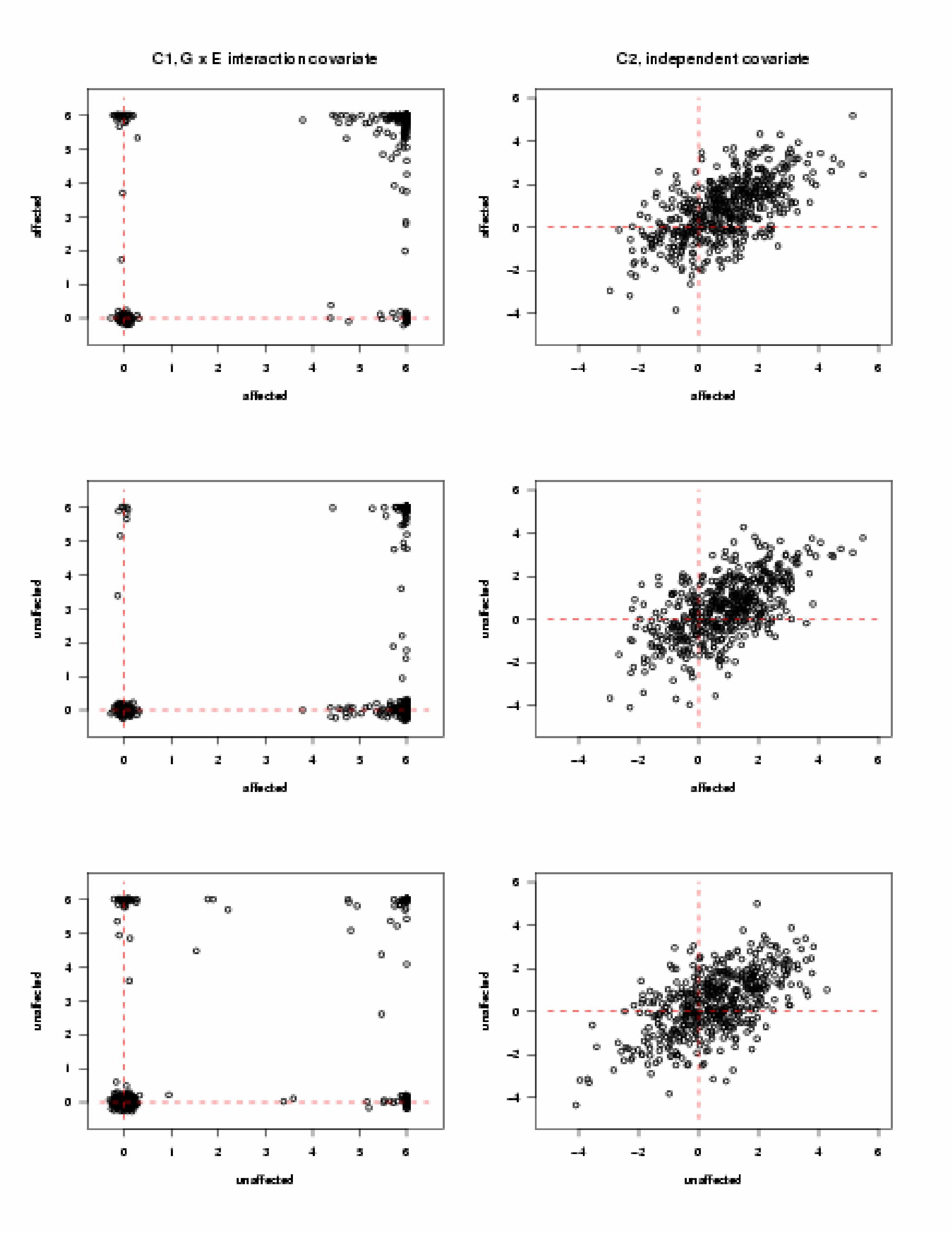


Figure 4.31 Covariate values in different sib pairs under Type III recessive model with disease allele frequencies equal to 0.3, based on 500 pedigrees

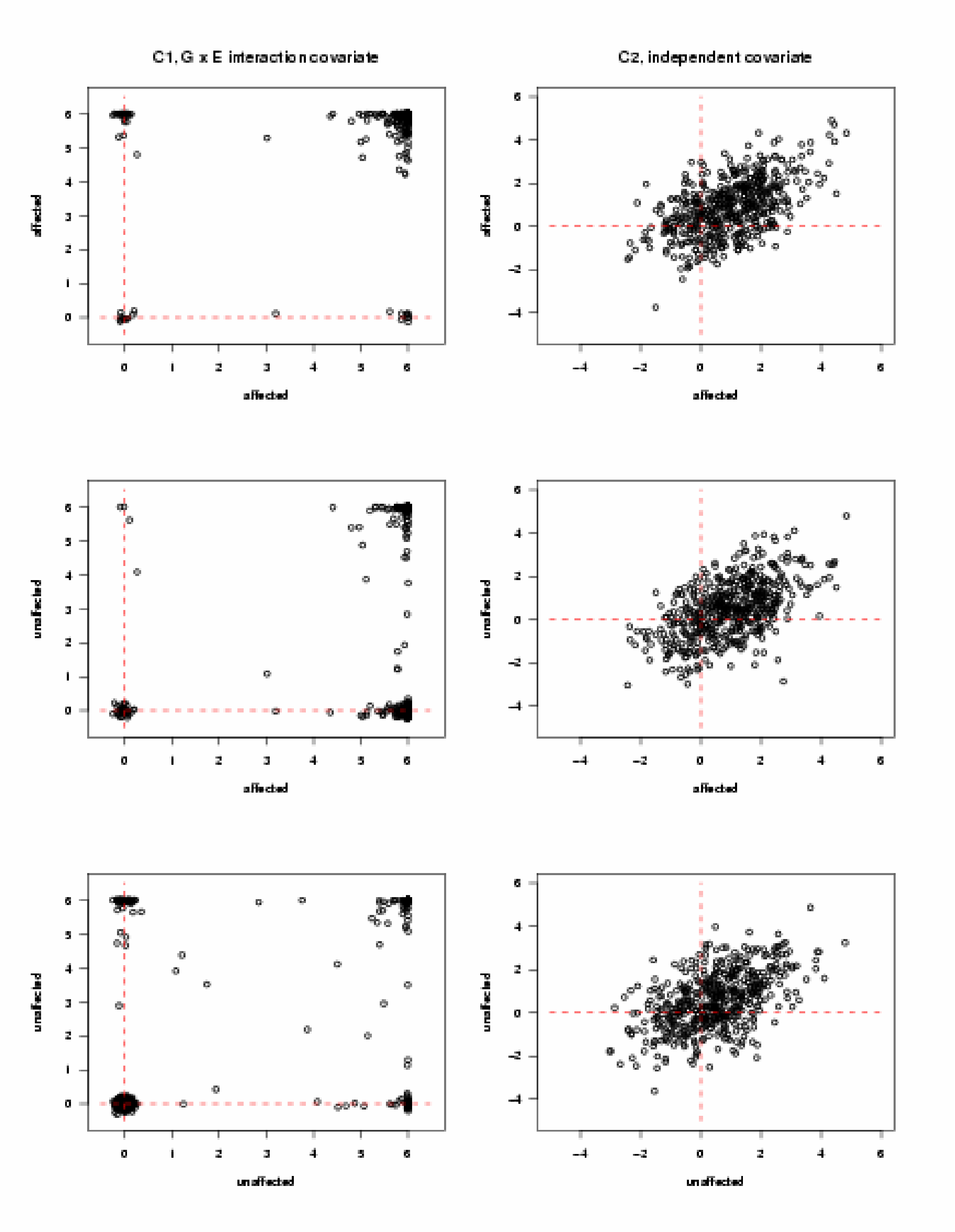


Figure 4.32 Covariate values in different sib pairs under Type III recessive model with disease allele frequencies equal to 0.4, based on 500 pedigrees

5. EMPIRICAL THRESHOLD CALCULATION

For estimating the empirical threshold for statistical significance, we first generate data under the null hypothesis, analyze the simulated data, and record the outputs. We then compute the "chromosome-wide" empirical thresholds for each statistic (Tables 5.2, 3, 4). The scale of the empirical thresholds is the LOD score scale, except for COVLINK, which reports the overall χ^2 values. The abbreviation of each statistic is defined in Table 5.1.

5.1. DATA SETS FOR THRESHOLD ESTIMATION

5.1.1. Covariate statistics

In order to calculate empirical thresholds, we generate three data sets under the null hypothesis of no linkage. The first set and the second set contain 8,000 replicates, respectively: 1,000 replicates per disease model under the Type I G x E interaction model; but C2 in the first set with familial correlation 0.8, and C2 in the second set without familial correlation. We then analyze these two sets by incorporating C2, separately. The third set has only 4,000 replicates: 1,000 replicates per model (2 dominant and 2 recessive models) under the Type I model. Likewise, we analyze the third set, but using C3.

Table 5.1 Abbreviation for each statistic

Method	Abbreviation
Mixture model	Mixture model
General conditional-logistic model	LODPAL
Multinomial logistic regression model under no dominance assumption	No-dom MLRM
Multinomial logistic regression model under no additive assumption	No-add MLRM
Multinomial logistic regression model using min-max restriction	Min-max MLRM
The extension of Maximum-Likelihood-Binomial linkage approach	MLB
Ordered-subsets analysis using rank order from high to low	H → L OSA
Ordered-subsets analysis using rank order from low to high	L → H OSA
Ordered-subsets analysis using optimal slice option	Optimal-slice OSA
Logistic regression modeling	COVLINK
Allele-sharing statistic using S_{all} score function	S_{all}
Allele-sharing statistic using S_{pairs} score function	S_{pairs}
Variance-component linkage analysis	VC
Regression-based quantitative-trait linkage analysis	RB

5.1.2. Model-free and QTL approaches

In addition to covariate statistics, we also calculate the empirical thresholds for the model-free methods and the QTL approaches. We simulate 8,000 replicates under the null hypothesis: 1,000 replicates per model under the Type I model. We analyze these replicates using the model-free methods, ignoring covariate information. For the QTL approaches, we treat three covariates as three quantitative traits, separately, and analyze the same 8,000 replicates.

5.2. THRESHOLDS FOR THE CORRESPONDING FALSE POSITIVE RATES

5.2.1. Thresholds for covariate statistics

We want to compute the empirical thresholds corresponding to the appropriate false positive rates. We first compare the thresholds estimated using C2 with and without familial correlation, respectively. The results indicate that the thresholds at 1%, 5% and 10% levels are similar using C2 with and without familial correlation (Table 5.2A). Moreover, we examine their distributions by generating Q-Q plots (with familial correlation vs. without familial correlation) for each covariate method. Based on Q-Q plots, these two distributions are almost identical (data not shown). Hence we pool all the simulated statistics using C2 (8,000 replicates with and 8,000 replicates without familial correlation), and calculate the overall empirical thresholds based on 16,000 replicates.

We also calculate the thresholds using C3, and compare with the thresholds using C2. According to the results, the thresholds using C2 are very close to those obtained using C3 (Table 5.2B). We compare these two distributions using Q-Q plots (C2 vs. C3). The results indicate the distributions are almost the same for most covariate statistics, except that those for the optimal-slice OSA method are only very close (data not shown). In addition, comparing with the analytical thresholds for some covariate methods, the thresholds using C2 are closer than

Table 5.2A Thresholds of the corresponding false positive rates using the environmental covariate C2 with and without familial correlation, based on 8,000 replicates

Method	10% level		5% level		1% level	
	With	Without	With	Without	With	Without
Mixture model	0.373	0.365	0.628	0.608	1.290	1.253
LODPAL	0.877	0.898	1.199	1.230	2.015	2.100
No-dom MLRM	1.000	0.998	1.298	1.298	2.000	1.999
No-add MLRM	0.976	0.978	1.279	1.272	1.991	1.956
Min-max MLRM	0.991	0.990	1.292	1.284	1.998	1.984
MLB	0.346	0.346	0.572	0.581	1.158	1.189
H → L OSA	0.788	0.783	1.197	1.175	2.003	1.951
L → H OSA	0.788	0.777	1.190	1.171	1.990	1.963
Optimal-slice OSA	1.639	1.593	2.110	2.060	2.923	2.870
COVLINK	6.135	6.101	7.664	7.624	11.263	11.095

those using C3 (Table 5.2B). We therefore apply the thresholds (computed using C2) in Table 5.2B to estimate the power of covariate statistics in Chapter 6.

Table 5.2B Thresholds of the corresponding false positive rates using the environmental covariate C2 or the random noise covariate C3 vs. analytical thresholds

Method	10% level			5% level			1% level		
	C2 [a]	C3 [b]	AT [c]	C2	C3	AT	C2	C3	AT
Mixture model	0.369	0.385	0.356	0.618	0.648	0.587	1.272	1.337	1.175
LODPAL	0.887	0.860	0.794	1.215	1.223	1.068	2.057	2.383	1.720
No-dom MLRM	0.999	0.963	1.000	1.298	1.241	1.301	1.999	1.898	2.000
No-add MLRM	0.977	0.763	1.000	1.276	1.079	1.301	1.974	1.786	2.000
Min-max MLRM	0.990	0.958	1.000	1.288	1.247	1.301	1.991	1.940	2.000
MLB	0.346	0.347	0.356	0.576	0.575	0.587	1.172	1.167	1.175
H → L OSA	0.785	0.802	NA	1.186	1.222	NA	1.977	2.038	NA
L → H OSA	0.782	0.808	NA	1.181	1.224	NA	1.976	2.083	NA
Optimal-slice OSA	1.616	1.733	NA	2.085	2.215	NA	2.898	3.040	NA
COVLINK	6.118	6.126	NA	7.644	7.719	NA	11.188	11.399	NA

Note:

a. Pooled data set with 16,000 replicates
b. Data set with 4,000 replicates
c. Analytical threshold at the corresponding FPR

5.2.2. “Chromosome-wide” thresholds vs. “point-specific” thresholds

We herein compute the “chromosome-wide” thresholds, instead of “point-specific” thresholds. Outputs of most covariate methods have 161 data points, which are approximately spaced every 1 cM along the chromosome, except the OSA methods have 153 points, which do not have information at the last 8 points, and COVLINK only has 33 points, which are spaced evenly every 5 cM. We calculate empirical thresholds using all points, and using a single point at 4 different positions (the first, middle, the last, and the disease locus location), separately. We then compare the thresholds obtained from both approaches. The results show that the “chromosome-wide” thresholds at 1%, 5% and 10% levels are almost the same as the “point-specific” thresholds (Tables 5.3A, B, C).

If there is no correlation between the adjacent points, all points along the chromosome are completely independent. In this situation, the “chromosome-wide” thresholds should be very close to the “point-specific” thresholds, because the “chromosome-wide” thresholds are the pools of all the independent “point-specific” thresholds. However, the data points are spaced approximately either every 1 cM or every 5 cM along the chromosome. Hence, data correlation issues cannot be neglected. But, in contrast to no correlation, if the adjacent points are

Table 5.3A Comparison of chromosome-wide and point-specific thresholds at 10% level, based on 16,000 replicates

Method	Empirical threshold at 10% level				
	All points	First point	Locus*	Middle	End point
Mixture model	0.369	0.363	0.373	0.379	0.365
LODPAL	0.887	0.883	0.885	0.903	0.873
No-dom MLRM	0.999	1.008	1.013	0.998	1.007
No-add MLRM	0.977	0.972	1.003	0.976	0.980
Min-max MLRM	0.990	0.992	1.001	0.988	0.996
MLB	0.346	0.343	0.330	0.350	0.344
H → L OSA	0.785	0.748	0.772	0.768	0.735
L → H OSA	0.782	0.717	0.761	0.794	0.726
Optimal-slice OSA	1.616	1.476	1.630	1.656	1.526
COVLINK	6.118	6.020	6.199	6.058	6.059

Note:

* The disease locus position is at 50 cM along the chromosome, but it is a "non-existent" locus in the null data.

Table 5.3B Comparison of chromosome-wide and point-specific thresholds at 5% level, based on 16,000 replicates

Method	Empirical threshold at 5% level				
	All points	First point	Locus	Middle	End point
Mixture model	0.618	0.624	0.631	0.656	0.607
LODPAL	1.215	1.195	1.229	1.220	1.203
No-dom MLRM	1.298	1.309	1.322	1.289	1.307
No-add MLRM	1.276	1.276	1.314	1.268	1.274
Min-max MLRM	1.288	1.276	1.318	1.287	1.294
MLB	0.576	0.553	0.558	0.578	0.563
H → L OSA	1.186	1.197	1.182	1.176	1.197
L → H OSA	1.181	1.172	1.172	1.201	1.188
Optimal-slice OSA	2.085	2.068	2.138	2.094	2.044
COVLINK	7.644	7.504	7.679	7.612	7.624

Table 5.3C Comparison of chromosome-wide and point-specific thresholds at 1% level, based on 16,000 replicates

Method	Empirical threshold at 1% level				
	All points	First point	Locus	Middle	End point
Mixture model	1.272	1.266	1.361	1.337	1.221
LODPAL	2.057	2.074	2.081	2.074	2.074
No-dom MLRM	1.999	2.017	2.080	2.000	1.992
No-add MLRM	1.974	1.954	1.959	1.926	1.990
Min-max MLRM	1.991	2.007	2.034	1.970	1.964
MLB	1.172	1.158	1.153	1.180	1.183
H → L OSA	1.977	2.017	2.052	1.961	2.034
L → H OSA	1.976	2.001	1.921	2.014	1.970
Optimal-slice OSA	2.898	2.895	3.004	2.884	2.906
COVLINK	11.188	10.997	11.193	11.113	11.411

completely correlated, the “chromosome-wide” thresholds then would be identical to the “point-specific” thresholds. When the data points are “completely correlated”, or “completely not correlated”, the thresholds obtained from both approaches should be the same, or almost the same. Hence, this assures that the “chromosome-wide” thresholds that we use for power evaluation are appropriate.

Unlike the other points, the IBD sharing probabilities at the end point (either the first or the last point) are estimated only based on the data between two points. One might expect that the thresholds estimated from the first or the last point are smaller than those from the other points. However, according to our results, we do not observe such end effects (Tables 5.3A, B, C).

5.2.3. Thresholds for model-free and QTL approaches

For the model-free methods (S_{all} and S_{pairs}) and the QTL approaches (VC and RB), we compute the empirical thresholds at 1%, 5% and 10% levels (Table 5.4). The empirical thresholds of S_{all}, S_{pairs} and RB (C1, C2 and C3) correspond to the expected analytical p-values, but not for VC.

We do not normalize covariate values when we analyze the data using VC. Because VC is very sensitive to departure from the normal distribution assumption, the empirical thresholds may be biased. Since C3 is generated from one of three different normal distributions (details provided in Section 3.2.2), the overall C3 distribution is not followed a normal distribution. Therefore, the empirical thresholds for VC using C3 as a quantitative trait are not close to the expected analytical thresholds (Table 5.4).

Table 5.4 Thresholds for the model-free methods and QTL approaches, based on 8,000 replicates

Method	10% level	p-value [a]	5% level	p-value	1% level	p-value
S_{all}	0.36	0.10	0.60	0.05	1.17	0.01
S_{pairs}	0.36	0.10	0.59	0.05	1.18	0.01
VC - C1 [b]	0.51	0.06	0.86	0.02	1.76	0.002
VC - C2	0.57	0.05	0.94	0.02	1.94	0.001
VC - C3	0.80	0.03	1.36	0.01	2.79	0.0002
RB - C1	0.386	0.09	0.659	0.04	1.34	0.006
RB - C2	0.371	0.10	0.623	0.05	1.26	0.008
RB - C3	0.368	0.10	0.613	0.05	1.27	0.008

Note:

a. Analytical p-value corresponding to the empirical threshold

b. C1, C2 and C3 are three different environmental covariates defined in Chapter 1

5.3. ANALYTICAL THRESHOLDS VS. EMPIRICAL THRESHOLDS

As we mentioned in Chapter 2, most papers of covariate methods describe the analytical distribution. Therefore we compared our empirical thresholds with the analytical thresholds at 1%, 5% and 10% significance levels (Table 5.5). Based on our results, the empirical thresholds of the mixture model, the MLRM approaches and MLB are close to the analytical thresholds, but the empirical threshold of the mixture model at 1% level is slightly higher than the analytical threshold. However, the analytical thresholds of LODPAL are markedly smaller than the empirical thresholds at all three levels. When we examine the results in Q-Q plots (empirical vs. analytical), as the distribution gets closer to the tail, the empirical thresholds of LODPAL deviate more from the analytical thresholds (Figure 5.1). One may easily reach false-positive conclusions if using the analytical p-values of LODPAL.

Table 5.5 Asymptotic distribution vs. empirical distribution in covariate methods

Method	10% level			5% level			1% level		
	ET [a]	p-value [b]	AT [c]	ET	p-value	AT	ET	p-value	AT
Mixture model	0.369	0.096	0.356	0.618	0.046	0.587	1.272	0.0078	1.175
LODPAL	0.887	0.086	0.794	1.215	0.039	1.068	2.057	0.0054	1.720
No-dom MLRM	0.999	0.100	1.000	1.298	0.050	1.301	1.999	0.0100	2.000
No-add MLRM	0.977	0.105	1.000	1.276	0.053	1.301	1.974	0.0106	2.000
Min-max MLRM	0.990	0.102	1.000	1.288	0.052	1.301	1.991	0.0102	2.000
MLB	0.346	0.103	0.356	0.572	0.052	0.587	1.172	0.0101	1.175
H → L OSA	0.785	NA	NA	1.186	NA	NA	1.977	NA	NA
L → H OSA	0.782	NA	NA	1.181	NA	NA	1.976	NA	NA
Optimal-slice OSA	1.616	NA	NA	2.085	NA	NA	2.898	NA	NA
COVLINK	6.118	NA	NA	7.644	NA	NA	11.188	NA	NA

Note:

a. Empirical threshold
b. Analytical p-value corresponding to the empirical threshold
c. Analytical threshold

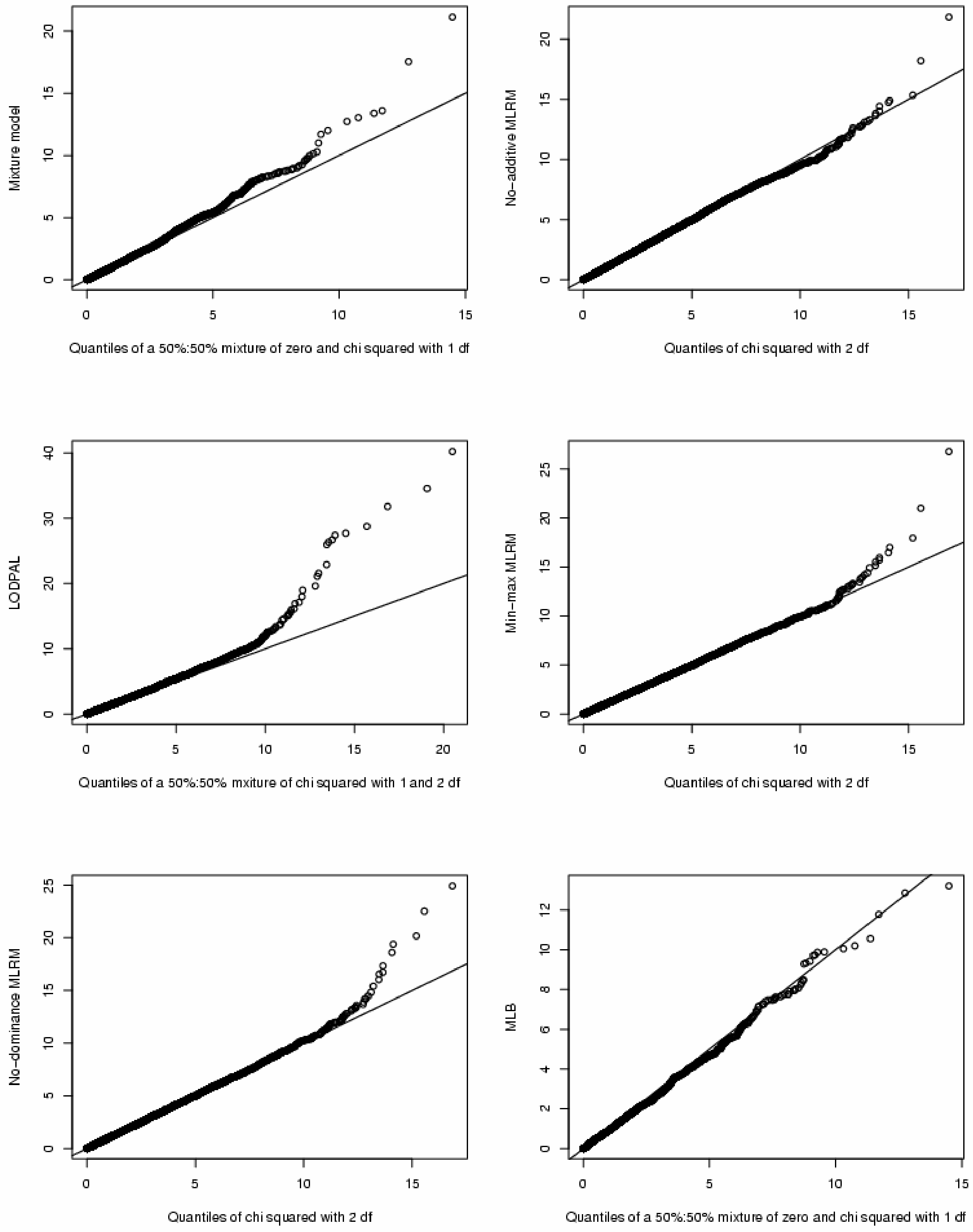


Figure 5.1 QQ plots of empirical distribution vs. analytical distribution for covariate methods

6. POWER EVALUATION

The empirical thresholds corresponding to false positive rates of 1%, 5% and 10% were described in the Chapter 5. The thresholds obtained using C2 at the 5% level are used to evaluate power. The power for each method is based on 500 replicates. We here present power of the covariate statistics, the model-free methods, and the QTL approaches (Figures 6.1 – 6.8). The three columns in each figure represent: (column 1) using the gene-environment interaction covariate, the "right" covariate C1; (column 2) using the independent environmental covariate, the "wrong" covariate C2; (column 3) using the random noise covariate, covariate C3. As described in Chapter 1, allele-sharing statistic, S_{all} score function, performs quite well on average across a wide variety of genetic models (Sengul et al. 2001). We therefore use S_{all} as the baseline to compare with the covariate methods. The number, symbol and abbreviation representing each statistic in the Figures are provided in Table 6.1. The recurrence risk ratio (λ_s) and the proportion of linked families, as described in Chapter 4, are provided in the Figures as well.

6.1. TYPE I G x E INTERACTION MODEL

6.1.1. Dominant models

Figure 6.1 shows the power and the corresponding 95% confidence interval (CI) of the statistics under the dominant Type I models. Except for the MLB approach and the L → H OSA method,

Table 6.1 Number, symbol and abbreviation of each statistic in texts and figures

Number on x axis	Symbol	Abbreviation	Statistics
1	▭	Mixture model	Mixture model
2	░	LODPAL	General conditional-logistic model
3	▨	No-dominance MLRM	Multinomial logistic regression model under dominance assumption
4	▨	No-additive MLRM	Multinomial logistic regression model under additive assumption
5	▨	Min-max MLRM	Multinomial logistic regression model using the min-max restriction
6	▒	MLB	The extension of Maximum-Likelihood-Binomial linkage approach
7	▓	H → L OSA	Ordered subset analysis; rank order from high to low
8	▓	L → H OSA	Ordered subset analysis; rank order from low to high
9	▓	Optimal-slice OSA	Ordered subset analysis; rank order using the optimal slice method
10	▥	COVLINK	Logistic regression for predicting the IBD sharing probability
11	░	S_{all}	Model-free method
12	░	S_{pairs}	Model-free method
13	⬚	VC	Variance-component linkage analysis
14	⬚	RB	Regression-based quantitative–trait linkage analysis

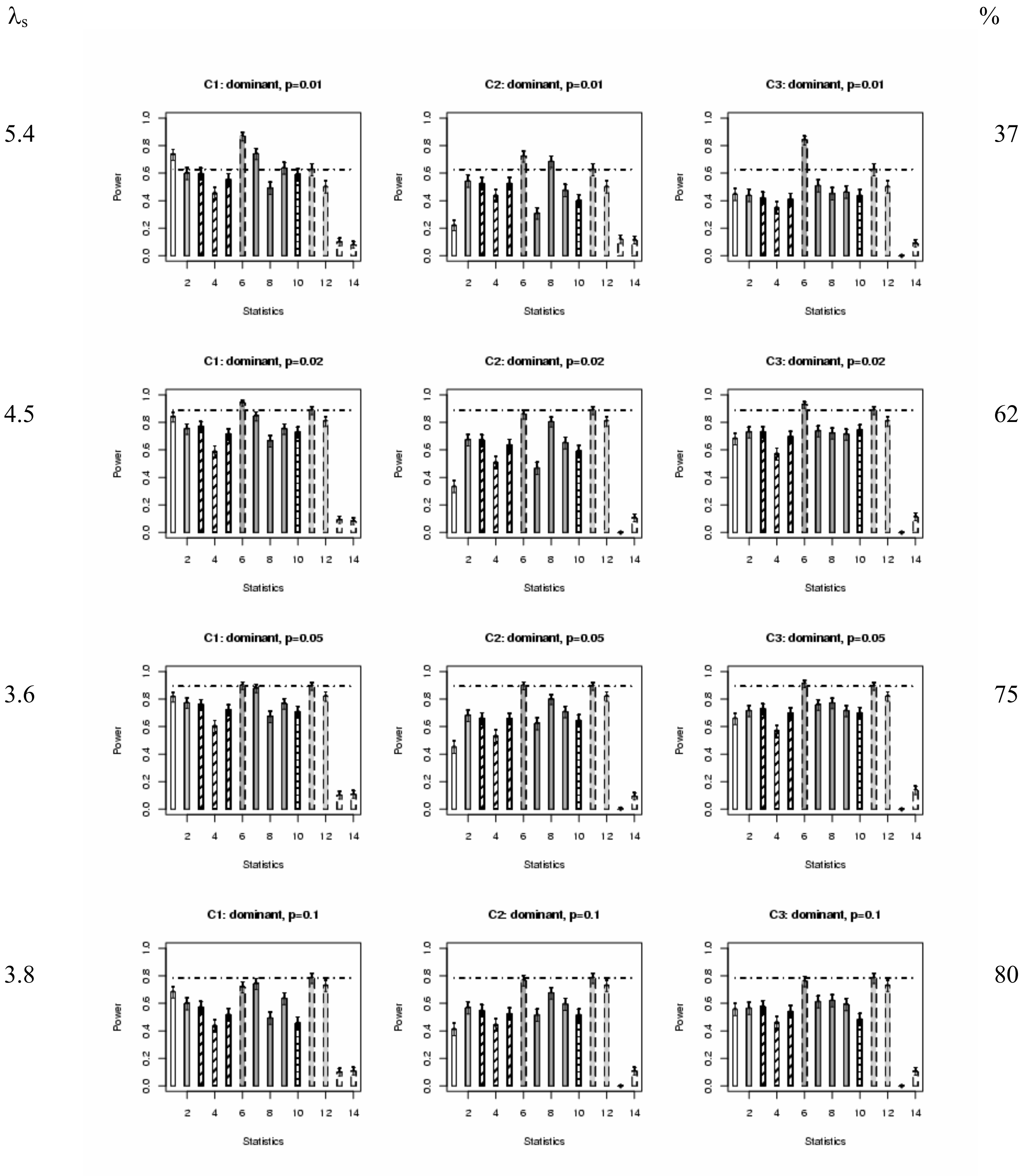
the covariate statistics perform better when using the “right” covariate C1 than when using the “wrong” covariate C2, but not necessarily better than when using the random noise covariate C3.

We then compare the performance of the covariate statistics with S_{all}. When C1 is used, the power of the mixture model, MLB and the H → L OSA method is significantly greater than S_{all}’s power for the rare dominant model (disease allele frequencies, d, = 0.01). The power of LODPAL, no-dominance MLRM, optimal-slice OSA and COVLINK is close to S_{all}’s power for the same model. However, except for the MLB and the H → L OSA method, the other covariate methods have lower power than S_{all} for the models with d = 0.05 and 0.1. When C2 or C3 is used, only the power of MLB is significantly better than S_{all} for the rare model (d = 0.01), and has similar power to S_{all} for the other models (Figure 6.1).

6.1.2. Recessive models

The power of the statistics under the recessive Type I G x E models is shown in Figure 6.2. Except for the L → H OSA method, the covariate statistics have higher power when using the "right" covariate C1 than when using the "wrong" covariate C2, but not necessarily better than when using the random noise covariate C3.

When C1 is used, the power of the mixture model and MLB is significantly higher than S_{all}'s power, and the power of LODPAL, the MLRM methods and the H → L OSA method is almost the same as S_{all}'s power for the model with d = 0.1. The mixture model, LODPAL and MLB have the similar power to S_{all} for the other three models (d = 0.2, 0.3 and 0.4). But the other covariate statistics have lower power than S_{all} in the common model (d = 0.4). When C2 is used, only the MLB method performs as well as S_{all} for all models. When C3 is used, LODPAL and MLB have comparable power to S_{all} for all models (Figure 6.2).


C1: dominant, p=0.01

C2: dominant, p=0.01

C3: dominant, p=0.01

C1: dominant, p=0.02

C2: dominant, p=0.02

C3: dominant, p=0.02

C1: dominant, p=0.05

C2: dominant, p=0.05

C3: dominant, p=0.05

C1: dominant, p=0.1

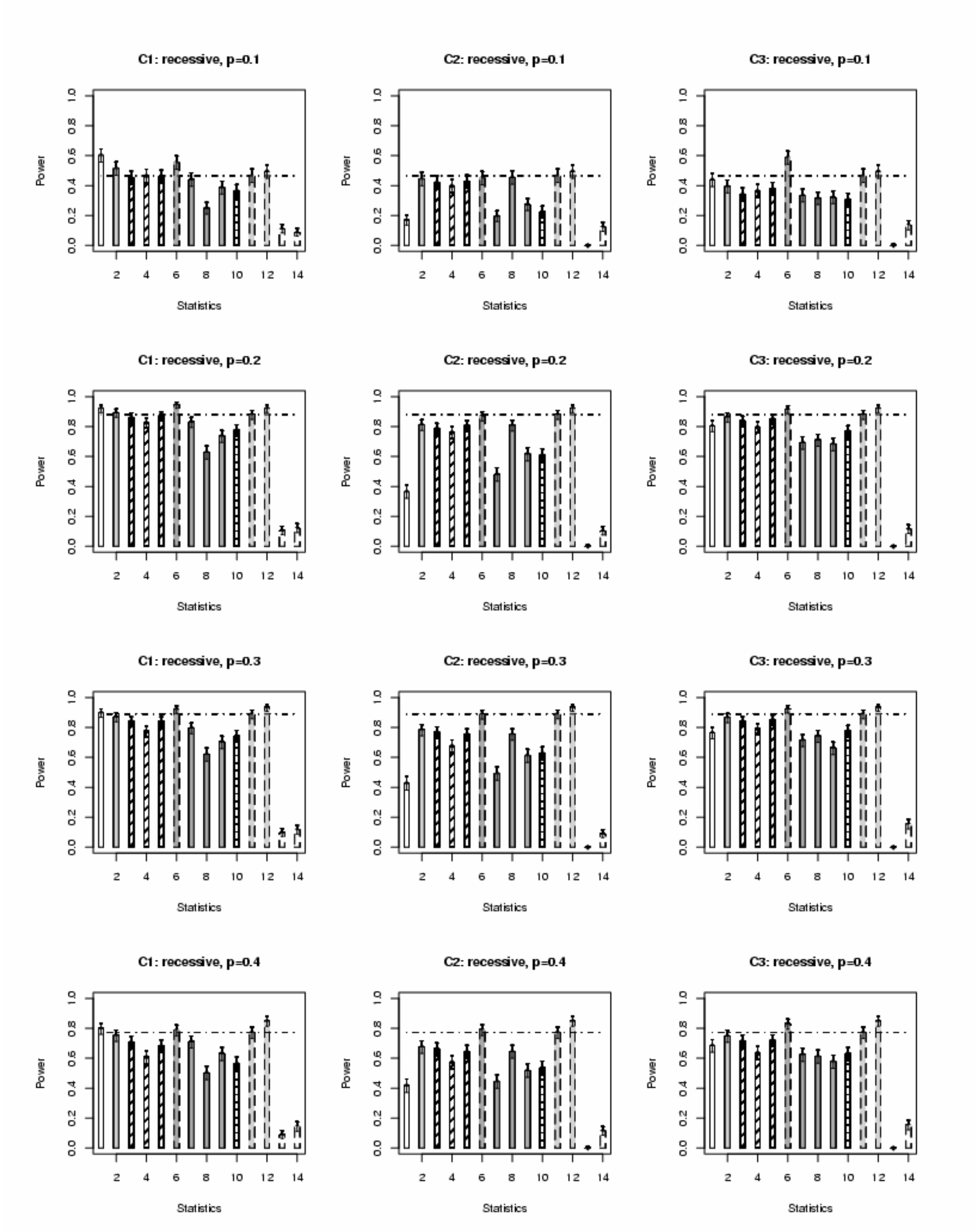
C2: dominant, p=0.1

C3: dominant, p=0.1

λ_s: recurrence risk ratio; %: percentage of linked families

Figure 6.1 Power and 95% CI for each statistic at the 5% level under different dominant Type I models, based on 500 replicates

λs	%
3.4	23
4.0	55
3.7	70
3.6	75



λs: recurrence risk ratio; %: percentage of linked families

Figure 6.2 Power and 95% CI for each statistic at the 5% level under different recessive Type I models, based on 500 replicates

6.2. TYPE II G x E INTERACTION MODEL

For the Type II model, we generate data with two different genetic variances (20% and 30%), separately. The results are discussed in the following sections.

6.2.1. Dominant models

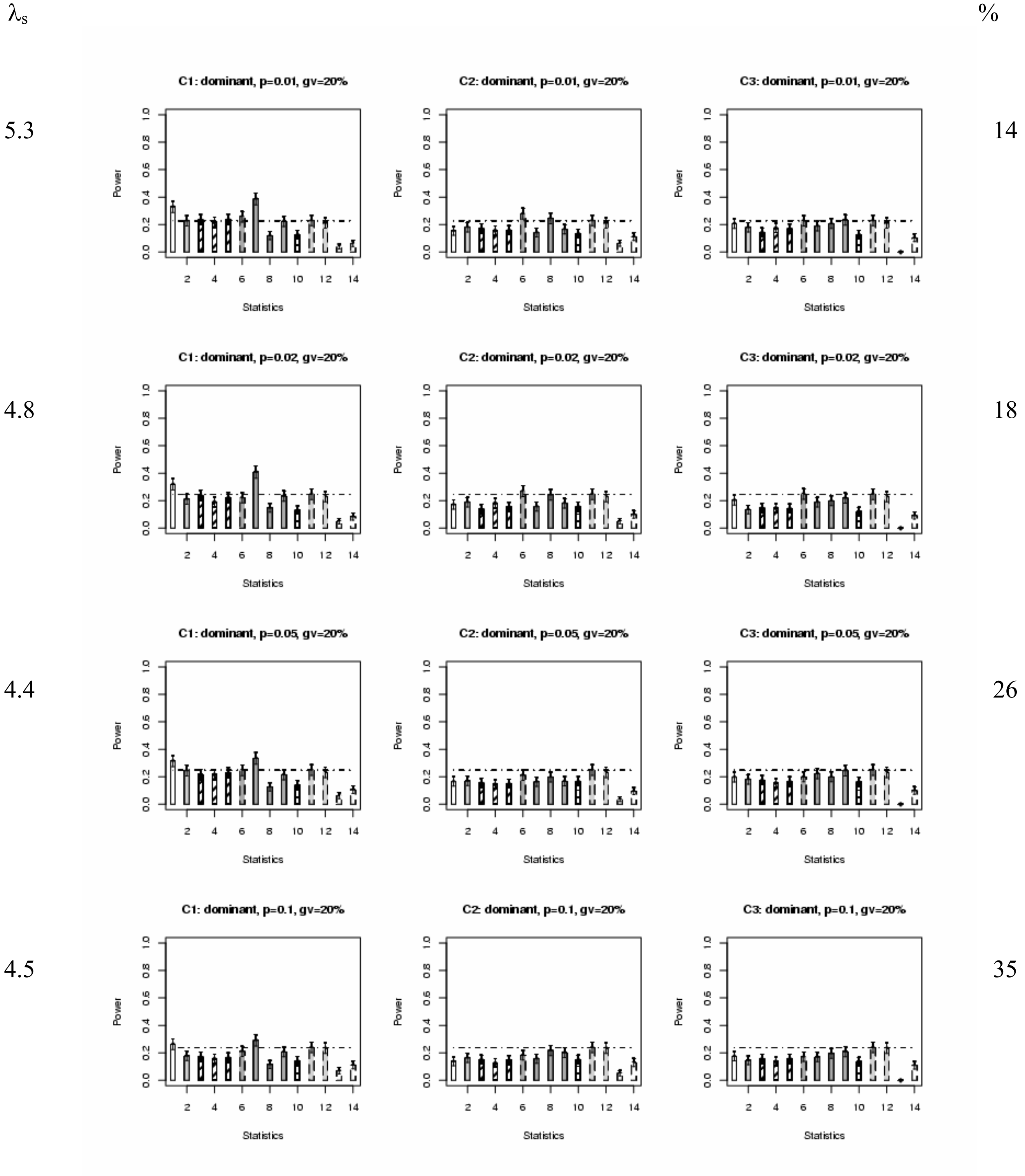
The power of each statistic under the dominant Type II models is presented in Figures 6.3 and 6.4. For both genetic variances (GV = 20% and 30%), the covariate methods have better power when using C1 than when using C2, but not for MLB, the L → H OSA method and COVLINK.

Compared to the S_{all}'s power, the mixture model and the H → L OSA method have significantly higher power than S_{all} for the models with d = 0.01, 0.02 and 0.05 (GV = 20% and 30%). The power of LODPAL, the MLRM approaches, MLB and the optimal-slice OSA method is equivalent to S_{all}'s power for the same models. But when using C2, only MLB and the L → H OSA method perform as well as S_{all} for all models (GV = 20% and 30%) (Figures 6.3, 6.4).

6.2.2. Recessive models

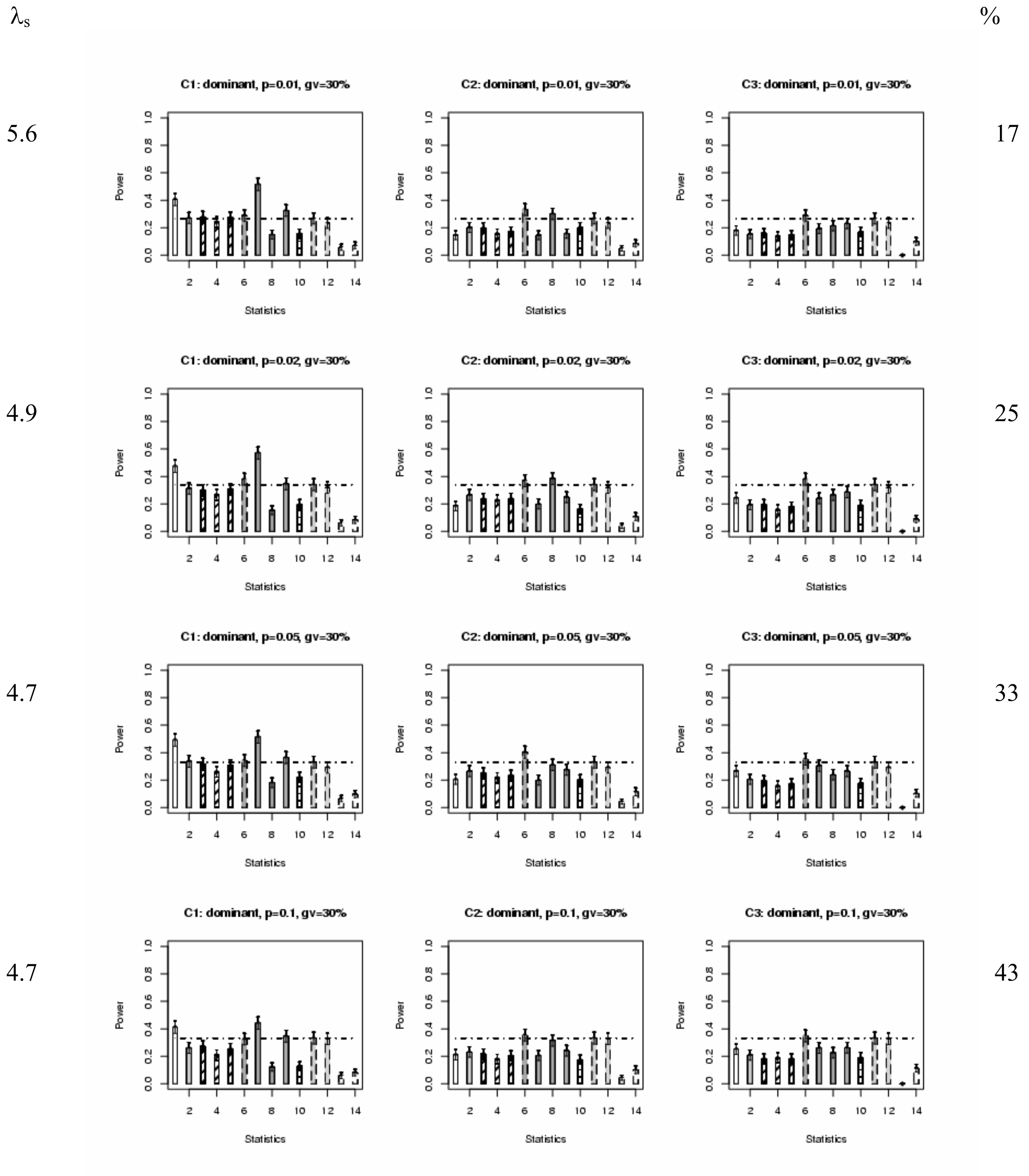
Figures 6.5 and 6.6 indicate the power under the recessive Type II models. Except for MLB and the L → H OSA method, the covariate statistics perform better for all models (with GV = 20% and 30%) when using C1 than when using C2 or C3.

When C1 is used, the mixture model and the H → L OSA method perform significantly better than S_{all} for all models (with GV = 20% and 30%). LODPAL and the MLRM approaches perform significantly better than S_{all} only for the model with d = 0.2 and GV = 30%. But they have the equivalent power to S_{all} for all models. When C2 is used, the power of LODPAL, MLB and the L → H OSA method is equivalent to S_{all}'s power for all models. However, when C3 is used, only MLB has the same power as S_{all} for all models (Figures 6.5, 6.6).



λ_s: recurrence risk ratio; %: percentage of linked families

Figure 6.3 Power and 95% CI for each statistic at the 5% level under different dominant Type II models with genetic variance equal to 20%, based on 500 replicates


C1: dominant, p=0.01, gv=30%

C2: dominant, p=0.01, gv=30%

C3: dominant, p=0.01, gv=30%

C1: dominant, p=0.02, gv=30%

C2: dominant, p=0.02, gv=30%

C3: dominant, p=0.02, gv=30%

C1: dominant, p=0.05, gv=30%

C2: dominant, p=0.05, gv=30%

C3: dominant, p=0.05, gv=30%

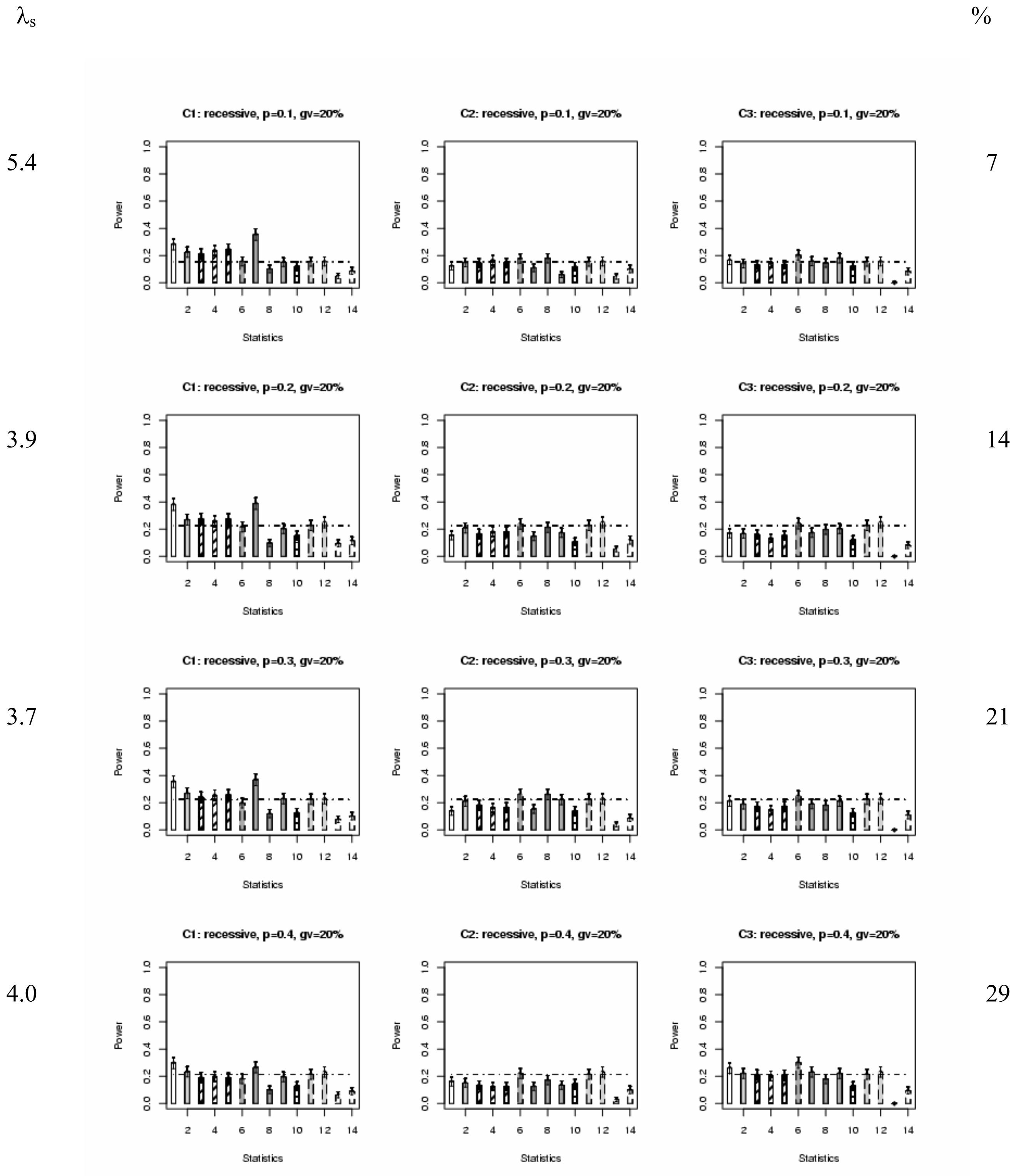
C1: dominant, p=0.1, gv=30%

C2: dominant, p=0.1, gv=30%

C3: dominant, p=0.1, gv=30%

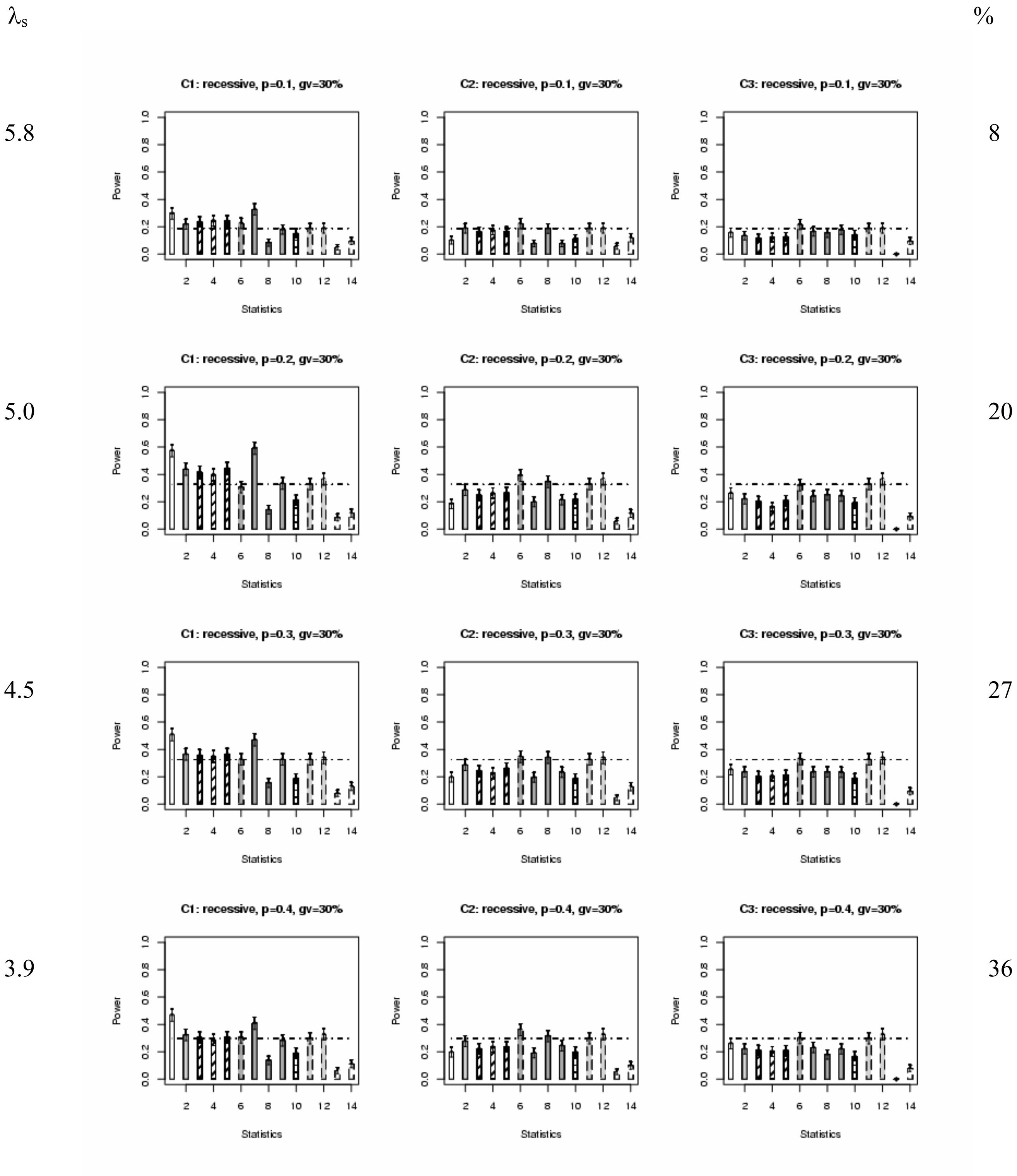
λ_s: recurrence risk ratio; %: percentage of linked families

Figure 6.4 Power and 95% CI for each statistic at the 5% level under different dominant Type II models with genetic variance equal to 30%, based on 500 replicates



λ_s: recurrence risk ratio; %: percentage of linked families

Figure 6.5 Power and 95% CI for each statistic at the 5% level under different recessive Type II models with genetic variance equal to 20%, based on 500 replicates



λs: recurrence risk ratio; %: percentage of linked families

Figure 6.6 Power and 95% CI for each statistic at the 5% level under different recessive Type II models with genetic variance equal to 30%, based on 500 replicates

6.3. TYPE III G x E INTERACTION MODEL

6.3.1. Dominant models

The power under the dominant Type III models is presented in Figure 6.7. Except for MLB and the L → H OSA method, when C1 is used, the covariate methods have higher power than when C2 or C3 is used.

Compared to S_{all}'s power, the power of the mixture model, LODPAL, no-dominance MLRM and min-max MLRM is significantly higher than S_{all}'s power for the rare model (d = 0.01), and the power is close to S_{all}'s power for the model with d = 0.02, when using C1. When C2 is used, the power of MLB is significantly better than S_{all}'s power for the model with d = 0.01, and is similar to S_{all}'s power for the other models (d = 0.02, 0.05, and 0.1). When C3 is used, MLB has the equivalent power to S_{all} for all models (Figure 6.7). But when C1 is used, MLB has extremely low power, in contrast with the results under the Type I and II models.

In contrast to the results under the Type I and II models, the power reaches almost 100% when treating C1 as a quantitative trait in the QTL approaches. But the power of the QTL approaches is very low when using C2 or C3.

6.3.2. Recessive models

The power under the recessive Type III models is shown in Figure 6.8. When C1 is used, the covariate methods have higher power than when C2 or C3 is used, but not for MLB and the L → H OSA. However, the power of the mixture model is lower when using C1 than when using C2 for the common model (d = 0.4), which is the only case in all G x E models.

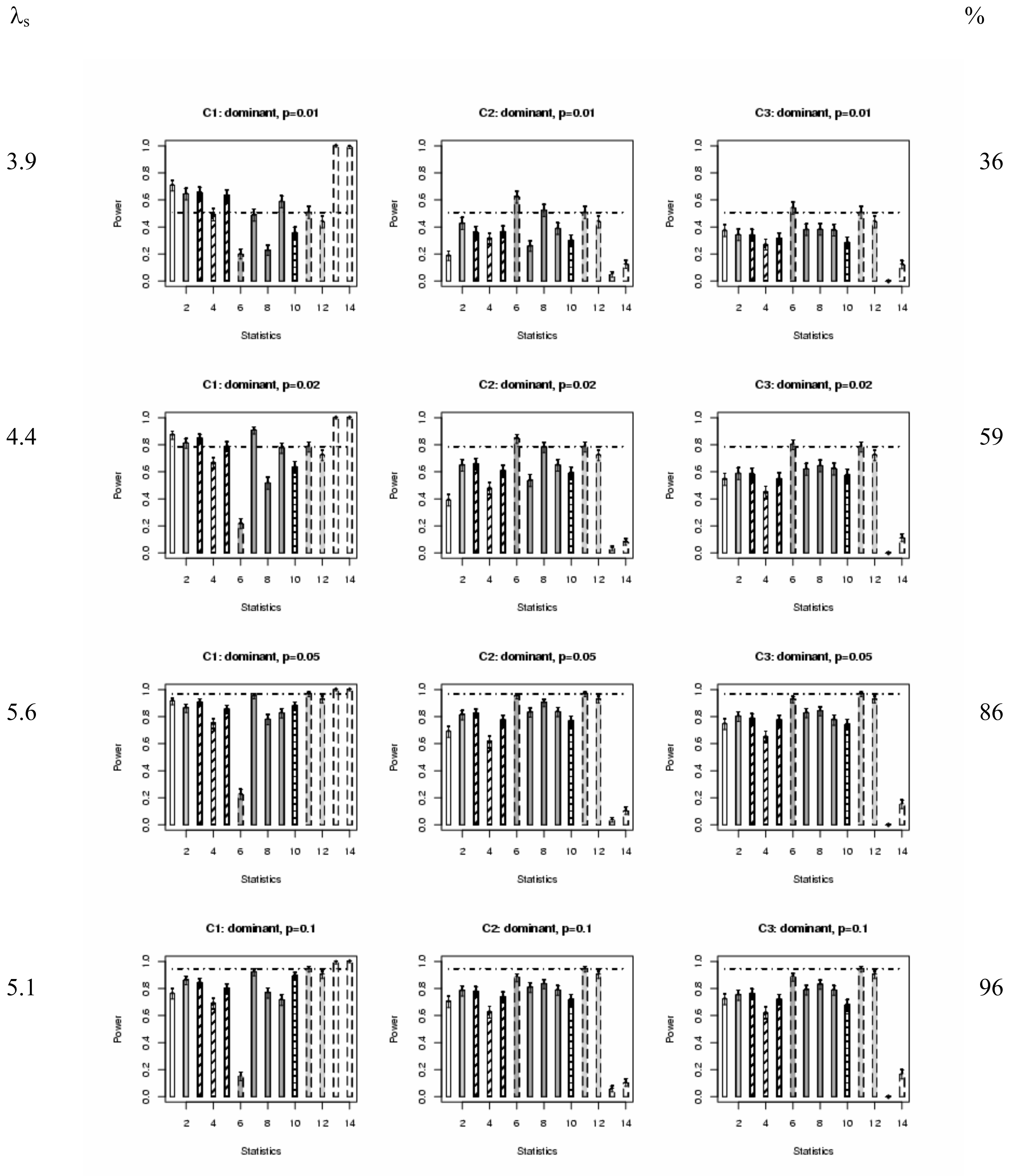
When using C1, the power of the mixture model, LODPAL, the MLRM methods, the H → L OSA is significantly higher than S_{all}'s power for the models with d = 0.1 and 0.2. The power of LODPAL, no-dominance MLRM, min-max MLRM and COVLINK is close to S_{all}'s power for the models with d = 0.3 and 0.4. When using C2 or C3, only the MLB's power is equivalent to S_{all}'s power for all models. Similar to dominant models, when C1 is used, the power of MLB is very low, in contrast with the results under the Type I and II models. The power almost reaches 100% when treating C1 as a quantitative trait, but drops significantly when using C2 or C3 as a quantitative trait.

6.4. SUMMARY

6.4.1. Using different covariates

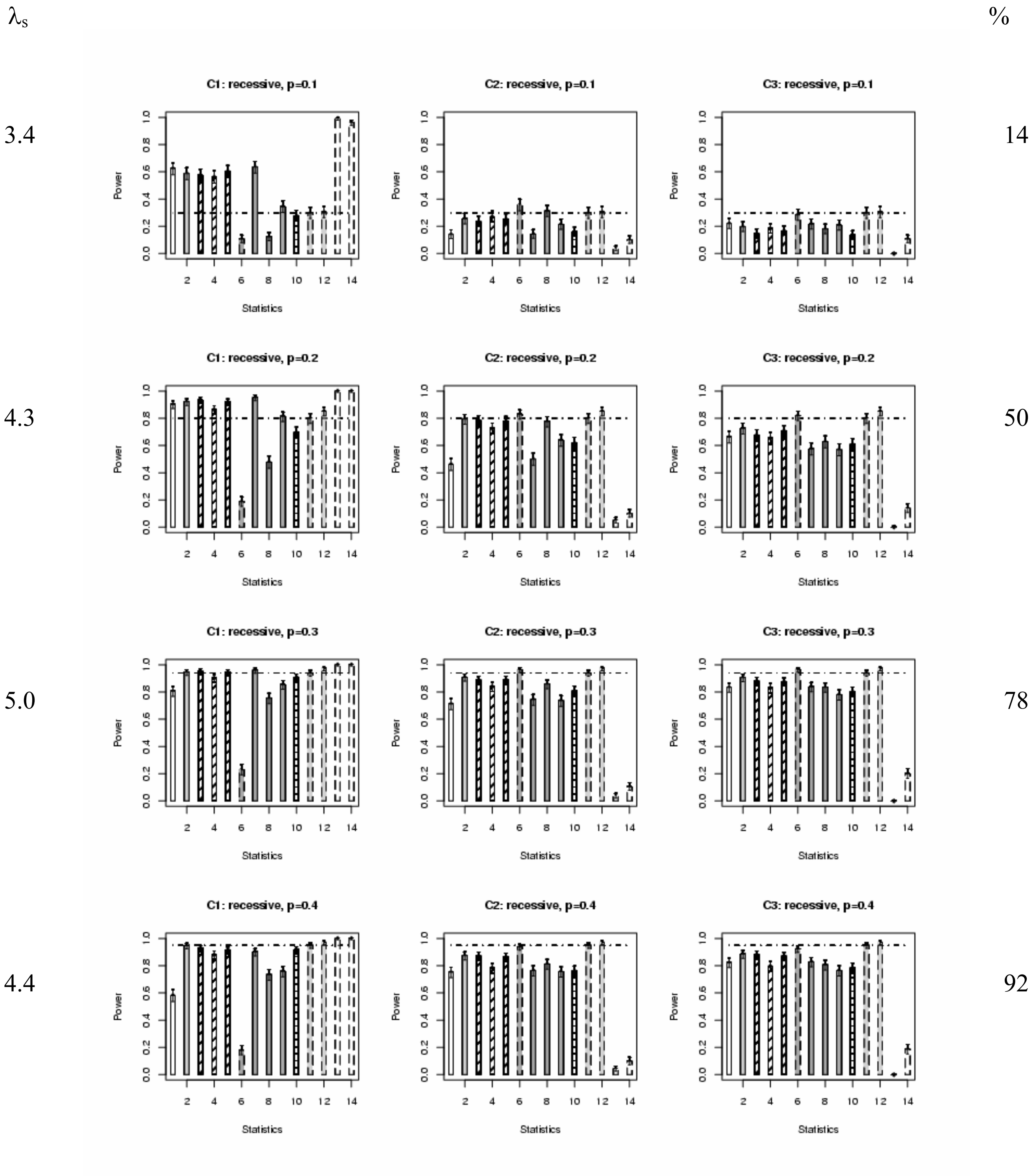
The performance of covariate statistics has been investigated under three G x E interaction scenarios (Figures 1.4, 6, 7). These G x E interaction models mimic G x E relations observed in the real cases (Ottman 1996). In terms of empirical thresholds, our results show that thresholds are quite consistent across different disease models, regardless of the covariate choice: C2 or C3 (Tables 5.2A, B). We analyze the data generated under the null hypothesis of no linkage. And sibship size only varies from 2 to 5 across the disease models. Hence, the thresholds should be independent of the disease models.

With respect to power, we find that the performance of the covariate statistics varies across the different G x E interaction models. In all three G x E models, most covariate methods perform better when using C1 than when using C2 or C3 (Figures 6.1 – 6.8). Overall, most covariate methods perform better (relative to S_{all}) for the rare models (both dominant and recessive traits). Moreover, when C3 is used, the mixture model and the H → L OSA method have better power than when C2 is used for almost all models. However, MLB and the L → H OSA method



λs: recurrence risk ratio; %: percentage of linked families

Figure 6.7 Power and 95% CI for each statistic at the 5% level under different dominant Type III models, based on 500 replicates



λs: recurrence risk ratio; %: percentage of linked families

Figure 6.8 Power and 95% CI for each statistic at the 5% level under different recessive Type III models, based on 500 replicates

behave differently from most covariate statistics. Especially when using C2 or C3, MLB provides much higher power than when using C1 for the Type III models (Figures 6.7, 6.8). The L → H OSA method generally performs better when using C2 than when using C1 for all models.

6.4.2. Covariate statistics vs. S_{all} and QTL approaches

We generated data under 32 different genetic models (8 disease models in the Type I and III models, respectively, and 16 disease models in the Type II model) and evaluated the percentage of times that the covariate statistics significantly increase the power relative to S_{all}'s power. We use all models except the ones with GV equal to 20% in the Type II model. Compared to the power of S_{all}, the percentage of the times that the covariate method performs significantly better than S_{all} when using C1 is as follow: 62.5% for the mixture model; 50% for the H → L OSA method; 25% for no-dominance MLRM; 20.83% for min-max MLRM; 16.67% for LODPAL, no-additive MLRM and MLB; 8.33% for the optimal-slice OSA method; zero for the L → H OSA method, and COVLINK (the greater details are provided in Table 6.2A).

With respect to the performance when C2 is used, MLB and the L → H OSA method yield significantly higher power than S_{all} (25% and 4.17%, individually). When C3 is used, only MLB yields significantly higher power than S_{all} (25%). The power of the mixture model and the H → L OSA method is always lower than S_{all}'s power across all models when C2 is used. When C3 is used, only the power of MLB is equivalent to S_{all}'s power for almost all models (the greater details are provided in Tables 6.2B, C).

Regarding the QTL approaches, our results indicate that the QTL approaches almost have 100% power for all Type III models when treating C1 as a quantitative trait, but not C2 or C3 (Figures 6.7, 6.8). They have extremely low power for all Type I and II models regardless of covariate choice. Hence, the covariate statistics only perform worse than the QTL approaches when using C1 under Type III model. Since the C1 values are dramatically different between the subjects carrying a high-risk genotype and those carrying a low-risk genotype, it is not surprising that the QTL approaches have extremely high power under the Type III models.

Table 6.2A Percentage of the time that the power of the covariate statistic is significantly higher than S_{all}'s power across three types of G x E interaction models, when using the "right" covariate C1

	Mixture model	LODPAL	No-dom MLRM	No- add MLRM	Min-max MLRM	MLB	H → L OSA	L → H OSA	Optimal-slice OSA	COVLINK
All disease models	62.50	16.67	25.00	16.67	20.83	16.67	50.00	0	8.33	0
All dominant models	58.33	8.33	16.67	0	8.33	16.67	50.00	0	16.67	0
All recessive models	66.67	25.00	33.33	33.33	33.33	16.67	50.00	0	0	0
Smallest frequency [a]	100	33.33	50.00	33.33	50.00	33.33	66.67	0	33.33	0
Largest frequency [b]	33.33	0	0	0	0	0	33.33	0	0	0
≥ 50% linked families [c]	25.00	8.33	16.67	8.33	8.33	16.67	16.67	0	0	0
< 50% linked families [d]	100	33.33	33.33	25.00	33.33	16.67	83.33	0	16.67	0

Note:

a. Models with the smallest disease allele frequency
b. Models with the largest disease allele frequency
c. Models with ≥ 50% linked families
d. Models with < 50% linked families

Table 6.2B Percentage of the time that the power of the covariate statistic significantly higher than S_{all}'s power across three types of G x E interaction models, when using the "wrong" covariate C2

	Mixture model	LODPAL	No-dom MLRM	No- add MLRM	Min-max MLRM	MLB	H → L OSA	L → H OSA	Optimal-slice OSA	COVLINK
All disease models	0	0	0	0	0	25.00	0	4.17	0	0
All dominant models	0	0	0	0	0	41.67	0	8.33	0	0
All recessive models	0	0	0	0	0	8.33	0	0	0	0
Smallest frequency	0	0	0	0	0	50.00	0	16.67	0	0
Largest frequency	0	0	0	0	0	16.67	0	0	0	0
≥ 50% linked families	0	0	0	0	0	8.33	0	0	0	0
< 50% linked families	0	0	0	0	0	41.67	0	8.33	0	0

Table 6.2C Percentage of the time that the power of the covariate statistic significantly higher than S_{all}'s power across three types of G x E interaction models, when using the random covariate C3

	Mixture model	LODPAL	No-dom MLRM	No- add MLRM	Min-max MLRM	MLB	H → L OSA	L → H OSA	Optimal-slice OSA	COVLINK
All disease models	0	0	0	0	0	25.00	0	0	0	0
All dominant models	0	0	0	0	0	16.67	0	0	0	0
All recessive models	0	0	0	0	0	33.33	0	0	0	0
Smallest frequency	0	0	0	0	0	33.33	0	0	0	0
Largest frequency	0	0	0	0	0	16.67	0	0	0	0
≥ 50% linked families	0	0	0	0	0	33.33	0	0	0	0
< 50% linked families	0	0	0	0	0	16.67	0	0	0	0

7. DISCUSSION AND FUTURE WORK

7.1. DISCUSSION

7.1.1. Why do covariate statistics not perform better than S_{all}?

To deal with genetic heterogeneity, the covariate statistics take covariate information into account. One would expect that covariate statistics should have increased power to detect signals when using more information. According to our results (Figures 6.1 – 6.8), most covariate statistics provide significantly better power than S_{all} only for few models, when using C1, the covariate with G x E interaction effect (Table 6.2A). In addition, when the environmental covariate C2 or the random noise covariate C3 is used, only MLB and the L → H OSA method yield significantly higher power than S_{all} (Tables 6.2B, C). Why do covariate statistics not perform better than S_{all} as one might expect? There are several caveats one should consider when interpreting the results.

First, the sample size of "linked families" varies across disease models. A "linked family" here is defined as a family with 2 or more affected children carrying the high-risk disease genotype. The results in Chapter 4 show as the disease allele frequency increases, more linked families are ascertained (Table 4.1). When the data set is more genetically homogeneous, S_{all} has adequate power. Therefore, we observe that most covariate methods perform better than S_{all} in rare disease models (d = 0.01 in dominant trait; d = 0.1 in recessive trait). But the power of covariate statistics becomes smaller than S_{all}'s power as the disease becomes more common. It may be because less linked families are ascertained in the rare models, and more linked families in the common models. When a majority of pedigrees is linked, S_{all} should have enough power. Hence, the covariate statistics can only increase the power a bit at best, since only few unlinked pedigrees can be “removed”. In addition, the covariate approaches use a degree of freedom to distinguish the linked and unlinked groups, which may not be worth while almost all pedigrees are linked.

Second, the covariate methods estimate more parameters than the model-free methods, so they have more degrees of freedom. This may lead to a loss of power if the covariate does not provide enough beneficial information. When C2, the covariate without G x E interaction effect, is used, the power decreases for most covariate statistics because use of C2 only increases degrees of freedom while adding no useful information to the model. Also, we observe similar results when using C3.

Third, we measure covariate information by taking the average covariate values across either the affected sib pairs or the whole family. Based on our G x E interaction models, the average covariate values in affecteds are higher than those in unaffecteds (Tables 4.3A, B), but the affecteds do not necessarily carry high-risk genotypes. Is taking the average the best way to capture covariate information for covariate statistics? If the measure of covariate information does not divide pedigrees into more homogeneous subgroups, the covariate methods may not increase power, even when using the “right” covariate.

Fourth, different types of G x E interactions influence the power of the covariate statistics. The genetic factor interacts with C1 in various ways in the types of G x E models. The genetic factor has a dramatically strong effect on C1 under the Type III model, and so the covariate methods have better power than S_{all}. But under the Type I model, the G x E interaction effect is moderate and under the Type II model, C1 only influences whether the genetic factor has an effect on disease risk. Adding the covariate information may provide little useful information and cannot influence power to a significant degree.

7.1.2. What is the "right" covariate?

Often, there are two main kinds of covariates involved in complex diseases (Hauser et al. 2003). The first kind is a covariate with a G x E interaction effect. One can enhance power to reveal a genetic effect by including such covariates in the model. Because this type of covariate interacts with the genetic factors, it can be viewed as an "effect modifier", to use terminology from epidemiological studies (Rothman and Greenland 1998). The second kind is a covariate with an environmental risk effect, independent of the genetic effect. One would prefer to remove the variability due to such covariates before proceeding with linkage analysis. This type of covariates can be thought of as a "confounder". The most critical difference between these two types of covariates is that one would remove the confounder effect prior to analysis, while one would consider the effect modifier as a finding to be reported rather than a bias to be avoided (or removed).

Therefore, we need to clarify what we mean by "biologically meaningful" covariates very carefully, and why we use the terminology: "right" and "wrong" covariates herein. The "right" covariate C1 represents a covariate with G x E interaction effect. The "wrong" covariate C2 indicates a covariate with an environmental risk effect. In a real study, we would not know whether or not the covariates interact with genetic factors. Hence, we are not only interested in the performance of covariate statistics, but also in detecting the covariates with G x E interaction. In some circumstances, accounting for the environmental risk factor can also increase power to detect the linkage signals since removing the confounding effect may make the data more genetically homogeneous. This would be a good issue to investigate.

Our results indicate that using different types of covariates (with or without G x E interaction) affects the power of covariate statistics. Some covariate methods, e.g. the mixture model and the H → L OSA method, have better power when using the G x E interaction covariate. But some covariate approaches, e.g. MLB, performs better when adjusting the environmental risk factor.

In a real case, one often collects several covariates, which are thought to be associated with a complex disease. Hence, how should one determine whether a covariate is an effect modifier or a confounder? It is still very challenging. In addition, if there are more than one "right" covariates, how many covariates should we simultaneously add into the model? Will adding in more "right" covariates give us higher power? How will the covariate methods perform if one includes one "right" and one "wrong" covariate simultaneously in the model? These are still open questions for investigation.

7.1.3. What is the best way to abstract covariate information?

As we described in Section 7.1.1, to include covariate information, we compute the mean covariate values across either the affected sib pairs (ASP) or the whole family (Table 7.1). The mixture model and the OSA methods use family-level covariate values. LODPAL, the MLRM methods and COVLINK require sib-pair-level covariate values. To account for the covariate information, taking the average is not the only way, and it may not be the optimal way.

Using the "Haseman-Elston regression" methods as an example, the original Haseman-Elston regression (Haseman and Elston 1972) measured quantitative trait information by taking the squared trait differences of sib pairs. Wright (1997) pointed out that this measure discards some useful information. He showed that if information from the trait sums of sib pairs are also

Table 7.1 Abstract covariate information in different covariate methods

	Family level		Sib-pair level		Individual level
	All member	Affected sibs	Affected sibs	All sibs	All members
Mixture model	√				
LODPAL			√		
MLRM			√		
MLB					√
OSA		√			
COVLINK				√	

included, a certain amount of power can be gained. Since then, various "revised Haseman-Elston" approaches have been suggested (Drigalenko 1998; Elston et al. 2000, Xu et al. 2000; Forrest 2001; Sham and Purcell 2001; Visscher and Hopper 2001). All of these revised approaches combine the squared trait sum and the squared trait difference in various ways, and increase the power successfully while preserving the robustness of the regression framework (Feingold 2002).

None of published articles indicates that taking the average is the best measure of covariate information. It is important to investigate what is the best (or appropriate) measure for covariate information. The LODPAL program in S.A.G.E. package offers several alternative options, such as taking difference of ASP covariate values, summing the ASP covariate values, using the maximum ASP covariate values, or using the minimum ASP covariate values. It would be of interest to evaluate the performance of the covariate methods using these alternative approaches.

7.1.4. What are the desirable features for covariate statistics?

The covariate statistics we investigate behave variously under different G x E models and when using different types of covariates. The unique features of the covariate statistics are described as follows.

Mixture model: the mixture model is sensitive to which covariate type is used. It performs very well when using C1, the covariate with G x E interaction. As long as C1 is used, it is robust across different types of G x E interaction models, and performs the best of the covariate statistics. The power of the mixture model is also significantly greater than S_{all}'s power for 62.5% of the models. But when using C2, the covariate does not interact with a genetic factor, the power of the mixture model is low. This can be viewed as a desirable property, in that it may help one determine which covariate out of many is relevant and interacts with the susceptibility genes determining the disease of interest.

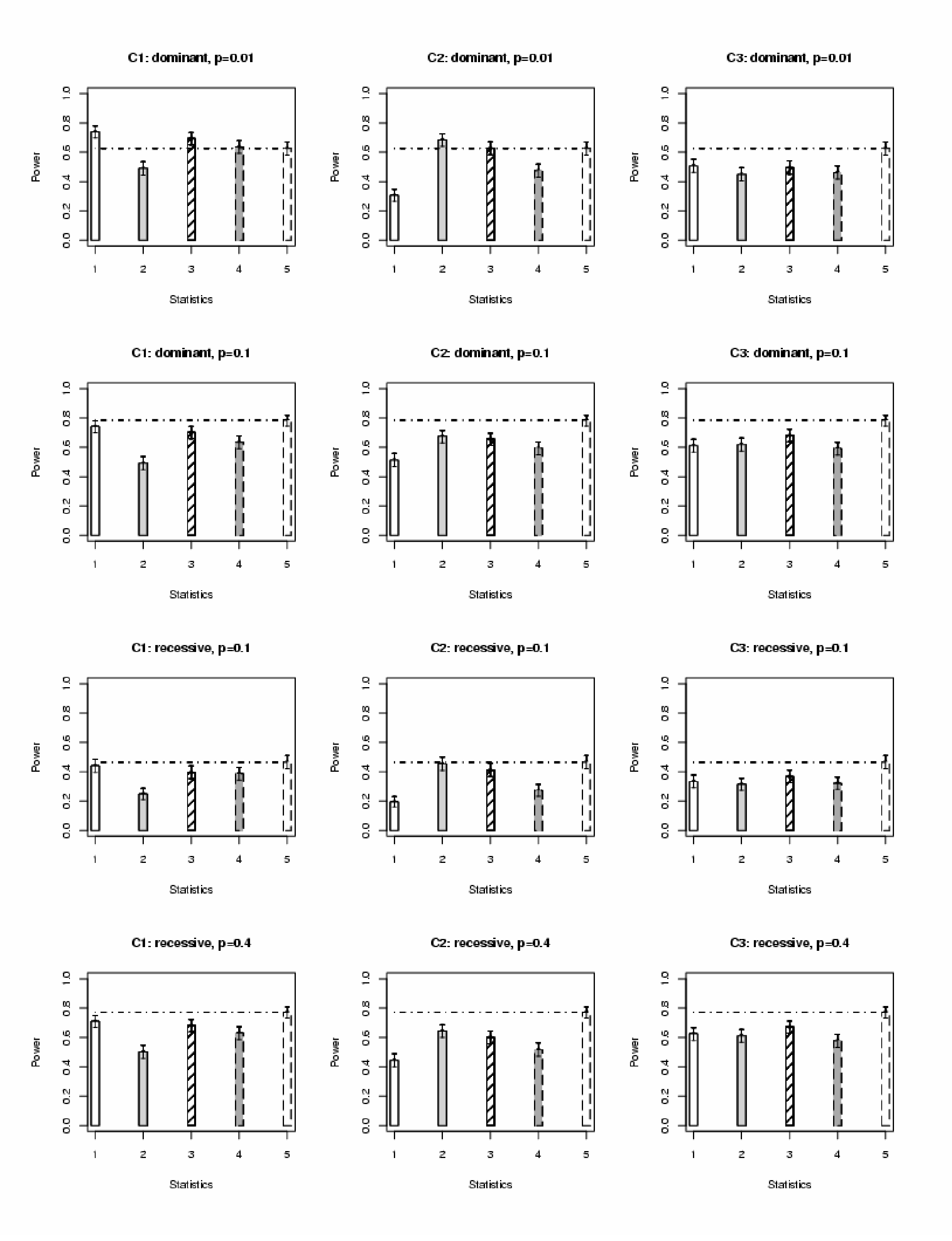
The OSA methods: we here run the OSA methods three times, based on the different rank orders (H → L, L → H and optimal-slice). When C1 is used, the H → L OSA method provides good power consistently across all three G x E interaction models. It has significantly greater power than S_{all} for 54.17% of all models. When C2 is used, the power of the L → H OSA method is higher than S_{all}'s power for 8.33% of all models. According to our simulated data, the "linked" families tend to have high C1 values and low C2 values. Hence, it is not surprising that the H → L OSA method has better power when C1 is used, and the L → H OSA method provides higher power when C2 is used.

However, whenever we run OSA, we always run it at least twice (H → L and L → H). Thus, we must take into account this multiple testing issue. One way to deal with this issue may be that one can run OSA twice in the H → L and L → H rank orders, and then only use the maximum value from these two tests to compute the empirical thresholds and evaluate the power. We apply this "max(H → L, L → H)" approach and the results indicate that the power of this alternative approach is often between the power of the H → L OSA and the L → H OSA methods (Figures 7.1, 7.2, 7.3). The power of this alternative approach is generally retained well across all models.

The MLB method: MLB performs well when using C1 in the Type I models, but not in the Type II or Type III models. However, it has significantly higher power than S_{all} for 37.5% of all Type II models and 25% of all Type III models when C2 is used. The MLB approach regresses out the covariate effect, takes the Pearson's residuals from the regression model, and then treats the residuals as a quantitative trait. From an epidemiological point of view, this approach would deal better with confounders than with effect modifiers. Since C1, a strong effect modifier, in the Type III models is highly correlated to the disease genotypes, regressing out C1 is equivalent to regressing out the genetic effect. This leads very low power for MLB.

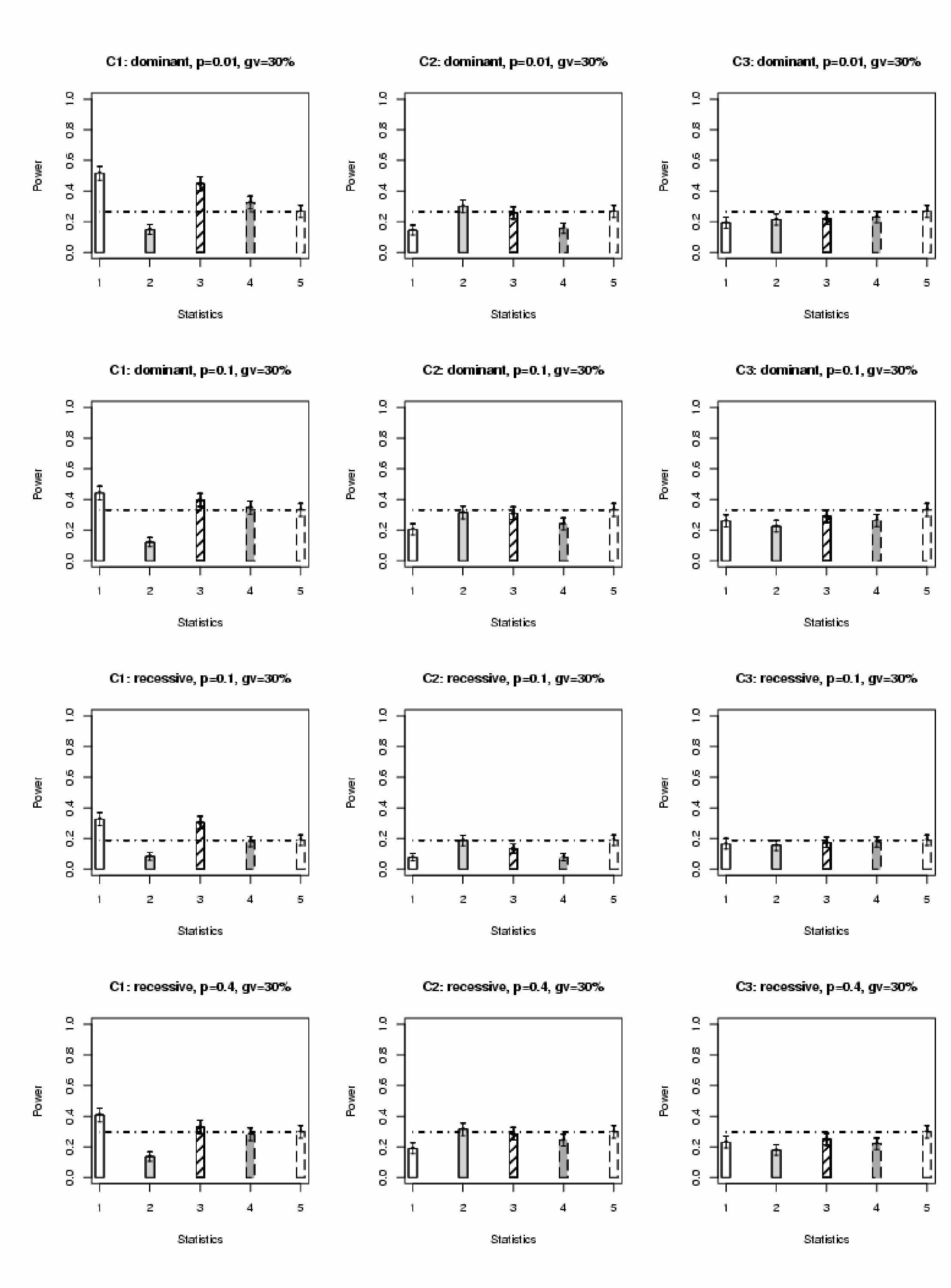
On the other hand, C1 in the Type I models is a relatively moderate effect modifier, compared with the Type III models. Hence, even though we regress individual's affection status on C1, a moderate level of genetic effect still remains in the residuals. Therefore, MLB has certain power when using C1 under the Type I models.

Alcaïs and Abel (2001) in their simulation study showed that power increases when the disease allele frequency is rare and when unaffected siblings are included in the analysis. Our results also indicate that MLB often performs well when the disease allele frequency is rare (Tables 6.1A, B, C). However, MLB has significantly higher power than S_{all} for only 16.67% of all



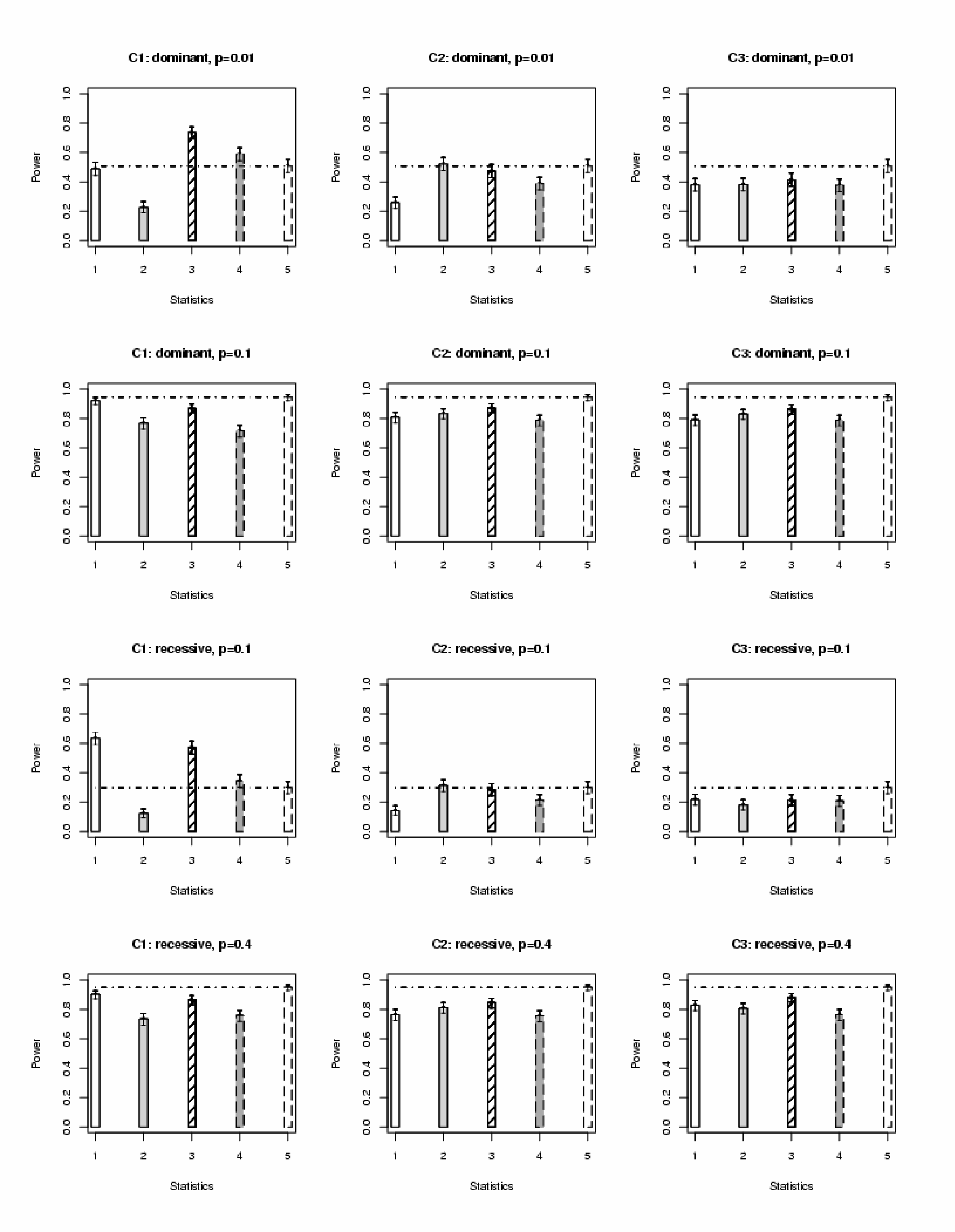
Number on x-axis: 1. H → L; 2. L → H; 3. max(H → L, L → H); 4. optimal-slice; 5. S_{all}

Figure 7.1 Power and 95% CI for the OSA methods and S_{all} at the 5% level under different Type I models, based on 500 replicates



Number on x-axis: 1. H → L; 2. L → H; 3. max(H → L, L → H); 4. optimal-slice; 5. S_{all}

Figure 7.2 Power and 95% CI for the OSA methods and S_{all} at the 5% level under different Type II models with genetic variance = 30%, based on 500 replicates



Number on x-axis: 1. H → L; 2. L → H; 3. max(H → L, L → H); 4. optimal-slice; 5. S_{all}

Figure 7.3 Power and 95% CI for the OSA methods and S_{all} at the 5% level under different Type III models, based on 500 replicates

models. Alcaïs and Abel (2001) only compared MLB using the whole sibship to MLB using affected sib pairs when accounting for the covariate effect. It would be interesting to compare the two MLB approaches to S_{all} as a baseline.

Logistic regression approaches: generally, logistic regression approaches (LODPAL, the MLRM methods and COVLINK) herein do not perform well across the models. One possible interpretation is that their performance highly depends on how we measure the covariate information. As we discussed in Section 7.1.3, we measure the covariate information by taking the average covariate values across sib pairs. Based on logistic regression approaches, if the average pair-level covariate values are close in three different IBD sharing patterns, the power may not increase, since we only increase degrees of freedom while no useful information is added. The sib pairs here may be affected because both siblings carry high-risk disease genotypes, but this is not necessarily the case. It could be that one carries a high-risk genotype, while the other carries a low-risk genotype. According to our simulated data, if the affecteds do not carry high-risk genotype, they tend to have high C2 values under the Type I and II models.

Another reason for COVLINK is that we only use the pairs with unambiguous IBD sharing probabilities. Therefore, the sample size in COVLINK is smaller than in the other covariate methods. Holmans (2002) suggested a conditional logistic regression for detecting gene-gene interactions using affected sib pair analysis (Holmans 2002). This approach is similar to COVLINK, but is applicable to partially informative IBD data. It would be of interest extending COVLINK using all IBD sharing data and re-evaluating the performance.

The articles that proposed LODPAL and the MLRM approaches describe the theoretical asymptotical distribution (Olson 1999; Greenwood and Bull 1999). Our results show that the empirical thresholds of the MLRM approaches are very close to the analytical thresholds, but the analytical thresholds of LODPAL are markedly smaller than the empirical thresholds in the tail of the distribution (Table 5.5 and Figure 5.1). Previous work compared the performance of LODPAL and GENEFINDER (Glidden et al. 2003) using the Genetic Analysis Workshop 13 (GAW13) simulation data (Hsu et al. 2003). Hsu et al. (2003) concluded that LODPAL has more reasonable power than GENEFINDER, but both approaches have high possibility of leading to false-positive conclusions. Therefore, one should avoid false-positive results based on the LODPAL approach by using the empirical p-values rather than analytical p-values.

With respect to the MLRM approaches, Greenwood and Bull (1999) concluded that: (1) the MLRM approach under no dominance assumption has the best power; (2) the min-max restriction approach has power between those under no dominance and no additive assumptions; (3) but the MLRM approach under no additive assumption tends to have very low power. Based on our G x E models, our results indicate the same conclusions.

Schaid et al. (2003) pointed out that LOD scores for testing the covariate influence should be identical between LODPAL and the MLRM under the min-max restriction, $z_1 = 0.335 + 0.58^* z_0$, where z_i is two of three IBD sharing probabilities (Table 7.2). We observe very close, but not identical results for these two methods. Two reasons may explain the very mild difference in these two methods.

First, as we described in Chapter 2, we use the “centered” average pair-level covariate value for LODPAL. But we use the average of the original pair-level covariate value for the MLRM approaches. Second, for some covariate values, the likelihood estimate obtained from LODPAL can be negative. In these situations, the LOD score is set to a very small positive value to avoid computational difficulty. We observe the similar situation in the MLRM approaches. In these cases, we set a very small positive value to the likelihood estimate instead of the LOD score.

7.1.5. How important are unaffecteds?

Does including unaffecteds increase the power? We here cannot reach a conclusion. According to our results, the mixture model and the H → L OSA method perform quite well across the models when C1 is used. We use all family members' data, both affecteds and unaffecteds, to cluster pedigrees for the mixture model, but only use affected siblings' information for the OSA methods. Although Alcaïs and Abel (2001) showed that using the whole sibships has better

Table 7.2 Logistic regression models for IBD sharing probabilities and covariates (From Schaid et al. 2003)

Model	$z_0(X;\beta)$ [a]	$z_1(X;\beta)$	$z_2(X;\beta)$
Trinomial	$\frac{1}{1+2e^{\beta_{1,0}+\beta_{1,1}X}+e^{\beta_{2,0}+\beta_{2,1}X}}$ [b]	$\frac{2e^{\beta_{1,0}+\beta_{1,1}X}}{1+2e^{\beta_{1,0}+\beta_{1,1}X}+e^{\beta_{2,0}+\beta_{2,1}X}}$	$\frac{e^{\beta_{2,0}+\beta_{2,1}X}}{1+2e^{\beta_{1,0}+\beta_{1,1}X}+e^{\beta_{2,0}+\beta_{2,1}X}}$
Logistic - Multiplicative	$\frac{1}{(1+e^{\beta_0+\beta_1X})^2}$	$\frac{2e^{\beta_0+\beta_1X}}{(1+e^{\beta_0+\beta_1X})^2}$	$\frac{(e^{\beta_0+\beta_1X})^2}{(1+e^{\beta_0+\beta_1X})^2}$
Logistic - Minimax	$\frac{1}{-1.634+5.634e^{\beta_0+\beta_1X}}$ [c]	$\frac{2e^{\beta_0+\beta_1X}}{-1.634+5.634e^{\beta_0+\beta_1X}}$	$\frac{3.634e^{\beta_0+\beta_1X}-2.634}{-1.634+5.634e^{\beta_0+\beta_1X}}$
	$\frac{e^{\beta_0+\beta_1X}}{1.5504+2.45e^{\beta_0+\beta_1X}}$ [d]	$\frac{0.5504+1.45e^{\beta_0+\beta_1X}}{1.5504+2.45e^{\beta_0+\beta_1X}}$	$\frac{1}{1.5504+2.45e^{\beta_0+\beta_1X}}$

Note:

a. z_i: three IBD sharing probabilities; X: a pair-specific covariate
b. LODPAL and MLRM approaches without constraints ($\beta_{1,0}$, $\beta_{1,1}$): coefficients for IBD = 1 vs. IBD = 0; ($\beta_{2,0}$, $\beta_{2,1}$): coefficients for IBD = 2 vs. IBD = 0
c. LODPAL under min-max restriction
d. MLRM under min-max restriction

power than using affected sibs only in their simulation study, we do not reach the same conclusion as theirs. Based on our results, when we use both affecteds and unaffecteds' data for MLB and COVLINK, they do not perform well. It may be because we generate simulated data under different G x E interaction scenarios. Hence, this issue is still unresolved.

7.1.6. What is the "right" study design?

"In the presence of strong (statistical) G-E interaction, there can be a gain in power to detect genetic linkage/ association… But, for discrete outcomes, gains in power are modest except when effects (and their interaction) are strong" [quoted from Clayton's slides]. Poisson et al. (2003) reported that when the genetic effect and environmental effect are independent, family-based association testing with covariate adjustment may reduce the power. Do those imply that accounting for covariate information in linkage analysis/ association studies for complex diseases can only increase power a little, or even reduce power? It requires more thorough investigation to address this issue. One straightforward way to enhance the power is selecting the right study design.

Our results indicate that the performance of the covariate statistics is influenced by the types of G x E models, and as well as the types of covariates. Guo (2000) reported the required sample size in three different G x E interaction examples: (1) genetic factors interact with exposures in an additive scale; (2) genetic factors increase disease risk in unexposed subjects, but decrease the risk in exposed subjects; (3) exposures have strong effect, but small G x E interaction effect. In the example 3, there is a strong genetic effect in unexposed subjects, but a small effect in exposed subjects. The results indicated that the required sample sizes vary under different G x E interaction scenarios (Table 7.3). Therefore, one should select appropriate study designs and suitable covariate methods for dealing with various G x E scenarios and covariates in complex diseases.

Table 7.3 Sample size required to reach 80% power using the affected-sib-pairs method (ASP) and the transmission disequilibrium test (TDT) in various study designs (modified from Guo 2000)

Design	ASP			TDT		
	Example 1	Example 2	Example 3	Example 1	Example 2	Example 3
Unexp-con [a]	95	2,466	28	15	183	8
Exp-con [b]	4,748	278	474,877	145	45	4,074
Concordant	442	477	13,080	45	51	1,131
Discordant [c]	452	241	707	27	1,542	18

Note:

a. Unexp-con: both affected sibs are unexposed
b. Exp-con: both affected sibs are exposed
c. Discordant: affected-unaffected sib pairs

7.2. FUTURE WORK

7.2.1. Use more than one covariate

Here we only apply the covariate statistics to one covariate at a time. Most covariate methods evaluated here can handle more than one covariate simultaneously. Often, in reality, several covariates interact with different disease genes, or several covariates interact with the same disease gene. So it is of interest to be able to use more than one covariate at a time. However, the drawback to using more covariates is it increases the degrees of freedom. But if one only adds “correct” covariates into the model, this may lessen the impact of increasing the degrees of freedom. In the future, we would like to investigate the performance of the covariate methods using two or more covariates simultaneously.

7.2.2. Generate more complicated G x E interaction model

We simulate the disease models under three different G x E interaction scenarios. However, only one genetic factor is included in the underlying liability. The major aim here is to evaluate whether covariate statistics increase power to detect signals. We therefore start from the simplest scenarios first. Our results offer valuable information regarding choosing appropriate covariate statistics and suitable types of covariates.

But these models do not completely capture the reality of complex diseases. It would be more realistic if we can inspect the performance of the covariate methods using more complicated G x E interaction models. In addition, the recurrence risk ratio for siblings, λ_s, is relatively high (3.4 – 5.8) across the models, which is not observed in complex traits usually. Hence, in the future, we plan to generate the genetic models with two genetic factors and two corresponding G x E interaction covariates and with lower λ_s value, and then analyze the data using the covariate statistics.

7.2.3. Empirical thresholds vs. number of covariates

We estimate the empirical thresholds at 1%, 5% and 10% level using the environmental risk covariate C2 or the random noise covariate C3, separately. But we only incorporate one covariate at a time. Our results indicate that using analytical p-values for some covariate statistics may lead to false-positive conclusions (Table 5.5 & Figure 5.1). We plan to compute the thresholds from the empirical distribution using more than one covariate, and then compare these with the analytical distribution.

7.2.4. Estimate genome-wide empirical thresholds

Because several covariate statistics apply optimization algorithms to maximize the likelihood, the simulations are computationally intensive and time-consuming. In average, it takes about 20 minutes to complete simulation and analysis for one replicate of 100 pedigrees on a 500 MHz processor, or around 12 minutes using a 1.53 GHz processor under the Linux operating system. Because of the computational limitations, the empirical thresholds herein are estimated by using 33 markers on one chromosome. Based on the chromosome-based approach, our simulation program can generate the marker data across the whole genome very efficiently. Ultimately, it will be of interest to obtain the genome-wide thresholds for the covariate statistics.

7.2.5. Measure covariate information in various ways

As we described in Section 7.1.3, taking the average across families or sib pairs may not be the most efficient measure of covariate information. We will evaluate the performance of the covariate methods using the alternative measures, e.g. taking difference, summing the covariate values or using the maximum (or minimum) covariate values.

7.2.6. Add discordant sib pairs into logistic regression approaches

Schaid et al. (2003) indicated that it is feasible to include both concordant (as well as unaffected pairs) and discordant sib pairs by using a dummy variable that codes concordance status. Although COVLINK takes into account unaffected sibs, it only uses unambiguous IBD sharing data, which limits the power. Insofar, LODPAL and the MLRM methods only include affected sib pairs. The power may be increased by adding in unaffected sibs. Therefore, it would be noteworthy to extend LODPAL and the MLRM methods to include concordant unaffected sib pairs and discordant sib pairs.

7.2.7. Compare with different approaches

Certainly, the covariate statistics is not the only choice for handling genetic heterogeneity. As we introduced in Chapter 2, the statistics based on parametric framework (e.g. admixture model) also account for heterogeneity and have been applied to linkage analysis. Tree-based recursive partitioning techniques have been also suggested for dealing with this issue (Shannon et al. 2001; Zhang et al. 2002). It will be of interest to compare the performance of the statistics, based on these different frameworks.

7.3. CLOSING REMARKS

The findings herein present: 1) a rapid data simulation program; 2) the performance of covariate statistics under various G x E interaction models and various disease models; 3) the comparison between covariate statistics, model-free methods and QTL approaches; 4) the effect of different types of covariates. In summary, covariate statistics can provide reasonable power under different types of G x E interaction scenarios. The prior knowledge of the relationship between genetic factors and the types of covariates cannot be neglected. Therefore, one should apply the appropriate covariate method and utilize the covariate information carefully in different situations. Moreover, it will be crucial to scrutinize the covariate statistics through applying genome-wide thresholds, employing more comprehensive G x E interaction models, adding more covariates in the model, measuring the covariate information in alternative ways, and comparing with different approaches.

APPENDIX A

OVERVIEW OF LINKAGE ANALYSIS

The aim of linkage analysis is to make statistical inference about the relative positions of different genetic loci according to the observed phenotypic and genotypic data on individuals in the same pedigree. In linkage analysis, the null hypothesis is no linkage and the alternative hypothesis is that there is linkage between two loci. Both parametric and nonparametric approaches have been widely employed in linkage analysis. On the basis of both approaches, many methods have been proposed to detect linkage between genetic loci. We briefly introduce these two approaches in the following sections.

A.1 PARAMETRIC LINKAGE ANALYSES

A.1.1. Characteristics and model specification

Parametric linkage analyses, also known as model-based methods, have been commonly used to map genetic loci that follow simple Mendelian inheritance (Morton 1955; Chotai 1984; Cleves and Elston 1997). To carry out parametric linkage analysis, one needs to collect phenotypic and genotypic data from families to estimate the overall pedigree likelihood. The samples may consist of collections of one or more nuclear families, or large pedigrees containing different degrees of relatives. The parameter primarily of interest is the recombination fraction, commonly denoted by θ, which is the probability that two loci, a putative disease locus and a marker locus, on the same chromosome will segregate to different gametes during meiosis. Consequently, the aim of linkage analysis is to estimate θ given an assumed underlying disease model. The parametric linkage approach generally provides good power if the underlying disease model is specified correctly.

The components in the likelihood equation are the population frequencies of the trait locus alleles, penetrances and transmission probabilities. We define $Prior(G_{founder})$ as population frequency, $Pen(X_i \mid G_i)$ as penetrance and $Trans(G_o \mid G_d, G_m)$ as transmission probability of genotype G_o. The overall likelihood then can be written as:

$$L = \sum_{G_1} \cdots \sum_{G_n} \prod_{founder} Prior\,(G_{founder}) \prod_{i} Pen\,(X_i \mid G_i) \prod_{\{o,d,m\}} Trans\,(G_o \mid G_d, G_m),$$

where $G_1 \ldots G_n$ are the n marker data, G_i is the genotyping data of the *i*th subject, X_i is the phenotypic data of the *i*th subject, and G_i runs over all members in the family (*o*: offspring, *d*: dad, *m*: mom).

A.1.2. Algorithms for likelihood calculation

The likelihood introduced in Section A.1.1. is computed by summing the products of founder probability, penetrance probability and transmission probability over all possible combinations of genotypes of all pedigree members. Two algorithms, the Elston-Stewart algorithm (Elston and Stewart 1971) and the Lander-Green algorithm (Lander and Green, 1987), have been proposed to compute the exact pedigree likelihood and implemented in various programs.

The Elston-Stewart algorithm prescribes an order for the iterated sum which minimizes the total number of additions and multiplications in the likelihood function calculation in such a way that the computational time of the algorithm is linear in pedigree size, but exponential in the number of markers. The likelihood function of a nuclear family with N children can be decomposed as:

$$L = \sum_{G_d} Pen(x_d \mid g_d)\Pr ior(g_d)\sum_{G_m} Pen(x_m \mid g_m)\Pr ior(g_m)$$

$$\prod_{i=1}^{N} Pen(x_i \mid g_i) Trans(g_i \mid g_d, g_m)$$

The genetic programs that implement the Elston-Stewart algorithm are LINKAGE (Lathrop et al. 1984), Mendel (Lange et al. 1988) and VITESSE (O'Connell and Weeks 1995).

Lander and Green (1987) proposed an alternative approach, the Lander-Green algorithm, to compute the pedigree likelihood function. The algorithm is based on a hidden Markov Model of the inheritance pattern at the ordered genetic loci (Sham 1997). The Lander-Green algorithm works as follows. Let $\mathbf{x}_L = (\mathbf{x}_{L1}, \mathbf{x}_{L2},\ldots, \mathbf{x}_{LN})$ denote the collection of phenotypes at locus *i*, and $\mathbf{g}_L = (\mathbf{g}_{L1}, \mathbf{g}_{L2},\ldots, \mathbf{g}_{LN})$ denote the collection of ordered genotypes at these loci for the individuals. Under the assumption of no crossover interference, the likelihood function can be defined as:

$$\sum_{g_{L_1} \in L_1} \cdots \sum_{g_{L_N} \in L_N} [\prod_i P(x_{L_i} \mid g_{L_i})] P(g_{L_N} \mid g_{L_{N-1}}) \ldots P(g_{L_2} \mid g_{L_1}) P(g_{L_1}).$$

The computational time of the Lander-Green algorithm is linear in the number of markers, but exponential in number of individuals in the pedigrees; consequently, the algorithm is complementary to that of Elston-Stewart. The algorithm is implemented in the programs Allegro (Gudbjartsson et al. 2000), GeneHunter (Kruglyak et al. 1996), Mendel and Merlin (Abecasis et al. 2002).

A.2 MODEL-FREE LINAKGE ANALYSES

In complex diseases, the mode of inheritance is often unclear. Power can drop significantly in parametric linkage methods if one assumes an incorrect disease model. As an alternative, model-free methods, also known as nonparametric methods, have been suggested and do not require explicit specification disease model (Penrose 1953; Green et al. 1983; Payami et al. 1985; Weeks and Lange 1988; Davis and Weeks 1997).

A.2.1. Identical-by-decent sharing pattern vs. relative pairs

The best known and most widely applied model-free approaches are the affected sib pair (ASP) and the affected relative pair (ARP) methods, which were introduced by Penrose (1935). The basic concept of these model-free methods is, if a specific genetic marker is not linked to a putative disease gene, then the identical-by-decent (IBD) sharing pattern among the relative pairs will depend only on their relationship and not on their disease status. However, if the marker is linked to the disease gene, one would expect an excess of IBD sharing among the affecteds over the null IBD sharing expectation. Taking sib pairs as an example, under the null hypothesis of no linkage, the probabilities of the pair sharing 0, 1, and 2 alleles IBD are ¼, ½, and ¼, respectively (Suarez et al. 1978). If a specific marker linked to the disease, one would expect an excess of IBD sharing in the affected sib pairs. Risch (1990a, b) provided a detailed description of the strategies, advantages and power of allele-sharing analysis to detect susceptibility genes in complex diseases.

During the past several years, varied extensions and modifications of the ASP/ARP approaches have been reported (Lange and Sobel 1991; Thomas and Cortessis 1992; Fulker and Cardon 1994; Kong and Cox 1997). Several studies have also pointed out that, for the diseases with high recurrence risk, using the discordant sib pair (DSP) approach provides substantial power to detect linkage as the ASP approaches (Risch and Zhang 1995; Rogus and Krolewski 1996).

A.2.2. Statistics

The statistics often used to test the null hypothesis in a model-free approach are introduced as follows. The first approach is the mean sharing statistic (Blackwelder and Elston 1985), which compares the mean number of alleles shared IBD with the expected sharing number under the null hypothesis, and can be written as:

Z = (Mean sharing – Expected sharing) / SD(sharing).

However, the mean sharing test is only applied to sibling data.

The second approach is the likelihood-ratio-test statistic. We first define Z_{null} as the vector of the IBD sharing probabilities in affected relative pairs under the null hypothesis. Let $\hat{Z}$ denote the estimated vector of the observed IBD sharing frequencies, and let n_i denote the number of affected relative pair with IBD sharing *i*. We then write the test as: $L(\hat{Z}) / L(Z_{null})$, "where $L(Z) \propto Z_0^{n_1} Z_1^{n_1} Z_2^{n_2}$ is the multinomial likelihood of the observed IBD sharing frequencies when the true IBD sharing probability is *Z*" (Shih and Whittemore 2001) and Z_i is the parameter corresponding to *i* IBD allele sharing. Under the H_0, twice the log-likelihood ratio has an asymptotic χ^2 distribution with two degrees of freedom. The likelihood-ratio test is usually applied to sibs.

The third approach is the allele-sharing statistic, which are often used to measure IBD sharing among affecteds within a pedigree (Whittemore and Halpern 1994; Kruglyak et al 1996). S_{pairs} and S_{all} are the two score functions most often used. The details are as follows.

We first denote the inheritance vector $v(x) = (p_1, m_1, p_2, m_2, \ldots, p_n, m_n)$, which captures an IBD sharing pattern between relatives at a given location *x*, where *n* is the number of non-founders, p_i represents the grandpaternal or grandmaternal allele from the paternal is transmitted to the individual *i*, m_i for the maternal is the similar definition as p_i. However, that a founder's alleles are inherited from the grandpaternal or grandmaternal side is often unobservable. Two identical-by-states are considered to be equivalent. The equivalent classes are called IBD configurations. Then the score function S_{pairs} can be defined as:

$$S_{pair}(\Pi) = \frac{2}{n(n-1)} \sum_{1 \le i \le j \le n} f_{ij}(v),$$

where ∏ is the IBD configuration, *n* is number of affected individuals, *v* is the inheritance vector in the IBD configuration ∏, and $f_{ij}(v)$ is one-fourth the number of alleles shared IBD between relative *i* and *j*.

The score function S_{all} can be written as:

$$S_{all}(\Pi) = \frac{1}{2^n} \sum_h \prod_{i=1}^{2f} c_i(h)!,$$

where ∏ is the IBD configuration, *n* is number of affected individuals, *f* is the number of founders, *h* is a collection of alleles obtained by choosing one allele from each affected individual, and $c_i(h)!$ is the number of times that founder allele *i* appears in the collection *h*.

Given a score function s_i on a location *x* in the *i*th pedigree, the test statistic is:

$$Z_i(x) = \frac{s_i(x, \Phi) - \mu_i}{\sigma_i},$$

where Φ is the phenotypic data, $\mu_i = E(s_i \mid H_0)$ and σ_i is the variance under the null hypothesis of no linkage for the *i*th pedigree. Note that Z_i has mean zero and variance one under H_0. The test statistic for N pedigrees is then:

$$Z = \frac{\sum_{i=1}^{N} r_i Z_i}{\sqrt{\sum_{i=1}^{N} r_i^2}},$$

where r_i is the weighting factor for the *i*th pedigree.

Various allele-sharing statistics have been incorporated into the programs Allegro (Gudbjartsson et al. 2000), GeneHunter (Kruglyak et al. 1996), GeneHunter-Plus (Kong and Cox 1997) and Merlin (Abecasis et al. 2002).

A.2.3. Constraints on the IBD sharing estimates

The likelihood-ratio-test statistic for ASP analysis (described in Section A.2.2) is a robust method for dealing with incomplete genotypic data. Moreover, Holmans (1993) introduced the "possible-triangle" method to improve power of the likelihood-ratio-test statistic by restricting maximization to the region that is consistent with a possible genetic model. He proved that the genetically-consistent allele-sharing estimates (Z_0, Z_1 and Z_2) fall within a possible triangle (Figure A.1).

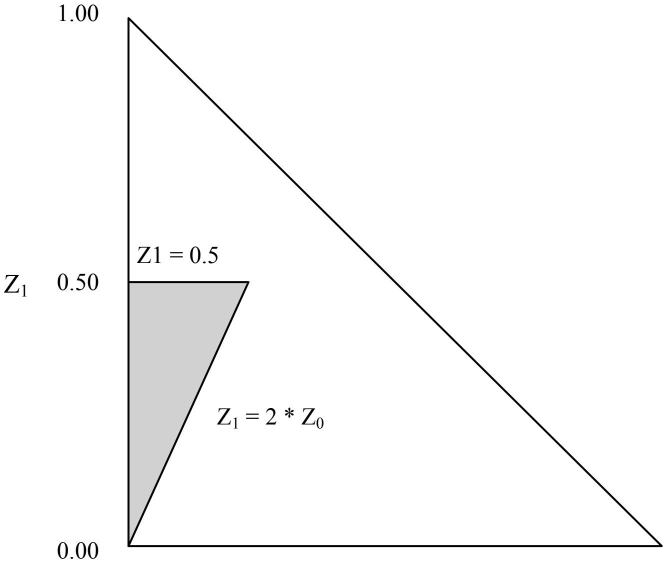


Figure A.1 The possible triangle of allele sharing estimates

The line $Z_1 = 0.5$ corresponds to the no dominance variance assumption, and the line $Z_1 = 2 * Z_0$ corresponds to the no additive variance assumption. Holmans (1993) showed that the likelihood-ratio test satisfying the possible triangle constraints has higher power than the general unrestricted likelihood-ratio test.

APPENDIX B

DATA SIMULATION PROGRAM

Introduction

The code in Appendix B will simulate data under a G x E model. In the Part I, the model-specifications are defined in order to generate the data under the corresponding disease model. The code in the Part II allows users to generate the genome-wide data under the model specified in the Part I. The examples of the R command and output file are provided in the Part III.

Part I: Parameter setting

```
# **************************
#  Model-specifications
# **************************

# LI = G1 + $\beta_{G1}$*C1 + C2 + PG + E
# model contains one gene, one interaction and one independent covariate
# proportion of var: 10% gene, 10% inter, 20% cov, 40% pg, 20% ran
# prevalance: 5%

mod <- function()
{
cat("model = Dm1 \n")

# model=(loci,a,d,-a,p)
model <- rbind(c(1,1,1,-1,0.01), c(2,1,1,-1,0.24))
colnames(model) <- c("loci","a","d","-a","p")

# beta=(NA,b1,b2,b3)
beta <- c(NA,1,1,0)
names(beta) <- c("NA","b1","b2","b3")

# mvr = (mu,var,rho)
mvr <- rbind(c(0,3.9196,0.8),c(0,0.156,0))
colnames(mvr) <- c("mu","var","rho")

# tri=(NA,mu1,mu2,mu3)
tri <- c(NA, 114.46, 0, -64.59)
names(tri) <- c("NA","mu1","mu2","mu3")
vtri <- 35.75

vpg <- 0.312
vran <- 0.156
threshold <- 0.4

return(list(model=model, beta=beta, mvr=mvr, tri=tri, vtri=vtri, vpg=vpg, vran=vran,
threshold=threshold))
}
```

Part II: Data simulation code

```
# Type I model: one gene, one G x E interaction and one independent covariate

library("MASS")

# --------------------------------------------------------
# Generate genotypes of disease loci for parents
# --------------------------------------------------------

# Input: pgeno(model,nloci)
# Output: return genotypes and allele types of each locus

pgeno <- function(model,nloci) {

allele <- NULL

gty <- function(m) {geno <- sample(2:4,1,replace=T, c(m[5]*m[5],2*m[5]*(1-m[5]),
        (1-m[5])*(1-m[5])));return(geno)}
gtype <- gty(model)

alle <- switch(paste(gtype), "1" = NULL, "2" = c(1,1), "3" = c(1,2), "4" = c(2,2))
allele <- rbind(allele,alle)

return(gtype,allele)
}
```

```
# -----------------------------------------------------
# Generate genotypes of disease loci for sibs
# Conditional on parents genotypes
# -----------------------------------------------------
# Input: sgeno(dalle,malle,nsib,nloci)
# Output: 1. return matrics w/ ONE allele randomly picked from parents
#           2. return sibs' genotypes of each locus
#           3. return the first two sibs' IBD sharing of each locus

sgeno <- function(dalle,malle,nsib,nloci) {

# Generate alleles of sibs from parents and track the origins of alleles

dibd.alle <- apply(dalle,1, function(alle) {
a <- sample(c(1,2),nsib,replace=T,c(0.5,0.5));return(a,alle[a])})
mibd.alle <- apply(malle,1, function(alle) {
a <- sample(c(1,2),nsib,replace=T,c(0.5,0.5));return(a,alle[a])})

dadibd <- NULL
dadall <- NULL
momibd <- NULL
momall <- NULL

for (i in 1:nloci) {
dadibd <- cbind(dadibd,dibd.alle[[i]][[1]])
dadall <- cbind(dadall,dibd.alle[[i]][[2]])
momibd <- cbind(momibd,mibd.alle[[i]][[1]])
momall <- cbind(momall,mibd.alle[[i]][[2]])
}

# Obtain genotype data from allele-type data

gtype <- dadall + momall

return(dadall,momall,gtype,dadibd,momibd)
}

# ----------------------------------------------------------------------
# Generate mean value for each genotype in each pedigree
# ----------------------------------------------------------------------
# Input: genom(nloci,gtype,model)
# gtype: genotype at a locus
# Output: return the mean value of each locus

genom <- function(nloci,gtype,model) {
```

```
pheno <- NULL
for (i in 1:nloci) {
  mg <- model[gtype[,i]]
  pheno <- cbind(pheno,mg)
}

return(pheno)
}

# -------------------------------------------------------------------------------
# Generate covariate values under multivariate normal distribution
# -------------------------------------------------------------------------------
# Input: covar(nper,mu,v,rho)
# nper: no. of persons in the family; mu: mean; v: var
# Output: return covariate values for each person

covar <- function(nper,mu,v,rho) {

sg <- matrix((rho*v),nper,nper)
diag(sg) <- (v)

return(mvrnorm(1,rep(mu,nper),sg))
}

# -------------------------------------------------------
# Generate genotype-specific covariate effect
# -------------------------------------------------------
# Input: eff(i,nloci,nper,tble,pedicov,b)
# i: indicator of the covarite w/ genotype-specific effect
# tble: current table contains the information of genotypes and covariate
# values; b: coefficient values in the 'beta' matrix
# Output: return the vaules of gene-covariate interaction for each person

eff <- function(i,nloci,nper,tble,pedicov,b) {

eff <- rep(NA,nper)
geno <- tble[,i]

# j is the indicator of persons in each pedigree

for (j in 1:nper) {
  eff[j] <- b[geno[j]]*pedicov[j,i]
}

return(eff)
```

```
}

# ---------------------------------------------------------------------------------------
# Generate marker genotypes (two-allele type) for parents
# ---------------------------------------------------------------------------------------
# Heterozygosity is 80% and 5 equal-freq alleles for each marker

pmark <- function(n.mar) {

par.mar <- vector("list", length(n.mar))
# names(par.mar) <- as.character(1:length(n.mar))

for(i in 1:length(n.mar)) {
  for(j in 1:n.mar[i]) {
    alle <- sample(1:5,2,replace=T,c(0.2,0.2,0.2,0.2,0.2))
    par.mar[[i]] <- cbind(par.mar[[i]],alle)
  }
colnames(par.mar[[i]]) <- NULL
}

return(par.mar)
}

# ----------------------------------------------------------
# Simulate no. of CE and place their positions
# ----------------------------------------------------------

cross <- function(len.chr,vpar,r) {

pos.chias <- vector("list", length(len.chr))
loc.chias <- vector("list", length(len.chr))
loc.cross <- vector("list", length(len.chr))
# vpar <- c(5,4,4.2,4.6,3.9,4.7,5.9,2.6,3.8,3.9,4.1,5.5,4.3,7.3,5,4.2,5.3,5,6.9,2.7,5.4,3.4)

for (i in 1:length(len.chr)) {

# Use rejection sampling to generate a sample from the first point
# distribution of a gamma renewal process with parameters (v, 2v).
# Here, v should be in the range of 2.6 to 7.3 in order for the chosen majorizing function to work.

# The Gamma distribution with parameters, `shape' = a, `scale' = s and 'rate' = r, has density:
# f(x)= 1/(s^a Gamma(a)) x^(a-1) e^-(x/s), for x > 0, a > 0 and s > 0.

if (vpar[i] < 2.6 || vpar[i] > 7.3) stop("ERROR: v out of range")

 repeat {
    # Generate uniform from (1,A), where A = area under h(x)
    x <- runif(1,min=0,max=(1.866666))
    # Invert to find y
    y <- log(1 - x*0.5357143)/(-1.5)
    # Generate U from the uniform distribution on the interval (0,1)
    u <- runif(1)
    # If u <= g(y)/h(y), deliver y
    pg <- 2*(1-pgamma(y,shape=vpar[i],rate=r[i]))
    ex <- 2.8*exp(-1.5*y)
    if (pg > ex) stop(paste("ERROR: pg = ",pg,">","ex =",ex))
    if (u <= pg/ex) break
  }

  pos.chias[[i]] <- y
  while (sum(pos.chias[[i]]) <= len.chr[i]) {
     pos.chias[[i]] <- c(pos.chias[[i]],rgamma(1,shape=vpar[i],rate=r[i]))
  }
  if (length(pos.chias[[i]]) >= 2) {
     pos.chias[[i]] <- pos.chias[[i]][1:(length(pos.chias[[i]])-1)]*100
  } else {
     pos.chias[[i]] <- 0
  }
  loc.chias[[i]][1] <- pos.chias[[i]][1]
  if (length(pos.chias[[i]]) >= 2) {
     for (j in 2:length(pos.chias[[i]])) {
        loc.chias[[i]][j] <- loc.chias[[i]][j-1]+pos.chias[[i]][j]
     }
  }
}

for(i in 1:22) {
   for (j in 1:length(loc.chias[[i]])) {
      x <- sample(1:2,1,replace=T)
      if (x == 1) {
         loc.cross[[i]] <- c(loc.cross[[i]],loc.chias[[i]][j])
      }
   }
   if (length(loc.cross[[i]])==0) {
       loc.cross[[i]] <- 0
   }
   names(loc.cross[[i]]) <- rep("C", length(loc.cross[[i]]))
}

return(loc.cross)
}
```

```
# ------------------------------------------------------------------------
# Generate map and decide the positions of markers
# ------------------------------------------------------------------------

map <- function(len, n.mar, eq.spacing = TRUE) {

n.chr <- length(n.mar)
map <- vector("list", n.chr)
names(map) <- as.character(1:n.chr)

  for (i in 1:n.chr) {
    if (eq.spacing) {
       map[[i]] <- seq(0, len[i]*100, length = n.mar[i])
    } else  {
       map[[i]] <- sort(c(map[[i]], runif(n.mar[i], 0, len[i]*100)))
    }
    names(map[[i]]) <- paste("D", names(map)[i], "M", 1:n.mar[i], sep = "")
  }

return(map)
}

# --  ---------------------------------------
# Check the number is odd or even
# ---  -------------------------------------

is.odd <- function (x) {

  if (is.numeric(x)) {
     if (x%%2 == 0) {
        FALSE
     } else {
        TRUE
     }
  } else {
    print("Warning: Input must be an integer value")
  }

}

is.even <- function (x) {

  if (is.numeric(x)) {
     if (x%%2 == 0) {
        TRUE
     } else {
        FALSE
      }
  } else {
     print("Warning: Input must be an integer value")
  }

}

# --------------------------------------------------------------------------------------------------------
# Generate the list of chromosome based on the locations of CE for UNLINKED chromosomes
# --------------------------------------------------------------------------------------------------------

list.unlink <- function(pos.mar, pos.cross) {

# Order positions of markers and CE in one vector

names(pos.mar) <- NULL
names(pos.cross) <- NULL
comb <- vector("list", 22)

for (i in 1:22) {
   comb[i] <- list(sort(c(unlist(pos.mar[i]),unlist(pos.cross[i]))))
}

# Generate the index list to pick which chromosome for data based on the CE

pick <- vector("list",1)

for (j in 10) {
  if (comb[[j]][2] != 0) {
     a <- rbind(pos.cross[[j]],pos.cross[[j]])
     b <- matrix(c(pos.mar[[j]][1],a,pos.mar[[j]][length(pos.mar[[j]])]), nrow=2)
     if (is.odd(ncol(b))) {
        b <- matrix(c(pos.mar[[j]][1],a,rep(pos.mar[[j]][length(pos.mar[[j]])],3)), nrow=2)
     }

     z <- sample(c(1,2),1,replace=T,c(0.5,0.5))
     for (k in 1:(length(b)/4)) {
        pick1 <- rep(z,length(comb[[j]][comb[[j]] > b[1,(2*k-1)] & comb[[j]] < b[1,2*k]]))
        pick2 <- rep((3-z),length(comb[[j]][comb[[j]] > b[2,(2*k-1)] & comb[[j]] < b[2,2*k]]))
        pick[[1]] <- c(pick[[1]],pick1,pick2)
     }
     pick[[1]] <- c(pick[[1]][1],pick[[1]],pick[[1]][length(pick[[1]])])
  } else {
     pick[[1]] <- rep(sample(c(1,2),1,replace=T,c(0.5,0.5)),length(pos.mar[[j]]))
  }
```

```
}

return(pick,comb[[10]],comb[[17]])
}

# ---------------------------------------------------------------------------------------------------
# Generate the list of chromosome based on
# odds or even no. of CE b/f disease loci for LINKED chromosomes (c10 & c17)
# ---------------------------------------------------------------------------------------------------

list.link <- function(pos.dis,chr10,len.dischr,dadibd,momibd,nloci,dischr,x) {

# Count the no. of CE b/f the positions of disease loci on c10 and c17
# "i" is for dad and mom

count.10 <- NULL
for (i in 1:2) {
  if (chr10[[i]][[2]] != 0) {
    cnt <- 0
    for (j in 1:len.dischr[1,i]) {
      if ((chr10[[i]][j] < pos.dis[1]) && (names(chr10[[i]][j]) == "C")) {
        cnt <- cnt + 1
      }
    }
  } else {
    cnt <- 0
  }
  count.10 <- c(count.10,cnt)
}

count <- count.10

# Pick the origin of chromosomes based on odds or even no. of CE
# x indicates the "x"th sib in the family

chrom.dad <- NULL
for (i in 1:nloci) {
 if (is.even(count[1])) {
   chr.dad <- dadibd[x,i]
 } else {
   chr.dad <- 3 - dadibd[x,i]
 }
 chrom.dad <- c(chrom.dad,chr.dad)
}

chrom.mom <- NULL
```

```
for (i in 1:nloci) {
   if (is.even(count[2])) {
      chr.mom <- momibd[x,i]
   } else {
      chr.mom <- 3 - momibd[x,i]
   }
   chrom.mom <- c(chrom.mom,chr.mom)
}

chrom <- rbind(chrom.dad,chrom.mom)
colnames(chrom) <- "c10"

# Generate a list containing the info of the vectors
# which indicate marker data coming from which chromosome

pick.dad <- vector("list", nloci)
pick.mom <- vector("list", nloci)
a <- list(chr10[[1]],chr10[[2]])

for (i in 1:nloci) {
   pick.dad[[i]] <- chrom[1,1]
   for (j in 2:len.dischr[1,1]) {
      if (names(a[[1]][j]) == "C") {
         chr <- 3 - pick.dad[[i]][j-1]
      } else {
         chr <- pick.dad[[i]][j-1]
      }
      pick.dad[[i]] <- c(pick.dad[[i]],chr)
   }

   pick.mom[[i]] <- chrom[2,1]
   for (j in 2:len.dischr[1,2]) {
      if (names(a[[i+1]][j]) == "C") {
          chr <- 3 - pick.mom[[i]][j-1]
      } else {
          chr <- pick.mom[[i]][j-1]
      }
      pick.mom[[i]] <- c(pick.mom[[i]],chr)
   }
}

index.dad <- vector("list",nloci)
index.mom <- vector("list",nloci)
list.dad <- vector("list",nloci)
list.mom <- vector("list",nloci)
```

```
for (i in 1:nloci) {
   idx.dad <- rbind(names(a[[i]]),pick.dad[[i]])
   idx.mom <- rbind(names(a[[i+1]]),pick.mom[[i]])
   index.dad[[i]] <- idx.dad
   index.mom[[i]] <- idx.mom

   list.dad[[i]] <- as.integer(index.dad[[i]][2, (index.dad[[i]][1,] != "C")])
   list.mom[[i]] <- as.integer(index.mom[[i]][2, (index.mom[[i]][1,] != "C")])
}

return(list.dad,list.mom)
}



# *********************************************
#  Main function generates the pedigree data
# *********************************************

pedi <- function(nped,model,beta,mvr,ncov,sdpgp,sdpgs,sdran,nloci,ncov.geno,threshold,n.mar,
            len.chr,dischr,pos.dis,pos.mar,vpar,r) {

tb1 <- NULL
tb2 <- NULL
family.unlink.data <- NULL
family.link.data <- NULL
# count.cross <- vector("list",22)

ped <- 1
total.nsib <- 0

while (ped <= nped) {
  tble1 <- NULL
  tble2 <- NULL

# Generate disease loci's genotypes of parents

  tble1 <- rbind(tble1,c(ped,1,0,0,1))
  dadg <- pgeno(model,nloci)
  tble2 <- rbind(tble2,dadg$gtype)
  tble1 <- rbind(tble1,c(ped,2,0,0,2))
  momg <- pgeno(model,nloci)
  tble2 <- rbind(tble2,momg$gtype)

# Generate disease loci's genotypes of kids conditional on parents
```

```
  nsib <- sample(2:5,1,replace=TRUE,prob=c(0.37551,0.29057,0.20237,0.13155))
  nper <- nsib+2

  gender <- sample(1:2, nsib, replace=T, prob=c(0.5,0.5))
  for(i in 1:nsib) {
    tble1 <- rbind(tble1,c(ped,(i+2),1,2,gender[i]))
  }
  sibg <- sgeno(dadg$allele,momg$allele,nsib,nloci)
  dadall <- sibg$dadall; momall <- sibg$momall
  sgtype <- matrix(sibg$gtype,nrow=nsib)
  dadibd <- sibg$dadibd; momibd <- sibg$momibd
  tble2 <- rbind(tble2,sgtype)

# Generate means of the major genes in a nper * nloci matrix for each pedigree

gmean <- genom(nloci,tble2,model)
 tble2 <- cbind(tble2,gmean)

# Generate genotype-specific covariate values within a family

  famcovg <- NULL
  pedicovg <- NULL
  for (i in 1:ncov.geno) {
    famcovg <- covar(nper,mvr[i,1],mvr[i,2],mvr[i,3])
    famcovg <- matrix(famcovg,nrow=(nper))
    pedicovg <- cbind(pedicovg,famcovg)
    tble2 <- cbind(tble2,famcovg)
  }

# Generate non-genotype-effect covariate values within a family

  famcov <- NULL
  pedicov <- NULL
  for (i in (ncov.geno+1):ncov) {
    famcov <- covar(nper,mvr[i,1],mvr[i,2],mvr[i,3])
    famcov <- matrix(famcov,nrow=(nper))
    pedicov <- cbind(pedicov,famcov)
    tble2 <- cbind(tble2,famcov)
  }

# Generate genotype-specific covariate effect

  inter <- NULL
  for (i in 1:ncov.geno) {
    coveff <- eff(i,nloci,nper,tble2,pedicovg,beta)
    coveff <- matrix(coveff,nrow=(nper))
```

```
      inter <- cbind(inter,coveff)
      tble2 <- cbind(tble2,coveff)
   }

# Generate polygenic-effect values

   dadp <- rnorm(1,0,sdpgp); momp <- rnorm(1,0,sdpgp)
   sibp <- NULL
   sibp <- c(sibp, rnorm(nsib,(dadp+momp)/2,sdpgs))
   poly <- c(dadp,momp,sibp)
   poly <- matrix(poly,nrow=(nper))
    tble2 <- cbind(tble2,poly)

# Generate random-effect values

   raneff <- rnorm(nper,0,sdran)
   raneff <- matrix(raneff,nrow=(nper))
   tble2 <- cbind(tble2,raneff)

# Define affection status based on LI and count the no. of affected sibs

   li <- gmean+pedicov+inter+poly+raneff

   affect <- li
   affect[li > threshold] <- 2; affect[li <= threshold] <- 1
   aff <- affect[3:nper]
   naff <- length(aff[aff==2])

   tble1 <- cbind(tble1,affect)
   tble2 <- cbind(tble2,li)

   if (naff >= 2) {
#   tble2 <- cbind(tble2,gmean,pedicovg,pedicov,inter,poly,raneff,li)
     tb1 <- rbind(tb1,tble1)
     tb2 <- rbind(tb2,tble2)

# Generate parent's data and offspring linked and unlinked marker data

     dad.mar <- pmark(n.mar); mom.mar <- pmark(n.mar)
     offs.mar <- vector("list",nsib)

     for (x in 1:nsib) {

# Generate no. and the location of CE
# pos.cross, pos.mar and list.chr are lists
# 2 here is for dad and mom

   list.final <- vector("list",2)
   chr10 <- vector("list",2)
   chr17 <- vector("list",2)

# "k" is for dad and mom

   for (k in 1:2) {
    pos.cross <- cross(len.chr,vpar,r)

    unlink.chr <- list.unlink(pos.mar, pos.cross)
    list.final[[k]] <- unlink.chr[[1]]

    link10 <- unlink.chr[[2]]; link17 <- unlink.chr[[3]]
    chr10[[k]] <- link10; chr17[[k]] <- link17

# Record CE position for all sibs in all simulated pedigrees

#   for (m in 1:22) {
#    if (pos.cross[[m]] != 0) {
#     count.cross[[m]] <- append(count.cross[[m]],pos.cross[[m]])
#    }
#    if (length(count.cross[[m]]) == 0) {
#     count.cross[[m]] <- -1
#    }
#   }

   }

   len.dischr <- rbind(c(length(chr10[[1]]),length(chr10[[2]])),
                   c(length(chr17[[1]]),length(chr17[[2]])))

# Generate offspings' unlinked marker data conditional on:
# 1. parents' marker data; 2. the locations of CE
# "1" is for dad; "2" is for mom

   offs.unlink <- vector("list", 1)

   for (i in 10) {
     dadt <- t(dad.mar[[i]]); momt <- t(mom.mar[[i]])
     dad.final <- t(rbind(c(1:length(list.final[[1]][[1]])),list.final[[1]][[1]]))
     mom.final <- t(rbind(c(1:length(list.final[[2]][[1]])),list.final[[2]][[1]]))
     offs.unlink[[1]] <- rbind(dadt[dad.final],momt[mom.final])
   }
```

```
# Generate offsprings linked marker data conditional on:
# 1. parents' marker data; 2. the location of disease loci
# 3. odds or even no. of CE before the position of disease loci

    list.offs <- list.link(pos.dis, chr10, len.dischr, dadibd, momibd, nloci, dischr, x)

    offs.link <- vector("list", nloci)

    for (i in 1:nloci) {
      for (j in 1:(n.mar[dischr[i],1])) {
        kid.linmar <- rbind(dad.mar[[dischr[i]]][list.offs$list.dad[[i]][j],j],
                            mom.mar[[dischr[i]]][list.offs$list.mom[[i]][j],j])
        offs.link[[i]] <- cbind(offs.link[[i]], kid.linmar)
      }
    }

    offs.mar[[x]] <- c(offs.unlink,offs.link)
   }

# Generate a LINKAGE-format matrix w/ all marker data

   dad.tb1 <- tble1[1,]; mom.tb1 <- tble1[2,]

   for (i in 10) {
     dad.tb1 <- c(dad.tb1,matrix(dad.mar[[i]],nrow=1))
     mom.tb1 <- c(mom.tb1,matrix(mom.mar[[i]],nrow=1))
   }

   dad.tb2 <- c(tble1[1,],dad.mar[[10]]); mom.tb2 <- c(tble1[2,],mom.mar[[10]])

   sib.tb1 <- vector("list",nsib); sib.tb2 <- vector("list",nsib)
   offs.tb1 <- NULL; offs.tb2 <- NULL

   for (i in 1:nsib) {
     sib.tb1[[i]] <- tble1[(i+2),]
     sib.tb2[[i]] <- c(tble1[(i+2),],matrix(offs.mar[[i]][[2]],nrow=1))

     for (j in 1) {
       sib.tb1[[i]] <- c(sib.tb1[[i]],matrix(offs.mar[[i]][[j]],nrow=1))
     }
     offs.tb1 <- rbind(offs.tb1,sib.tb1[[i]]); offs.tb2 <- rbind(offs.tb2,sib.tb2[[i]])
   }

   family.unlink.data <- rbind(family.unlink.data, dad.tb1,mom.tb1,offs.tb1)
   family.link.data <- rbind(family.link.data, dad.tb2,mom.tb2,offs.tb2)
```

```
    total.nsib <- total.nsib + nsib
    ped <- ped + 1
  }
}

# Generate the title names for tables

colnames(tb1) <- c("ped","per","dad","mom","sex","affect")

name.gene <- NULL
for(i in 1:nloci) {
  name.gene <- c(name.gene,paste("geno",i,sep=""))
}
for (i in 1:nloci) {
  name.gene <- c(name.gene,paste("mg",i,sep=""))
}

name.cov <- NULL
for (i in 1:ncov) {
  name.cov <- c(name.cov,paste("cov",i,sep=""))
}

name.int <- NULL
for (i in 1:ncov.geno) {
  name.int <- c(name.int,paste("gc",i,sep=""))
}

colnames(tb2) <- c(name.gene,name.cov,name.int,"pg","e","li")
rownames(tb2) <- NULL

# Convert one genotype into two alleles and replace the genotype data in tb2

allele1 <- NULL
allele2 <- NULL

for (i in 1:nloci) {
  for (j in 1:(length(tb2)/length(colnames(tb2)))) {
    alle <- switch(paste(tb2[j,i]), "1" = NULL, "2" = c(1,1), "3" = c(1,2), "4" = c(2,2))
    allele1 <- rbind(allele1,alle)
  }
  allele2 <- cbind(allele2,allele1)
  allele1 <- NULL
}

name.alle <- NULL
for (i in 1:nloci) {
  for (j in 1:2) {
    name.alle <- c(name.alle,paste("g",i,j,sep=""))
  }
}

# Generate a matrix w/ pedigree structure, info of disease loci, interaction and covariates

cov.data <- cbind(tb1,round(tb2,digits=3))

# colnames(allele2) <- name.alle
# rownames(allele2) <- NULL
rownames(family.unlink.data) <- NULL
rownames(family.link.data) <- NULL
colnames(cov.data) <- c(colnames(tb1),colnames(tb2))
rownames(cov.data) <- NULL

# group <- vector("list",22)
# count.rf <- vector("list",22)

# Summarize the counts of recombination fraction in each interval of markers

# for (i in 1:22) {
#   group[[i]] <- cut(count.cross[[i]],pos.mar[[i]])
#   count.rf[[i]] <- tapply(count.cross[[i]],group[[i]],function (x) {length(x)/(total.nsib*2)})
# }

# tb2 <- cbind(allele2,tb2[,(nloci+1):(length(colnames(tb2)))])

colnames(family.unlink.data) <- NULL
colnames(family.link.data) <- NULL

return(family.unlink.data,family.link.data,cov.data)
}


# ***************************************************
#   Start main program to generate pedigrees and replicates
# ***************************************************

sim <- function(nped) {

# Adjust the values of parameters in the model

model.value <- mod()
model <- model.value$model
beta <- model.value$beta
```

```
mvr <- model.value$mvr
sdpgp <- sqrt(model.value$vpg)
sdpgs <- sqrt(model.value$vpg/2)
sdran <- sqrt(model.value$vran)
threshold <- model.value$threshold

nloci <- length(model)/5
ncov.geno <- length(beta)/4
ncov <- length(mvr)/3

print(mod())
cat("No. of loci \n"); print(nloci)
cat("No. of covariates \n"); print(ncov)
cat("No. of covariates with genotype-specific effect \n"); print(ncov.geno)

if (nloci <= 0) stop("No. of loci must be at least one")
if (nloci < ncov.geno)
stop("No. of loci must be more than no. of covaraites w/ genotype-specific effect")
if (ncov < ncov.geno)
stop("No. of covariates must be more than no. of covariates w/ genotype-specific effect")

# Specify the total lengths of chromosomes and the corresponding Gamma paramters

len.chr <- matrix(c(2.84,2.62,2.19,2.07,1.98,1.89,1.79,1.64,1.6,1.69,1.46,1.69,1.15,
            1.28,1.17,1.29,1.26,1.25,1.01,0.96,0.5,0.57),nrow=1)
vpar <- c(5,4,4.2,4.6,3.9,4.7,5.9,2.6,3.8,3.9,4.1,5.5,4.3,7.3,5,4.2,5.3,5,6.9,2.7,5.4,3.4)
r <- 2*vpar

# Generate no. of markers on a chromosome

n.mar <- apply(len.chr,1, function(a) {as.integer(a*100/5)})
# cat(total.mar,"\n")

# Generate the positions and labels of marker data

pos.mar <- map(len.chr, n.mar)

# Specify the position of disease loci

dischr <- c(10,17)
pos.dis <- c(50, 30)
cat("Location of disease loci are on the chromosome:", dischr,"\n")

pedigree <- pedi(nped,model,beta,mvr,ncov,sdpgp,sdpgs,sdran,nloci,
              ncov.geno,threshold,n.mar,len.chr,dischr,pos.dis,pos.mar,vpar,r)
```

```
write.matrix(pedigree$family.unlink.data,file=paste("pedin_all.",nped,sep=""))
write.matrix(pedigree$family.link.data,file=paste("pedin_dis.",nped,sep=""))
write.table(pedigree$cov.data,file=paste("covariate.",nped,sep=""),col.names=T,
            row.names=F,sep="  ",quote=F)

# return(pedigree$count.rf)
}
```

Part III: Examples

For running the data simulation program, users need to prepare a model-specification R code, which is provided in the Part I. After evoking R environment, the R commands are as follows:

```
# call in the model-specification R code, which is named as ("Dm1.R") here
source("Dm1.R")

# call in the data simulation R code, which is named as ("main_1g1cov1int.R ") here
source("main_1g1cov1int.R")

# use sim() function in the data simulation R code to generate 100 pedigrees
sim(100)
```

The program puts its results in three output files: “covariate.*”, “pedin_all.*” and “pedin_dis.*”, * here means number of pedigrees in the simulated data set. The “covariate.*” file contains the covariate information. The simulated marker data without and with linking to the disease locus are in the “pedin_all.*” and “pedin_dis.*” files, separately.

APPENDIX C

STATISTICAL PROGRAM

Introduction

The code in Appendix C will analyze the data using four covariate methods: mixture model, LODPAL, MLRM and COVLINK, and write out the results into the “summary.txt” file. Before running this program, users need to obtain the IBD sharing probabilities by running the GeneHunter program. The statistical code is provided in the Part I. The examples of the R command and output file are provided in the Part II.

Part I: Statistical program

```
# This code implements four covariate methods: mixture model, LODPAL, MLRM and
COVLINK

stat <- function(nped,repli,chr)
{

library(MASS)
library(mclust)
source("chiglm.R")

# Compute LR(alpha, lamda); lrwa: LR with alpha values
# function for the mixture model

lrwa <- function(lamda,data)
{
  z0 <- data[,8]
  z1 <- data[,9]
  z2 <- data[,10]
  alpha <- data[,11]

return(-2*sum(log(alpha*(z0/lamda+z1+z2*(2-(1/lamda)))+(1-alpha))))
}

# Compute LR(alpha=1, lamda); lraone: LR with alpha=1
# function for the mixture model

lraone <- function(lamda,data)
{
  z0 <- data[,8]
  z1 <- data[,9]
  z2 <- data[,10]

return(-2*sum(log(z0/lamda+z1+z2*(2-(1/lamda)))))
}

# IBD sharing z0, z1 and z2 under the null for sib pairs are 0.25, 0.5, 0.25
# alp.mode is a mode of inheritance parameter in LODPAL in SAGE package
# deminant -> alp.mode = 1; recesive -> alp.mode = 10 or go to Inf
# we use default alp.mode value: 2.634 for minmax model
# we specify w = 1 in the model -> w: weights corresponding to the pairs

lodpal <- function(param,data)
{

beta <- param[1]
delta <- param[2]

posit <- data[1,2]

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.cen <- data[,8]
est.par <- exp(beta+delta*cov.cen)

lr <- (z0 + z1*est.par + z2*(3.634*est.par-2.634))/
        (0.25 + 0.5*est.par + 0.25*(3.634*est.par-2.634))

if (lr <= 0 | lr == "NaN") {
  warning("LR is negative or NaN")
}

lr[lr=="NaN"] <- 1e-20
lr[lr<=0] <- 1e-20

return(sum(log10(lr)))
}

# MLRM under no dominance assumption

epart.1 <- function(beta00,beta01,data)
{

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]

ez0 <- exp(beta00+beta01*cov.pair)/(2+2*exp(beta00+beta01*cov.pair))
ez2 <- 0.5-ez0

denom <- 4*z0*ez0 + z1 + 4*z2*ez2

zi0 <- (4*z0*ez0) / denom
zi1 <- z1 / denom
zi2 <- (4*z2*ez2) / denom

return(zi0,zi1,zi2)
}

mpart.1 <- function(param,data)
{

beta00 <- param[1]
beta01 <- param[2]

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]
```

```
est.z0 <- exp(beta00+beta01*cov.pair)/(2+2*exp(beta00+beta01*cov.pair))

lod <- z0*log(est.z0)+z2*log(0.5-est.z0)

if (lod == "NaN") {
  warning("lod in model 1 is NaN")
}

lod[lod == "NaN"] <- 1e-20

return(sum(lod))
}

# MLRM under no additive assumption

epart.2 <- function(beta00,beta01,data)
{

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]

ez0 <- exp(beta00+beta01*cov.pair)/(3+3*exp(beta00+beta01*cov.pair))
ez1 <- 2*ez0
ez2 <- 1-ez0-ez1

denom <- 4*z0*ez0 + 2*z1*ez1 + 4*z2*ez2

zi0 <- (4*z0*ez0) / denom
zi1 <- (2*z1*ez1) / denom
zi2 <- (4*z2*ez2) / denom

return(zi0,zi1,zi2)
}

mpart.2 <- function(param,data)
{

beta00 <- param[1]
beta01 <- param[2]
```

```
# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]

est.z0 <- exp(beta00+beta01*cov.pair)/(3+3*exp(beta00+beta01*cov.pair))

lod <- z0*log(est.z0)+z1*log(2*est.z0)+z2*log(1-3*est.z0)

if (lod == "NaN") {
  warning("lod in model 2 is NaN")
}

lod[lod == "NaN"] <- 1e-20

return(sum(lod))
}

# MLRM under min-max restriction

epart.3 <- function(beta00,beta01,data)
{

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]

ez0 <- 0.645*exp(beta00+beta01*cov.pair)/(1.58+1.58*exp(beta00+beta01*cov.pair))
ez1 <- 0.355+0.58*ez0
ez2 <- 1-ez0-ez1

denom <- 4*z0*ez0 + 2*z1*ez1 + 4*z2*ez2

zi0 <- (4*z0*ez0) / denom
zi1 <- (2*z1*ez1) / denom
zi2 <- (4*z2*ez2) / denom

return(zi0,zi1,zi2)
}
```

```
mpart.3 <- function(param,data)
{

beta00 <- param[1]
beta01 <- param[2]

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]

est.z0 <- 0.645*exp(beta00+beta01*cov.pair)/(1.58+1.58*exp(beta00+beta01*cov.pair))

lod <- z0*log(est.z0)+z1*log(0.335+0.58*est.z0)+z2*log(0.645-1.58*est.z0)

if (lod == "NaN") {
  warning("lod in model 3 is NaN")
}

lod[lod == "NaN"] <- 1e-20

return(sum(lod))
}

# Compute LOD score for MLRM under no dominance assumption

lod.score1 <- function(beta00,beta01,data)
{

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]

ez0 <- exp(beta00+beta01*cov.pair)/(2+2*exp(beta00+beta01*cov.pair))
ez1 <- 0.5
ez2 <- 0.5-ez0

lod <- log10(4*z0*ez0+2*z1*ez1+4*z2*ez2)
```

```
return(sum(lod))
}

# Compute LOD score for MLRM under no additive assumption

lod.score2 <- function(beta00,beta01,data)
{

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]

ez0 <- exp(beta00+beta01*cov.pair)/(3+3*exp(beta00+beta01*cov.pair))
ez1 <- 2*ez0
ez2 <- 1-ez0-ez1

lod <- log10(4*z0*ez0+2*z1*ez1+4*z2*ez2)

return(sum(lod))
}

# Compute LOD score for MLRM under min-max restriction

lod.score3 <- function(beta00,beta01,data)
{

# estimated prob. of IBD sharing

z0 <- data[,5]
z1 <- data[,6]
z2 <- data[,7]

cov.pair <- data[,8]

ez0 <- 0.645*exp(beta00+beta01*cov.pair)/(1.58+1.58*exp(beta00+beta01*cov.pair))
ez1 <- 0.355+0.58*ez0
ez2 <- 1-ez0-ez1

lod <- log10(4*z0*ez0+2*z1*ez1+4*z2*ez2)

return(sum(lod))
```

```
}

# *********************************************************
# Start reading covariate information and preparing proper mean values
# *********************************************************

# Read in covariate information

fam.cov <- read.table(file=paste("covariate.",nped,sep=""),header=T)
attach(fam.cov)
names(fam.cov)

# Figure out the no. of members in each family
# Compute the family-level average using all members data

fam <- as.factor(fam.cov$ped)
lev.fam <- levels(fam)
ped <- unique(fam.cov$ped)
nmemb <- tapply(fam.cov$per,fam,max)
lev.cov1 <- tapply(fam.cov$cov1,fam,mean)

# Read in IBD sharing file, which is generated from GH

ibd <- read.table(paste("ibd.c",chr,sep=""),skip=1)
attach(ibd)
names(ibd) <- c("pos","ped","pair","pz0","pz1","pz2","z0","z1","z2")

# Compute pair-level covariate mean using all sibs

pair.mean <- NULL

for (i in 1:length(lev.fam)) {
  b <- fam.cov[fam.cov$ped==lev.fam[i],]
  for (j in 3:(nmemb[i]-1)) {
    for (k in (j+1):nmemb[i]) {
      pmean <- cbind(i,j,k,b$affect[j],b$affect[k],
              mean(c(b$cov1[j],b$cov1[k])),mean(c(b$cov2[j],b$cov2[k])),
              mean(c(b$ran.cov[j],b$ran.cov[k])))
      pair.mean<- rbind(pair.mean,pmean)
    }
  }
}

# Compute pair-level covariate mean using affected sib-pair only

ap.mean <- pair.mean[pair.mean[,4]==2 & pair.mean[,5]==2,]
ap.mean <- data.frame(ap.mean,fac=(paste(ap.mean[,2],",",ap.mean[,3],sep="")))

aspm <- cbind(ap.mean[,1],ap.mean[,9],ap.mean[,6:8])
colnames(aspm) <- c("ped","pair","mc1","mc2","mc3")
aspm$key <- paste(aspm$ped, aspm$pair)

# Center the mean (x-xbar)

ap.mean <- cbind(ap.mean,ap.mean[,6]-mean(ap.mean[,6]),
               ap.mean[,7]-mean(ap.mean[,7]),ap.mean[,8]-mean(ap.mean[,8]))
ap.mean <- cbind(ap.mean[,1],ap.mean[,9:12])
colnames(ap.mean) <- c("ped","pair","cmcov1","cmcov2","cmcov3")
ap.mean$key <- paste(ap.mean$ped, ap.mean$pair)

# Cluster families by the covariate values and estimate the alpha value of each family

  hcTree <- hcE(as.matrix(lev.cov1))
  clust.fam <- hclass(hcTree,2)
  cf <- as.data.frame(cbind(clust.fam,lev.cov1))
  names(cf) <- c("clust.fam","lev.cov1")
  alp <- qda(clust.fam~lev.cov1,data=cf,method='mle')
  i <- NULL
  i[alp$mean[1] > alp$mean[2]] <- 1
  i[alp$mean[1] <= alp$mean[2]] <- 2
  alpha <- as.vector(predict.qda(alp,cf)$posterior[,i])
  fam.alp <- as.data.frame(cbind(ped,alpha))

# Merge IBD sharing with alpha for all sib pairs in each pedigree

  sibd <- ibd[ibd$pz0==0.25 & ibd$pz1==0.5 & ibd$pz2==0.25,]
  sibd$key <- paste(sibd$ped, sibd$pair)
  sibd.alpha <- merge(sibd,fam.alp,by='ped')
  sibd.alp <- merge(sibd.alpha,ap.mean,by='key')

#*******************
#  The mixture model
#*******************

  pos <- unique(sibd$pos)

  res.mix <- NULL

  for (i in 1:length(pos))
  {
    pos.ibd <- sibd.alp[sibd.alp$pos==pos[i],]
```

```
   a <- optim(1.2,lrwa,method = "L-BFGS-B",lower=1,upper=30,data=pos.ibd)

   lam.hat <- a$par
   LOD <- log10(exp(-1*a$value/2))
#    pval <- 0.5 - 0.5*pchisq(-1*a$value,1)
   pos.alp <- cbind(pos[i],LOD)
   res.mix <- rbind(res.mix,pos.alp)
  }

# Generate the subset of the 'sibd' object

sibd <- cbind(sibd[,1:3],sibd[,7:10])

# Merge IBD-sharing prob w/ pair-specific mean for each pair in the same family
# the mean covariate values are centered

ap.mean <- ap.mean[,3:6]
sibd.cmean <- merge(sibd, ap.mean, by='key')

#**************
#  LODPAL
#**************

res.lodpal <- NULL

  for (i in 1:length(pos))
  {
   pos.ibd <- sibd.cmean[sibd.cmean$pos==pos[i],]

   a <- optim(c(0.002,0.002),lodpal,method="L-BFGS-B",lower=c(0,-5),
          upper=c(5,5),data=pos.ibd,control=list(fnscale=-1))

   beta.hat <- round(a$par[1],digits=3)
   delta.hat <- round(a$par[2],digits=3)
   LOD <- round(a$value,digits=5)
   pos.lod <- cbind(pos[i],LOD)
   res.lodpal <- rbind(res.lodpal,pos.lod)
  }

# Creat two dummy variables for AA, AU, UU pairs

aff1 <- NULL
aff1[pair.mean[,4]==2 & pair.mean[,5]==2] <- 1
aff1[pair.mean[,4]==1 & pair.mean[,5]==2] <- 1
aff1[pair.mean[,4]==2 & pair.mean[,5]==1] <- 1
aff1[pair.mean[,4]==1 & pair.mean[,5]==1] <- 0
```

```
aff2 <- NULL
aff2[pair.mean[,4]==2 & pair.mean[,5]==2] <- 1
aff2[pair.mean[,4]==1 & pair.mean[,5]==2] <- 0
aff2[pair.mean[,4]==2 & pair.mean[,5]==1] <- 0
aff2[pair.mean[,4]==1 & pair.mean[,5]==1] <- 0


# Compute pair-specific covariate mean using all sib pairs

clink.mean <- data.frame(pair.mean,fac=(paste(pair.mean[,2],",",pair.mean[,3],sep="")))
clink.mean <- cbind(clink.mean[,1],clink.mean[,9],clink.mean[,6:8],aff1,aff2)
colnames(clink.mean) <- c("ped","pair","mcov1","mcov2","mcov3","aff1","aff2")

# Merge IBD-sharing prob w/ pair-specific mean for each pair in the same family

clink.mean$key <- paste(clink.mean$ped, clink.mean$pair)
clink.mean <- clink.mean[,3:8]
sibd.mean <- merge(sibd, clink.mean, by='key')

# Extract the data w/ unambiguous IBD sharing

unamb.ibd <- sibd.mean[(sibd.mean$z0==1 & sibd.mean$z1==0 &
sibd.mean$z2==0) | (sibd.mean$z0==0 & sibd.mean$z1==1 & sibd.mean$z2==0) |
(sibd.mean$z0==0 & sibd.mean$z1==0 & sibd.mean$z2==1),]

ibd.zero <- cbind(unamb.ibd[unamb.ibd$z0==1,],0)
names(ibd.zero) <- c(names(unamb.ibd), "cnt")
ibd.one.1 <- cbind(unamb.ibd[unamb.ibd$z1==1,],0)
names(ibd.one.1) <- c(names(unamb.ibd), "cnt")
ibd.one.2 <- cbind(unamb.ibd[unamb.ibd$z1==1,],1)
names(ibd.one.2) <- c(names(unamb.ibd), "cnt")
ibd.two <- cbind(unamb.ibd[unamb.ibd$z2==1,],1)
names(ibd.two) <- c(names(unamb.ibd), "cnt")

mer.ibd <- rbind(ibd.zero,ibd.zero,ibd.one.1,ibd.one.2,ibd.two,ibd.two)

#*************
# COVLINK
#*************

una.pos <- sort(unique(mer.ibd$pos))
res.covlink <- NULL

  for (i in 1:length(una.pos))
  {
```

```
  pos.mer <- mer.ibd[mer.ibd$pos==una.pos[i],]

  clink <- glm(pos.mer$cnt~pos.mer$aff1+pos.mer$aff2+pos.mer$mcov1,
         data=pos.mer,family=binomial)

  chi <- round(chiglm(clink)[[1]],digits=3)
  clink <- cbind(una.pos[i],chi)
  res.covlink <- rbind(res.covlink,clink)
  rownames(res.covlink) <- NULL
 }

# Merge IBD-sharing prob w/ pair-specific mean for each pair in the same family
# the mean values are not centered or standardized

aspm <- aspm[,3:6]
sib.mean <- merge(sibd, aspm, by='key')


#***********************
# The MLRM approaches
#***********************

# Start the EM steps for the model under no dominance assumption

res.bull <- NULL
miter <- 30

 for (i in 1:length(pos))
 {
  pos.ibd <- sib.mean[sib.mean$pos==pos[i],]
  iter <- 1
  convg <- F

# Initialization of the EM steps

  beta00 <- 0.02
  beta01 <- 0.02
  zi0 <- 0.25
  zi1 <- 0.5
  zi2 <- 0.25
  lod.mstep <- 1e+7

  while (iter < miter) {

# Set the zi's values in the previous E-step to 'zi*.old'
```

```
    zi0.old <- zi0
    zi1.old <- zi1
    zi2.old <- zi2
    lod.mstepold <- lod.mstep

# Run E-step to compute zi's values

    a <- epart.1(beta00,beta01,pos.ibd)
    zi0 <- a[[1]]
    zi1 <- a[[2]]
    zi2 <- a[[3]]

# Run M-step to estimate beta00  and beta01

    b <- optim(c(beta00,beta01),mpart.1,method="BFGS",data=pos.ibd,control=list(fnscale=-1))

    beta00 <- b$par[1]
    beta01 <- b$par[2]
    lod.mstep <- b$value

# Check if the program converges

    if (abs(lod.mstep-lod.mstepold) < 1e-10) {
      convg <- T
      break
    }

    iter <- iter + 1
    }

# Compute the LOD scores based on the estimated parameters

    LOD <- round(lod.score1(beta00,beta01,pos.ibd),digits=5)
    pos.lod <- cbind(pos[i],LOD)
    res.bull <- rbind(res.bull,pos.lod)
  }

# Record the results to the object -- m1

res.m1 <- res.bull
# colnames(m1) <- c("pos","beta00","beta01","LOD","iter","convg","meth")

# Start the EM steps for the model under no additive assumption

res.bull <- NULL
miter <- 30
```

```
 for (i in 1:length(pos))
 {
  pos.ibd <- sib.mean[sib.mean$pos==pos[i],]
  iter <- 1
  convg <- F

# Initialization of the EM steps

  beta00 <- 0.02
  beta01 <- 0.02
  zi0 <- 0.25
  zi1 <- 0.5
  zi2 <- 0.25
  lod.mstep <- 1e+7

  while (iter < miter) {

# Set the zi's values in the previous E-step to 'zi*.old'

  zi0.old <- zi0
  zi1.old <- zi1
  zi2.old <- zi2
  lod.mstepold <- lod.mstep

# Run E-step to compute zi's values

  a <- epart.2(beta00,beta01,pos.ibd)
  zi0 <- a[[1]]
  zi1 <- a[[2]]
  zi2 <- a[[3]]

# Run M-step to estimate beta00 and beta01

  b <- optim(c(beta00,beta01),mpart.2,method="BFGS",data=pos.ibd,control=list(fnscale=-1))

  beta00 <- b$par[1]
  beta01 <- b$par[2]
  lod.mstep <- b$value

# Check if the program converges

  if (abs(lod.mstep-lod.mstepold) < 1e-10) {
    convg <- T
    break
  }
```

```
    iter <- iter + 1
    }

# Compute the LOD scores based on the estimated parameters

    LOD <- round(lod.score2(beta00,beta01,pos.ibd),digits=5)
    pos.lod <- cbind(pos[i],LOD)
    res.bull <- rbind(res.bull,pos.lod)
  }

# Record the results to the object -- m2

res.m2 <- res.bull
# colnames(m2) <- c("pos","beta00","beta01","LOD","iter","convg","meth")

# Start the EM steps for the min-max restriction model

res.bull <- NULL
miter <- 30

  for (i in 1:length(pos))
  {
    pos.ibd <- sib.mean[sib.mean$pos==pos[i],]
    iter <- 1
    convg <- F

# Initialization of the EM steps

    beta00 <- 0.02
    beta01 <- 0.02
    zi0 <- 0.25
    zi1 <- 0.5
    zi2 <- 0.25
    lod.mstep <- 1e+7

    while (iter < miter) {

# Set the zi's values in the previous E-step to 'zi*.old'

    zi0.old <- zi0
    zi1.old <- zi1
    zi2.old <- zi2
    lod.mstepold <- lod.mstep

# Run E-step to compute zi's values
```

```
    a <- epart.3(beta00,beta01,pos.ibd)
    zi0 <- a[[1]]
    zi1 <- a[[2]]
    zi2 <- a[[3]]

# Run M-step to estimate beta00 and beta01

    b <- optim(c(beta00,beta01),mpart.3,method="BFGS",data=pos.ibd,control=list(fnscale=-1))

    beta00 <- b$par[1]
    beta01 <- b$par[2]
    lod.mstep <- b$value

# Check if the program converges

    if (abs(lod.mstep-lod.mstepold) < 1e-10) {
      convg <- T
      break
    }

    iter <- iter + 1
    }

# Compute the LOD scores based on the estimated parameters

    LOD <- round(lod.score3(beta00,beta01,pos.ibd),digits=5)
    pos.lod <- cbind(pos[i],LOD)
    res.bull <- rbind(res.bull,pos.lod)
  }

# Record the results to the object – min-max restriction model

res.m3 <- res.bull
# colnames(m3) <- c("pos","beta00","beta01","LOD","iter","convg","meth")

# res.bull <- rbind(m1,m2,m3)

# Fill in missing value -99 for the position w/ data for COVLINK

covlink <- NULL
j <- 1
for (i  in 1:length(pos)) {
  if (pos[i]==res.covlink[j,1]) {
    tcov <- c(res.covlink[j,2])
    j <- j + 1
    } else {
      tcov <- c(-99)
      j <- j
    }
  covlink <- cbind(covlink,tcov)
}

# Merge all outputs and organize by the position

mix <- rbind(res.mix[,2])
lodpal <- res.lodpal[,2]
m1 <- res.m1[,2]
m2 <- res.m2[,2]
m3 <- res.m3[,2]

res.all <- round(rbind(mix,lodpal,m1,m2,m3,covlink),digits=3)
res.all <- cbind(rep(repli,6),c(1,2,3.1,3.2,3.3,4),res.all)

# Write out all outputs and the header (stat + position)

write.table(res.all,file="summary.txt",quote=F,col.names=F,row.names=F)
}
```

Part II: Examples

Before running this program, users need to obtain the IBD sharing probabilities by running the GeneHunter program. After evoking R environment, the R commands are as follows:

```
# run the code for mix, lodpal, mlrm and covlink
# this code needs two input files: covariate.* and ibd.c$
# *: no. of pedigrees in the data set; $: chromosome no.
# call in the statistical program, which is named as ("four_stat.R ") here

source("four_stat.R")

# use stat() function to run these four statistics
# *: no. of pedigrees in the data set; !: replicate no.; $: chromosome no.

stat(*,!,$)
```

The program puts its results in the output file: “summary.txt”, which reports the LOD score from the covariate statistics: mixture model, LODPAL, MLRM and COVLINK.

GLOSSARY

Allele: different forms of a locus or gene.

Allele frequency: the rate of an allele occurring in the general population.

Ascertainment: the pedigree recruitment through an affected sibling, proband.

Complex trait: a trait whose mode of inheritance does not follow the Medelian laws.

Dominant: a disease is transmitted in a dominant way (only one copy of the disease allele is required to develop disease).

Gene: a region of DNA sequences coding for a protein product.

Genotype: the type of alleles found at a locus

Identical by descent (IBD): alleles in an individual or in two people are identical because they have been transmitted from the same common ancestor.

Identical by state (IBS): coincidental possession of identical alleles in an individual or in two people.

Locus: a region of chromosome. It can be a gene or a marker.

LOD score: a measure of the likelihood that there is linkage between two loci.

Marker: any polymorphism determined by segregation at a locus in a known fashion.

Model-free analysis: linkage analysis in which no mode of inheritance needs to be specified.

Parametric analysis: linkage analysis in which a mode of inheritance needs to be specified.

Penetrance: the probability that an individual carrying the disease allele(s) will develop the disease phenotype.

Phenotype: the observable manifestations of a gene.

Recessive: a disease is transmitted in a recessive way (two disease alleles are required to develop disease).

Recombination fraction: the probability of occurrence of a recombinant event (usually denoted as θ).

Segregation analysis: a statistical methodology for investigating the mode of inheritance of a phenotype from family data.

Transmission probabilities: the probability that an allele is transmitted from one generation to the next generation.

BIBLIOGRAPHY

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97-101.

Abel L, Alcais A, Mallet A (1998) Comparison of four sib-pair linkage methods for analyzing sibships with more than two affecteds: interest of the binomial maximum likelihood approach. Genet Epidemiol 15:371-390.

Abel L, Muller-Myhsok B (1998) Robustness and power of the maximum-likelihood-binomial and maximum- likelihood-score methods, in multipoint linkage analysis of affected-sibship data. Am J Hum Genet 63:638-647.

Alcais A, Abel L (1999) Maximum-Likelihood-Binomial method for genetic model-free linkage analysis of quantitative traits in sibships. Genet Epidemiol 17:102-117.

Alcais A, Abel L (2001) Incorporation of covariates in multipoint model-free linkage analysis of binary traits: how important are unaffecteds? Eur J Hum Genet 9:613-620.

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198-1211.

Bhat A, Heath SC, Ott J (1999) Heterogeneity for multiple disease loci in linkage analysis. Hum Hered 49:229-231.

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85-97.

Boehnke M (1986) Estimating the power of a proposed linkage study: a practical computer simulation approach. Am J Hum Genet 39:513-527.

Bonney GE (1984) On the statistical determination of major gene mechanisms in continuous human traits: Regressive models. Am J Med Genet 18:731-749.

Broman KW, Weber JL (2000) Characterization of human crossover interference. Am J Hum Genet 66:1911-1926.

Bull SB, Greenwood CM, Mirea L, Morgan K (2002) Regression models for allele sharing: analysis of accumulating data in affected sib pair studies. Stat Med 21:431-444.

Chotai J (1984) On the lod score method in linkage analysis. Ann Hum Genet 48 (Pt 4):359-378.

Cleves MA, Elston RC (1997) Alternative test for linkage between two loci. Genet Epidemiol 14:117-131.

Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M (1995) Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type I diabetes. Am J Hum Genet 57:920-934.

Davis S, Weeks DE (1997) Comparison of nonparametric statistics for detection of linkage in nuclear families: Single-marker evaluation. Am J Hum Genet 61:1431-1444.

Devlin B, Bacanu SA, Klump KL, Bulik CM, Fichter MM, Halmi KA, Kaplan AS, Strober M, Treasure J, Woodside DB, Berrettini WH, Kaye WH (2002a) Linkage analysis of anorexia nervosa incorporating behavioral covariates. Hum Mol Genet 11:689-696.

Devlin B, Jones BL, Bacanu S-A, Roeder K (2002b) Mixture models for linkage analysis of affected sibling pairs and covariates. Genet Epidemiol 22:52-65.

Devlin B, Bacanu S-A, Jones BL, Roeder K (2002c) Reply to Olson. Genet Epidemiol 23:449-455.

Drigalenko E (1998) How sib pairs reveal linkage. Am J Hum Genet 63:1242-1245.

Eaves LJ (1984) The resolution of genotype x environment interaction in segregation analysis of nuclear families. Genet Epidemiol 1:215-228.

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523-542.

Elston RC (2000) Introduction and overview. Statistical methods in genetic epidemiology. Stat Methods Med Res 9:527-541.

Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. Genet Epidemiol 19:1-17.

Elston RC (2001) Statistical Analysis for Genetics Epidemiology, Release 4.0 (http://darwin.cwru.edu/pub/sage.html), Computer program package from the Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, Cleveland.

Faraway JJ (1993) Distribution of the admixture test for the detection of linkage under heterogeneity. Genet Epidemiol 10:75-83.

Feingold E (2002) Regression-based quantitative-trait-locus mapping in the 21st century. Am J Hum Genet 71:217-222.

Fisher RA, Lyon MF, Owen ARG (1947) The sex chromosome in the house mouse. Heredity 1:335-365.

Forrest WF (2001) Weighting improves the "new Haseman-Elston" method. Hum Hered 52:47-54.

Foss E, Lande R, Stahl FW, Steinberg CM (1993) Chiasma interference as a function of genetic distance. Genetics 133:681-691.

Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait loci. Am J Hum Genet 54:1092-1103.

Fulker DW, Cherny SS (1996) An improved multipoint sib-pair analysis of quantitative traits. Behav Genet 26:527-532.

Gauderman WJ, Siegmund KD (2001) Gene-environment interaction and affected sib pair linkage analysis. Hum Hered 52:34-46.

Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, et al. (2000) The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to Type 2 diabetes. Am J Hum Genet 67:1174-1185.

Glidden DV, Liang KY, Chiu YF, Pulver AE (2003) Multipoint affected sibpair linkage methods for localizing susceptibility genes of complex diseases. Genet Epidemiol 24:107-117.

Goddard KA, Witte JS, Suarez BK, Catalona WJ, Olson JM (2001) Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. Am J Hum Genet 68:1197-1206.

Goldin LR (1992) Detection of linkage under heterogeneity: comparison of the two-locus vs. admixture models. Genet Epidemiol 9:61-66.

Green JR, Low HC, Woodrow JC (1983) Inference on inheritance of disease using repetitions of HLA haplotypes in affected siblings. Ann Hum Genet 47 Pt 1:73-82.

Greenwood CM, Bull SB (1997) Incorporation of covariates into genome scanning using sib-pair analysis in bipolar affective disorder. Genet Epidemiol 14:635-640.

Greenwood CM, Bull SB (1999) Analysis of affected sib pairs, with covariates--with and without constraints. Am J Hum Genet 64:871-885.

Gudbjartsson DF, Jonasson K, Frigge ML, Kong A (2000) Allegro, a new computer program for multipoint linkage analysis. Nat Genet 25:12-13.

Guo SW (2000) Gene-environment interactions and the affected-sib-pair designs. Hum Hered 50:271-285.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King M-C (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684-1689.

Hanson RL, Knowler WC (1998) Analytic strategies to detect linkage to a common disorder with genetically determined age of onset: diabetes mellitus in Pima Indians. Genet Epidemiol 15:299-315.

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3-19.

Hauser ER, Watanabe RM, Duren WL, Boehnke M (1998) Stratified linkage analysis of complex genetic traits using related covariates. Am J Hum Genet 63:A45.

Hauser ER, Hsu FC, Daley D, Olson JM, Rampersaud E, Lin JP, Paterson AD, Poisson LM, Chase GA, Dahmen G, Ziegler A (2003) Effects of covariates: A summary of Group 5 contributions. Genet Epidemiol 25 Suppl 1:S43-49.

Hodge SE, Vieland VJ, Greenberg DA (2002) HLODs remain powerful tools for detection of linkage in the presence of genetic heterogeneity. Am J Hum Genet 70:556-557.

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362-374.

Holmans P (2002) Detecting gene-gene interactions using affected sib pair analysis with covariates. Hum Hered 53:92-102.

Hsu FC, Hetmanski JB, Li L, Markakis D, Jacobs K, Shugart YY (2003) Comparison of significance level at the true location using two linkage approaches: LODPAL and GENEFINDER. BMC Genetics 4(Suppl 1): S46.

Huang J, Vieland VJ (2001) Comparison of 'model-free' and 'model-based' linkage statistics in the presence of locus heterogeneity: single data set and multiple data set applications. Hum Hered 51:217-225.

Karlin S, Liberman U (1978) Classification and comparisons of multilocus recombination distributions. Proc Natl Acad Sci USA 75:6332-6336.

Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of genetic epidemiology. Oxford University Press, New York.

Kong A, Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179-1188.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241-247.

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugen 12:172-175.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347-1363.

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363-2367.

Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5:471-472.

Lange K, Sobel E (1991) A random walk method for computing genetic location scores. Am J Hum Genet 49:1320-1334.

Lathrop GM, Lalouel J-M (1984) Easy calculations of lod scores and genetic risks on small computers. Am J Hum Genet 36:460-465.

Leal SM, Ott J (2000) Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. Am J Hum Genet 66:567-575.

Liang K-Y, Zeger S (1986) Longitudinal data analysis using general linear models. Biometrika 73:12-22.

McPeek MS, Speed TP (1995) Modeling interference in genetic recombination. Genetics 139:1031-1044.

Mirea L, Briollais L, Bull S (2004) Tests for covariate-associated heterogeneity in IBD allele sharing of affected relatives. Genet Epidemiol 26:44-60.

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277-318.

Morton NE (1956) The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. Am J Hum Genet 8:80-96.

O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402-408.

Olson JM (1999) A general conditional-logistic model for affected-relative-pair linkage studies. Am J Hum Genet 65:1760-1769.

Olson JM, Goddard KA, Dudek DM (2002a) A second locus for very-late-onset Alzheimer disease: a genome scan reveals linkage to 20p and epistasis between 20p and the amyloid precursor protein region. Am J Hum Genet 71:154-161.

Olson JM, Song Y, Dudek DM, Moser KL, Kelly JA, Bruner GR, Downing KJ, Berry CK, James JA, Harley JB (2002b) A genome screen of systemic lupus erythematosus using affected-relative-pair linkage analysis with covariates demonstrates genetic heterogeneity. Genes Immun 3 Suppl 1:S5-S12.

Olson JM (2002c) Mixture models for linkage analysis of affected sibling pairs with covariates. Genet Epidemiol 23:444-448.

Olson JM (2002d) Rejoinder. Genet Epidemiol 23:456-457.

Ott J (1999) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore.

Ott J, Terwilliger JD (1992) Assessing the evidence for linkage in psychiatric genetics. In Mendlevicz J, Hippius H (eds) Genetic Research in Psychiatry. New York, Sprinder.

Ottman R (1990) An epidemiologic approach to gene-environment interaction. Genet Epidemiol 7:177-185.

Ottman R (1996) Gene-environment interaction: definitions and study designs. Prev Med 25:764-770.

Payami H, Thomson G, Motro U, Louis EJ, Hudes E (1985) The affected sib method. IV. Sib trios. Ann Hum Genet 49 (Pt 4):303-314.

Penrose LS (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugen 6:133-138.

Penrose LS (1953) The general purpose sib-pair linkage test. Ann Eugen 18:120-124.

Poisson LM, Rybicki BA, Coon SW, Barnholtz-Sloan JS, Chase GA (2003) Susceptibility scoring in family-based association testing. BMC Genetics 4(Suppl 1): S49.

Rice JP, Rochberg N, Neuman RJ, Saccone NL, Liu KY, Zhang X, Culverhouse R (1999) Covariates in linkage analysis. Genet Epidemiol 17:S691-695.

Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, McLeod RS, Griffiths AM, Green T, Brettin TS, Stone V, Bull SB, Bitton A, Williams CN, Greenberg GR, Cohen Z, Lander ES, Hudson TJ, Siminovitch KA (2000) Genomewide search in Canadian families with inflammatory bowel disease reveals two novel susceptibility loci. Am J Hum Genet 66:1863-1870.

Ripley BD (1987) Stochastic simulation. J. Wiley, New York.

Risch N, Lange K (1979) An alternative model of recombination and interference. Ann Hum Genet 43:61-70.

Risch N (1988) A new statistical test for linkage heterogeneity. Am J Hum Genet 42:353-364.

Risch N (1990a) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-228.

Risch N (1990b) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229-241.

Risch N, Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584-1589.

Rogus JJ, Krolewski AS (1996) Using discordant sib pairs to map loci for qualitative traits with high sibling recurrence risk. Am J Hum Genet 59:1376-1381.

Rothman KJ, Greenland S (1998) Modern epidemiology. Lippincott-Raven, Philadelphia.

Saccone NL, Rochberg N, Neuman RJ, Rice JP (2001) Covariates in linkage analysis using sibling and cousin pairs. Genet Epidemiol 21 Suppl 1:S540-545.

Schaid DJ, McDonnell SK, Thibodeau SN (2001) Regression models for linkage heterogeneity applied to familial prostate cancer. Am J Hum Genet 68:1189-1196.

Schaid DJ, Olson JM, Gauderman WJ, Elston RC (2003) Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. Hum Hered 55:86-96.

Sengul H, Weeks DE, Feingold E (2001) A survey of affected-sibship statistics for nonparametric linkage analysis. Am J Hum Genet 69:179-190.

Sham P (1997) Statistics in human genetics. Arnold ; Wiley, London New York.

Sham PC, Purcell S (2001) Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. Am J Hum Genet 68:1527-1532.

Sham PC, Purcell S, Cherny SS, Abecasis GR (2002) Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 71:238-253.

Shannon WD, Province MA, Rao DC (2001) Tree-based recursive partitioning methods for subdividing sibpairs into relatively more homogeneous subgroups. Genet Epidemiol 20:293-306.

Shao Y, Cuccaro ML, Hauser ER, Raiford KL, Menold MM, Wolpert CM, Ravan SA, Elston L, Decena K, Donnelly SL, Abramson RK, Wright HH, DeLong GR, Gilbert JR, Pericak-Vance MA (2003) Fine mapping of autistic disorder to chromosome 15q11-q13 by use of phenotypic subtypes. Am J Hum Genet 72:539-548.

Shih MC, Whittemore AS (2001) Allele-sharing among affected relatives: non-parametric methods for identifying genes. Stat Methods Med Res 10:27-55.

Smith CAB (1959) Some comments on the statistical methods used in linkage investigations. Am J Hum Genet 11:289-304.

Strachan T, Read AP (1996) Human molecular genetics. BIOS Scientific Publishers ; Wiley-Liss, Oxford, New York.

Suarez BK, Rice JP, Reich T (1978) The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87-94.

Terwilliger JD, Speer M, Ott J (1993) Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet Epidemiol 10:217-224.

Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins University Press, Baltimore.

Terwilliger JD, Weiss KM (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 9:578-594.

Thomas DC, Cortessis V (1992) A Gibbs sampling approach to linkage analysis. Hum Hered 42:63-76.

Tiret L, Abel L, Rakotovao R (1993) Effect of ignoring genotype-environment interaction on segregation analysis of quantitative traits. Genet Epidemiol 10:581-586.

Vieland VJ, Wang K, Huang J (2001) Power to detect linkage based on multiple sets of data in the presence of locus heterogeneity: comparative evaluation of model-based linkage methods for affected sib pair data. Hum Hered 51:199-208.

Visscher PM, Hopper JL (2001) Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. Ann Hum Genet 65:583-601.

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61:439-447.

Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315-326.

Weeks DE, Ott J, Lathrop GM (1990) SLINK: a general simulation program for linkage analysis. Am J Hum Genet 47:A204.

Weeks DE, Harby LD (1995) The affected-pedigree-member method: power to detect linkage. Hum Hered 45:13-24.

Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118-127.

Whittemore AS (1996) Genome scanning for linkage: an overview. Am J Hum Genet 59:704-716.

Whittemore AS, Tu I-P (1998) Simple, robust linkage tests for affected sibs. Am J Hum Genet 62:1228-1242.

Whittemore AS, Halpern J (2001) Problems in the definition, interpretation, and evaluation of genetic heterogeneity. Am J Hum Genet 68:457-465.

Wright FA (1997) The phenotypic difference discards sib-pair QTL linkage information. Am J Hum Genet 60:740-742.

Xu X, Weiss S, Wei LJ (2000) A unified Haseman-Elston method for testing linkage with quantitative traits. Am J Hum Genet 67:1025-1028.Haldane JBS (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299-309.

Zhang H, Leckman JF, Pauls DL, Tsai CP, Kidd KK, Campos MR (2002) Genomewide scan of hoarding in sib pairs in which both sibs have Gilles de la Tourette syndrome. Am J Hum Genet 70:896-904.