Power Studies of Regression-Based Linkage Methods for Selected Sibpairs in the Presence of Epistasis

A Dissertation Presented

by

Chengrui Huang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

August 2010

Stony Brook University

The Graduate School

Chengrui Huang

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Nancy R. Mendell - Dissertation Advisor

Professor, Applied Mathematics and Statistics

Stephen J. Finch – Chairperson of Defense

Professor, Applied Mathematics and Statistics

Wei Zhu

Professor, Applied Mathematics and Statistics

Tao Wang

Assistant Professor, Department of Epidemiology and Population Health,

Albert Einstein College of Medicine, Yeshiva University

This dissertation is accepted by the Graduate School

Lawrence Martin

Dean of the Graduate School

Abstract of the Dissertation

Power Studies of Regression-Based Linkage Methods for Selected Sibpairs in the Presence of Epistasis

by

Chengrui Huang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2010

Although the ubiquitousness of epistasis, or gene-gene interactions, is widely acknowledged, many commonly used quantitative-trait-locus (QTL) linkage analysis methods have been developed without explicitly modeling any dominance or epistasis effects. The power of regression-based linkage methods was investigated in this paper under a range of two-locus models of various degrees and types of epistasis.

A quantitative trait is studied usually because of its association with some complex disease of interest. Therefore we introduced selection through disease affected probands, which has commonly been used in qualitative trait studies, into our QTL analysis, and compared it to random selection and selection based on individuals having abnormal quantitative trait values.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my deepest appreciation toward my dear advisor, Prof. Mendell. Without her insightful guidance, generous support, and timely encouragement I would not have accomplished my academic goals. Her agreeable personality also inspires me to be a better person: optimistic, easy-going, and considerate.

I would also like to thank my committee members, Dr. Finch, Dr. Zhu, and Dr. Wang, for their invaluable comments and advice.

I am also greatly indebted to my family and friends. It was their generous support helped me go through this stressful thesis writing and job hunting period. First and foremost, I want to thank my dearest grandma. She has always blessed me from above. I am sure she would be so happy to see what I have achieved today!

Thank you all: dear mom and dad, Uncle Brain, Uncle Shu, Aunt Song, Justin Seyster, Dake Feng, Rong Chen, Zhuying Huang, Soyoun Shin, Shuwei Chang, Donghyung Lee, Paichuan Chen, Kathryn Sharpe, Songjie Li, Jane Rong, Lu Wang, Qilong Yuan, Ming Zhao, Astrid and Wolfgang Wander, Cora Walter, Shirley Leong, Xiaohui Peng, Mathangi Ramesh, Ying Wu, Rose St. Fleur, Yang Yang, Heejong Sung, Adrienne Chu, Jin Gao, Rong Cai, Xiao Wu, Jessica Li, Minyoung Lee, Tingting He, Yifan Wang, Xiawei Tu, everyone in the research group of Prof. Mendell and Prof. Finch, and so many people in my life that have given me rich and wonderful memories. Thank you all for being part of my life!

Chapter 1 Introduction

One of the main purposes of genetic studies is to identify, or to map, genes that have effects on diseases, or disease-related quantitative traits, of interest. Linkage analysis is one of the statistical methods of gene mapping. By examining the patterns of allele-transmission from parent to offspring, or the patterns of allele-sharing by relatives, we can detect the cosegregation of the disease or trait gene and the marker gene with known position hence infer the relative position of the disease or trait gene (Sham 1998).

In this thesis, a quantitative trait, together with a complex disease with which the trait is associated, are modeled under a general two-locus bi-allele epistatic and pleiotropic genetic framework. Two existing regression-based method, the original Haseman-Elston method and the modification by Sham and Purcell, are extended to detect epistasis. Simulated genotypes at three types of marker loci and simulated phenotypes, the quantitative trait values, are generated under random sampling and two selected samplings. The power to detect epistasis is compared across different linkage methods, types of marker, and sampling methods.

The outline of this paper is as follows: background information and literature review, along with description of important concepts, definitions of terms, and motivation and goal of our study, will be given this Chapter. In Chapter 2, we will describe the methodologies used in this paper. In Chapter 3, we will present the details of our simulation and study design. We will then analyze and interpret the results from our simulation study and draw conclusions in Chapter 4. Finally the limitations of our study and possible directions for future studies will be discussed in Chapter 5.

1.1 Genetic Epistasis

Most statistical tools were developed assuming an additive relationship between genotype and phenotype, and are effective in the analysis of simple, Mendelian diseases, such as sickle cell anemia, cystic fibrosis, Huntington's disease, and early-onset Alzheimer's disease (Culverhouse et al. 2004). But for most common human diseases with complex etiology, the search for susceptibility loci has been less successful. For example, although diabetes, depression, and schizophrenia are known to have large genetic components, traditional methods of genetic analysis in these diseases, however, have resulted in conflicting findings (Cordell 2002). One reason may be that so many genes are involved in a complex disease or trait that individual genes have effects that are too small to be detected even with a large sample. Another explanation is that there are substantial interactions between relatively few genes and between a major gene and environmental factors.

In this paper we will focus on gene-gene interaction, or epistasis. Studies in model organisms, such as fruit fly and yeast, have confirmed the ubiquitousness of epistasis. In addition, recent studies in human and animals have identified loci that interact significantly but have little or no effect at the margins (Evan et al. 2006). Therefore we can expect the power of the single-locus tests would be greatly reduced in this situation. However most of the widely-used genome-wide linkage or association analysis methods still assume single-locus model; i.e., the epistatic effect is negligible. One of the reasons that single-locus genome-wide searches are still a predominant practice lies in the practical difficulty in conducting even the pairwise genetic search. Namely a search involve *n* markers would entail $_nC_2$ = n!/[2!(n-2)!] pairwise comparisons. An intuitive alternative to lessen the calculation burden is to conduct the search in two stages: only the markers that have passed the threshold in single-locus tests are further investigated for two-locus interaction. However, if the epistatic loci have little or no marginal effect, they cannot to be detected with this strategy. Therefore the two-stage genome-wide search is still not as powerful as the exhaustive two-locus probe when loci interact (Evans et al. 2006).

In this paper, we will examine the power of regression-based methods in detecting epistasis under several two-locus epistatic models. The goals are to see, firstly, if single-locus analysis can detect epistatic loci, and secondly how much, if any, statistical power can we gain if we use two-locus analysis, and thirdly to evaluate the power to detect epistasis between loci.

1.2 Sampling Schemes

The theoretical basis of a statistical analysis usually assumes that a sample of families is selected from a particular population at random. In practice however it is more common to first select a random sample of individuals in the population called probands that are affected with a particular disease or have extreme quantitative trait values, and then select the probands' family members.

By selecting families through probands, we hope the sample can contain more information relevant to the disease or the trait of interest and therefore increase the statistical power in detecting the underlying genes.

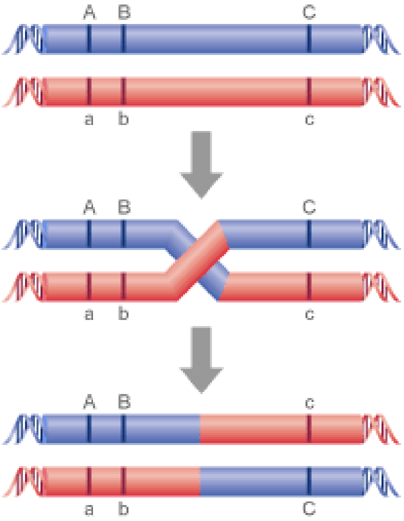
Many studies have shown that various methods for sample selection on the basis of trait values increase power over random sampling for detection of linkage with quantitative trait loci (QTLs). One could select sibling pairs in which one sibling having extremely high or low trait values (Carey and Williams 1991), or one could select relative pairs having discordant and/or concordant extreme phenotypes (Zhang and Risch 1996; Szatkiewicz and Feingold 2005).

Usually we are interested in some quantitative trait because it is a critical measurement for the diagnosis of a disease (for example time to onset for breast cancer and Alzeimer's disease) or a risk factor for some complex disease but with simpler etiology (for example hypertension is a major contributing factor in development of kidney failure) (Amos and de Andrade 2001). Because of these connections the trait and the disease may share some genetic components in common. Therefore we expect selection based on disease affected probands can yield higher power over random sampling for detecting QTLs and may have power comparable to selection based on trait value (Huang et al. 2009).

Affected sib pair (ASP) data is commonly used in linkage analysis to detect disease loci (Elston et al. 2005), but is rarely applied on quantitative traits related to some diseases. Some promising applications of ASP in linkage analysis of QTLs can be found in the literature (e.g. Huang et al. 2007), but an extensive comparison of power under different conditions is needed. This sampling method is called **Disease Selected Sampling** instead of ASP in this paper to avoid confusion.

1.3 Gene-Model-Free Methods for Genetic Linkage Analysis

Before we introduce the methods for linkage analysis we will first use the following diagram taken from the internet (http://genome.wellcome.ac.uk/doc_WTD020778.html) to explain the process of crossing over in meiosis and the principle of linkage analysis.



The top part of this diagram shows a pair of chromosomes in a cell before meiosis, one inherited from the father (blue) and another inherited from the mother (red). During the meiosis a phenomenon called crossing over would happen where the genetic information contained in paternal and maternal chromosomes is shuffled, or recombined. This is shown in the middle panel of the diagram. Finally the bottom panel shows the two haploid reproductive cells, or gametes, generated at the end of meiosis.

Three genes, labeled *A*, *B*, and *C*, are also shown in the diagram. The capital letters represent the paternal alleles and the lower case letters represent the maternal alleles. Note that, when only *A* and *B* loci are considered, the haplotypes of the gamete are the same as the haplotype of the parental gametes (*AB* and *ab*). Such gametes are defined as non-recombinants with respect to the *A* and *B* loci. When the loci *A* and *C* are considered, however, the haplotypes of the gamete are neither one of the parental haplotypes (*AC* and *ac*) but a new combination of alleles (*Ac* and *aC*). Such gametes are referred to as recombinants with respect to the *A* and *C* loci. The recombination fraction between two loci, denoted as θ, is defined as the probability that a gamete is a recombinant. (Sham 1998)

If two loci are far apart on a chromosome, one is equally likely to have the recombinant or the non-recombinant gametes. Therefore the recombination fraction equals ½. When two loci are physically close to each together, however, the crossover between them is less likely to occur. Hence we would expect the recombination fraction to be less than ½. The smaller the recombination fraction, the more tightly linked are the two loci. If the two loci are in "complete linkage", then the recombination fraction equals zero.

We assume that locus *A* in the above diagram is the disease or trait gene whose position is unknown and loci *B* and *C* are two marker genes whose positions are known. Based on our observation that the recombination occurs more frequently between *A* and *B* then between *A* and *C*, we can map the position of locus *A* relative to markers *B* and *C*.

There are two different categories of linkage analysis methods, gene-model-based (or “parametric”) methods and gene-model-free (or “non-parametric”) methods. The first type of method relies heavily on the knowledge of the mode of inheritance and the specification of parameters of interest in a genetic model such as genotype mean and variance or penetrance values and the recombination fraction between two loci. Since the mode of inheritance and parameters of interest are often hard to determine, and model misspecification can have detrimental effects on gene-model-based methods, in this paper we focus on the gene-model-free linkage methods.

One of the key concepts in gene-model-free linkage analysis is allele-sharing, or specifically identity-by-descent. A pair of relatives are said to share one allele identical-by-descent (IBD) when that particular allele can be traced to a common ancestor. Since humans are diploid, i.e. we have a pair of each type of chromosome. The three possible numbers of alleles shared IBD at one locus are 0, 1, or 2; equivalently the IBD proportion can be 0, ½, or 1. Usually not all family members are available for study. Therefore the exact number of alleles shared IBD cannot be deduced unequivocally. But the probabilities of sharing 0, 1, and 2 alleles IBD can be deduced, and the expected IBD proportion can be estimated in this case.

In this section, we give background information about some of the most commonly used gene-model-free linkage methods. We first introduce regression-based methods and then likelihood-based methods.

1.3.1 Regression-Based Linkage Methods

One of the most commonly used tests for linkage is the regression method proposed by Haseman and Elston in 1972 (Haseman and Elston 1972). This approach is conceptually simple and computationally convenient and is ideal for a fast preliminary linkage scan.

The sample unit is a pair of siblings, or a "sibpair". Let x_{ij} be the observed trait values for the i^{th} sibling in sibpair j in a sample of n sibpairs, $Y_{Dj} = (x_{1j} - x_{2j})^2$ be the squared trait difference for sibpair j, and π_j be the estimated expected proportion of genes IBD at the marker locus for sibpair j ($i = 1, 2; j = 1, 2, \dots n$), where π takes the values of 0, ½, or 1 if the marker is fully informative but can take intermediate values otherwise.

The Haseman-Elston method regresses Y_D on π and tests for significant negative slope as indication of linkage. The interpretation for a negative slope is very intuitive if the tested marker is linked with the locus influencing the trait. The more genetic information shared in a sibpair the less difference in trait values we should expect in the sibpair. By assuming linkage equilibrium and no interaction between genes or between genetic and environmental factors, Haseman and Elston proved that the expectation of the slope if there is no dominance is given by

$$E(\beta \mid \pi) = -2(1-2\theta)^2 V_G . \qquad (1\text{-}1)$$

Here θ is the recombination fraction between the trait locus and the marker locus, and V_G is the genetic variance. The same result will hold asymptotically when dominance is present. If the marker locus is in completely linkage with the trait locus, then $\theta = 0$. If the marker locus is unlinked to the trait locus, $\theta = ½$. Therefore a one-sided t-test of negative slope is a test for the presence of both linkage (H_0: $\theta = ½$ vs H_1: $\theta < ½$) and a nonzero genetic variance component (H_0: $V_G = 0$ vs H_1: $V_G > 0$).

The original Haseman-Elston method is generally criticized as having low power (Feingold 2002). One of the explanations lies in the fact that it discards the information in the squared trait sum in sibpair. Many investigators have tried to increase the power by incorporating this information into the regression framework. For example Drigalenko (1998) and Elston et al. (2000) changed the dependent variable to the mean-corrected trait product, which is the un-weighted sum of the squared trait difference and the mean-corrected squared trait sum, $Y_{Sj} = (x_{1j}+x_{2j}-2\mu)^2$, where x_{ij} is the observed trait values for the i^{th} sibling in sibpair j in a sample of n sibpairs and μ is the population mean. A number of other modifications involve using different weighted average forms as dependent variable (Forrest 2001; Visscher and Hopper 2001; Xu et al. 2000; Sham and Purcell 2001).

Cuenco and her colleagues conducted a comprehensive comparison of seven regression-based statistics, in addition to the original Haseman-Elston method, in terms of type I error, power, and robustness to parameter value misspecification, selected sampling, and violation of distributional assumptions (Cuenco et al. 2003). The modification proposed by Sham and Purcell was found consistently to have correct type I error, relatively high power, and more importantly, robustness to trait parameter misspecification in selected samples. Therefore it was recommended.

In addition to the notation defined at the beginning of this section, let μ be the mean of the trait values in the population, σ^2 be the variance in the population, z_{ij} be the standardized trait values for the i^{th} sibling in sibpair j, i.e. $z_{ij} = \dfrac{x_{ij}-\mu}{\sigma}$, and Y_{Sj} =

$(z_{1j} + z_{2j})^2$ be the squared standardized trait sum for sibpair j. Sham and Purcell showed that the variances of Y_D and Y_S can be expressed as functions of the sibling trait correlation, ρ, under the normality assumption (Sham and Purcell 2001). They proposed to use the value of

$$Y^* = \frac{Y_S}{(1+\rho)^2} - \frac{Y_D}{(1-\rho)^2} \tag{1-2}$$

as the dependent variable, which is regressed on the IBD proportion, π. Then one should use the t statistic to test for linkage. Here a null hypothesis of a slope equals to zero is tested against an alternative that the slope is positive.

Although these methods are derived under the assumption of normally distributed residual variance, the regression framework makes them robust to distributional assumption in the sense that the asymptotic null distributions of the test statistics do not depend on the distribution of the trait values or dependent variable used. These methods can also be easily extended to accommodate gene-gene and/or gene-environment interactions.

1.3.2 Likelihood-Based Linkage Methods

Another category of methods for linkage analysis is based on the likelihood function. Almsay and Blangero (1998) proposed a method to partition the phenotypic variance into its components arise from genetic, polygenetic, and environmental factors and so on and used the likelihood ratio statistic to test for linkage. This method has much higher power than that of the Haseman-Elston method under ideal conditions, but it is very sensitive to sampling selection and violation of distributional assumptions (Allison et al. 1999). In addition, it is much more computationally intensive since the genetic variance component has to be estimated under the null hypothesis of no linkage and under the alternative hypothesis.

Recently many methods have been developed based on score statistics in an attempt to retain to the robustness of the Haseman-Elston method and the high power of the likelihood ratio test under normality (Bhattacharjee et al. 2008).

It would certainly be interesting to extend the existing methods based on the score test to incorporate epistasis. However, in this paper we choose to focus on regression-based methods for the following reasons in addition to the advantages listed in the previous Section. First of all, the original Haseman-Elston method and various modified methods are very intuitive. These methods are all based on the idea that the more alleles shared identical-by-descend (IBD) at the marker locus in a relative pair the smaller the squared difference of trait values it should be, if the marker is in linkage with the quantitative trait locus (QTL). Secondly the distribution of the *t* statistic used to test for linkage is well-characterized. Finally regression is much easier and faster to be implemented than estimating variance components from likelihood functions.

1.4 Ascertainment Correction

With the exception of the original Haseman-Elston method, all of the above mentioned methods use some or all of the trait parameters, namely the population values for the mean and variance of the quantitative trait, and correlation between siblings. Although these are nuisance parameters, with respect to the test of linkage, they greatly influence the power (Bhattacharjee et al. 2008). They can be estimated fairly well in random samples. However since random sampling does not favor pedigrees with more linkage information, selected samples are more commonly used in practice. It is difficult to obtain unbiased and consistent estimates of these parameters from selected samples. Although it is possible to use population estimates from previous studies, we cannot be sure that populations in different studies have the same parameter values (Peng and Siegmund 2006). An alternative is to use the maximum likelihood estimates corrected for ascertainment.

There are two ascertainment corrections to the maximum likelihood estimation of trait parameters. Elston and Sobel (1979) suggested using likelihood function conditioning on the event of ascertainment. Hopper and Mathews (1982) proposed to condition on the exact values of the probands. These corrections were proven to be asymptotically equivalent in both simulation and analytic studies (Andrade and Amos 2000; Peng and Siegmund 2006). However since the first correction involves calculation of the probability of ascertainment, sometimes it is hard to make sure the procedure is well-defined. Hence Peng and Siegmund suggest using the correction proposed by Hopper and Mathews, especially when the ascertainment procedure is unknown or ill-defined. The details of this method will be present in the next chapter.

Chapter 2 Methodologies

In this chapter, we present the details on our extensions of two regression-based methods, and the ascertainment correction used in our paper to obtain conditional maximum likelihood estimates (CMLE) of the trait parameter values.

2.1 Two-Locus Pleiotropic Epistatic Model

Let X be the quantitative trait variable. We assume the following general model

$$X = G + e_s + e_{ns} \tag{2-1}$$

where G denotes the genetic effect, e_s denotes the environmental effects that are shared within family members, or shared residual component, and e_{ns} denotes the non-shared environmental effects. The polygene, a group of genes that together influence the trait, can be considered as a part of the shared environment within the family. Here we assume that there is no interaction between genetic and environmental factors, since it is not the focus of this paper.

Let μ be the overall mean of the trait values in the population, and σ^2 be the overall variance. We can partition the mean and variance into genetic and environmental components (Falconer 1981), but first we have to make several assumptions and describe the notation we use.

We assume in this paper that there are only two main trait loci, locus ***A*** and locus ***B***, and that they are **unlinked** to each other, **bi-allelic**, **pleiotropic**, and **epistatic**. This means that (1) these two genes segregate independently; (2) they each have two alleles; (3) at least one allele affects more than one phenotype, in this case both the value of a quantitative trait and the status of a disease; and (4) the effects of these two genes on the quantitative trait are not additive. Furthermore, random mating, Hardy-Weinberg equilibrium, and linkage equilibrium are assumed throughout.

Let $\boldsymbol{A_1}$ be the minor allele, the less common allele, at trait locus A, and $\boldsymbol{A_2}$ be the more common allele, with allele frequencies $\boldsymbol{p_1}$ and $\boldsymbol{q_1} = 1 - p_1$ respectively. Analogously the two alleles at trait locus B are denoted $\boldsymbol{B_1}$ and $\boldsymbol{B_2}$ with allele frequencies $\boldsymbol{p_2}$ and $\boldsymbol{q_2} = 1 - p_2$ respectively. The allele A_1 and B_1 are modeled to be two of the alleles that increase both the trait value and the chance of developing the disease in this paper.

Let $\boldsymbol{g_i}$ be the genotypes at locus A with i copies of minor allele A_1, $\boldsymbol{g_j}$ be the genotypes at locus B with j copies of minor allele B_1, and $\boldsymbol{g_{ij}}$ be the two-locus genotypes (i, j = 0, 1, 2). The genotype frequencies at locus A under Hardy-Weinberg equilibrium are as follow:

$$\begin{aligned} Pr(G_A = g_2 = A_1A_1) &= p_1^2 \\ Pr(G_A = g_1 = A_1A_2) &= 2p_1q_1 \\ Pr(G_A = g_0 = A_2A_2) &= q_1^2 \end{aligned} \tag{2-2}$$

The genotype frequencies at locus B can be calculated analogously. The two-locus genotype frequencies are given by

$$Pr(G_{AB} = g_{ij}) = Pr(G_A = g_i)Pr(G_B = g_j) \tag{2-3}$$

Since the two trait loci are also the disease genes in our paper, we also present how we model the disease. Let $\boldsymbol{f_{ij}}$ be the penetrance, or the conditional probability of developing the disease in individuals with genotype g_{ij}, i.e.,

$$f_{ij} = Pr(D \mid G_{AB} = g_{ij}) \tag{2-4}$$

To simplify our analysis we assume the allele A_1 and B_1 have equal and additive effects on the disease. Specifically we assume a baseline penetrance f_b, and that having more copies of minor allele A_1 and/or B_1 linearly increases the chance of developing the disease by factor of f_l. We impose two constraints, $0 < f_l \leq 0.25$ and $0 \leq f_b \leq 1-4f_l$, to ensure $0 \leq f_{ij} \leq 1$ for all i and j. All nine penetrance values under this assumption are listed in the following table.

Table 2-1 Penetrance Values for the Disease Assuming Equal Additive Allelic Effects

G_A \ G_B	B_1B_1	B_1B_2	B_2B_2
A_1A_1	$f_{22} = f_b + 4f_l$	$f_{21} = f_b + 3f_l$	$f_{20} = f_b + 2f_l$
A_1A_2	$f_{12} = f_b + 3f_l$	$f_{11} = f_b + 2f_l$	$f_{10} = f_b + f_l$
A_2A_2	$f_{02} = f_b + 2f_l$	$f_{01} = f_b + f_l$	$f_{00} = f_b$

The disease prevalence or the probability of having the disease in the population is then obtained as

$$Pr(D) = \sum_{i=0}^{2}\sum_{j=0}^{2} f_{ij}\, Pr(G_{AB} = g_{ij}) \tag{2-5}$$

If we assume that the minor allele frequencies are the same, i.e. $p_1 = p_2 = p$, then the following equation about the disease prevalence, $Pr(D)$, holds

$$Pr(D) = f_b + 4pf_l \tag{2-6}$$

Furthermore if we set the baseline penetrance, f_b, to be zero, we assume that susceptible genotypes are necessary to be affected and that individuals without any copy of minor allele at both disease genes have no chance in developing the disease. The disease prevalence in this case equals to $4pf_l$.

By applying Bayes Theorem we can get the conditional probability of one's genotype given that this individual is disease affected

$$Pr(G_{AB} = g_{ij} \mid D) = \frac{f_{ij}\, Pr(G_{AB} = g_{ij})}{Pr(D)} \tag{2-7}$$

If we set $f_b = 0$ and assume equal minor allele frequencies, then the conditional probabilities, $Pr(G_{AB}|D)$, does not depend on the disease penetrance at all, but solely on the allele frequencies as shown in the following table.

Table 2-2 Conditional Probabilities of Genotype of Disease Affected Proband Assuming No Phenocopy and Equal Minor Allele Frequencies

G_B / G_A	B_1B_1	B_1B_2	B_2B_2
A_1A_1	p^3	$3p^2q/2$	$pq^2/2$
A_1A_2	$3p^2q/2$	$2pq^2$	$q^3/2$
A_2A_2	$pq^2/2$	$q^3/2$	0

Comparing these conditional probabilities, $Pr(G_{AB}|D)$, with the unconditional ones in random samples, $Pr(G_{AB})$, we can see that selecting samples through disease affected probands increases the probabilities of obtaining individuals with genotypes having at least one copy of a minor allele from either one of the pleiotropic loci, by factors of $1/4p$ to $1/p$. Therefore we expect this selection scheme would result in a higher power over random sampling as the minor allele frequency p decreases.

Following the notation of Tiwari and Elston (1997) and Purcell and Sham (2004), let

$\boldsymbol{a_1}$, $\boldsymbol{a_2}$ = the additive effects at locus A and locus B respectively;

$\boldsymbol{d_1}$, $\boldsymbol{d_2}$ = the dominant effects at locus A and locus B respectively;

aa = the interaction between the additive effects at the two loci;

ad = the interaction between the additive effect at locus A and dominant effect at locus B;

da = the interaction between the dominant effect at locus A and additive effect at locus B;

dd = the interaction between the dominant effects at the two loci;

and ***m*** be the mean effect. We can assume the residual components (e_s and e_{ns}) to have mean of 0 and, for convenience, follow normal distribution with variances σ_s^2 and σ_{ns}^2 respectively.

Let $\boldsymbol{\mu_{ij}}$ be the expected trait value given that the two-locus genotype is g_{ij} (i, j = 0, 1, 2). The expressions of these conditional trait means in terms of additive, dominance, and epistatic effects are given in the following table:

Table 2-3 Partition of Expected Trait Values under Two-Locus Epistatic Model

G_B \ G_A	B_1B_1	B_1B_2	B_2B_2
A_1A_1	$m+a_1+a_2+aa$	$m+a_1+d_2+ad$	$m+a_1-a_2-aa$
A_1A_2	$m+d_1+a_2+da$	$m+d_1+d_2+dd$	$m+d_1-a_2-da$
A_2A_2	$m-a_1+a_2-aa$	$m-a_1+d_2-ad$	$m-a_1-a_2+aa$

Source: Table II in Purcell and Sham (2004) Epistasis in quantitative trait locus linkage analysis: interaction or main effect? *Behav Genet* 34: 143-152

The overall trait mean can be calculated by

$$E(X) = \mu = \sum_{i=0}^{2}\sum_{j=0}^{2} \mu_{ij}\, Pr(G_{AB} = g_{ij}) \qquad (2\text{-}8)$$

The overall variance of the quantitative trait value can be partitioned into three parts: genetic, shared residual between siblings, and non-shared residual

$$Var(X) = \sigma^2 = V_T = V_G + V_S + V_{NS} \tag{2-9}$$

The genetic variance under this bi-allelic two-locus model can be further divided into components due to additive, dominant, and epistatic effects

$$V_G = V_A + V_D + V_I = (V_{a1} + V_{a2}) + (V_{d1} + V_{d2}) + (V_{aa} + V_{ad} + V_{da} + V_{dd}) \tag{2-10}$$

or more specifically, variance due to the additive effect at locus *A*

$$\begin{aligned} V_{a1} = 2p_1q_1[&a_1 - (p_1 - q_1)d_1 + (p_2 - q_2)aa - (p_1 - q_1)(p_2 - q_2)da \\ &+ 2p_2q_2ad - 2(p_1 - q_1)p_2q_2dd]^2 \end{aligned} \tag{2-11}$$

variance due to the additive effect at locus *B*

$$\begin{aligned} V_{a2} = 2p_2q_2[&a_2 - (p_2 - q_2)d_2 + (p_1 - q_1)aa - (p_1 - q_1)(p_2 - q_2)da \\ &+ 2p_1q_1ad - 2p_1q_1(p_2 - q_2)dd]^2 \end{aligned} \tag{2-12}$$

variance due to the dominant effect at locus *A*

$$V_{d1} = 4p_1^2q_1^2[d_1 - (p_2 - q_2)da + 2p_2q_2dd]^2 \tag{2-13}$$

variance due to the dominant effect at locus *B*

$$V_{d1} = 4p_2^2q_2^2[d_2 - (p_1 - q_1)ad + 2p_1q_1dd]^2 \tag{2-14}$$

variance due to the interaction of additive effects at the two loci

$$V_{aa} = 4p_1q_1p_2q_2[aa - (p_2 - q_2)ad - (p_1 - q_1)da + (p_1 - q_1)(p_2 - q_2)dd]^2 \tag{2-15}$$

variance due to the interaction between the additive effect at locus *A* and dominant effect at locus *B*

$$V_{ad} = 8p_1q_1p_2^2q_2^2[ad - (p_1 - q_1)dd]^2 \tag{2-16}$$

variance due to the interaction between the dominant effect at locus *A* and additive effect at locus *B*

$$V_{da} = 8 p_1^2 q_1^2 p_2 q_2 [da - (p_2 - q_2) dd]^2 \qquad (2\text{-}17)$$

and variance due to the interaction between the dominant effects at the two loci (Tiwari and Elston 1997)

$$V_{dd} = 16 p_1^2 q_1^2 p_2^2 q_2^2 dd^2 \qquad (2\text{-}18).$$

For two unlinked trait loci, the covariance of trait values between two siblings is given by

$$Cov(X_1, X_2) = \frac{V_{a1}}{2} + \frac{V_{a2}}{2} + \frac{V_{d1}}{4} + \frac{V_{d1}}{4} + \frac{V_{aa}}{4} + \frac{V_{ad}}{8} + \frac{V_{da}}{8} + \frac{V_{dd}}{16} + V_S \qquad (2\text{-}19)$$

Furthermore, if we assume equal variance for the trait values in a sibpair, then we can obtain the population correlation by

$$\rho = \frac{Cov(X_1, X_2)}{Var(X)} \qquad (2\text{-}20)$$

2.2 Extension of Haseman-Elston Method to Analysis of Two Marker Loci

Let x_s be the observed trait values for the s^{th} sibling in a sibpair (s = 1, 2), $Y_D = (x_1 - x_2)^2$ be the squared trait difference, $\boldsymbol{h_{ij}}$ be the probability, conditional on all the marker information available in the family, that two siblings share i marker alleles IBD at marker locus j, and π_j be the expected proportion of genes IBD at the marker locus j, then $\hat{\pi}_j = h_{1j} + \frac{1}{2} h_{2j}$ (i = 0, 1, 2; j = 1, 2), where $\hat{\pi}_j$ takes the values of 0, ½, or 1 if the marker is fully informative but can take intermediate values otherwise. Let

θ_j denote the recombination fraction between the quantitative trait locus j and the marker linked to it, and define $\psi_j = \theta_j^2 + (1-\theta_j)^2$.

Tiwari and Elston (1997) derived the expectation of the squared difference given the bi-allelic two-locus epistatic model

$$\begin{aligned} &E(Y_D \mid \hat{\pi}_1, \hat{\pi}_2, h_{11}, h_{12}) = \\ &\alpha + \beta_1\hat{\pi}_1 + \beta_2\hat{\pi}_2 + \delta_1 h_{11} + \delta_2 h_{12} + \gamma_1\hat{\pi}_1\hat{\pi}_2 + \gamma_2\hat{\pi}_1 h_{12} + \gamma_3 h_{11}\hat{\pi}_2 + \gamma_4 h_{11}h_{12} \end{aligned} \tag{2-21}$$

where

$$\begin{aligned}
\beta_1 &= 2(1-2\psi_1)[V_{a1} + V_{d1} + (1-\psi_2)(V_{aa} + V_{da}) + (1-\psi_2)^2(V_{ad} + V_{dd})] \\
\beta_2 &= 2(1-2\psi_2)[V_{a2} + V_{d2} + (1-\psi_1)(V_{aa} + V_{ad}) + (1-\psi_1)^2(V_{da} + V_{dd})] \\
\delta_1 &= (1-2\psi_1)^2[V_{d1} + (1-\psi_2)V_{da} + (1-\psi_2)^2 V_{dd}] \\
\delta_2 &= (1-2\psi_2)^2[V_{d2} + (1-\psi_1)V_{ad} + (1-\psi_1)^2 V_{dd}] \\
\gamma_1 &= -2(1-2\psi_1)(1-2\psi_2)[V_{aa} + V_{ad} + V_{da} + V_{dd}] \\
\gamma_2 &= -(1-2\psi_1)(1-2\psi_2)^2[V_{ad} + V_{dd}] \\
\gamma_3 &= -(1-2\psi_1)^2(1-2\psi_2)[V_{da} + V_{dd}] \\
\gamma_4 &= -\frac{1}{2}(1-2\psi_1)^2(1-2\psi_2)^2 V_{dd}
\end{aligned}$$

If all of the dominance and epistatic variances are negligible, then all coefficients but β_j would be close to zero, and β_j would reduce to

$$2(1-2\psi_j)V_A = -2(1-2\theta_j)^2 V_G$$

as given in the original Haseman and Elston regression (1972).

If the variances due to any dominance effect at either loci (V_{d1}, V_{d2}, V_{ad}, V_{da}, and V_{dd}) are negligible, then all of coefficients of the terms that involve h_{1j}, namely δ_1 δ_2 γ_2 γ_3 and γ_4, would be zero. That is the epistatic model would reduce to

$$E(Y_D \mid \hat{\pi}_1, \hat{\pi}_2) = \alpha + \beta_1\hat{\pi}_1 + \beta_2\hat{\pi}_2 + \gamma_1\hat{\pi}_1\hat{\pi}_2 \tag{2-22}$$

where

$$\begin{aligned}\beta_1 &= 2(1-2\psi_1)[V_{a1}+(1-\psi_2)V_{aa}]\\ \beta_2 &= 2(1-2\psi_2)[V_{a2}+(1-\psi_1)V_{aa}]\\ \gamma_1 &= -2(1-2\psi_1)(1-2\psi_2)V_{aa}\end{aligned}$$

Note that $\psi_j = 1$ if and only if $\theta_j = 0$ (the marker locus and the trait locus are in complete linkage), and $\psi_j = 0$ if and only if $\theta_j = ½$ (the marker locus is unlinked to the trait locus). So if both marker loci are unlinked to the trait loci, then all of these coefficients would be zero.

Also note that γ_1 can be rewritten as $-2(1-2\theta_1)^2(1-2\theta_2)^2V_I$ where V_I is total two-locus epistatic variance. Therefore a one-sided *t* test for this negative slope is a combined test for $\theta_1 = ½$, $\theta_2 = ½$, and $V_I = 0$.

It is also important to note that the coefficient of the IBD proportion at a single locus (β_j) consists of not only that variance arising from the main effects (additive and dominance) at that locus but also a proportion of variance attributable to epistatic effects when the marker locus is not in complete linkage with the trait locus. Therefore it is possible for the analysis of a single locus to detect QTL with epistasis effects even without explicitly modeling them. If that is the case then the additional power gain by fitting a two-locus model would be limited. (Purcell and Sham 2004)

In this paper we regress the squared trait difference on the IBD proportions at two marker loci and their product ($\hat{\pi}_1\hat{\pi}_2$). If a significant negative *t* value corresponding to the estimated regression coefficient $\hat{\gamma}_1$ is obtained then the following combined null hypothesis is rejected:

$$H_0:\ \theta_1 = 1/2\ \ \theta_2 = 1/2\ \ and\ \ V_I = 0 \tag{2-23}$$

We also conduct a set of single regressions with only one IBD proportion in the model ($\hat{\pi}_1$ or $\hat{\pi}_2$) to see how well the single-locus analysis performs in detecting epistatic genes.

2.3 Extension of Sham-Purcell Method to Analysis of Two Marker Loci

In addition to the original Haseman-Elston method, we also want to extend the modified regression method proposed by Sham and Purcell (2001). Specifically we regress the variable defined in (1-2) on the IBD proportions at two marker loci and their product; i.e.

$$Y_{SP} \equiv \frac{Y_S}{(1+\rho)^2} - \frac{Y_D}{(1-\rho)^2} = \alpha + \beta_1\hat{\pi}_1 + \beta_2\hat{\pi}_2 + \gamma_1\hat{\pi}_1\hat{\pi}_2 + \varepsilon$$

where Y_S is the squared sum of trait values in a sibpair and Y_D is the squared trait difference. The null hypothesis is the same as defined in (2-23). However, a significant *positive t* value is needed to reject the null hypothesis. A set of single-locus regression analyses is also performed.

As explained in Section 1.4, population trait parameters, such as mean, variance, and correlation, usually are unknown and need to be estimated. Therefore we conduct our regression using two sets of Y_{SP}, one with actual parameter values and the other with estimated parameter values. However, since this method is robust to trait parameter misspecification, we do not expect the power to be reduced substantially by using estimated values.

2.4 Conditional Maximum Likelihood Estimates (CMLEs) of Trait Parameter Values

As explained in Section 1.4 we need to adjust our estimation for trait parameters in selected samples. In this section we derive the maximum likelihood estimators of the trait parameters conditional on the exact trait values of the probands (CMLEs).

Without loss of generality, let X_1 be the trait value of the proband and X_2 be the trait value of the proband's sibling. Assume that X_1 and X_2 follow a bivariate normal distribution with equal means and equal variances:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right)$$

where μ, σ^2, and ρ are the overall trait mean, variance, and sibpair correlation in the population respectively, as defined in Section 2.1.

The conditional distribution of the sibling's trait value, X_2, given the proband's trait value, X_1, is then a normal distribution with mean $\mu + \rho(x_1 - \mu)$ and variance $\sigma^2(1-\rho^2)$, i.e.

$$X_2 \mid X_1 \sim N(\mu + \rho(x_1 - \mu), \sigma^2(1-\rho^2))$$

If we collect trait values from a sample of n independent (i.e. unrelated) sibpairs then the conditional likelihood is given by

$$L_C = (2\pi)^{-n}[\sigma^2(1-\rho^2)]^{-n/2} \exp\{-\frac{\sum_{i=1}^{n}[x_{2i} - \rho x_{1i} - \mu(1-\rho)]^2}{2\sigma^2(1-\rho^2)}\} \qquad (2\text{-}24)$$

and its natural logarithm form is simply

$$\ln L_C = -n\ln(2\pi) - \frac{n}{2}\ln[\sigma^2(1-\rho^2)] - \frac{\sum_{i=1}^{n}[x_{2i} - \rho x_{1i} - \mu(1-\rho)]^2}{2\sigma^2(1-\rho^2)} \qquad (2\text{-}25)$$

To obtain the CMLEs we take the partial derivatives of the logarithm of the conditional likelihood function with respect to μ, σ^2, and ρ and solve the derivative equations for μ, σ^2, and ρ respectively. The equations to be solved are thus

$$\frac{\partial \ln L_C}{\partial \mu} = \frac{\sum_{i=1}^{n}[x_{2i} - \hat{\rho} x_{1i} - \hat{\mu}(1-\hat{\rho})]}{\hat{\sigma}^2(1+\hat{\rho})} = 0 \qquad (2\text{-}26)$$

$$\frac{\partial \ln L_C}{\partial \sigma^2} = \frac{-n\hat{\sigma}^2(1-\hat{\rho}^2) + \sum_{i=1}^{n}[x_{2i} - \hat{\rho} x_{1i} - \hat{\mu}(1-\hat{\rho})]^2}{2\hat{\sigma}^4(1-\hat{\rho}^2)} = 0 \qquad (2\text{-}27)$$

$$\frac{\partial \ln L_C}{\partial \rho} = \frac{n\hat{\rho}}{1-\hat{\rho}^2} + \frac{(1+\hat{\rho}^2)\sum(x_{1i}-\hat{\mu})(x_{2i}-\hat{\mu}) - \hat{\rho}[\sum(x_{1i}-\hat{\mu})^2 + \sum(x_{2i}-\hat{\mu})^2]}{\hat{\sigma}^2(1-\hat{\rho}^2)^2} = 0 \qquad (2\text{-}28)$$

Here $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\rho}$ denote the CMLEs of μ, σ^2, and ρ respectively.

From Equations (2-26) and (2-27) we can get the close-form solutions for the CMLE of μ and σ^2:

$$\hat{\mu} = \frac{\overline{x}_2 - \hat{\rho}\overline{x}_1}{1-\hat{\rho}}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}[x_{2i} - \hat{\rho} x_{1i} - \mu(1-\hat{\rho})]^2}{n(1-\hat{\rho}^2)}$$

However there is no close-form solution for the CMLE of ρ. We therefore have to use numerical method to obtain the CMLEs of all three parameters.

In this paper we used the Broyden-Fletcher-Goldfarb-Shanno (BFGS) minimization algorithm executed in one of the GSL functions. “This is a quasi-Newton method which builds up an approximation to the second derivatives of the function using the difference between successive gradient vectors. By combining the first and second derivatives the algorithm is able to take Newton-type steps towards the function minimum assuming quadratic behavior in that region." (cf. GSL 1.12 reference manual for more information) We inverted the sign of the log likelihood function to perform the maximization.

2.5 Types of Marker Loci Evaluated

In this paper we are interested in the power of the regression-based linkage analysis methods in the presence of epistasis when the marker loci are in complete linkage with the trait loci, i.e. zero recombination fractions between marker and trait loci. Three types of marker loci are studied in this paper: (1) Marker loci (bi-allelic) that are in fact the trait loci themselves. (2) Marker loci that are bi-allelic and in complete linkage with the trait loci. Examples of such loci are the single nucleotide polymorphisms (SNPs). (3) Marker loci that are multi-allelic and in complete linkage with the trait loci.

Since the IBD number can be more easily deduced when the markers are highly variable than when they are bi-allelic, we expect the third type of markers would result in higher variability in the explanatory variable and therefore has higher power than that of the second type. The first type of markers intuitively is expected to result in the highest power of all.

2.6 Null Distribution of Test Statistic and Power Calculation

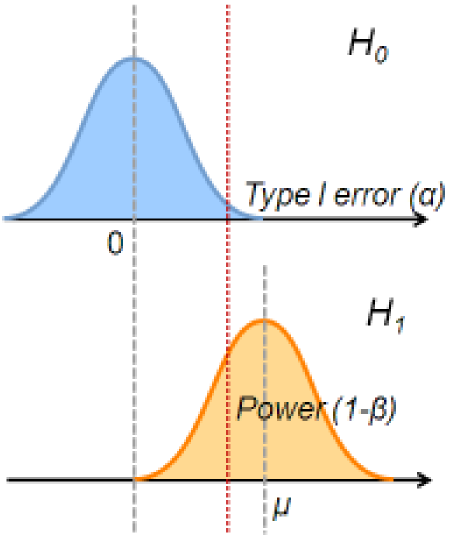
Both the Haseman-Elston method and the Sham-Purcell method use the *t* statistic to test for a nonzero regression coefficient for $\hat{\pi}$, the estimated expected IBD proportion, and therefore test for linkage. However, if there are major QTLs, the distribution of the quantitative trait value is a mixture of normal distributions. The genetic modeling framework described in Section 2.1 is an example. The dependent variables used result in residuals that are not normally distributed. Thus we are not sure if the test statistic still has an asymptotically *t* distribution or standard normal distribution in large sample. One possible alternative would be to use a permutation test to obtain *p*-value for the observed test statistic value and test for linkage. The rationale is as follows:

If the null hypothesis is true, a marker locus is not linked with a QTL, the IBD proportion at the marker loci would not be associated with the trait value. Then if we permute the order of the trait values, keep the order of the genotypes at the marker loci, and perform linkage analysis, then any significant linkage result should be obtained by chance. Hence we can obtain an empirical null distribution of the test statistic by sampling from the possible permutations and repeating the analysis on these permuted data. The empirical null distribution then can be used to estimate the empirical *p*-value for the test statistic value obtained from the original data. Details can be found in Section 3.2.

Wan et al. (1997) used permutation tests on the Haseman-Elston method under single-locus model to test for linkage. The power values they obtained by using the conventional *t*-test were not significantly different than those obtained by using permutation test. In another words, the null distribution of the *t* statistic in Haseman-Elston method is well approximated by *t* distribution, or standard normal distribution in large samples.

We conjecture that the distribution of the test statistic used in this study, *t*, under both the null and alternative hypotheses, is characterized by normal distribution especially in large samples. Specifically under the null hypothesis of no linkage we expect to observe the standard normal distribution. Under the alternative hypothesis it is expected to approximate the normal distribution with a non-zero mean and **unit standard deviation**. We will demonstrate in the Results Chapter that our conjectures are correct under the two-locus epistatic model.

As illustrated in the following figure, once the alternative mean is known and the type I error rate, or significance level, is specified, one can readily calculate the expected power. Therefore instead of reporting power under a given significance level, we based most of our power analyses on the mean of the observed *t* statistic values in this paper.



Chapter 3 Simulation

3.1 Data Generation

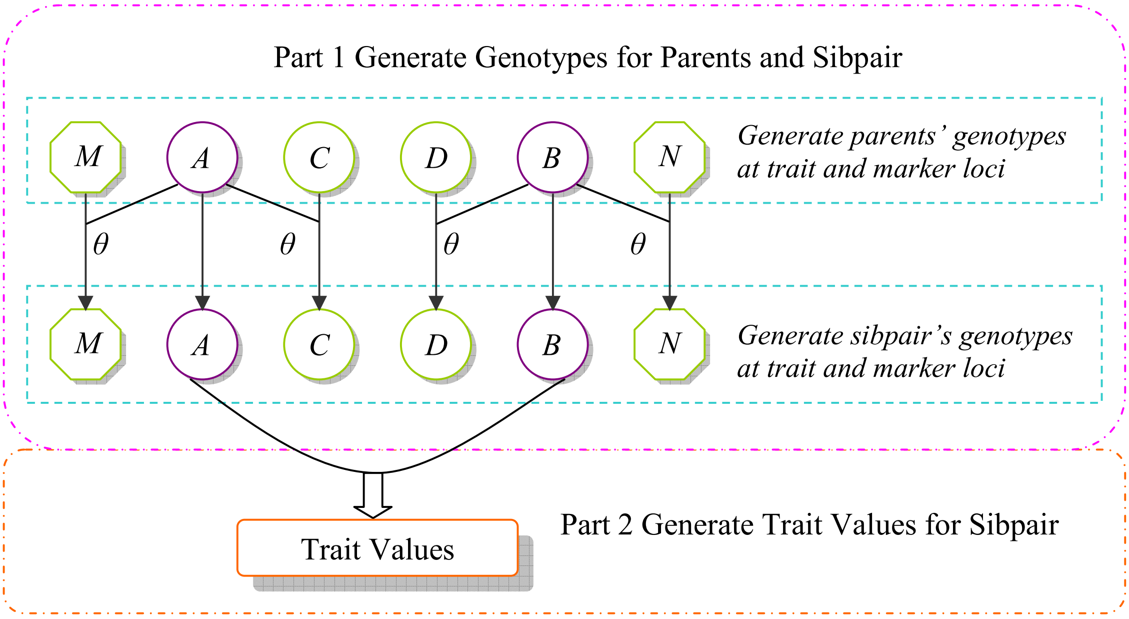
We conducted a simulation study to compare the performance of the Haseman-Elston methods for sibpair data. Since the genotypes of the sibpairs are related, we also generated their parents' genotypes to ensure this relationship. Hence the sampling unit in our paper is an independent nuclear family with both parents and a pair of siblings (i.e., a sibpair). The simulation procedure for one sampling unit, a nuclear family, is shown in Figure 3-1.

The genotypes of all family members at two unlinked, bi-allelic, pleiotropic, and epistatic trait loci, *A* and *B*, were first simulated. The quantitative trait values of the sibpairs were then generated based on the simulated genotypes at trait loci.

In addition to the trait loci we also generated two markers in linkage with each trait gene, where one is bi-allelic and the other is highly polymorphic. The bi-allelic marker linked with locus *A* was denoted as *C* and the polymorphic one was denoted as *M*. Their counterparts that linked with locus *B* were denoted as *D* and *N* respectively. We simulated a situation where there is no allelic association between the trait and marker loci, i.e. they are in linkage equilibrium, which means the frequency of a gamete having haplotype, say A_iC_j, is equal to the product of the two allele frequencies ($h_{ij} = p_iq_j$).

We will explain the specific procedure for generating genotypes and trait values in Section 3.1.1 and 3.1.2 respectively and other details about the simulation in Section 3.3. The described procedure were repeated for $\boldsymbol{n_f}$ (number of families in a sample, or sample size) multiplied by $\boldsymbol{n_s}$ (number of simulations or replications) times to obtain the final dataset.

Figure 3-1 Flowchart of Simulation Procedure for One Nuclear Family


Part 1 Generate Genotypes for Parents and Sibpair
M
A
C
D
B
N
Generate parents' genotypes at trait and marker loci
θ
θ
θ
θ
M
A
C
D
B
N
Generate sibpair's genotypes at trait and marker loci
Trait Values
Part 2 Generate Trait Values for Sibpair

Notation: loci *A* and *B* are quantitative trait loci (QTLs); loci *C* and *D* are bi-allelic markers linked to *A* and *B* respectively; loci *M* and *N* are multi-allelic markers linked to *A* and *B* respectively.

3.1.1 Generating Genotypes at Trait and Marker loci

The exact procedure varies a little depending on the sampling method. The three sampling methods considered in this paper are: (1) **random sampling**; (2) **trait**

truncated sampling, in which each family was selected through a proband that has an abnormally high quantitative trait value; (3) **disease selected sampling**, in which each family was selected through a disease affected proband.

Generating Random Samples

First we generated both parents' genotypes at the trait loci and the marker loci as independent events since we assume linkage equilibrium. Each genotype can be considered as two independent parts: allele at paternal gamete and allele at maternal gamete. Therefore all the alleles can be assigned according to their population allele frequencies. For example to obtain the father's genotype at trait locus A, we first generated a uniformly distributed random number, say w, in the range [0, 1). If $0 \leq w < p_1$ then allele A_1 was assigned. Otherwise $p_1 \leq w < p_1 + q_1 = 1$, then allele A_2 was assigned. This random-number-based assignment method was used throughout this paper.

We assume that markers M and N are highly polymorphic. That is, there could be a few dozen or even a few hundreds of different form of alleles at these loci. Unrelated individuals, such as the parents, are very unlikely to have the same alleles. Therefore I used $\boldsymbol{M_1}$ to denote the father's allele at locus M in his paternal gamete regardless of the actual form of the allele, $\boldsymbol{M_3}$ to denote the father's allele at locus M in his maternal gamete, $\boldsymbol{M_2}$ and $\boldsymbol{M_4}$ to denote the two analogous maternal alleles. The alleles $\boldsymbol{N_1}$, $\boldsymbol{N_2}$, $\boldsymbol{N_3}$, and $\boldsymbol{N_4}$ were defined in the analogous way for alleles at marker locus N.

Then we generated the sibpair's haploid genotypes, or haplotypes, based on their parents' genotypes, first at the trait loci and then at the marker loci.

For example, to obtain the paternal allele at trait locus A, the allele inherited from the father, we referred to the father's simulated genotype $A_{(f)}A_{(m)}$, where $A_{(f)}$ denoted his paternal allele which can take the form of either A_1 or A_2 and $A_{(m)}$ denoted his maternal allele. The offspring can inherit either one of the alleles with equal probabilities.

Once we randomly picked an allele at trait locus, the offspring would get the alleles in that gamete at the marker loci linked to the trait locus unless crossover occurred (see Section 1.3). For example if the offspring's paternal allele was generated to be $A_{(m)}$, then s/he would obtain $C_{(m)}$ at locus C if the gamete s/he received was non-recombinant and $C_{(f)}$ if the gamete was recombinant, with probabilities $1 - \theta$ and θ respectively.

However if we cannot deduce whether the proband's allele at the trait locus was inherited from the parent's paternal or maternal gamete, *i.e.* if the parent's genotype at trait locus was homozygous, then we assigned either one of the two alleles at the linked marker locus with equal probability.

Generating Trait Truncated Samples

The trait truncated samples were generated similar to the random samples except that only families with at least one offspring's simulated trait value greater than a certain value were kept in the dataset. We kept simulating families until we obtained n_f such families in each replicate.

The threshold value used to select the trait truncated samples is set to be 1.645, the 95th percentile for the standard normal distribution, or 1.645 standard deviation (SD) away from the population mean.

CHAPTER 3. SIMULATION

Generating Disease Selected Samples

The order of simulation for disease selected samples differed from that of the random samples. We first generated the two-locus genotype at the trait/disease genes for the disease proband. We then generated the parents' genotypes at the trait loci based on the proband's genotype. The parents' genotypes at marker loci were simulated as independent events as in the random samples, as were the genotypes for the proband's sibling. Finally, based on the recombination fraction, the parents' genotypes at the trait and marker loci, and the proband's genotypes at the trait loci, we obtained the proband's genotypes at that marker loci.

The two-locus genotype for the proband at the trait loci was generated conditional on the disease status being "affected". The conditional probability, $Pr(G_{AB}|D)$, was given in Equation (2-7) in Section 2.1.

Based on the procedure described for random samples we can easily deduce the conditional probabilities of both parents' genotypes at each trait locus (denoted as *M*) given proband genotypes. The conditional probabilities at locus *A* are tabulated in Table 3-1. The conditional probabilities at locus *B* can be done analogously with allele frequencies p_2 and q_2 replacing p_1 and q_1 in the table.

Table 3-1 The Probability of Parents' Genotype at Locus *A* Given Proband's Genotype

$Pr(M \mid G_A = A_1A_1)$

G_M / G_F	A_1A_1	A_1A_2	A_2A_2
A_1A_1	p_1^2	p_1q_1	0
A_1A_2	p_1q_1	q_1^2	0
A_2A_2	0	0	0

$Pr(M \mid G_A = A_1A_2)$

G_M G_F	A_1A_1	A_1A_2	A_2A_2
A_1A_1	0	$\frac{1}{2}\, p_1^2$	$\frac{1}{2}\, p_1q_1$
A_1A_2	$\frac{1}{2}\, p_1^2$	p_1q_1	$\frac{1}{2}\, q_1^2$
A_2A_2	$\frac{1}{2}\, p_1q_1$	$\frac{1}{2}\, q_1^2$	0

$Pr(M \mid G_A = A_2A_2)$

G_M G_F	A_1A_1	A_1A_2	A_2A_2
A_1A_1	0	0	0
A_1A_2	0	p_1^2	p_1q_1
A_2A_2	0	p_1q_1	q_1^2

Once we generated the parents' haplotypes at the trait and marker loci, we can generate the sibling's haplotype as described earlier with the given recombination fractions between two loci.

For the proband, since we had generated his/her genotype at trait loci, we would generate his/her genotype at marker loci as described for random samples.

Calculating IBD proportions

Once we generated all the genotypes, we could calculate the proportion of alleles shared identical by descent (IBD), or IBD proportions, at the trait and marker loci.

Since the four alleles at each polymorphic marker locus can be distinguished from each other, the number of alleles shared IBD in a sibpair can be obtained by simply counting the matching alleles. The IBD proportion at these markers (denoted as π_M and π_N) then can be obtained by dividing the IBD number by 2.

However, since the IBD proportion cannot be deduced at a bi-allelic locus directly, the estimated expected values had to be calculated based on the given parents' genotypes (or the mating type) and the given sibpair's genotypes (or the sibpair type). The estimated value for $\boldsymbol{\pi}$ was then calculated as

$$\hat{\pi} = E(\pi) = 0 * Pr(\pi = 0) + 0.5 * Pr(\pi = 0.5) + 1 * Pr(\pi = 1) \qquad (3\text{-}1)$$

In this paper the observable IBD proportion for polymorphic marker and the expected value for bi-allelic locus are used synonymously.

Since the trait locus is bi-allelic, there are four types of mating for the parents. Analogously, there were four types of sibpairs.

I. Both parents were homozygous at the trait locus, and the parents' genotypes were the same: {11, 11} or {22, 22};

II. Both parents were homozygous at the trait locus, but the parents' genotypes were different: {11, 22};

III. Both parents were heterozygous at the trait locus: {12, 12};

IV. One parent was heterozygous and the other was homozygous: {12, 11} or {12, 22}.

Given the parents' and the sibpair's genotypes, the probability of sibpair sharing a specific number of alleles IBD and the expected IBD proportion then can be unequivocally calculated using Table 3-2 below. The estimated IBD proportion at a bi-allelic locus can take either one of the five values (0, 0.25, 0.5, 0.75, or 1). Note that only heterozygous parents result in variability in the expected IBD proportion.

Table 3-2 Probability and Expectation of the IBD Proportion Based on the Mating and the Sibpair Type

Parent's Genotypes (Mating Type)	**Sibpair's Genotypes (Sibpair Type)**	**Pr(π= 0)**	**Pr(π= 0.5)**	**Pr(π= 1)**	$\hat{\pi}$
I. {11, 11} or {22, 22}	I. {11, 11} or {22, 22}	0.25	0.5	0.25	0.5
II. {11, 22}	III.{12, 12}	0.25	0.5	0.25	0.5
III. {12, 12}	I. {11, 11} or {22,22}	0	0	1	1
	II. {11, 22}	1	0	0	0
	III. {12, 12}	0.5	0	0.5	0.5
	IV. {12, 11} or {12, 22}	0	1	0	0.5
IV. {12, 11} or {12, 22}	I. {11, 11} or {22, 22}	0	0.5	0.5	0.75
	III. {12, 12}	0	0.5	0.5	0.75
	IV. {12, 11} or {12, 22}	0.5	0.5	0	0.25

3.1.2 Generating Quantitative Trait Values

The procedure for generating the sibpair's quantitative trait values is the same regardless of the sampling scheme. The genetic model framework described in Section 2.1 was used throughout this paper. Specifically the trait value was generated from one fixed genetic effect and two random environmental (including polygenic) effects. A fixed value μ_{ij} was assigned if the simulated two-locus genotype of the offspring was g_{ij}, i.e. having i copies of A_1 allele and j copies of B_1 allele (i, j = 0, 1, 2). A random number drawn from the normal distribution with mean zero and shared residual variance σ_s^2 was assigned to both siblings. Two different random numbers were assigned to the siblings, both of which follow normal distribution with mean zero and non-shared residual variance σ^2_{ns}. The summation of these three parts became the trait value of an offspring, *X*.

The actual mean, the variance, and the correlation of the quantitative trait values in the population can be calculated using Equations (2-8) through (2-20) defined in Section 2.1. The population mean and variance then can be used to standardize the trait values

$$Z = \frac{X - \mu}{\sqrt{\sigma^2}}$$

The dependent variables in regression, namely squared difference in original Haseman-Elston method, Y_{HE} and the one proposed by Sham and Purcell, Y_{SP}, then can be calculated as follow

$$Y_{HE} = (Z_1 - Z_2)^2 \tag{3-2}$$

$$Y_{SP} = \frac{(Z_1 + Z_2)^2}{(1+\rho)^2} - \frac{(Z_1 - Z_2)^2}{(1-\rho)^2} \tag{3-3}$$

where Z_i is the standardized trait value of i^{th} sibling in a sibpair (i = 1, 2). This order is of no importance for random samples. For convenience we called the proband the first sibling and the other offspring the second sibling in trait truncated and disease selected samples.

As explained in Section 1.4 we may not know the population trait parameter values. Therefore we also calculated $\tilde{Y}_{SP}$ based on CMLEs defined in Section 2.4

$$\tilde{Y}_{SP} = \frac{(\tilde{Z}_1 + \tilde{Z}_2)^2}{(1+\hat{\rho})^2} - \frac{(\tilde{Z}_1 - \tilde{Z}_2)^2}{(1-\hat{\rho})^2} \tag{3-4}$$

where

$$\tilde{Z}_i = \frac{X_i - \hat{\mu}}{\sqrt{\hat{\sigma}^2}} \ (i = 1, 2).$$

Here $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\rho}$ are the CMLEs obtained by numerically solving Equations (2-26) through (2-28) explained in Section 2.4.

3.2 Permutation Tests

Since we only simulated data under the alternative hypothesis that the marker loci were closely linked to the epistatic trait loci, we used a permutation test to verify indirectly the distributional assumption about the test statistic under the null hypothesis.

Specifically, for each replicate, we kept the order of the IBD proportions at the marker loci and randomly shuffled the trait values between families. Then Haseman-Elston and Sham-Purcell regressions were performed on the permuted data and the *t* statistic value was recorded. For each replicate, 1000 permutations were done. Therefore, for each *t* value obtained from the original dataset, there were 1000 corresponding *t* values obtained from the permuted dataset. These “permuted *t* values” were then used as the empirical null distribution to obtain the empirical *p*-value for the original *t* value in this replicate.

The empirical *p*-value was calculated as the percentage of permuted *t* values at least as extreme as the original *t* value. For the Haseman-Elston method, we test for a negative regression coefficient; i.e., lower tail test, “as extreme as” was defined to be “smaller”. For Sham-Purcell method we test for a positive regression coefficient; i.e., upper tail test, “as extreme as” was then defined to be “larger”.

3.3 Study Design and Simulation Settings

Now that we have introduced our analysis methods and simulation procedures, we can summarize all the factors involved in our study design and present the specific settings and other details of the simulation.

The following seven factors are modeled to affect the statistical power to detect linkage: [1] Epistatic Model (expected mean values given two-locus genotype) (M) [2] Minor Allele Frequency (P = 0.05, 0.10, 0.15) [3] Regression-Based Linkage Method (R = he: original Haseman-Elston, sp_tru: Sham-Purcell modification with actual trait parameter values, sp_est: with estimated parameter values) [4] Type of Analysis used to identify QTL (A = s: single-locus analysis, i: two-locus interaction analysis) [5] Type of Markers (L = t: trait gene, b: bi-allelic marker, m: multi-allelic marker) [6] Sampling Method (S = d: disease selected sampling, t: trait truncated sampling, r: random sampling) [7] Sample Size (N = 250, 500, 1000). Noted that factors R and L shared common samples.

All of my simulations were generated under one genetic modeling framework described in the Section 2.1. By changing μ_{ij}, the expected trait value given two-locus genotype being g_{ij}, different genetic models can be tested. Six models were considered in this paper. Specification of these models is shown in Table 3-3. Each matrix corresponds to the nine genotypes shown in Table 2-3 in Section 2.1. Note that except for Model 2 and Model 4 (in italic) all other models are symmetric, i.e., the effects of trait locus *A* and *B* are equal. In Model 2 and Model 4 the genetic effect (additive, dominance, and epistatic) of locus *A* is designed to be greater than that of locus *B*.

Table 3-3 Specification of Expected Trait Values Before Standardization

Model	1	*2*	3	*4*	5	6
μ_{ij}	**1 1 0**	***1 1 0***	**1 0 0**	***1 1 1***	**1 1 0**	**0 1 1**
	1 1 0	***0 0 0***	**0 0 0**	***1 0 0***	**1 0 0**	**1 0 0**
	0 0 0	***0 0 0***	**0 0 0**	***0 0 0***	**0 0 0**	**1 0 0**

Like many other QTL studies, we focus on this set of models with limited trait mean values, namely 0 or 1, to allow easy comparison with binary disease models. There are $2^9 = 512$ possible two-locus binary models in total, 50 of which are unique (Li and Reich 2000). Here we considered the six that have occurred in a biological examples (cf. Neuman and Rice 1992).

In Model 1 (dominant-dominant), a dominant minor allele must be present at both loci in order to increase the trait value. A binary trait example for this epistatic model would be the production of chlorophyll in corn plants.

In Model 2 (recessive-dominant), the minor allele is recessive at locus *A* and dominant at locus *B*. The plumage color of chickens could be modeled by this epistasis.

In Model 3 (recessive-recessive), a recessive minor allele must be present at both loci. This model can be used to explore the prelingual deafness.

In Model 4, one locus has modifying effect on the other locus. All A_1A_1 genotypes are at risk. However, the A_1A_2 genotype is only at risk only if B_1B_1 genotype is present.

In Model 5, at least 3 minor alleles in total at two loci have to present to increase the trait value. This model can be used as a model for the kernel color in wheat. The trait takes value of 1 if kernel color is darkest red.

In Model 6, the minor alleles at loci *A* and *B* are recessive. The epistasis is manifested by having the genotype $A_1A_1B_1B_1$ result in the same mean trait value as those without the minor allele. This is the two-locus epistasis model which has been proposed for handedness.

As mentioned in Section 2.1, the two QTLs were also modeled to determine the status of a disease related to the trait. We assume that the baseline penetrance for the disease equals zero and that the penetrance increases linearly at the rate of 0.2 per disease susceptible allele, so that the disease penetrances of the nine genotypes at these two loci range from 0 to 0.8. The minor allele frequencies at four bi-allelic loci *A* through *D* (*A* and *B* being trait loci and *C* and *D* being loci linked to *A* and *B* respectively) were all set to be equal to a single value p and took values of 0.05, 0.1, or 0.15. These particular values were chosen to ensure that the disease prevalence is less than 15%, specifically 4% (when $p = 0.05$), 8% (when $p = 0.1$), or 12% (when $p = 0.15$).

The population mean, variance, and correlation between siblings of the quantitative trait then can be calculated using Equations (2-8) through (2-20) defined in Section 2.1, and were used to standardize the trait value. Furthermore the proportion of the total trait variance accounted for by the two QTLs combined, the shared residual effect, and the non-shared residual effect was the same: 1:2:7 in all cases.

CHAPTER 3. SIMULATION

The regression-based linkage methods, the types of analysis, and the types of markers were explained in Chapter 2. The sampling methods were explained in Section 1.2 and in Section 3.1.

For each of the three sampling methods, I generated 1000 simulations or replicates. Each batch of 1000 simulation consists of a sample of 250, 500, or 1000 independent nuclear families (sample size N = 250, 500, 1000).

The data generation and the Haseman-Elston regression were done using a C program that I wrote. Some functions in the GNU Scientific Library (GSL) 1.12 were used as well. The random number generator MT19937 by Makoto Matsumoto and Takuji Nishimura (described in the GSL reference manual) was used. The analysis was performed mostly in SAS and partly in R.

Chapter 4 Results

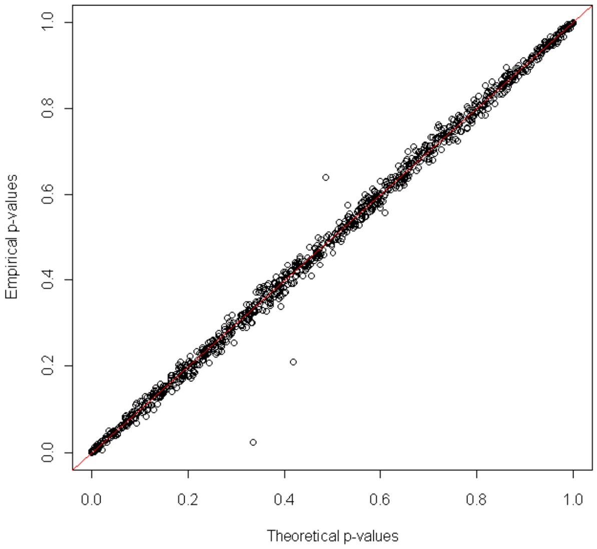
4.1 Null and Alternative Distributions of the Test Statistic

Before we conducted our power analyses, we sought to verify our conjectures stated in Section 2.6 about the distribution of the *t* statistic. First we will present the results from the permutation test described in Section 3.2.

Specifically, we want to compare the p-values obtained by using *t* test (theoretical p-values) with the p-values obtained by using permutation tests (empirical p-values). The former assumes that the null distribution of the test statistic is standard normal. The latter requires no distributional assumptions about the test statistic. As an example, we show the results in Figure 4-1 with the following setting: number of simulations = 1000, number of permutations (only for the empirical p-values) = 1000, number of sibpairs per simulation = 500, epistatic model M = 3, minor allele frequency P = 0.05, sampling method S = trait truncated sampling, regression-based linkage method R = Sham-Purcell method with estimated trait parameter values, type of marker loci = trait gene itself, and type of analysis = two-locus interaction.

We can see from the figure that the scatterplot of the p-values obtained by using *t* test is almost identical to that of the empirical p-values. The two-sample *t* test result also suggests that the difference is not significant (p-value = 0.9581). This conclusion also applied to other settings (results not shown).

Figure 4-1 Scatterplot of Theoretical p-values obtained by using *t* test vs. Empirical p-values obtained by using permutation test



Our results are consistent with the findings of Wan et al. (1997). They also concluded that the *t* test approximates the permutation test very well. Further we considered a number of complex two-locus epistatic models in this paper. These results thus indirectly verified our conjecture about the distribution of the test statistic under the null hypothesis of no linkage.

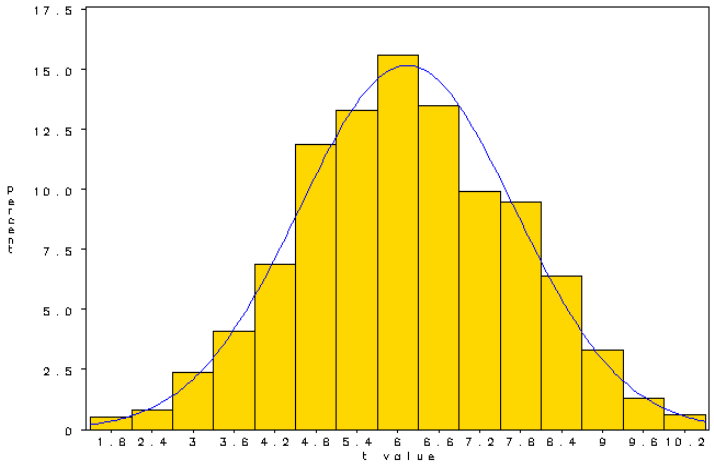
To demonstrate that our conjecture about the alternative distribution of the test statistic is also correct, we conducted tests of normality on the *t* values obtained by using genotypic and quantitative trait data generated under the alternative hypothesis. Because, for the seven factors considered in this paper, there are over a thousand combinations. We show here the results for only one situation. The general conclusion applied for the rest of the cases (data not shown).

CHAPTER 4. RESULTS

The results from all three normality tests, the Kolmogorow-Smirnow test, the Cramer-von Mises test, and the Anderson-Darling test (performed by UNIVARIATE Procedure in SAS) suggested that the normal distribution fits the observed *t* values fairly well (all p-values > 0.15). We can also see this from the histogram shown in Figure 4-2.

The mean and standard deviation of the *t* values in this case are 6.14 and 1.58 respectively. The standard deviation is a little bigger than what we conjectured in Section 2.6. However, if we consider all of the cases except for Epistatic Model 3 with Minor Allele Frequency equals to 0.05[1], then we can obtain that the median, mean, and standard deviation of the standard deviation of *t* values are 1.01, 1.19, and 0.51 respectively. Thus a normal distribution with standard deviation of 1 is overall a good assumption for our data.

Figure 4-2 Histogram of *t* value under Epistatic Model 1 with Minor Allele Frequency equals to 0.05, Sample Size equals to 500, and using Trait Loci as Markers, Trait Truncated Sampling Method, Sham-Purcell Method, and Two-Locus Interaction Analysis



[1] We exclude this setting because it produced two extreme outliers.

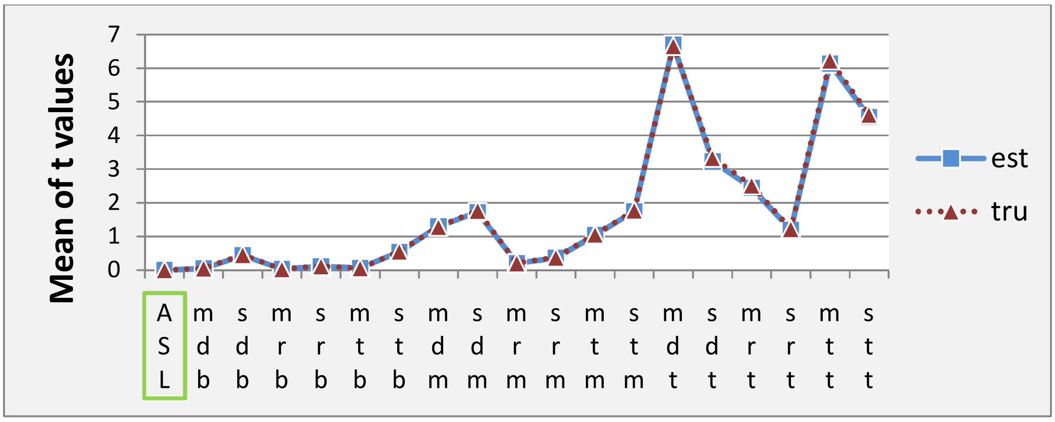
4.2 Robustness of the Sham-Purcell Method to Bias in Estimates of Trait Parameter Values

In this paper we used the maximum likelihood method conditional on the proband's trait value, discussed in Section 1.4 and 2.4, to estimate the trait parameter values for the Sham-Purcell method. Regardless of how well this method of estimation performs, we do not expect it to affect the power of the Sham-Purcell method under our two-locus epistatic and pleiotropic model for the large samples used in this paper ($N \geq 250$).

As explained in Section 2.6, given constant standard deviation with respect to model etc., the shift in the mean of the *t* statistic values under these alternative hypotheses and study designs can be translated to power once the significance level is specified. Therefore all power analyses in this chapter focus on the (shifted) mean of the *t* values.

As an example, we compared the mean of the observed *t* values using the Sham-Purcell method based on estimated parameter values with that observed using the Sham-Purcell method based on true parameter values for analyses. We consider analyses based on three types of marker locus (L = b: bi-allelic; m: multi-allelic; t: trait gene), three sampling schemes (S = d: disease selected sampling; t: trait truncated sampling; r: random sampling), and two types of analysis (A = i: two-locus interaction; s: single-locus) under the epistatic Model 1 with minor allele frequency equals to 0.05, and sample size equals to 500. The results are summarized in Figure 4-3.

Figure 4-3 Mean of *t* Values Obtained by using the Sham-Purcell Method under Epistatic Model 1 when Minor Allele Frequency equals to 0.05 and Sample Size equals to 500: Comparison of Results Using Estimates of Parameters with the True Parameter Values



Notation: R: Regression-Based Linkage Method (R = est: Sham-Purcell method with estimated parameter values; tru: Sham-Purcell method with true parameter values); A: Type of Analysis (A = s: single-locus; i: two-locus interaction); S: Sampling Method (S = d: disease selected sampling, t: trait truncated sampling, r: random sampling); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself).

As we can see the mean of the *t* statistic using the estimated parameter values (blue line) almost completely overlaps with the mean obtained using the true parameter values in the analysis (red line) regardless of the sampling method (S = d, t, r). In another words, using estimated trait parameter values does not affect the power of Sham-Purcell method at all even in selected samples (S = d or t).

In fact this conclusion can be applied to all of the situations examined in this paper (results not shown in graph). ANOVA analysis with all seven factors listed in Section 3.3 also confirmed this conclusion from visual examination (F = 1.09, df = 1, p = 0.30). See detailed SAS output in Appendix A.

Although we cannot obtain unbiased parameter estimates from selected samples (trait truncated samples, t, or disease selected samples, d), the statistical power appears to be unaffected in the cases we have considered.

4.3 Results for Six Epistatic Models

Because the underlying genetic model is not a variable we can observe or control in real life, we conduct our analysis separately for each of the six models studied in this paper. To visualize the relative power under different settings we first draw a comparative graph of the mean of *t* statistic values for each genetic model (Figure 4-4 to Figure 4-9). As we explained in Section 2.6, a higher mean of *t* indicates higher statistical power. For ease of visual examination, we only show results obtained by using the Sham-Purcell method. The results obtained by using Haseman-Elston method were included in the analyses shown later in this chapter.

Also shown in these figures are some percentages of variance components relative to the total two-locus genetic variance (V_G) which, as we described in Section 3.3, accounts for **10%** of the total trait variance (V_T) in all cases. All of the variance components were calculated by using Equations (2-10) through (2-18) defined in Section 2.1. The complete list of the percentages can be found in Appendix B. Note that the variance components are determined solely by two factors, the epistasis model and the minor allele frequency.

4.3.1 Model 1 (dominant-dominant)

In Model 1 the majority of the genetic variance consists of variance due to the interaction of additive effects at the two trait loci (*Vaa*) especially when the minor allele frequencies (P) are small.

We can see from Figure 4-4 that for all three sampling methods (S = d: disease selected sampling; t: trait truncated sampling; r: random sampling) we have six almost monotonically increasing lines as the sample size increases (N/250 = 1, 2, 4) and as the type of markers changes (L = b: bi-allelic; m: multi-allelic; t: trait gene) for the 2 types of analysis (A = i: two-locus interaction; s: single-locus) and the 3 minor allele frequencies (P/0.05 = 1, 2, 3).

The effect of using different types of marker is noticeably large. Both bi-allelic and multi-allelic markers results in a mean *t* of less than 2 with all but few exceptions. This translates to a power lower than 50% at any significance level smaller than 0.023, even though these markers are modeled to be in complete linkage (i.e. $\theta = 0$) with the trait genes. On the other hand, when we use the trait genes themselves, the mean *t* can soar up to as high as 9, or 100% power, for both selected sampling methods.

In most of the cases under Model 1, the disease selected samples (d) result in higher mean *t* values than those obtained by using the random samples (r), and almost the same as those obtained by using the trait truncated samples (t).

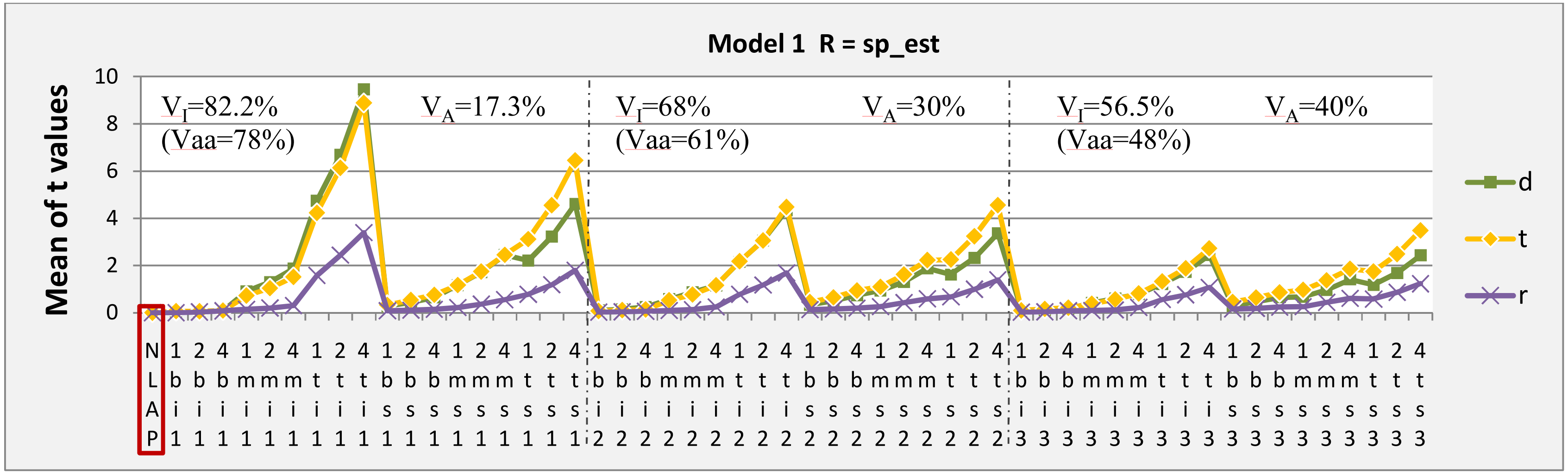
We also notice that the single-locus analysis results in mean of *t* comparable to those using two-locus multiple regression analysis in most of the cases except when minor allele frequencies equal to 0.05.

We can infer from the figure that there are some significant interactions between the explanatory factors. Therefore an ANOVA test including all six factors and their two-way interactions was conducted to quantify these relations. The notation we used for these six factors are: P = Minor Allele Frequencies (0.05; 0.10;

0.15); R = Regression Method (he: original Haseman-Elston method; sp_est: Sham-Purcell method using estimated parameter values); L = Type of markers (b: bi-allelic; m: multi-allelic; t: trait gene); A = Type of Analysis (s: single-locus; i: two-locus interaction); S = Sampling Method (d: disease selected sampling; t: trait truncated sampling; r: random sampling); N = Sample Size (250; 500; 1000). The full output can be found in Appendix C.

Five factors, except for A, type of analysis, have significant main effect (p<0.0001). All six factors have significant two-way interactions with at least one other factor (p<0.0001). I noted firstly that R, the regression-based linkage method, only interacts with L, the type of markers. Secondly L interacts with all five other factors and its main effect explains a majority of variance of the mean of *t* statistic values in terms of the total sum of square (43%). Finally all significant factors and their two-way interactions have p-value less than 0.0001.

Figure 4-4 Mean of *t* Values of Three Sampling Method (d: disease selected sampling, t: trait truncated sampling, r: random sampling) under Model 1 (dominant-dominant)



Notation: R: Regression-Based Linkage Method (R = sp_est: Sham-Purcell method with estimated parameter values); P: Minor Allele Frequencies (P/0.05 = 1; 2; 3); A: Type of Analysis (A = s: single-locus; i: two-locus interaction); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); N: Sample Size (N/250 = 1; 2; 4). V_A: Additive Variance; V_I: Two-Locus Epistatic Variance; Vaa: Variance due to interaction of additive effects at two loci.

Note: all variance component values are shown in percentage of the total two-locus genetic variance (V_G) which account for 10% of the total variance of the trait (V_T).

4.3.2 Model 2 (recessive-dominant)

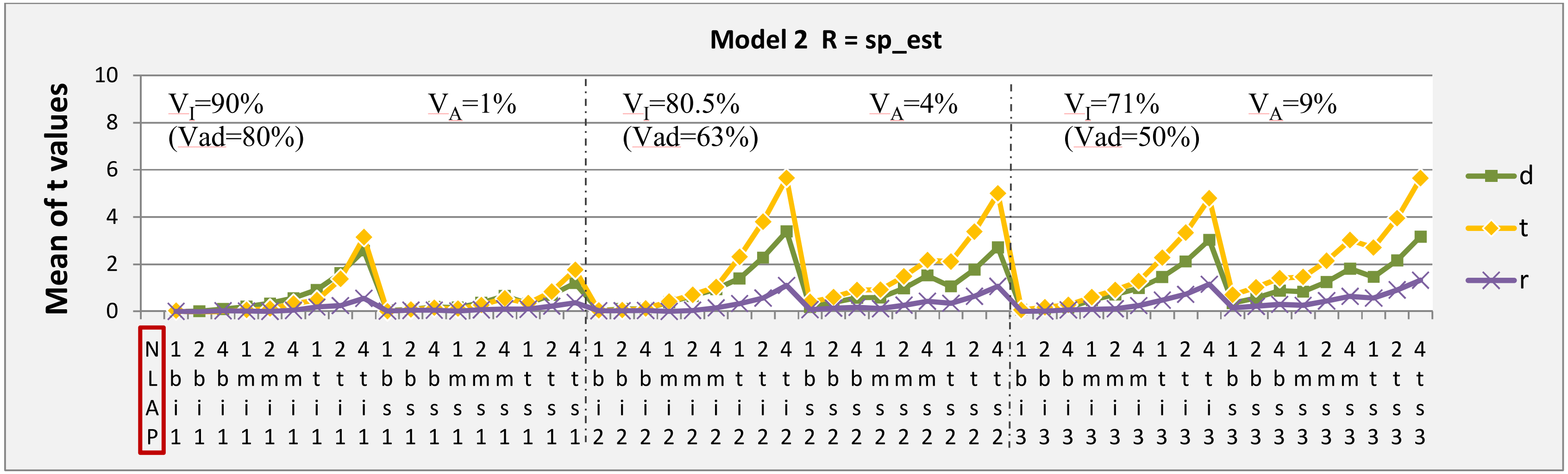
In Model 2 the majority of the two-locus genetic variance consists of the epistatic variance due to the interaction of additive effect at one locus and dominant effect at the other locus (*Vad* ranges from 50% to 80% of V_G).

Overall we can see that the pattern shown in Figure 4-5 is somewhat similar to that in Figure 4-4. There are six almost monotonically increasing lines for each sampling method. Use of larger sample size and multi-allelic marker or trait loci increases power. Selected samples generally result in a higher mean *t* value than random samples.

All main effects and two-way interactions are highly significant ($p<0.0001$) except the interaction between type of analysis and sample size (A*N), the interaction between type of analysis and regression-based linkage method (A*R), and interaction between sample size and minor allele frequency (N*P).

We also notice that although the total epistasis (V_I) also accounts for over 80% of the two-locus genetic variance when minor allele frequencies equal 0.05, the mean of *t* values is generally lower than 3 (Figure 4-5) which in turn is much lower than what we observed in Figure 4-4.

Figure 4-5 Mean of *t* Values of Three Sampling Method (d: disease selected sampling, t: trait truncated sampling, r: random sampling) under Model 2 (recessive-dominant)



Notation: R: Regression-Based Linkage Method (R = sp_est: Sham-Purcell method with estimated parameter values); P: Minor Allele Frequencies (P/0.05 = 1; 2; 3); A: Type of Analysis (A = s: single-locus; i: two-locus interaction); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); N: Sample Size (N/250 = 1; 2; 4). V_A: Additive Variance; V_I: Two-Locus Epistatic Variance; Vad: Variance due to interaction of additive effect at one locus and dominant effect at the other locus.

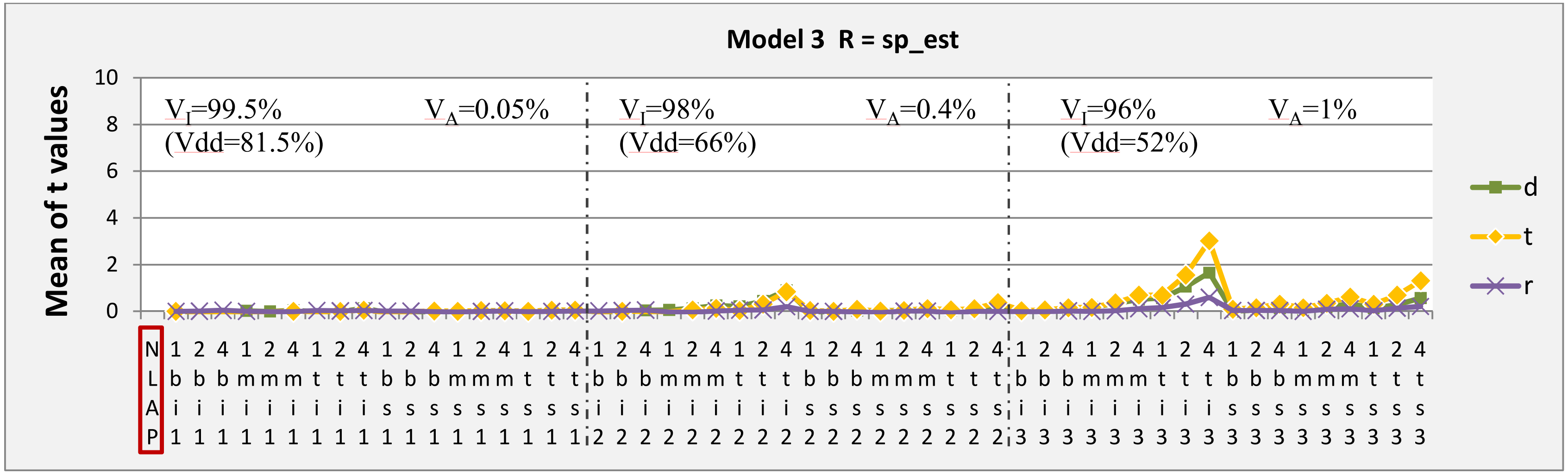
Note: all variance component values are shown in percentage of the total two-locus genetic variance (V_G) which account for 10% of the total variance of the trait (V_T).

4.3.3 Model 3 (recessive-recessive)

In Model 3, the majority of the two-locus genetic variance consists of the variance due to the interaction of dominant effects at the two loci (*Vdd* ranges from 50% to 80% of V_G in the cases considered).

Although the total epistasis accounts for over 95% of the two-locus genetic variance, a great proportion of which is from the interaction between dominant effects at the two loci. The regression-based linkage methods have virtually no power at all in detecting this type of epistasis.

Figure 4-6 Mean of *t* Values of Three Sampling Method (d: disease selected sampling, t: trait truncated sampling, r: random sampling) under Model 3 (recessive-recessive)



Notation: R: Regression-Based Linkage Method (R = sp_est: Sham-Purcell method with estimated parameter values); P: Minor Allele Frequencies (P/0.05 = 1; 2; 3); A: Type of Analysis (A = s: single-locus; i: two-locus interaction); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); N: Sample Size (N/250 = 1; 2; 4). V_A: Additive Variance; V_I: Two-Locus Epistatic Variance; Vdd: Variance due to interaction of dominant effects at two loci.

Note: all variance component values are shown in percentage of the total two-locus genetic variance (V_G) which account for 10% of the total variance of the trait (V_T).

4.3.4 Model 4 (one locus has modifying effect on the other)

In Model 4, the majority of the two-locus genetic variance consists of the variance due to the dominant effect at two loci combined (V_D ranges from 58% to over 80% of V_G).

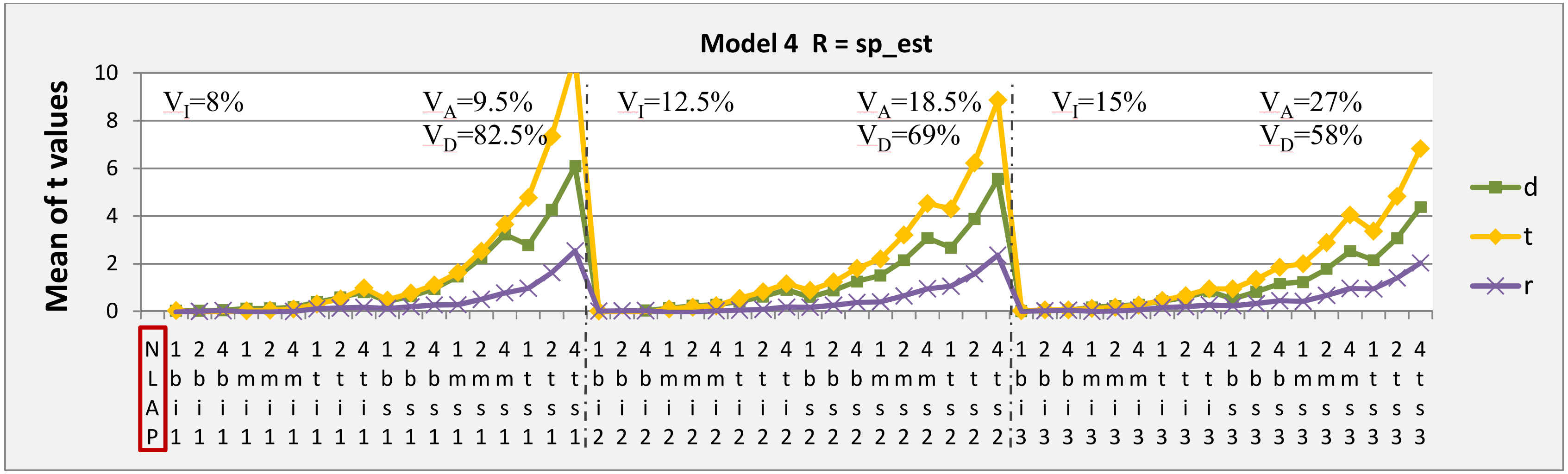
Unlike Figure 4-4 or Figure 4-5, we can only see three almost monotonically increasing lines for each sampling method.

Since the overall epistasis accounts for less than 15% of V_G, i.e., less than 1.5% of the total variance of the trait, the two-locus interaction analysis naturally has no power in detecting epistasis.

The single-locus analysis, on the other hand, results in very high mean of *t* statistic values, especially when the minor allele frequencies are low (P = 0.05), even though the additive variance is relatively small (less than 10%) in this case.

All main effects are significant except for factor P, the minor allele frequency. Factor P and factor R, the regression-based linkage method, interact with one other.. Factor S, the sampling method, and N, the sample size, have significant two-way interactions with three other factors. Factors L, the type of markers, and A, the type of analysis, have significant two-way interactions with four other factors. See Appendix C for details.

Figure 4-7 Mean of *t* Values of Three Sampling Method (d: disease selected sampling, t: trait truncated sampling, r: random sampling) under Model 4 (one locus has modifying effect on the other)



Notation: R: Regression-Based Linkage Method (R = sp_est: Sham-Purcell method with estimated parameter values); P: Minor Allele Frequencies (P/0.05 = 1; 2; 3); A: Type of Analysis (A = s: single-locus; i: two-locus interaction); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); N: Sample Size (N/250 = 1; 2; 4). V_A: Additive Variance; V_D: Dominant Variance; V_I: Two-Locus Epistatic Variance.
Note: all variance component values are shown in percentage of the total two-locus genetic variance (V_G) which account for 10% of the total variance of the trait (V_T).

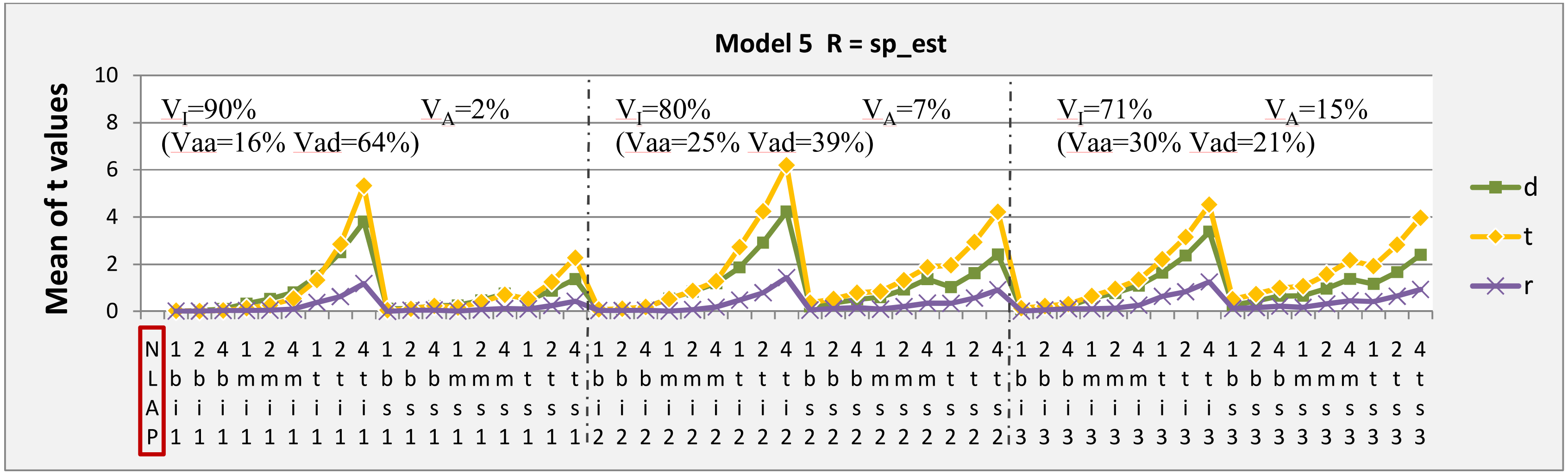
4.3.5 Model 5 (at least 3 minor alleles in total at two loci)

In Model 5, the majority of the two-locus genetic variance consists of the variance due to the interaction of the additive effect at one locus and the dominant effect at the other locus (*Vad*) and the variance due to the additive effects at two loci (*Vaa*).

The proportions of *Vaa* and *Vad* in this model, are between Model 1, where the majority is *Vaa*, and Model 3, where the majority is *Vad*. Interestingly the pattern of the mean of *t* values we observed in Figure 4-8 is also close to those in Figure 4-4 and Figure 4-5.

Again the six factors have significant main effects and some two-way interaction effects on the mean of *t* statistic values. See Appendix C for details.

Figure 4-8 Mean of *t* Values of Three Sampling Method (d: disease selected sampling, t: trait truncated sampling, r: random sampling) under Model 5 (at least 3 minor alleles in total at two loci)



Notation: R: Regression-Based Linkage Method (R = sp_est: Sham-Purcell method with estimated parameter values); P: Minor Allele Frequencies (P/0.05 = 1; 2; 3); A: Type of Analysis (A = s: single-locus; i: two-locus interaction); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); N: Sample Size (N/250 = 1; 2; 4). V_A: Additive Variance; V_I: Two-Locus Epistatic Variance; Vaa: Variance due to interaction of additive effects at two loci; Vad: Variance due to interaction of additive effect at one locus and dominant effect at the other locus.

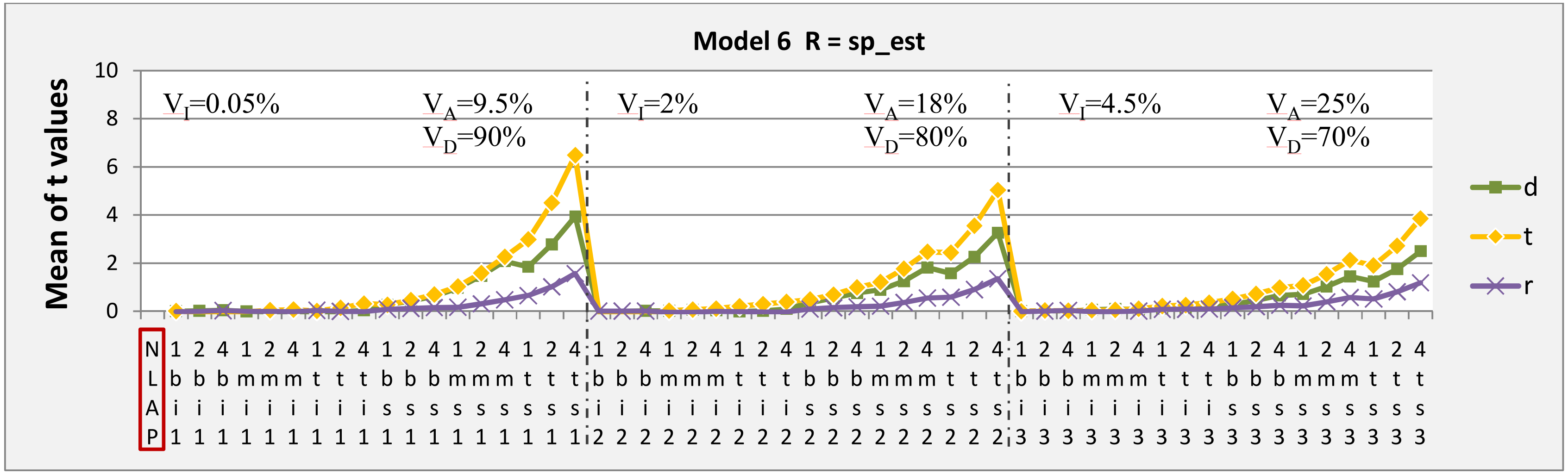
Note: all variance component values are shown in percentage of the total two-locus genetic variance (V_G) which account for 10% of the total variance of the trait (V_T).

4.3.6 Model 6 (recessive with no expression for $A_1A_1B_1B_1$)

In Model 6, the majority of the two-locus genetic variance consists of the variance due to the dominant effect at the two loci combined (V_D ranges from 70% to 90% of V_G).

The epistasis, in this model, only accounts for less than 5% of the two-locus genetic variance, or less than 0.5% of the total variance of the trait. Thus it is not surprising to observe pattern similar to that of Model 4. The two-locus interaction analysis virtually has no power, and the single-locus analysis combined with selected sampling gives much higher power even though the additive variance is relatively small.

Figure 4-9 Mean of *t* Values of Three Sampling Method (d: disease selected sampling, t: trait truncated sampling, r: random sampling) under Model 6 (recessive with no expression for $A_1A_1B_1B_1$)



Notation: R: Regression-Based Linkage Method (R = sp_est: Sham-Purcell method with estimated parameter values); P: Minor Allele Frequencies (P/0.05 = 1; 2; 3); A: Type of Analysis (A = s: single-locus; i: two-locus interaction); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); N: Sample Size (N/250 = 1; 2; 4). V_A: Additive Variance; V_D: Dominant Variance; V_I: Two-Locus Epistatic Variance.
Note: all variance component values are shown in percentage of the total two-locus genetic variance (V_G) which account for 10% of the total variance of the trait (V_T).

4.4 Modeling the Mean of the Test Statistic Values in Terms of the Epistatic Variance Component

Based on the visual examination described in the previous section, we can see that all of the factors discussed in this paper affect the power to detect epistatic genes. These factors have significant main effects and/or two-way interactions on the mean of *t* values.

We also noted that the regression-based linkage methods considered in this paper have much higher power to detect epistatic variance due to interaction of additive effects (*Vaa*) than to detect epistatic variance due to interaction of dominant effects (*Vdd*).

In this section, we will try to model the mean of *t* values obtained by using two-locus interaction analysis in terms of the following factors: (1) the epistatic variance due to interaction of additive effects, Vaa, (2) the type of markers, L, (3) the sampling method, S, (4) the regression-based linkage method, R, and (5) the sample size, N.

The information contained in the two factors, the genetic model, M, and the minor allele frequency, P, appears to be represented by the variance component, Vaa. Hence in this section the variance component Vaa would be used in lieu of the factors M and P to model the mean of *t* values.

As we can see from the previous section, the factor L (type of markers studied) has highly significant main effect and many significant two-way interactions with other factors. Therefore we decided to fit separate regression lines for three types of markers evaluated in this paper.

We first conduct the ANCOVA (analysis of covariance) including all four factors and all of the possible interactions. The epistatic variance component and the sample size were treated as quantitative variables, the remaining two factors were considered as categorical variables. The results indicated that the four-way and most of the three-way interactions are not significant. In the case of the analysis based on bi-allele markers, the main effects of Vaa, N, and S, and two-way interactions Vaa*N and N*S are significant. For the multi-allelic markers, the main effects of Vaa and N, the two-way interactions Vaa*N and Vaa*S, and the three-way interaction Vaa*N*S are significant. For the QTLs, the main effects of Vaa and N, and the two-way interactions Vaa*N and N*S are significant. We noted that the epistatic variance component, the sample size, and their interaction are significant for all three types of markers evaluated. We also noticed that the regression-based linkage method and its higher order interactions are not a significant factor in all three cases. The significance level of 0.10 was used for the above conclusions. Detailed SAS output can be found in Appendix D.

Next we fit a general linear model with only significant main effects and interactions to the mean of *t* values for each type of markers. The regression coefficients are summarized in Table 4-1. The detailed SAS output can be found in Appendix E.

By using the fitted general linear models we can obtain the fitted values for the mean of *t* values. Then, if a significance level is specified, we can calculate two types of expected power, one obtained by using the observed mean *t* (denoted as Expected Power), another obtained by using the fitted mean *t* (denoted as Predicted Power). These two types of expected power then can be compared with the power observed from our simulation (denoted as Observed Power). Specifically, the observed power is obtained by counting, out of 1000 replications, the number of significant *t* values and then dividing by 1000.

Table 4-1 Estimates of Regression Parameters in the fitted General Linear Models of the Mean of *t* Values for Three Types of Markers[2]

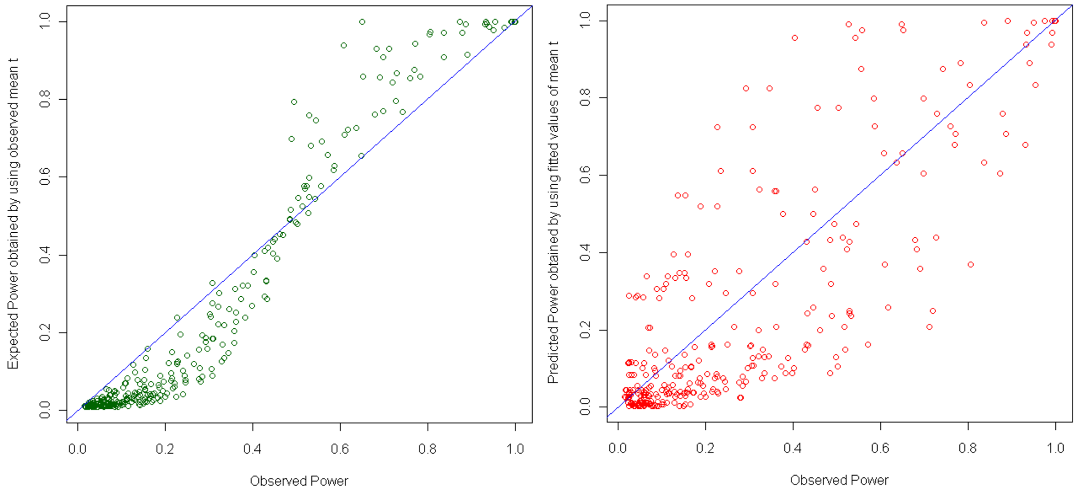
L	S	Intercept	Vaa	$N^{\dagger}$	$Vaa*N^{\dagger}$
t	d	0.1276	2.2297	0.2476	0.9650
t	t	0.1276	2.2297	0.4056	0.9650
t	r	0.1276	2.2297	-0.1479	0.9650
m	d	0.0316	0.6763	0.0402	0.2967
m	t	0.0316	0.5628	0.0402	0.3014
m	r	0.0316	0.0177	0.0402	0.0051
b	d	-0.0001	0.0407	0.0157	0.0265
b	t	0.0326	0.0407	0.0044	0.0265
b	r	-0.0168	0.0407	0.0078	0.0265

Notation: L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene); S: Sampling Method (S = d: disease selected sampling, t: trait truncated sampling, r: random sampling); Vaa: Epistatic Variance Component shown in proportion of the two-locus genetic variance (V_G) ($0 \leq Vaa \leq 1$); $N^{\dagger}$: Sample Size divided by 250 ($N^{\dagger} \equiv N/250 = 1, 2, 4$).

[2] To ensure that the three continuous variables, the mean of *t* values, the epistatic variance component, and the sample size, are in the same scale, we use the proportion, Vaa/V_G, ranging from 0 to 1, as the epistatic variance component value, and the sample size divided by250 as $N^{\dagger}$. This should not affect the significance of any factor.

For example, if we choose a significance level of 0.01 and use the general linear model for the trait loci only in predicting the power, we can acquire three columns of power values. The results are summarized in Figure 4-10.

Figure 4-10 Scatterplots of Observed Power vs Expected Power Obtained by using Mean of *t* Values (in green, on the left) and Observed Power vs Predicted Power Obtained by using Fitted Values of Mean of *t* Values (in red, on the right)



We see from Figure 4-10 that the green dots in the left plot lay more closely around the diagonal line than the red dots in the right plot do. By using the observed mean of *t* values we can obtain a very good estimate of the power value. The mean and standard error of the difference of the observed power minus the expected power are 3.78% and 0.52% respectively. If we use the fitted values from the estimated general linear model we would get a poorer estimate. The mean and standard error of the difference between the observed power and the predicted power are 3.48% and 1.00% respectively. However, considering that this model only has four factors and two two-way interactions, and that we only use one epistatic genetic variance component (Vaa) to summarize the information contained in the genetic models, this result is fairly reasonable and is within expectation.

We also noticed that the expected power values obtained by using the observed mean of *t* values are a little conservative when the observed power values are less than 50%, and the expected power values are a little liberal when the observed power values are greater than 50%.

4.5 Conclusions

The power of two regression-based linkage methods in detecting epistatic QTLs was evaluated in this dissertation. A two-locus epistatic and pleiotropic modeling framework was used in our simulation study. The two major QTL not only accounts for 10% of the total variance of a quantitative trait, but also increases the probability of developing a complex disease associated with the quantitative trait. We proposed a sampling method that selects families through the disease affected probands for QTL linkage analysis. Random sampling and another selected sampling based the extreme trait values were also included for comparison.

Three types of markers closely linked to the epistatic QTLs were considered: bi-allele markers, multi-allelic markers, and QTL genes themselves. We simulated a situation where there is no allelic association, or no linkage disequilibrium, between the trait and marker loci. We thus evaluated the power to identify the QTLs in epistasis (single-locus analysis) as well as establishing the epistasis itself.

The disease selected sampling method resulted in much higher power over random sampling, and, in some cases, even attained high power comparable to the trait truncated sampling, especially when the underlying epistasis is mostly due to the interaction of additive effects at the two QTLs.

CHAPTER 4. RESULTS

The results of this simulation study suggested that by using the bi-allelic or the multi-allelic markers one has very little or no power in detecting epistasis even though these markers are in complete linkage with the QTLs ($\theta = 0$). This may due to the fact that we simulated no allelic association (linkage disequilibrium) between the marker loci and the trait loci at all, or it may simply indicate low power for linkage analysis even to closely linked markers when genetic variance is only 10% of the total variance of the quantitative trait.

We also noticed that the single-locus analysis, in most of the situations considered in this paper, can detect epistatic genes even when the additive effects are very small.

Not surprisingly, a bigger sample size resulted in a greater power. A sample with 500 sibpairs would generally give acceptable power for epistasis detection, especially when selected sampling methods were adopted.

Although we expected the power of the two-locus interaction analysis to be affected by the total epistatic variance (V_I), we observed that the linkage methods are much more powerful in detecting epistasis due to the interaction of additive effects (*Vaa*) than in detecting epistasis due to the interaction of dominant effects (*Vdd*). Therefore we used only *Vaa* to summarize all the information in genetic models, and modeled the power with this epistatic variance component, together the sample size, the sampling method, and a few two-way interactions. The predicted power values had large variance, but gave unbiased estimates overall, compared with the observed power values.

Chapter 5 Future Work

We realize that there are some limitations to our study which could be the possible directions for future research on linkage analysis of epistatic QTLs.

First of all, the two regression-based linkage methods evaluated in this paper, the Haseman-Elston method and the Sham-Purcell method, can only be applied on sibpair data. To better analyze genetic information contained in larger sibship and/or general pedigree, it is important to also extend methods, such as those based on the score statistics, for epistasis models. In fact, there is some literature, for example, Tang and Siegmund (2002) and Wang (2003), addressing this issue. However, current studies had only focused on epistasis due to the interaction of additive effects (*Vaa*). It would be interesting to find methods that can better detect epistasis due to the interaction of dominant effects (*Vdd*).

Secondly, we only used one set of disease penetrance values in our simulation. Since the disease selected sampling gave satisfactory results in most of the cases considered in this paper, it would be worth investigating to try different disease patterns and see how this would affect linkage analysis for QTLs.

Thirdly, in this thesis, I fixed the ratio of two-locus genetic variance, shared residual variance, and non-shared residual variance to be 1:2:7, i.e. only 10% of the total variance can be explained by the two epistatic genes. This may account for the low power for the closely linked markers. Therefore a reasonable next step could be to examine other situations where different variance proportions are adopted.

Last but not least, we could also try to extend and evaluate the two regression-based linkage methods to incorporate gene-environment interaction.

References

[1] Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J (1999) Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait-loci-mapping procedure. *Am J Hum Genet* 62:1198-1211

[2] Almsay L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62:1198-1211

[3] Amos CI, de Andrade M (2001) Genetic linkage methods for quantitative traits. *Statistical Methods in Medical Research* 10:3-25

[4] Andrade M, Amos CI (2000) Ascertainment issues in variance component models. *Genet Epidemiol* 19:333-344

[5] Bhattacharjee S, Kuo CL, Mukhopadhyay N, Brock GN, Weeks DE, Feingold E (2008) Robust score statistics for QTL linkage analysis. *Am J Hum Genet* 82:567-582

[6] Carey G, Williams J (1991) Linkage analysis of quantitative traits: Increase power by using selected samples. *Am J Hum Genet* 49:786-796

[7] Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Molecul Genet* 11:2463-2468

[8] Cuenco KT, Szatkiewicz JP, Feingold E (2003) Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. *Am J Hum Genet* 73:863-873

[9] Culverhouse R, Klein T, Shannon W (2004) Detecting epistatic interactions contributing to quantitative traits. *Genetic Epidemiol* 27:141-152

[10] Drigalenko E (1998) How sib pairs reveal linkage. *Am J Hum Genet* 63:1242-1245

[11] Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. *Genet Epidemiol* 19:1-17

[12] Elston RC, Sobel E (1979) Sampling considerations in the gathering and analysis of pedigree data. *Am J Hum Genet* 31:62-69

[13] Elston RC, Song D, Iyengar SK (2005) Mathematical assumptions versus biological reality: myths in affected sib pair linkage analysis. *Am J Hum Genet* 76:152-156

REFERENCES

[14] Evans DM, Marchini J, Morris AP, Cardon LR (2006) Two-stage two-locus models in genome-wide association. *PLOS Genet.* 2:1424-1432

[15] Falconer DS (1981) Introduction to quantitative genetics (second edition). New York: Longman

[16] Feingold E (2002) Regression-based quantitative-trait-locus mapping in the 21st century. *Am J Hum Genet* 71:217-222

[17] Forrest W (2001) Weighting improves the “new Haseman-Elston” method. *Hum Hered* 52:47-54

[18] Galassi M, Davies J, Theiler J, Gough B, Jungman G, Alken P, Booth M, Rossie F (2009) GNU Scientific Library Reference Manual (3rd Ed) *ISBN 0954612078.*

[19] Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2:3-19

[20] Hopper JL, Mathews JD (1982) Extensions to multi-variate normal models for pedigree analysis. *Am J Hum Genet* 46:373-383

[21] Huang C, Ke L, Saint Fleur R, Chang SW, Choi SH, Shen T, Shin SY, Finch SJ, Mendell NR (2009) Family-based analysis of a myocardial infarction endophenotype: comparison of sampling designs. *BMC Proc* 3(Suppl 7):S120

[22] Huang Y, Bartlett CW, Segre AM, O’Connell JR, Mangin L, Vieland VJ (2007) Exploiting gene x gene interaction in linkage analysis. *BMC Proc* 3(Suppl 1):S64

[23] Huang Y, Bartlett CW, Segre AM, O’Connell JR, Mangin L, Vieland VJ (2007) Exploiting gene x gene interaction in linkage analysis. *BMC Proc* 1(Suppl 1):S64

[24] Kraja AT, Culverhouse R, Daw EW, Wu J, Brunt AV, Province MA, Borecki IB (2009) The Genetic Analysis Workshop 16 Problem 3: simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study. *BMC Proc* 3(Suppl 7):S4

[25] Li W, Reich J (2000) A complete enumeration and classification of two-locus disease models. *Hum Hered* 50:334-349

[26] Neuman RJ, Rice JP (1992) Two-locus models of disease. *Genet Epidemiol* 9:347-365

[27] Peng J, Siegmund D (2006) QTL mapping under ascertainment. *Ann Hum Genet* 70:867-881

REFERENCES

[28] Purcell S, Sham PC (2004) Epistasis in quantitative trait locus linkage analysis: interaction or main effect? *Behav Genet* 34:143-152

[29] Sham PC, Purcell S (2001) Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am J Hum Genet* 68:1527-1532

[30] Sham PC (1998) *Statistics in human genetics*. London: Arnold

[31] Sung H, Ji F, Levy DL, Matthysse S, Mendell NR (2009) The power of linkage analysis of a disease-related endophenotype using asymmetrically ascertained sib pairs. *Comp Stat Data Analy* 53:1829–1842

[32] Szatkiewicz JP, Feingold E (2005) QTL mapping with discordant and concordant sibling pairs: New statistics and new design strategies. *Genet Epidemiol* 28:326-340

[33] Tang HK, Siegmund D (2002) Mapping multiple genes for quantitative or complex traits. *Genet Epidemiol* 22:313-327

[34] Tiwari HK, Elston RC (1997) Linkage of multilocus components of variance to polymorphic markers. *Ann Hum Genet* 61:253-261

[35] 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

[36] Wan Y, Cohen J, Guerra R (1997) A permutation test for the robust sib-pair linkage method. *Ann. Hum. Genet.* 61:79-87

[37] Wang K (2003) Score tests for epistasis models on quantitative traits using general pedigree data. *Genet Epidemiol* 25:314-326

[38] Xu X, Weiss S, Xu X, Wei LJ (2000) A unified Haseman-Elston method for testing linkage with quantitative traits. *Am J Hum Genet* 67:1025-1028

[39] Zhang H, Risch N (1996) Mapping quantitative-trait loci in humans by use of extreme concordant sib pairs: selected sampling by parental phenotypes. *Am J Hum Genet* 59:951-957

Appendices

A. ANOVA Output from SAS

```
                              Class Level Information

                       Class         Levels    Values
                       M                  6    1 2 3 4 5 6
                       P                  3    0.05 0.1 0.15
                       N                  3    250 500 1000
                       R                  2    sp_est sp_tru
                       L                  3    b m t
                       A                  2    i s
                       S                  3    d r t

                       Number of Observations Used        1944

Dependent Variable: t_mean

                                         Sum of
 Source                      DF         Squares     Mean Square    F Value    Pr > F
 Model                       15     1721.277024      114.751802     124.86    <.0001
 Error                     1928     1771.861700        0.919015
 Corrected Total           1943     3493.138724

                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.492759      109.3696      0.958653       0.876526

 Source                      DF        Anova SS     Mean Square    F Value    Pr > F
 M                            5     272.2397130      54.4479426      59.25    <.0001
 P                            2       1.1542254       0.5771127       0.63    0.5338
 N                            2     148.1561597      74.0780798      80.61    <.0001
 R                            1       0.9989247       0.9989247       1.09    0.2973
 L                            2     814.7846471     407.3923235     443.29    <.0001
 A                            1     136.9403338     136.9403338     149.01    <.0001
 S                            2     347.0030203     173.5015102     188.79    <.0001
```

Notation: M: Genetic Model; P: Minor Allele Frequencies; N: Sample Size; R: Regression-Based Linkage Method (R = sp_est: Sham-Purcell method using estimated parameter values; sp_tru: SP method using true parameter values); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); A: Type of Regression Analysis (A = s: single-locus; i: two-locus interaction); S: Sampling Method (S = d: disease selected sampling, t: trait truncated sampling, r: random sampling).

B. Percentages of Variance Components Relative to the Total Two-Locus Genetic Variance (*Vg*) of Six Genetic Models

Model	*P*	V_A*(%)*	V_D*(%)*	V_I*(%)*	*Vaa(%)*	*Vad(%)*	*Vdd(%)*
1 (dominant-dominant)							
1 1 0	*0.05*	17.3121	0.4556	82.2324	78.0694	4.1089	0.0541
1 1 0	*0.10*	30.2521	1.6807	68.0672	61.0908	6.7879	0.1886
0 0 0	*0.15*	39.9217	3.5225	56.5558	47.7564	8.4276	0.3718
2 (recessive-dominant)							
1 1 0	*0.05*	1.1464	8.8073	90.0463	8.3559	79.6014	2.0890
0 0 0	*0.10*	4.1953	15.4620	80.3427	13.8389	63.0440	3.4597
0 0 0	*0.15*	8.6239	20.0681	71.0681	17.0363	49.7727	4.2591
3 (recessive-recessive)							
1 0 0	*0.05*	0.0475	0.4513	99.5013	0.9025	17.1476	81.4511
0 0 0	*0.10*	0.3600	1.6202	98.0198	3.2403	29.1629	65.6166
0 0 0	*0.15*	1.1481	3.2529	95.5990	6.5058	36.8662	52.2271
4 (one locus has modifying effect on the other)							
1 1 1	*0.05*	9.5787	82.5678	7.8535	0.6694	6.4382	0.7459
1 0 0	*0.10*	18.5051	68.9636	12.5313	1.7783	8.5024	2.2506
0 0 0	*0.15*	26.8902	57.8830	15.2268	2.6126	8.7620	3.8522
5 (at least 3 minor alleles in total at two loci)							
1 1 0	*0.05*	2.0052	8.0267	89.9681	16.0535	64.3898	9.5248
1 0 0	*0.10*	7.1193	12.7005	80.1802	25.4010	38.7607	16.0185
0 0 0	*0.15*	14.1824	14.8467	70.9710	29.6933	21.1860	20.0917
6 (recessive with no expression for $A_1A_1B_1B_1$)							
0 1 1	*0.05*	9.4761	90.0227	0.5013	0.0045	0.0864	0.4103
1 0 0	*0.10*	17.8146	80.1655	2.0200	0.0668	0.6010	1.3522
1 0 0	*0.15*	24.8867	70.5122	4.6011	0.3131	1.7744	2.5137

Notation: P: minor allele frequencies; V_A: Additive Variance; V_D: Dominant Variance; V_I: Two-Locus Epistatic Variance; *Vaa*: Variance due to interaction of additive effects at two loci; *Vad*: Variance due to interaction of additive effect at one locus and dominant effect at the other locus; *Vdd*: Variance due to interaction of dominant effects at two loci.

Note: all variance components are shown in percentage of the total two-locus genetic variance (V_G) which accounts for 10% of the total variance of the trait (V_T).

C. ANOVA output from SAS

Notation: P: Minor Allele Frequencies (P = 0.05; 0.10; 0.15); R: Regression-Based Linkage Method (R = he: original Haseman-Elston method; sp_est: Sham-Purcell method using estimated parameter values); L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); A: Type of Analysis (A = s: single-locus; i: two-locus interaction); S: Sampling Method (S = d: disease selected sampling, t: trait truncated sampling, r: random sampling); N: Sample Size (N = 250; 500; 1000).

```
------------------------------------- Model=1 -------------------------------------

                                The ANOVA Procedure

Dependent Variable: t_mean

                                              Sum of
 Source                         DF           Squares     Mean Square    F Value    Pr > F
 Model                          51       629.7384426      12.3478126      64.27    <.0001
 Error                         272        52.2599214       0.1921321
 Corrected Total               323       681.9983641


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.923372      37.58369      0.438329       1.166274


 Source                         DF          Anova SS     Mean Square    F Value    Pr > F

 P                               2        39.1512191      19.5756095     101.89    <.0001
 R                               1         5.6721561       5.6721561      29.52    <.0001
 L                               2       293.5144991     146.7572495     763.84    <.0001
 A                               1         0.0236740       0.0236740       0.12    0.7258
 S                               2        72.7768251      36.3884126     189.39    <.0001
 N                               2        36.2485664      18.1242832      94.33    <.0001
 P*R                             2         0.9914944       0.4957472       2.58    0.0776
 P*L                             4        57.9402383      14.4850596      75.39    <.0001
 P*A                             2         9.6761508       4.8380754      25.18    <.0001
 P*S                             4         8.0394200       2.0098550      10.46    <.0001
 P*N                             4         3.4922101       0.8730525       4.54    0.0014
 R*L                             2         3.5009620       1.7504810       9.11    0.0001
 R*A                             1         0.1438375       0.1438375       0.75    0.3877
 R*S                             2         1.4432963       0.7216482       3.76    0.0246
 R*N                             2         0.5821676       0.2910838       1.52    0.2217
 L*A                             2        21.9402610      10.9701305      57.10    <.0001
 L*S                             4        39.5333462       9.8833365      51.44    <.0001
 L*N                             4        23.7258060       5.9314515      30.87    <.0001
 A*S                             2         5.7363738       2.8681869      14.93    <.0001
 A*N                             2         0.0063167       0.0031584       0.02    0.9837
 S*N                             4         5.5996221       1.3999055       7.29    <.0001
```

APPENDICES

```
-------------------------------------- Model=2 --------------------------------------

                                 The ANOVA Procedure

Dependent Variable: t_mean

                                           Sum of
 Source                      DF           Squares     Mean Square    F Value    Pr > F
 Model                       51       305.1851106       5.9840218      88.70    <.0001
 Error                      272        18.3503987       0.0674647
 Corrected Total            323       323.5355093


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.943282      34.16859      0.259740       0.760171


 Source                      DF          Anova SS     Mean Square    F Value    Pr > F

 P                            2        24.2449975      12.1224988     179.69    <.0001
 R                            1         4.5963055       4.5963055      68.13    <.0001
 L                            2       111.9403402      55.9701701     829.62    <.0001
 A                            1         1.0157935       1.0157935      15.06    0.0001
 S                            2        54.7735979      27.3867990     405.94    <.0001
 N                            2        21.2706526      10.6353263     157.64    <.0001
 P*R                          2         1.7352473       0.8676237      12.86    <.0001
 P*L                          4         7.4113875       1.8528469      27.46    <.0001
 P*A                          2         4.2003528       2.1001764      31.13    <.0001
 P*S                          4         7.9070319       1.9767580      29.30    <.0001
 P*N                          4         1.1978710       0.2994677       4.44    0.0017
 R*L                          2         2.3116521       1.1558260      17.13    <.0001
 R*A                          1         0.2834418       0.2834418       4.20    0.0414
 R*S                          2         1.5684785       0.7842392      11.62    <.0001
 R*N                          2         1.2384895       0.6192448       9.18    0.0001
 L*A                          2         5.3712214       2.6856107      39.81    <.0001
 L*S                          4        33.8083036       8.4520759     125.28    <.0001
 L*N                          4        12.9977719       3.2494430      48.17    <.0001
 A*S                          2         1.9625775       0.9812888      14.55    <.0001
 A*N                          2         0.0503690       0.0251845       0.37    0.6888
 S*N                          4         5.2992275       1.3248069      19.64    <.0001
```

APPENDICES

```
-------------------------------------- Model=3 --------------------------------------

                                 The ANOVA Procedure

Dependent Variable: t_mean

                                         Sum of
 Source                      DF         Squares     Mean Square    F Value    Pr > F
 Model                       51     28.68788337      0.56250752       7.24    <.0001
 Error                      272     21.12060278      0.07764927
 Corrected Total            323     49.80848615


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.575964      224.4957      0.278656       0.124125


 Source                      DF        Anova SS     Mean Square    F Value    Pr > F

 P                            2      6.22269067      3.11134534      40.07    <.0001
 R                            1      0.25138836      0.25138836       3.24    0.0731
 L                            2      3.26214661      1.63107330      21.01    <.0001
 A                            1      0.38917423      0.38917423       5.01    0.0260
 S                            2      1.41925581      0.70962790       9.14    0.0001
 N                            2      2.28786797      1.14393398      14.73    <.0001
 P*R                          2      0.13482449      0.06741224       0.87    0.4209
 P*L                          4      4.53640084      1.13410021      14.61    <.0001
 P*A                          2      1.12967998      0.56483999       7.27    0.0008
 P*S                          4      2.94363131      0.73590783       9.48    <.0001
 P*N                          4      0.48022655      0.12005664       1.55    0.1891
 R*L                          2      0.27826736      0.13913368       1.79    0.1686
 R*A                          1      0.09265022      0.09265022       1.19    0.2757
 R*S                          2      0.09086940      0.04543470       0.59    0.5577
 R*N                          2      0.04609162      0.02304581       0.30    0.7434
 L*A                          2      0.73350893      0.36675447       4.72    0.0096
 L*S                          4      0.76422931      0.19105733       2.46    0.0457
 L*N                          4      1.92591889      0.48147972       6.20    <.0001
 A*S                          2      0.24013372      0.12006686       1.55    0.2149
 A*N                          2      0.44918059      0.22459029       2.89    0.0572
 S*N                          4      1.00974651      0.25243663       3.25    0.0126
```

APPENDICES

```
-------------------------------------- Model=4 --------------------------------------

                                The ANOVA Procedure

Dependent Variable: t_mean

                                            Sum of
 Source                      DF            Squares     Mean Square    F Value    Pr > F
 Model                       51        752.0137482      14.7453676      66.75    <.0001
 Error                      272         60.0873915       0.2209095
 Corrected Total            323        812.1011397


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.926010      44.62019      0.470010       1.053358


 Source                      DF           Anova SS     Mean Square    F Value    Pr > F

 P                            2          0.9903141       0.4951570       2.24    0.1083
 R                            1          7.3845988       7.3845988      33.43    <.0001
 L                            2        144.5275743      72.2637872     327.12    <.0001
 A                            1        240.5695903     240.5695903    1089.00    <.0001
 S                            2         90.8300631      45.4150316     205.58    <.0001
 N                            2         31.3445904      15.6722952      70.94    <.0001
 P*R                          2          0.1452661       0.0726331       0.33    0.7201
 P*L                          4          4.9972857       1.2493214       5.66    0.0002
 P*A                          2          1.4054487       0.7027243       3.18    0.0431
 P*S                          4          0.6492202       0.1623050       0.73    0.5690
 P*N                          4          0.1844448       0.0461112       0.21    0.9335
 R*L                          2          2.9937852       1.4968926       6.78    0.0013
 R*A                          1          5.8239483       5.8239483      26.36    <.0001
 R*S                          2          2.5961592       1.2980796       5.88    0.0032
 R*N                          2          0.9508288       0.4754144       2.15    0.1182
 L*A                          2         75.6426567      37.8213284     171.21    <.0001
 L*S                          4         36.2094687       9.0523672      40.98    <.0001
 L*N                          4         12.7931609       3.1982902      14.48    <.0001
 A*S                          2         64.4314350      32.2157175     145.83    <.0001
 A*N                          2         20.5460999      10.2730500      46.50    <.0001
 S*N                          4          6.9978091       1.7494523       7.92    <.0001
```

APPENDICES

```
------------------------------------- Model=5 -------------------------------------

                                The ANOVA Procedure

Dependent Variable: t_mean

                                             Sum of
 Source                          DF         Squares     Mean Square    F Value    Pr > F
 Model                           51     310.0553112       6.0795159      86.90    <.0001
 Error                          272      19.0287005       0.0699585
 Corrected Total                323     329.0840117


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.942177      33.51097      0.264497       0.789284


 Source                          DF        Anova SS     Mean Square    F Value    Pr > F

 P                                2       8.3860159       4.1930080      59.94    <.0001
 R                                1       5.0089345       5.0089345      71.60    <.0001
 L                                2     132.2340066      66.1170033     945.09    <.0001
 A                                1       2.5990093       2.5990093      37.15    <.0001
 S                                2      53.0230364      26.5115182     378.96    <.0001
 N                                2      21.1893163      10.5946582     151.44    <.0001
 P*R                              2       0.5476569       0.2738284       3.91    0.0211
 P*L                              4       2.1985356       0.5496339       7.86    <.0001
 P*A                              2       2.2239862       1.1119931      15.90    <.0001
 P*S                              4       2.6488987       0.6622247       9.47    <.0001
 P*N                              4       0.2899681       0.0724920       1.04    0.3889
 R*L                              2       2.9544369       1.4772184      21.12    <.0001
 R*A                              1       0.0025167       0.0025167       0.04    0.8497
 R*S                              2       1.5330993       0.7665497      10.96    <.0001
 R*N                              2       1.1716042       0.5858021       8.37    0.0003
 L*A                              2      18.4559325       9.2279662     131.91    <.0001
 L*S                              4      34.3460240       8.5865060     122.74    <.0001
 L*N                              4      14.2904506       3.5726127      51.07    <.0001
 A*S                              2       1.6470082       0.8235041      11.77    <.0001
 A*N                              2       0.3756208       0.1878104       2.68    0.0701
 S*N                              4       4.9292534       1.2323133      17.61    <.0001
```

APPENDICES

```
-------------------------------------- Model=6 --------------------------------------

                                 The ANOVA Procedure

Dependent Variable: t_mean

                                         Sum of
 Source                      DF         Squares     Mean Square    F Value     Pr > F
 Model                       51     273.4530398       5.3618243      60.71     <.0001
 Error                      272      24.0211509       0.0883131
 Corrected Total            323     297.4741907


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.919250      50.08308      0.297175       0.593364


 Source                      DF        Anova SS     Mean Square    F Value     Pr > F

 P                            2      0.68690468      0.34345234       3.89     0.0216
 R                            1      2.17810947      2.17810947      24.66     <.0001
 L                            2     43.55148909     21.77574454     246.57     <.0001
 A                            1     96.86983053     96.86983053    1096.89     <.0001
 S                            2     28.93735568     14.46867784     163.83     <.0001
 N                            2      9.81613701      4.90806850      55.58     <.0001
 P*R                          2      0.06665721      0.03332861       0.38     0.6860
 P*L                          4      2.05991031      0.51497758       5.83     0.0002
 P*A                          2      1.43855450      0.71927725       8.14     0.0004
 P*S                          4      0.39867342      0.09966835       1.13     0.3433
 P*N                          4      0.13849420      0.03462355       0.39     0.8143
 R*L                          2      0.87263446      0.43631723       4.94     0.0078
 R*A                          1      1.90541987      1.90541987      21.58     <.0001
 R*S                          2      0.83284620      0.41642310       4.72     0.0097
 R*N                          2      0.27013557      0.13506779       1.53     0.2185
 L*A                          2     34.40523501     17.20261751     194.79     <.0001
 L*S                          4     12.23575058      3.05893764      34.64     <.0001
 L*N                          4      3.78879081      0.94719770      10.73     <.0001
 A*S                          2     22.68602198     11.34301099     128.44     <.0001
 A*N                          2      8.09282697      4.04641349      45.82     <.0001
 S*N                          4      2.22126227      0.55531557       6.29     <.0001
```

D. ANCOVA output from SAS

Notation: L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); N: Sample Size (N = 250; 500; 1000); S: Sampling Method (S = d: disease selected sampling, t: trait truncated sampling, r: random sampling); R: Regression-Based Linkage Method (R = he: original Haseman-Elston method; sp_est: Sham-Purcell method using estimated parameter values); Vaa: Epistatic Variance due to the interaction of the additive effects at the two trait loci, shown in percentage relative to the total two-locus genetic variance (V_G).

```
------------------------------------ Marker=b -------------------------------------

                                    The GLM Procedure

Dependent Variable: t_mean

                                                   Sum of
 Source                          DF          Squares     Mean Square    F Value    Pr > F
 Model                           23       0.45146719      0.01962901      10.35    <.0001
 Error                          282       0.53482325      0.00189654
 Corrected Total                305       0.98629045


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.457743      95.14335      0.043549       0.045772


 Source                          DF      Type III SS     Mean Square    F Value    Pr > F

 Vaa                              1       0.00581712      0.00581712       3.07    0.0810
 N                                1       0.02488054      0.02488054      13.12    0.0003
 Vaa*N                            1       0.01717461      0.01717461       9.06    0.0029
 S                                2       0.01470424      0.00735212       3.88    0.0218
 Vaa*S                            2       0.00226695      0.00113348       0.60    0.5508
 N*S                              2       0.00955894      0.00477947       2.52    0.0823
 Vaa*N*S                          2       0.00333360      0.00166680       0.88    0.4164
 R                                1       0.00151160      0.00151160       0.80    0.3727
 Vaa*R                            1       0.00117784      0.00117784       0.62    0.4313
 N*R                              1       0.00268514      0.00268514       1.42    0.2351
 Vaa*N*R                          1       0.00000006      0.00000006       0.00    0.9957
 S*R                              2       0.00392691      0.00196345       1.04    0.3565
 Vaa*S*R                          2       0.00069859      0.00034929       0.18    0.8319
 N*S*R                            2       0.00121222      0.00060611       0.32    0.7267
 Vaa*N*S*R                        2       0.00020506      0.00010253       0.05    0.9474
```

APPENDICES

```
------------------------------------ Marker=m -------------------------------------

                                 The GLM Procedure

Dependent Variable: t_mean

                                            Sum of
 Source                          DF        Squares     Mean Square    F Value    Pr > F
 Model                           23    29.95648644      1.30245593      52.51    <.0001
 Error                          282     6.99417151      0.02480203
 Corrected Total                305    36.95065795


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.810716      53.73368      0.157487       0.293087


 Source                          DF    Type III SS     Mean Square    F Value    Pr > F

 Vaa                              1     0.68630908      0.68630908      27.67    <.0001
 N                                1     0.46537980      0.46537980      18.76    <.0001
 Vaa*N                            1     0.99209359      0.99209359      40.00    <.0001
 S                                2     0.07241106      0.03620553       1.46    0.2340
 Vaa*S                            2     0.16023691      0.08011846       3.23    0.0410
 N*S                              2     0.09157168      0.04578584       1.85    0.1598
 Vaa*N*S                          2     0.26215461      0.13107730       5.28    0.0056
 R                                1     0.00004553      0.00004553       0.00    0.9659
 Vaa*R                            1     0.00672379      0.00672379       0.27    0.6030
 N*R                              1     0.02118832      0.02118832       0.85    0.3561
 Vaa*N*R                          1     0.00694712      0.00694712       0.28    0.5971
 S*R                              2     0.00456470      0.00228235       0.09    0.9121
 Vaa*S*R                          2     0.00181563      0.00090781       0.04    0.9641
 N*S*R                            2     0.00756687      0.00378343       0.15    0.8586
 Vaa*N*S*R                        2     0.00184382      0.00092191       0.04    0.9635
```

APPENDICES

```
------------------------------------ Marker=t -------------------------------------

                                  The GLM Procedure

Dependent Variable: t_mean

                                           Sum of
 Source                      DF          Squares     Mean Square    F Value    Pr > F
 Model                       23      574.6120481      24.9831325      42.01    <.0001
 Error                      282      167.6876983       0.5946372
 Corrected Total            305      742.2997463


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.774097      57.36114      0.771127       1.344337


 Source                      DF      Type III SS     Mean Square    F Value    Pr > F

 Vaa                          1      17.42591647     17.42591647      29.31    <.0001
 N                            1       8.15551348      8.15551348      13.72    0.0003
 Vaa*N                        1      22.84752885     22.84752885      38.42    <.0001
 S                            2       0.42099759      0.21049879       0.35    0.7022
 Vaa*S                        2       2.54935313      1.27467657       2.14    0.1191
 N*S                          2       3.20823828      1.60411914       2.70    0.0691
 Vaa*N*S                      2       2.45749875      1.22874937       2.07    0.1286
 R                            1       0.07340331      0.07340331       0.12    0.7256
 Vaa*R                        1       0.23020270      0.23020270       0.39    0.5343
 N*R                          1       0.63311351      0.63311351       1.06    0.3030
 Vaa*N*R                      1       0.20169922      0.20169922       0.34    0.5608
 S*R                          2       0.14676943      0.07338472       0.12    0.8839
 Vaa*S*R                      2       0.07448037      0.03724019       0.06    0.9393
 N*S*R                        2       0.31928953      0.15964476       0.27    0.7647
 Vaa*N*S*R                    2       0.02427037      0.01213519       0.02    0.9798
```

E. Fitted General Linear Models

Notation: L: Type of Markers (L = b: bi-allelic; m: multi-allelic; t: trait gene itself); S: Sampling Method (S = d: disease selected sampling, t: trait truncated sampling, r: random sampling); $N^{\dagger}$: Sample Size divided by 250 ($N^{\dagger} \equiv N/250 = 1; 2; 4$); Vaa: Epistatic Variance due to the interaction of the additive effects at the two trait loci, shown in proportion of the total two-locus genetic variance (V_G) ($0 \leq \text{Vaa} \leq 1$).

```
------------------------------------ Marker=b -------------------------------------

                                  The GLM Procedure

Dependent Variable: t_mean

                                             Sum of
 Source                      DF         Squares     Mean Square    F Value    Pr > F
 Model                        7      0.39051280      0.05578754      27.90    <.0001
 Error                      298      0.59577765      0.00199925
 Corrected Total            305      0.98629045


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.395941      97.68589      0.044713       0.045772


 Source                      DF     Type III SS     Mean Square    F Value    Pr > F

 Vaa                          1      0.00581712      0.00581712       2.91    0.0891
 N†                           1      0.02488054      0.02488054      12.44    0.0005
 S                            2      0.02859395      0.01429698       7.15    0.0009
 Vaa*N†                       1      0.01717461      0.01717461       8.59    0.0036
 N† *S                        2      0.01063751      0.00531876       2.66    0.0716


                                                   Standard
        Parameter           Estimate                  Error    t Value    Pr > |t|

        Intercept       -.0167904228 B           0.01036695      -1.62      0.1064
        Vaa             0.0407374277             0.02388216       1.71      0.0891
        N†              0.0078103719 B           0.00391834       1.99      0.0471
        S         d     0.0166921429 B           0.01328175       1.26      0.2098
        S         t     0.0493738198 B           0.01328175       3.72      0.0002
        S         r     0.0000000000 B            .                .           .
        Vaa*N†          0.0264565951             0.00902661       2.93      0.0036
        N† *S     d     0.0078815710 B           0.00502003       1.57      0.1175
        N† *S     t     -.0034060063 B           0.00502003      -0.68      0.4980
        N† *S     r     0.0000000000 B            .                .           .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was
      used to solve the normal equations.  Terms whose estimates are followed by the
      letter 'B' are not uniquely estimable.
```

APPENDICES

```
------------------------------------- Marker=m --------------------------------------

                                  The GLM Procedure

Dependent Variable: t_mean

                                        Sum of
 Source                      DF        Squares     Mean Square    F Value    Pr > F
 Model                        7    27.92274307      3.98896330     131.67    <.0001
 Error                      298     9.02791488      0.03029502
 Corrected Total            305    36.95065795


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.755676      59.38661      0.174055       0.293087


 Source                      DF    Type III SS     Mean Square    F Value    Pr > F

 Vaa                          1     0.68630908      0.68630908      22.65    <.0001
 N†                           1     0.46537980      0.46537980      15.36    0.0001
 Vaa*N†                       1     0.99209359      0.99209359      32.75    <.0001
 Vaa*S                        2     0.53613479      0.26806740       8.85    0.0002
 Vaa*N† *S                    2     0.78001536      0.39000768      12.87    <.0001


                                               Standard
        Parameter           Estimate              Error    t Value    Pr > |t|

        Intercept       0.0316358664         0.02715762       1.16      0.2450
        Vaa             0.0176537786 B       0.13814533       0.13      0.8984
        N†              0.0402309993         0.01026461       3.92      0.0001
        Vaa*N†          0.0051246663 B       0.05221403       0.10      0.9219
        Vaa*S      d    0.6940417965 B       0.17698655       3.92      0.0001
        Vaa*S      t    0.5804546113 B       0.17698655       3.28      0.0012
        Vaa*S      r    0.0000000000 B        .                .         .
        Vaa*N† *S   d   0.2915838293 B       0.06689463       4.36      <.0001
        Vaa*N† *S   t   0.2962793991 B       0.06689463       4.43      <.0001
        Vaa*N† *S   r   0.0000000000 B        .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was
      used to solve the normal equations.  Terms whose estimates are followed by the
      letter 'B' are not uniquely estimable.
```

APPENDICES

```
------------------------------------- Marker=t --------------------------------------

                                 The GLM Procedure

Dependent Variable: t_mean

                                          Sum of
 Source                       DF         Squares     Mean Square    F Value    Pr > F
 Model                         5     512.6020843     102.5204169     133.90    <.0001
 Error                       300     229.6976620       0.7656589
 Corrected Total             305     742.2997463


                 R-Square     Coeff Var      Root MSE    t_mean Mean
                 0.690559      65.08926      0.875019       1.344337


 Source                       DF     Type III SS     Mean Square    F Value    Pr > F

 Vaa                           1      17.4259165      17.4259165      22.76    <.0001
 N†                            1       8.1555135       8.1555135      10.65    0.0012
 Vaa*N†                        1      22.8475289      22.8475289      29.84    <.0001
 N† *S                         2     116.0802980      58.0401490      75.80    <.0001


                                                 Standard
        Parameter           Estimate                Error    t Value    Pr > |t|

        Intercept        0.127647707           0.13652863       0.93      0.3506
        Vaa              2.229653578           0.46736613       4.77      <.0001
        N†              -0.147909726 B         0.05811857      -2.54      0.0114
        Vaa*N†           0.964961920           0.17664779       5.46      <.0001
        N† *S      d     0.395479562 B         0.04631094       8.54      <.0001
        N† *S      t     0.553496874 B         0.04631094      11.95      <.0001
        N† *S      r     0.000000000 B          .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was
      used to solve the normal equations.  Terms whose estimates are followed by the
      letter 'B' are not uniquely estimable.
```