Stony Brook University

The Power

of Linkage Analysis of a Quantitative Endophenotype

A Dissertation Presented

by

Zhuying 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

May 2008

Stony Brook University

The Graduate School

Zhuying 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 & Statistics

Stephen J. Finch, Chairperson of Defense
Professor, Applied Mathematics & Statistics

Wei Zhu
Professor, Applied Mathematics & Statistics

Deborah L. Levy
Associate Professor, Department of Psychiatry
Harvard Medical School

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

The Power of Linkage Analysis of a Quantitative Endophenotype

by

Zhuying Huang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2008

In our study, we consider a complex disease with low frequency for which we are studying a disease related trait (endophenotype) instead of the disease itself. We assume that there is a pleiotropic gene that determines both the disease and disease related trait. First we developed software to simulate a quantitative endophenotype and linked markers in sib-pairs ascertained through a single disease affected proband. We simulated data sets of 100 sib-pairs under 210 genetic parameter values for disease allele frequency, disease penetrance, quantitative trait heritability with variable dominance and evaluated the power of the Haseman-Elston test for linkage. We then evaluated these results to estimate the effects of the parameter values and their interactions on power.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Introduction of Complex Disease

A disease is called complex if it is caused by multiple genes and environmental factors and their interactions. Examples of complex disease are schizophrenia and bipolar disease, for most complex diseases, there is no known direct relationship between genotype at a single locus and phenotype. In most cases, individuals who have the disease associated allele are more likely to have the disease than those who do not. However, individuals who do not have the associated allele have a non-zero probability the disease as well. However, it is sometimes possible to find major genes that determine a trait related to the disease. Since complex diseases are usually affected by many genes, they may have many different phenotypes associated with them too. Therefore, we need carefully choose the endophenotype/disease related trait since different traits may result in different outcomes.

1.2 Endophenotype

1.2.1 Definition

For most complex diseases, the same genotype may affect a large range of the phenotype and the same phenotype may be determined by many genotypes, environment factors and their interaction. However, it's possible to find the gene(s) if they have a major effect on some less complex related traits.

An endophenotype (EndoP), also called a disease related trait (DRT), is a trait which is associated with the disease as a result of being determined by a factor also involved in the disease. Researchers can get a better understanding of the biological components of a disease by identifying the endophenotype; also we can get a more accurate prediction and more effective prevention through an endophenotype.

For most complex disease genetic analysis, endophenotypes are biological markers between genotype and external phenotype that may indicate disease susceptibility loci. Researchers are interested in endophenotypes because the endophenotype might have simpler etiology than disease itself. A good measurable endophenotype might be more biological than a clinical diagnosis and more directly tied to the gene expression, it has higher penetrance than the disease itself. Therefore, the endophenotype may play an important role in studying the complex disease.

1.2.2 Criteria for an Endophenotype

Due to the complexity of the transmission of complex diseases, more and more researchers are paying attention to endophenotypes instead of disease itself. Since for complex disease, they are affected by many genes and have various phenotypes, how to choose the endophenotype becomes a big issue. Normally, endophenotype should meet the following criteria (Gottesman and Gould 2003):

"(1). It is associated with illness in the population.
(2). It is heritable.
(3). It is primary state-independent (manifests in an individual whether or not illness is active).
(4). within families, endophenotype and illness co-segregate.
(5). the endophenotype found in affected family members is found in non-affected family members at a higher rate than in the general population. "

We define an endophenotype as an abnormality if it appears more frequently in cases (diseased individuals) than controls and it has a higher frequency in unaffected siblings of cases than in controls.

There are many different explanations of the relationship between the disease and endophenotype. One popular explanation is that the traits are determined by a pleiotropic gene, a gene which controls more than one trait. For instance, a single gene mutation may cause an enzyme deficiency, which in turn may affect more than one tissue in one individual. Or pleiotropic effects may cause both a disease and an endophenotype, in turn the disease/DRT will or will not affect endophenotype/disease level. In our model, we assume that there is a pleiotropic gene which effect more than one trait.

1.2.3 Applications of Endophenotype

Recently, researchers have had many discussions on endophenotypes. Of interest are the possibilities of prevention and implications for the prediction of schizophrenia (Gottesman and Gould 2003). The endophenotypes of schizophrenia include eye tracking dysfunction (Levy et al.1993), thought disorder (Holzman et al. 1997) and working memory (Goldberg TE and Green MF, 2002). Also researchers have worked on language deficit in studying autism (Alarcon et al. 2002) and plasma cholesterol levels in studying coronary heart disease (Sing and Boerwinkle, 1987).

Even through an endophenotype has higher penetrance than the disease itself, an individual with the disease of interest may not always have the endophenotype. Conversely, having the endophenotype does not mean that one has disease either. That's why we sometimes call the endophenotype a risk factor. For example, people would like to consider hypertension and/or cholesterol level as an endophenotype in studying the genetics of coronary heart disease. However, sometimes the endophenotype is benign and unnoticeable.

1.3 Quantitative Genetics

1.3.1 Introduction

Quantitative genetics, founded by R.A. Fisher, is the study of continuous traits such as height or blood pressure. Using quantitative genetic analysis, one can predict the response to selection given data on the phenotype and relationships of individuals based on combined effect of the many underlying genes results in a continuous distribution of phenotypic values. In other words, the variation is quantitative, not qualitative.

Analysis of quantitative trait loci or QTL is a more recent addition to the study of quantitative genetics. A QTL is a region in the genome that affects the trait or traits of interest. OTL approach requires accurate phenotypic, pedigree and genotypic data from a large number of individuals.

Quantitative genetics can be applied to all traits determined by many genes, is not limited to continuous traits. The traits are: 1) Continuous traits are quantitative traits with a continuous phenotypic range. They are often polygenic, and may also be influenced significantly by environmental effects. 2)Traits or other ordinal numbers are expressed in whole numbers, such as number of offspring, or number of bristles on a fruit fly. These traits can be either treated as approximately continuous traits or as threshold traits; 3) Some qualitative traits can be treated as if they have an underlying quantitative basis, expressed as a threshold trait (or multiple thresholds). Some human diseases (such as, schizophrenia) have been studied in this manner.

1.3.2 Quantitative Mean and Variance

We know that by the gene frequency and genotype frequency we can express the genetic properties of the population. But it not enough for quantitative trait, in quantitative genetics, we need to have a new concept to show how the character is measured. All observations, including mean, variance and covariance, must clearly be based on the measurement of phenotypic value.

The phenotype is mainly determined by the genotype and the environment. We may think that the genotype having a certain value on the individual and the environment causing a deviation from this. We assume the mean of the environment deviation in the population is 0, and then the mean of the phenotypic value is equal to the genotypic values. Also for simplicity we assume that the environmental effect remains constant from generation to generation, so the population mean is constant in the process where no genetic changes. Assume a two allele trait locus with allele frequencies p and q=1-p, and

$$\begin{cases} g_i = -a \text{ if genotype is AA} \\ g_i = d \text{ if genotype is AB} \\ g_i = -a \text{ if genotype is BB} \end{cases}$$

We can calculate the population mean. Table 1.1 showed the population mean.

Table 1.1: Population Mean

Genotype	Freq	Value	Freq*Value
AA	p^2	+a	p^2a
AB	2pq	d	2pqd
BB	q^2	-a	$-q^2a$
Mean= $a(p^2-q^2)+2dpq$			

1.3.3 Quantitative Trait Heritability

The heritability of a trait is defined as the proportion of the total variance which is genetic. Let V_P denote is the variance of total phenotype, V_G the genetic variance, and V_E the environment variance, V_A the variance due to additive genetic effects and V_D the dominant effects. Then we have the following equation:

$$V_P = V_G + V_E = V_A + V_D + V_E$$

$$Heritability(Broad) = \frac{V_G}{V_p}$$

$$Heritability(narrow) = \frac{V_A}{V_p}$$

We should realize that the term heritability of trait is different from the mode of the inheritance. The mode of inheritance is a fixed property of a trait, for example, autosomal dominance, polygenic etc. But the heritability of trait may change. For example, in different social circumstances, the heritability of IQ will be not the same.

1.4 Pleiotropic Model

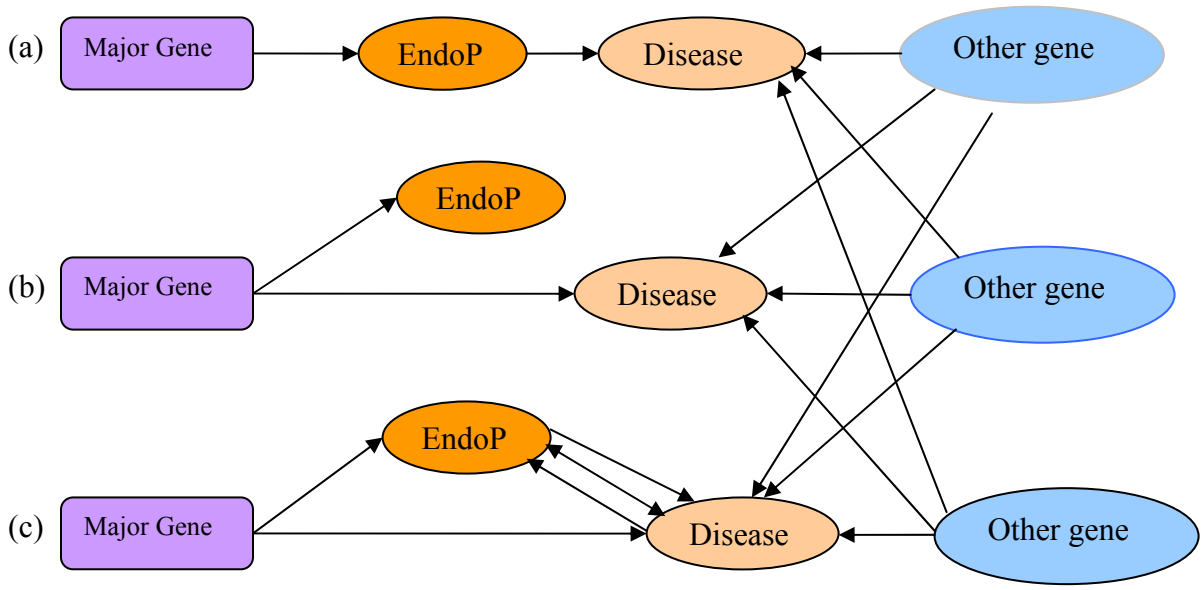
The term pleiotropy is from Greek pleio, which means “many” and trepein, which means “influencing”. Pleiotropy is a property of a gene which has more than one phenotypic effect. It describes the genetic effects of a single gene on multiple pleiotropic traits. The underlying mechanism is that the gene codes for a product that is for example used by various cells, or has a signaling function on various targets (Williams, G.C. 1957). For example, a single gene mutation may cause an enzyme deficiency. For instance, human disease PKU (phenylketonuria) can cause many other characteristics such as mental retardation and reduced hair and skin pigmentation. It can be caused by any of a large number of mutations in a single gene. This single gene is codes for an enzyme. It may change one amino acid to another amino acid. Then change the concentrations of the particular amino acid concentration. Then increase the toxic levels which will cause damage at several locations in a body. In genetics, we call a single gene a pleiotropic gene when the single gene influences multiple phenotypic traits and this genetic model is called pleiotropic model.

Of course, it could come out the same effect as a pleiotropic gene if several genes are clustered tightly in a single region. It is very important to distinguish between the pleiotropic effects of a single locus influencing all traits and separate tightly clustered loci that each influences a single trait (Almasy et al. 1997).

Pleiotropy may cause the correlation of two or more traits. Two or more characteristics may be affected by the same gene if the degree of the correlation arising from pleiotropy. Some of the genes may tend to increase the values for all the traits while other genes may increase one and reduce others.

The following Figure 1.1 shows various possible relationships among a major gene, endophenotype (EndoP) and a complex disease. In case a, the endophenotype causes the disease. This case is very simple, if we want to know if there is a disease, we only need to check if there is an endophenotype. The other two cases are a bit more complicated, the major gene in b and c has the pleiotropic effect which causes both the endophenotype and the disease at the same time. In these two cases, even we know there is endophenotype; we still can not 100% sure if there is a disease. In our study, we are focusing on case b where the pleiotropic effects may cause both the endophenotype and a disease, but there is no interaction between the disease and the endophenotype.

Figure 1.1: The relationship between gene and disease (Based on Sung, 2005)



1.5 Types of Phenotype

The phenotype describes physical appearance or a specific manifestation of a trait, such as height, sex, or behavior that varies between individuals. The difference between phenotype and genotype was proposed by Wilhelm Johannsen in 1911. The phenotype is composed of traits or chararcteristics. Some phenotypes are derterminded totally by the individual's genotypes. For example, people's blood type is determined when he is born. Others are determined by genes and significantly affected by enviroment. For example, almost all people have the ability to speak and understand language, but to learn and speak a particular language is significantly affected by the evironment. There are two types of phenotypes. One is a qualitative phenotype (discrete traits) and the other one is the quantitative phenotype (continuous traits). Very often several genotypes will result in the same phenotype and conversely one genotype can result in more than one phenotype.

1.5.1 Qualitative Phenotype

This is also referred to as a nominal trait. Outcomes of qualitative phenotype are descriptions such as color, pattern, sex and blood type. Many qualitative phenotypes are mainly determined by genotypes and usually are less affected by environmental factors than the quantitative phenotypes. For instance, the individual's blood type is determined when he is born. It will not be changed by the environment. Qualitative traits usually have two or more outcomes. For example, if father's blood type is A and mother's blood type is A, then their children's blood type is A or O. We can predict the offspring's phenotype from the parents' genotype and we also can predict the genotype of the parents if we know offspring's genotype.

In most cases, we treat diseases as dichotomous traits, i.e. there are two phenotypic classes – affected or unaffected with the disease. In the case of one locus with 2 alleles and 3 genotypes, then there are at least two genotypes that share the same phenotype. If the disease gene A is dominant, then people who have genotype Aa or AA are affected. If the disease gene B is recessive, then only people who have genotype BB are affected.

1.5.2 Quantitative Phenotype

Quantitative traits are a bit more complicated and are more likely to be affected by the environment than qualitative traits. Quantitative traits do not fall into discrete class. These traits could have any values within a range and they can be measured, i.e. weight, height and IQ. When we analyze a population, we will find a continuous distribution of phenotype. For example, blood pressure, or cholesterol level are phenotypes which can be measured and usually represented by a continuous distribution.

If the trait values have continuous distribution in the population and are controlled by multiple genes with small, equal and additive effects, then we may have one distribution in the population like the following Figure 1.2.

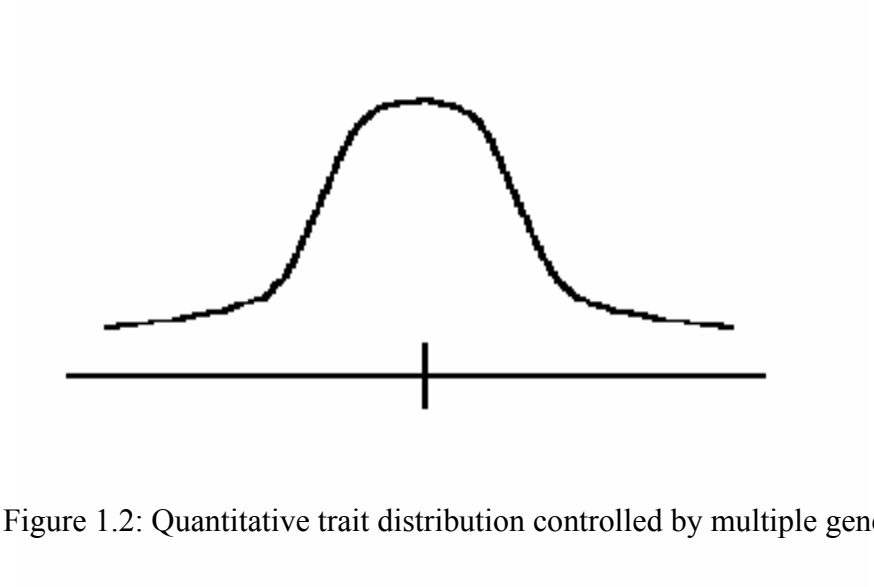


Figure 1.2: Quantitative trait distribution controlled by multiple genes

If the trait values are determined principally by the genotype at a single locus, then we called such a locus a quantitative trait locus (QTL). We usually assume that the trait distribution conditional on genotype is a normal distribution and would observe a mixture of normal distribution in the population. Examples are thought disorder scores, eye tracking disorder measures and blood pressure. The genetic model for QTL determines the distribution of the quantitative trait. In our study, we denote that there is a major gene which is related to a disease. We denote that the allele for abnormal level of the trait is A and for normal level is B, so there are 3 types of genotypes at the QTL gene locus, AA, AB and BB. We assume the quantitative phenotype X follows a normal distribution with a variance and means being genotype dependent, i.e. f (X|AA) ~ N (μ_2, σ_2^2), f (X | AB) ~ N (μ_1, σ_1^2) and f (X|BB) ~ N (μ_0, σ_0^2). The phenotype distribution for each genotype and the whole population is illustrated in the following Figure 1.3. In Figure 1.3, we set $\mu_0=0$, $\mu_1=1$, $\mu_2=4$ and $\sigma_0=\sigma_1=\sigma_2=\sigma$, so the population phenotype distribution is a mixture of normal distributions with unequal means and equal within group variance.

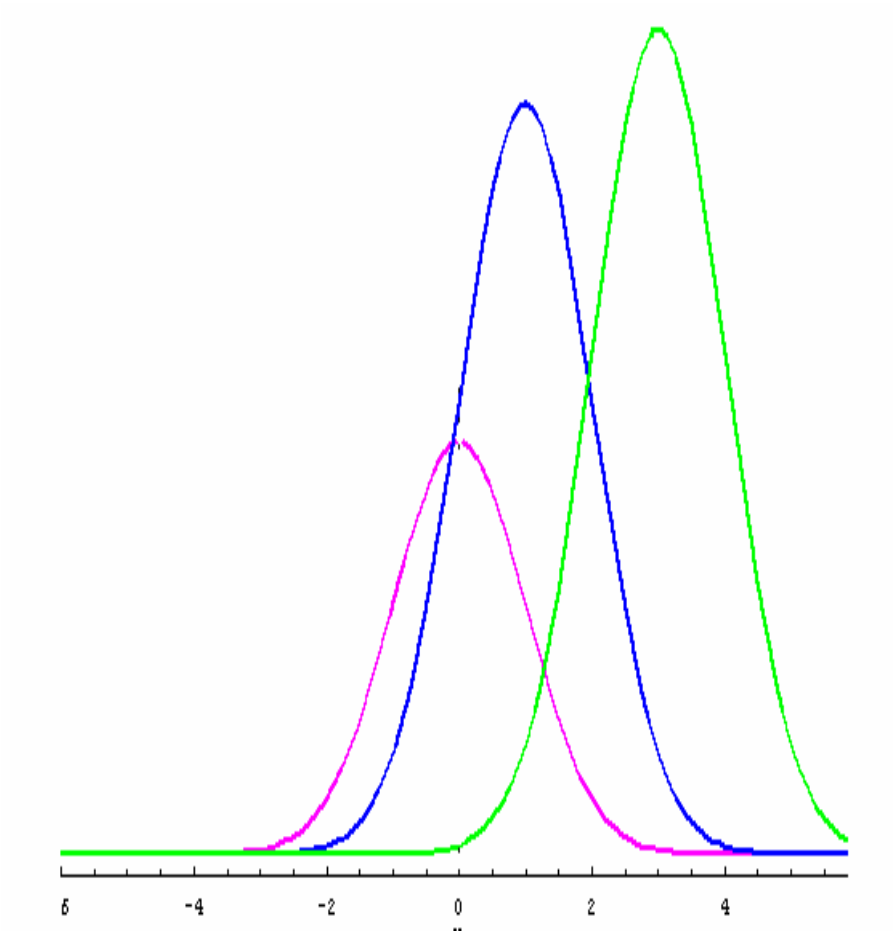


Figure 1.3: Graph of f (X^*|Genotype), $x^* = \frac{x - u_0}{\sigma}$

1.6 Methods for Genetic Linkage Analysis

We know that one way to find genetic control for a trait is to show it is linked to a known marker; that is, the two traits tend to be inherited together more often than it would be inherited by alone. There are two general methods that are commonly used in genetic linkage analysis: one is the classic model-based method and the other one is model-free methods. There are two important terminologies that are often used in linkage analysis. One is recombination fraction and the other one is identity by descent (IBD) number.

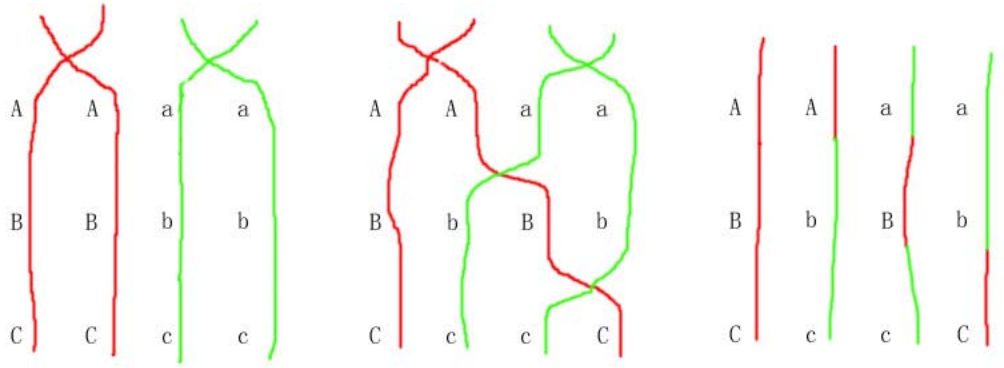
1.6.1 Terminology

1.6.1.1 Recombination Fraction

Genes are transmitted from parents to offspring through meiosis. Genetic recombination is commonly happened in sexual reproduction during in the process of meiosis. In the process separation, the two chromosomes cross over at different points. By crossing over, the alleles of gene are exchanged between the homologous chromosomes. Then the offspring can get set of genes which are different from either parent's, being a combination of genes from both parents. We could be able to see these cross over directly if we know the cross over directly, but as marker are typically available only at discrete intervals, all we can tell is whether two consecutive markers derive from the same parental chromosome. Since the recombination can be happened with small probability at any location in the chromosome, the frequency of recombination between two locations dependents on their distance.

In Figure 1.4, we illustrated the recombination and nonrecombination resulting from 1 and 2 crossovers occurring between 3 loci respectively.

Figure 1.4: Example of Cross-over



The proportion for recombination and nonrecombination is expected to be equal (1/2 and ½) when two loci are inherited independently (θ=1/2). In the case when the

recombination fraction is less than ½, the ratio of recombinant to no-recombinant will differ from 1:1. Also the recombination fraction is related to the genetic distance. The bigger in genetic distance, the more chance there is a recombination.

In many cases, the linkage analysis recombination fraction is sex dependent, which means recombination may different in male and female.

1.6.1.2 IBD Number

Allele sharing is an important concept in non-parametric linkage analysis. Identity by state (IBS) and identity by descent (IBD) are two different measurements of allele sharing. IBS are two alleles which have same form. If two alleles not only have same form, but they are both copies of the same gene in a common ancestor called IBD. IBD is a more important tool in genetic analysis. Figure 1.5 shows allele sharing in a nuclear family.

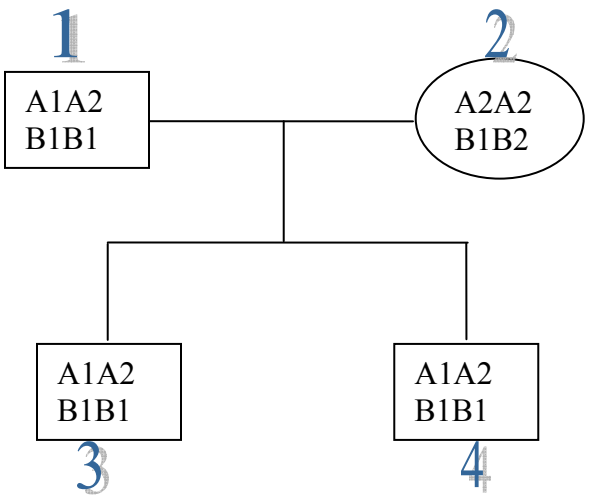


Figure 1.5: Example of allele transmission in a nuclear family
(Sham, P. 1998. Statistic in Human Genetics)

Here, we consider the nuclear family with two loci A and B, which are linked tightly (the theta is close to 0), then there is no recombination between A and B. Studying on locus A, from the above we can see that the two siblings (individual 3 and 4) have the same genotype at locus A as A1A2. From the definition, the number of shared IBS is 2 at locus A since both of them have the same allele states. But to get the number

of shared IBD is a bit more complicated than IBS. The two A1 alleles in sibling's are transmitted from father's (individual 1) A1 allele since there is no A1 allele in mother's (individual 2) alleles. So these two A1 alleles (one for each sibling) are IBD. Therefore the A2 alleles for both siblings should be from mother (individual 2). Since the mother has two A2 alleles, we do not know if the two A2 alleles in sibling are from same allele. Also, the mother transmitted B1 to sibling 3 and transmitted B2 to sibling 4. So if there is a recombination between A and B, the two A2 alleles of siblings are IBD. But we assume that the loci A and B are linked tightly, a recombination event can rare happen. So in this case, the number of shared IBD is 1, only A1 allele at locus A is IBD.

From above, we can see that there is a close relationship between the recombination and IBD status at two genetic loci in linkage analysis. The concept of IBD is important in genetic linkage analysis because if we know that an allele is shared IBD, then we can also pretty sure that a small region around this allele was transmitted IBD to both relatives as all because recombination in close region is rare and the region has same ancestral origin. So we can use this information to find out whether the unobserved disease gene is located in the vicinity of the marker locus. If we only know that an allele is shared IBS, then we can not sure that surrounding region has similar structure in both relatives. Since it's possible that the allele is shared IBS while the adjacent genetic sequence was inherited from different founders. So IBD is more relevant than IBS in linkage analysis.

1.6.2 Model Based Genetic Linkage Analysis

Model-based methods often are referred to as parametric methods. The classic LOD score method is a model-based method under the assumption of a disease inheritance model which specifies a gene frequency and the relationship between the disease genotype and phenotype distribution. The concept of LOD score was introduced by Morton in 1955. The term LOD score is simply refers to the logarithm of the likelihood ratio, but take to base 10 rather than base e. In linkage analysis, the ratio that we are interested in is between the likelihood at the null value $\theta = 0.5$ and its maximum at $\hat{\theta}$. The two likelihoods in the ratio are the likelihood of observing the specific marker genotypes and trait phenotypes in the family giving linkage at a particular recombination fraction $\theta<0.5$; otherwise, there is no evidence of linkage. We usually use the value 3 as a statistically significant evidence for linkage at the test recombination fraction. For LOD score equals 3, the significant level α is around 0.001. χ^2 = (2ln10).LOD. LOD score value equals 3 means that the linkage hypothesis with recombination fraction θ is 10^3 times more likely than the null of no-linkage. The problem with this method in the case of a quantitative trait is it requires that we know u_0, u_1, u_2 σ^2 and gene frequency.

Model-based linkage analysis is the most powerful genetic analysis and usually more complicated in calculation whenever the mode of inheritance is known (Ott 1997). When the model of inheritance is unknown, we can improve the power by considering a large number of models of inheritance parameter values and maximizing the LOD scores over multiple models (Ulgen et al. 2001).

1.6.3 Model Free Genetic Linkage Analysis

Gene model free methods of genetic analysis also are referred to as nonparametric methods. There are several nonparametric methods for linkage analysis. The main difference between model based and model free method is that in model free methods do not require any assumptions about the mode of inheritance of the traits.

Penrose (1938) proposed a test for genetic linkage in humans using random independent sib pairs for the situation when one or both traits are quantitative. Jayakar (1970) suggested testing for linkage between a known marker locus and a locus for a quantitative trait by comparing trait variability among marker genotypes to variability within marker genotypes.

Recently, two of the major approaches for linkage analysis with quantitative traits in humans including Haseman-Elston regression [Haseman and Elston, 1972] and the Variance Components method (Amos, 1994) are widely used.

1.6.3.1 Haseman-Elston Method

The Haseman-Elston method (1972) used random sib pairs to study quantitative traits and detect linkage between a quantitative trait and a genetic marker. This is the earliest proposed method of using π_0, π_1 and π_2 (the proportion of alleles shared IBD) at marker locus to the values of a quantitative phenotype in a sib-pair. This method is based on the model that the greater proportion of genes at a marker locus that are identical by descent for a pair of sib, the smaller the squared difference between the sibs' trait values should be if the trait value is affected by a locus linked to the marker locus.

The original Haseman-Elston method is based on regression of squared trait difference on the estimated proportion of alleles shared IBD at a marker locus. That is, let d^2 denote the squared difference of the quantitative trait values observed in a pair of siblings, i.e. $d^2 = (y_1-y_2)^2$, where y1 and y2 are measured quantitative trait values for a sib-pair. The quantity d^2 is regressed on the proportion of marker alleles shared IBD, π. Haseman-Elston method assumes random mating, linkage equilibrium and no epitasis. The statistics hypothesis test is:

H_0: $\beta = 0$

H_a: $\beta < 0$

Note that β is the regression coefficient.

If there is no linkage between the marker and the trait locus, then the regression coefficient equals 0. Otherwise, if the estimated regression coefficient is significantly less than 0, it indicates that there is a linkage of the marker to the QTL.

Performing Haseman-Elston Method, we will get the following regression equation:

$$E[(y_1-y_2)^2] = \alpha + \beta\,\pi$$

$$\alpha = \sigma_e^2 + 2\psi\,\sigma_g^2 + 2\,\psi\,(1-2\psi)\,\sigma_d^2$$

$$\beta = 2\,\psi\,(1-2\psi)\,\sigma_g^2$$

Where $\psi = \theta^2 + (1-\theta)^2$

The recombination fraction θ must satisfy $0 \leq \theta \leq \frac{1}{2}$, so we will have $\beta \leq 0$. If there is no linkage ($\theta=1/2$), the quantity $(1-2\psi)$ is 0 then β equals 0 and there is no regression.

The Haseman-Elston method is widely used because of its simplicity and its robustness against departures from normality of the phenotypes. Blackwelder (1977) has shown the power and robustness of Haseman-Elston Method for genetic linkage between a marker locus and quantitative trait locus by comparing it to Penrose's test. This method can be applied on both qualitative and quantitative (Rao and Li, 2000) univariate and multivariate traits (Amos at al. 1990). Recently, many researchers have worked on revised regression based methods to improve the power, such as the revised Haseman-Elston method, weighted Haseman-Elston Method (Xu et al. 2000).

The revised Haseman-Elston method (Elston et al. 2000) uses the cross-product deviations of quantitative trait values from the population mean instead of square of differenced as the dependent variable since the regression of square of differences does not captive all the information of linkage(Wright 1997). We can get the additional evidence by the regression of square of sums.

Xu et al. (2000) suggested a unified Haseman-Elston method that uses a linear combination of the estimate of the proportion of phenotypic variance from square of sums and square of differences. The weight is determined by the overall trait correlation between the sibs in the population. Researchers (Sham and Purcell, 2001) recently have demonstrated the equivalence between the weighted Haseman-Elston method (Xu et al. 2000) and variance-components methods.

1.6.3.2 Variance- Components Method

R.A. Fisher introduced variance components in 1918. Amos (1994) developed variance components method for linkage analysis in order to increase the power of the original Haseman-Elston regression based linkage analysis. Variance components can be applied to both univariate and multivariate trait linkage analysis. It is an alternative method for investigating linkage between a marker and a QTL. This approach models the quantitative traits in terms of its genetic variance under a multivariate normal assumption. The variance component method is based on likelihood test and extends to arbitrary pedigrees by Blangero (1995). It is widely used (Arya et al. 2002; Kraft et al. 2002; Olswold and de Andrade, 2002; Pankratz et al. 2002). To perform variance components methods, people need to know the parameters (σ, ρ and μ) for the quantitative trait. Here ρ denoted the correlation of the quantitative trait in sib pairs (or in the case of an arbitrary pedigree, all of the relative pairs involved in the analysis).

1.7 Purpose: Model Free Methods for Endophenotype Analysis

Power studies have been done for both the Haseman-Elston method and the Variance Components method. These have focused on random samples of sib pairs and samples in which at least one sibling has a quantitative value greater than 90^{th} percentile. It has been observed that the latter sampling method gives good power using feasible (n=200) sample size in the situation. No power studies of linkage analysis of pairs of siblings ascertained through a disease affected individual have been done. On the other hand, quantitative traits are studied with some complex disease in mind and many genetic studies involves samples of families with at least one person is affected. We investigate the power of the regression based method - Haseman-Elston method using the sib pairs where at least one sibling with a disease for which the quantitative trait is a risk factor. Based on the regression, we can calculate the power of the proposed analysis for setting different genetic parameters.

Chapter 2

Model-free Methods for Linkage Analysis Based on Proband /Sibling Pairs

2.1 Proposed Study Design

We know that the complex disease is affected by many genes, environmental factors and their interactions. Our study is based on some simplifying assumptions. We assume that the disease is determined in part by a bi-allelic disease gene, we also assume this gene is in the Hardy- Weinberg equilibrium and this gene is a pleiotropic gene which causes both a quantitative trait and disease.

2.2 The Sampling Unit

The sampling unit in our study is a sib pair. In each sib pair one is the proband which is affected and the other one is random chosen sibling which may be affected or unaffected. First, we assigned the proband's (affected sib) genotype. From the proband's genotype, we got the parents' genotype. We also assign parent's marker genotype. Then from the parent's genotype and marker genotype, we got both proband's and sibling's marker genotype and sibling's pleiotropic locus genotype. Our study is based on the sib ships' data. Since we are sampling only sib ships with at least one disease affected individual, we are much more likely to observe more endophenotype positive than we would in a random sample of the sib pairs.

2.3 The Proposed Data

In our study, the proposed data requires quantitative trait values and marker data on both sibs and marker data on their parents. We will observe the number of alleles shared IBD at a marker locus based on the sibs' marker data and their parents' marker data.

Table 2.1: Brief description of the study design in endophenotype analysis

familyID	individualID	GM[0]	GM[1]	PT	GD[f2]	GM[f2]
1	3	43	12	3.353604	12	41
1	4	43	12	2.839293	12	42
2	3	23	41	3.451288	12	21
2	4	23	41	1.130917	12	24
3	3	32	41	1.923809	12	31
3	4	32	41	5.033534	12	31
4	3	13	24	1.664957	12	14
4	4	13	24	1.596932	12	12
5	3	34	12	2.420756	12	31
5	4	34	12	-0.89957	22	42
6	3	31	24	5.204206	12	32
6	4	31	24	-0.22507	22	12
7	3	42	31	2.613266	12	43
7	4	42	31	3.082184	12	41
8	3	24	31	2.325471	12	23
8	4	24	31	4.077521	12	23
9	3	24	13	3.776028	12	23
9	4	24	13	0.957283	22	41
10	3	24	31	2.838258	12	21
10	4	24	31	2.191085	12	21

Note that GM[0] and GM[1] are parents' marker genotype, GM[f2] is sib's marker genotype, GD[f2] sib's genotype and PT is sib's quantitative trait value.

2.4 The Proposed Genetic Generating Model

In our study, we proposed a genetic generating model in which there is a pleiotropic gene and this pleiotropic gene determines two traits: a quantitative endophenotype and presence or absence of a disease. The pleiotropic gene influences both the endophenotype and disease at the same time.

2.5 The Proposed Analysis Methods

The regression based method is applied in our study. Particularly, we used Haseman-Elston method. We investigate the genetic linkage between a marker locus and a locus affecting the quantitative trait. This method is based on regression of squared trait difference on the proportion of alleles shared IBD at a marker locus in sib pairs. The null hypothesis of the test is β=0 and the alternative hypothesis is β<0. If we reject the null hypothesis, we conclude that there is a linkage between the marker locus and pleiotropic locus.

Table 2.2: Data for Haseman-Elston method

Square of diff	IBD number	Proportion of IBD
3.44487	2	1
4.375351	1	0.5
1.337481	1	0.5
14.06116	1	0.5
8.44397	2	1
11.94665	0	0
3.732203	2	1
40.64755	1	0.5
2.874225	1	0.5
4.491975	1	0.5
0.010285	2	1
0.006611	1	0.5
2.123598	1	0.5
0.674629	2	1
0.678953	1	0.5
6.340722	1	0.5
0.443448	2	1
14.08093	1	0.5
0.137931	2	1

Then we use this data to apply Haseman-Elston regression method. The dependent variable in this model is squared difference of quantitative trait values and the independent variable is the proportion of shared marker IBD.

2.6 Goal

The goal of the study is to assess the power of linkage analysis between the marker data and the quantitative trait locus.

In our study, we set the alpha level equals 0.05, and then we reject the null hypothesis (Rh=1, there is a linkage between the QTL and marker locus) if t-value is less than -1.65, otherwise Rh=0.

$$\begin{cases} \text{Rh=1} & \text{if t-value} \le \text{ -1.65} \\ \text{Rh=0} & \text{otherwise} \end{cases}$$

$$Power = \frac{\#Rh = 1}{N}$$

For example, based on the following model p=0.01, f_{2D}=0.5, f_{1D}=0.25 and f_{0D}=0 and z_0=0, z_1=2 and z_2=4, we applied the Haseman-Elston regression method and then get the results of T-value. Appendix Table A1.1 shows the simulation example.

We set that:

$$\begin{cases} \text{Rh=1} & \textit{if } \text{ t-value} \le -1.65 \\ \text{Rh=0} & \text{Otherwise} \end{cases}$$

The following is what we got:

$$\begin{cases} \text{n=6} & \text{Rh=0} \\ \text{n=94} & \text{Rh=1} \end{cases}$$

So the power is 94% in for this set of genetic model parameter values.

Chapter 3

Simulation of Quantitative Pedigrees under Pleiotropic Endophenotype Generating Model

3. 1 Calculations for the Simulations

In this part, we develop the probabilities that govern the distribution of endophenotype and marker allele sharing under the conditions of linkage between a pleiotropic trait locus and a marker locus. These probability values will be in the simulation.

As mentioned before, we denote the disease allele frequency as p, and the normal allele frequency as q, where q = 1-p. The population probability of genotype G_i, assuming Hardy-Weinberg equilibrium is straight forward.

$$\mathrm{P}(G_i) = \begin{cases} \mathrm{P}(AA) = p^2 \\ \mathrm{P}(AB) = 2pq, i = 1,2,3 \\ \mathrm{P}(BB) = q^2 \end{cases} \qquad (3.1.1)$$

The marginal probability of parents' pleiotropic locus (PL) genotype $(M_{j\otimes k})$ with one parent genotype j and the other genotype k is given as follows:

$$\mathrm{P}(M_{j\otimes k}) = \mathrm{P}(G_j)\cdot\mathrm{P}(G_k) \quad \text{Where j, k=1, 2, 3} \qquad (3.1.2)$$

Assuming random mating, the probability of parents' PL genotypes in the population is giving in Table 3.1.

Table 3.1: The probability of parents' PL genotypes

Parents' PL genotypes	Probability
AA*AA	p^4
AA*AB	$4p^3q$
AA*BB	$2p^2q^2$
AB*AB	$4p^2q^2$
AB*BB	$4pq^3$
BB*BB	q^4

Let f_{G_i} denote the disease penetrance associated with pleiotropic locus genotype G_i, G_i=AA, AB, BB.

$$f_{G_i} = \begin{cases} f_{AA} \\ f_{AB} \\ f_{BB} \end{cases} \quad \text{i=1, 2, 3} \qquad (3.1.3)$$

So the probability of D+ in the population is:

$$P(D+) = f_{AA} \cdot p^2 + f_{AB} \cdot 2pq + f_{BB} \cdot q^2 \qquad (3.1.4)$$

The conditional probability of a pleiotropic locus genotype given the phenotype is D+ is calculated using Bayes Theorem. This is the probability of a disease affected proband has each genotype.

$$\mathrm{P}(G_i \mid D+) = \begin{cases} \mathrm{P}(AA \mid D+) = \dfrac{p^2 f_2}{p^2 f_2 + 2pqf_1 + q^2 f_0} \\ \mathrm{P}(AB \mid D+) = \dfrac{2pqf_1}{p^2 f_2 + 2pqf_1 + q^2 f_0}, i = 1,2,3 \\ \mathrm{P}(BB \mid D+) = \dfrac{q^2 f_0}{p^2 f_2 + 2pqf_1 + q^2 f_0} \end{cases} \qquad (3.1.5)$$

The following equation gives the probability of parents' mating type given proband genotype:

$$P(M_{j\otimes k} \mid G_i) = \frac{P(G_i \mid M_{j\otimes k})P(M_{j\otimes k})}{\sum P(G_i \mid M_{j\otimes k})P(M_{j\otimes k})} \qquad \text{i = 1, 2, 3} \qquad (3.1.6)$$

So the probability of parent's mating type conditional on proband genotype is as follows:

$$\mathrm{P}(\mathrm{M}_{\mathrm{j}\otimes\mathrm{k}}|\mathrm{AA}) = \begin{array}{c} \\ \mathrm{AA} \\ \mathrm{AB} \\ \mathrm{BB} \end{array} \begin{array}{c} \begin{array}{ccc} AA & AB & BB \end{array} \\ \begin{pmatrix} p^2 & 2pq & 0 \\ 2pq & q^2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \qquad (3.1.7)$$

$$\mathrm{P}(\mathrm{M}_{\mathrm{j}\otimes\mathrm{k}}|\mathrm{AB}) = \begin{array}{c} \\ \mathrm{AA} \\ \mathrm{AB} \\ \mathrm{BB} \end{array} \begin{array}{c} \begin{array}{ccc} AA & AB & BB \end{array} \\ \begin{pmatrix} 0 & p^2 & pq \\ p^2 & pq & q^2 \\ pq & q^2 & 0 \end{pmatrix} \end{array} \qquad (3.1.8)$$

$$\mathrm{P}(\mathrm{M}_{\mathrm{j}\otimes\mathrm{k}}|\mathrm{BB}) = \begin{array}{c} \\ \mathrm{AA} \\ \mathrm{AB} \\ \mathrm{BB} \end{array} \begin{array}{c} \begin{array}{ccc} AA & AB & BB \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & p^2 & 2pq \\ 0 & 2pq & q^2 \end{pmatrix} \end{array} \qquad (3.1.9)$$

The following matrix gives the probability of having an offspring with disease for a given parents' mating type.

$$\mathrm{P}(D+ \mid M_{j\otimes k}) = \begin{array}{c} \\ AA \\ AB \\ BB \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{AA} & \mathrm{AB} & \mathrm{BB} \end{array} \\ \begin{pmatrix} f_{AA} & \frac{1}{2}(f_{AA}+f_{AB}) & f_{AB} \\ \frac{1}{2}(f_{AA}+f_{AB}) & \frac{1}{2}f_{AB}+\frac{1}{4}(f_{AA}+f_{BB}) & \frac{1}{2}(f_{AB}+f_{BB}) \\ f_{AB} & \frac{1}{2}(f_{AB}+f_{BB}) & f_{BB} \end{pmatrix} \end{array}$$

j, k=1, 2, 3. (3.1.10)

Then the probability of a parents' PL genotype given one of offspring is affected is calculated, again using Bayes Theorem as:

$$\mathrm{P}(M_{j\otimes k} \mid D+) = \frac{\mathrm{P}(D+ \mid M_{j\otimes k}) \cdot \mathrm{P}(M_{j\otimes k})}{\sum_{m,n=1}^{3} \mathrm{P}(D+ \mid M_{m\otimes n}) \cdot \mathrm{P}(M_{m\otimes n})} \qquad (3.1.11)$$

Where $P(D+ \mid M_{m\otimes n})$ is given in equation 3.1.10 and $P(M_{j\otimes k})$ is given in equation 3.1.2.

The probability of the offspring has G_i given the parents' PL genotypes are given in the following matrices:

$$\mathrm{P}(\mathrm{Sib}_{G_i}=\mathrm{AA}|\mathrm{M}_{j\otimes k}) = \begin{array}{c} \\ \mathrm{AA} \\ \mathrm{AB} \\ \mathrm{BB} \end{array} \begin{array}{c} \begin{array}{ccc} AA & AB & BB \end{array} \\ \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 0.25 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \qquad (3.1.12)$$

$$\mathrm{P}(\mathrm{Sib}_{G_i}=\mathrm{AB}|\mathrm{M}_{j\otimes k}) = \begin{array}{c} \\ \mathrm{AA} \\ \mathrm{AB} \\ \mathrm{BB} \end{array} \begin{array}{c} \begin{array}{ccc} AA & AB & BB \end{array} \\ \begin{pmatrix} 0 & 0.5 & 1 \\ 0.5 & 0.5 & 0.5 \\ 1 & 0.5 & 0 \end{pmatrix} \end{array} \qquad (3.1.13)$$

$$\mathrm{P}(\mathrm{Sib}_{G_i}=\mathrm{BB}|\mathrm{M}_{j\otimes k}) = \begin{array}{c} \\ \mathrm{AA} \\ \mathrm{AB} \\ \mathrm{BB} \end{array} \begin{array}{c} \begin{array}{ccc} AA & AB & BB \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.25 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix} \end{array} \qquad (3.1.14)$$

The distribution of quantitative trait values of the proband and sibling are obtained as:

$$f(T_{G_i}) \sim N(t_{G_i}, 1) \qquad \text{i=1, 2, 3} \qquad (3.1.15)$$

G_i is individual's genotype and is defined in equation 3.1.5 for proband and 3.1.12, 3.1.13, 3.1.14 for sibling.

The pdf of quantitative trait values in the population is defined as:

$$f(T) = \sum_{i=1}^{3} p_i \cdot f(T_{G_i}) \qquad \text{i=1, 2, 3} \qquad (3.1.16)$$

$$\text{Where} \begin{cases} p_1 = p^2 \\ p_2 = 2pq \\ p_3 = q^2 \end{cases} \qquad (3.1.17)$$

Without loss of generality, we set $t_{AA} \geq t_{AB} \geq t_{BB}$ in the simulation. We first set the trait heritability range from 0.1 to 0.5, and then we can calculate the trait values given the disease allele frequency and trait model dominance (dominant, additive or recessive).

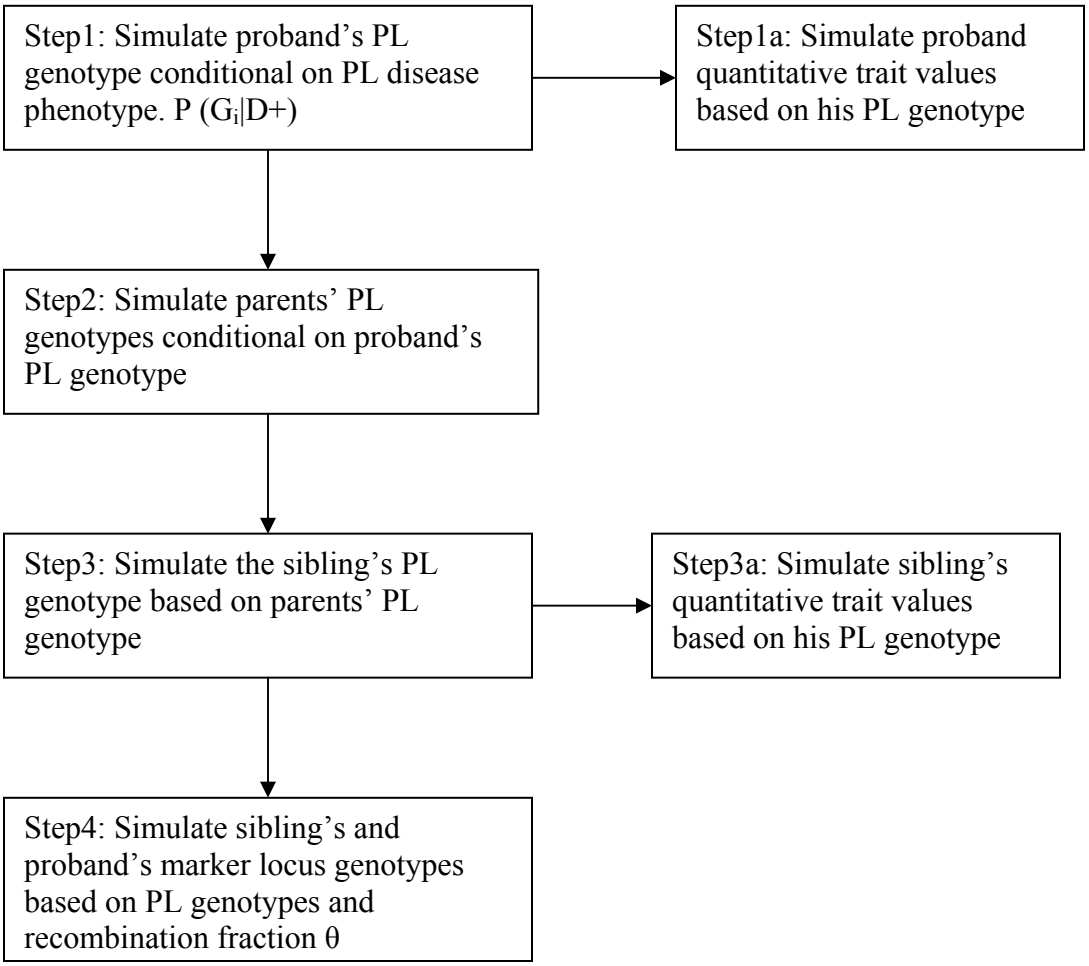
3.2. Simulation Programming Strategies

Here, we are interested in simulating the joint distribution of a disease and a disease related endophenotype (DRT). Also we are interesting in simulating the quantitative trait values and using the Haseman-Elston method to test the power of linkage analysis of the endophenotype.

We are interested in particular in the case where we sample only nuclear families that have at least one disease affected individual. We also consider a case where both disease and disease related trait are affected by a major pleiotropic gene which we refer to as a pleiotropic trait locus (PL). We simulate marker genotype and the Endophenotype trait locus genotype for each individual in the family based on the proband's (affected individual) genotype. The Endophenotype trait values are based on individuals' trait locus genotype only.

The simulations were done as described in the flow chart on the next page. A detailed description of each step of the simulation and sample results follow.

Figure 3.1: Flowchart of Simulation Procedure of Pleiotropic Locus (PL) and marker Locus (ML) Genotypes and Phenotypes



In Step 1: Simulate proband's PL genotype conditional on PL disease phenotype $P(G_i|D+)$.

Let P (G_i) denote the probability of genotype G_i (G_1=AA, G_2=AB, G_3=BB) in the population and as given in Equation 3.1.1. And let P ($D+|G_i$) denote the penetrance for genotype G_i (AA, AB and BB). Then we can calculate the conditional probability that proband (who is affected) has genotype G_i as follows:

$$P(G_i \mid D+) = \frac{P(D+ \mid G_i)P(G_i)}{\sum_{i=1}^{3} P(D+ \mid G_i)P(G_i)}, (G_1=AA, G_2=AB, G_3=BB), \qquad (3.2.1)$$

In the simulation, the genotypes of the proband (G_i) were assigned with probability P ($G_i|D+$) by first sampling a random variable Y from U (0, 1). We then assigned the proband's PL genotype.

$$\begin{cases} proband \quad genotype \quad = \quad G_1, \quad if \quad 0 < y < P(G_1 \mid D) \\ proband \quad genotype \quad = \quad G_2, \quad if \quad P(G_1 \mid D) < y < P(G_1 \mid D) + P(G_2 \mid D) \\ proband \quad genotype \quad = \quad G_3, \quad if \quad P(G_1 \mid D) + P(G_2 \mid D) < y < 1 \end{cases}$$

(3.2.2)

As an example consider the case where the gene frequency of A(disease allele) p=0.01 and f_{AA}=P($D+|G_1$)=0.5, f_{AB}=P($D+|G_2$)=0.25, f_{BB}= P($D+|G_3$)=0 and z_0=0, z_1=2, z_2=4. Then applying the equation 3.1.1, we got

$$\begin{cases} P(G_1) = 0.0001 \\ P(G_2) = 0.0198 \\ P(G_3) = 0.9801 \end{cases}$$

and

$$\begin{cases} P(G_1 \mid D+) = 0.01 \\ P(G_2 \mid D+) = 0.99 \\ P(G_3 \mid D+) = 0 \end{cases}$$

Table 3.2 shows the simulation results and expected results. The results based on simulating100*100, 100 simulations of samples of 100 families. From the 10,000 simulation of "proband", 9885 probands' genotypes are AB and 115 are AA.

Table 3.2: Population and simulation results on probability of genotype

genotype(p=.01, q=.99)	Population results	Simulation results (N=10,000 probands)
AA	1%	1.15%
AB	99%	98.85%
BB	0	0

From Table 3.2 the simulation results, we got 1.15% proband whose genotype is assigned as AA and 98.85% proband whose genotype is assigned as AB based on this model (p=0.01, q=0.99, f_{AA}=0.5, f_{AB}=0.25, f_{BB}=0, z_0=0, z_1=2, z_2=4) . Here, the probability of genotype BB is 0 because we set the penetrance for genotype BB at 0 in this model. Also the table above showed the probability of proband genotype on population based on this model. In population the probability of genotype AA is 1% and genotype AB is 99%.

In Step 1a: Simulate Proband's Quantitative Trait Values Based on his PL Genotype.

In this step, we assign the quantitative trait values to the proband based on proband genotype. We assume that the distribution of the quantitative trait values conditional on genotype is normal distribution. For a given proband genotype, the quantitative trait value of the proband is assigned by sampling using a random number generator according to proband's genotype. The generator generates a random number from normal distribution. Based on this model (p=0.01, q=0.99, f_{AA}=0.5, f_{AB}=0.25, f_{BB}=0, z_0=0, z_1=2, z_2=4), we generate a random number from N (2, 1) for proband's genotype is AB and we generate a random number from N (4, 1) for proband's genotype is AA. Below is the schematic plot based on this simulation model, from the simulation, we got quantitative trait values fitting a normal distribution with mean 4.15 and standard deviation 0.98 (standard error 0.09 skewness -0.21 and kurtosis -0.51) for proband genotype AA. The 95% confidence interval (CI) is $4.15\pm1.96.\sigma/\sqrt{n}$ =4.15 ± 0.18 = [3.97, 4.33]. And we also see that the quantitative trait values fit normal distribution with mean 1.99 and standard deviation 1.01 (standard error 0.01 skewness 0.04 and kurtosis -0.008) for proband genotype AB. The 95% confidence interval (CI) is $1.99\pm1.96.\sigma/\sqrt{n}$ =1.99 ± 0.02 = [1.97, 2.01]. From the simulation results, we can conclude that our simulation program to simulation the quantitative trait data works correct. Figure 3.2 showed the quantitative trait distribution of probands given genotypes.

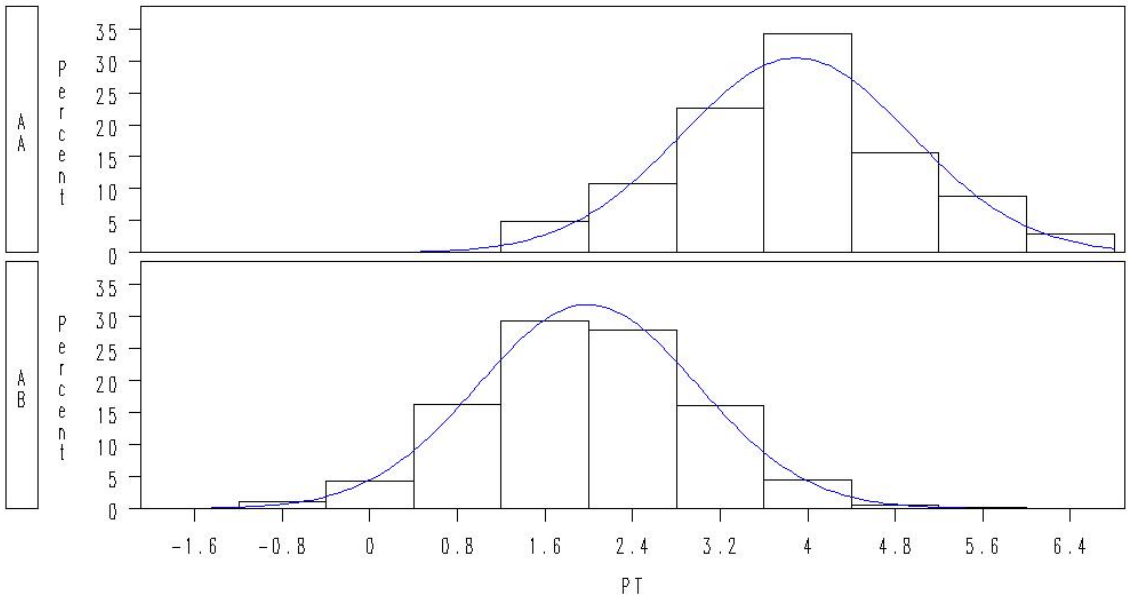


Figure 3.2 Quantitative trait distributions given the proband's genotype

In Step 2: Simulate the Parents' PL Genotypes Based on the Proband's PL Genotype

We calculate the probability of the parents' PL genotype conditional on the genotype of the proband, using following equation 3.2.3. The genotype of the proband, G_{ip}, was obtained in Step 1.

$$P(M_{j\otimes k} \mid G_{ip}) = \frac{P(G_{ip} \mid M_{j\otimes k})P(M_{j\otimes k})}{\sum_{i,j,k=1}^{3} P(G_{ip} \mid M_{j\otimes k})P(M_{j\otimes k})} \qquad (3.2.3)$$

G_{ip} denotes the genotype of proband, i =1, 2 and $P(M_{j\otimes k})$ is defined by equation 3.1.2. $P(G_{ip} \mid M_{j\otimes k})$ is defined by equation 3.1.8, 3.1.9 and 3.1.10.

Let G_{jf}, G_{jm} denote father's genotype and mother's genotype respectively, j=1, 2 and 3. The genotypes of G_{jf}, G_{jm} are assigned with the probability $P(M_{j\otimes k}|G_i)$ by sampling using a random number generator. The generator generates a random variable Y from U(0,1).If G_{pro}=AA, then

$$\begin{cases} G_{jf}=AA \quad and \quad G_{jm}=AA \quad if \quad 0<y<P(M_{AA\otimes AA} \mid G_{ip}=AA) \\ G_{jf}=AA \quad and \quad G_{jm}=AB \quad if \quad P(M_{AA\otimes AA} \mid G_{ip}=AA)<y<P(M_{AA\otimes AA} \mid G_{ip}=AA)+P(M_{AA\otimes AB} \mid G_{ip}=AA) \\ G_{jf}=AB \quad and \quad G_{jm}=AB \quad if \quad P(M_{AA\otimes AA} \mid G_{ip}=AA)+P(M_{AA\otimes AB} \mid G_{ip}=AA)<y<1) \end{cases} \qquad (3.2.4)$$

Use the same strategies for the cases of G_{pro}=AB and G_{pro}=BB. Below are the results based on the model (p=0.01, q=0.99, f_{AA}=0.5, f_{AB}=0.25, f_{BB}=0, z_0=0, z_1=2, z_2=4) in the population.

$$P(M_{j\otimes k}|AA) = \begin{array}{c} \\ AA \\ AB \\ BB \end{array} \begin{pmatrix} AA & AB & BB \\ 0.0101\% & 0.9999\% & 0 \\ 0.9999\% & 98.99\% & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad (3.2.5)$$

$$P(M_{j\otimes k}|AB) = \begin{array}{c} \\ AA \\ AB \\ BB \end{array} \begin{pmatrix} AA & AB & BB \\ 0 & 0.0099\% & 0.9803\% \\ 0.0099\% & 1.9606\% & 97.0492\% \\ 0.9803\% & 97.0492\% & 0 \end{pmatrix} \quad (3.2.6)$$

$$P(M_{j\otimes k}|BB) = \begin{array}{c} \\ AA \\ AB \\ BB \end{array} \begin{pmatrix} AA & AB & BB \\ 0 & 0 & 0 \\ 0 & 0.0101\% & 0.9999\% \\ 0 & 0.9999\% & 98.99\% \end{pmatrix} \quad (3.2.7)$$

And the probability of mating type given the proband genotype in the simulation results is as following:

$$P(M_{j\otimes k}|AA) = \begin{array}{c} \\ AA \\ AB \\ BB \end{array} \begin{pmatrix} AA & AB & BB \\ 0 & 2.61\% & 0 \\ 2.61\% & 97.39\% & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (3.2.8)$$

$$P(M_{j\otimes k}|AB) = \begin{array}{c} \\ AA \\ AB \\ BB \end{array} \begin{pmatrix} AA & AB & BB \\ 0 & 0.01\% & 1.04\% \\ 0.01\% & 2.19\% & 96.76\% \\ 1.04\% & 96.76\% & 0 \end{pmatrix} \quad (3.2.9)$$

$$P(M_{j\otimes k}|BB) = \begin{array}{c} \\ AA \\ AB \\ BB \end{array} \begin{pmatrix} AA & AB & BB \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (3.2.10)$$

From the results, we can see that there are 96.76% parents' mating type is BBXAB given proband genotype AB based on model (p=0.01, q=0.99, f_{AA}=0.5, f_{AB}=0.25, f_{BB}=0, z_0=0, z_1=2, z_2=4) and there are 97.39% parents' mating type is ABXAB given proband genotype AB. Also we calculated the population mating type given genotype. We found that for large samples (i.e. for proband genotype AB, we have total 9885 individuals from the simulation) we got very similar probabilities (0.01%, 0.98%, 1.96% and 97.05% compare with 0.01%, 1.04%, 2.19% AND 96.76%). But the small sample size (total 115 for given proband genotype AA), we got a slightly different probabilities (0.01%, 1.0% and 98.99% in population and 0%, 2.61% and 97.39% in simulation).

In Step 3: **Simulate Sibling's Genotype Giving Parents' PL Genotype**

We calculate the probability of sibling's genotype given parents' PL genotype by using the equation 3.1.12, 3.1.13, 3.1.14. This step is very simple, the probability of sibling's genotype for given parent's genotype is independent of penetrance and gene frequency.

If mating type is AA*AA, then sibling's genotype is always AA.

If the mating type is AB*AB, then generate the random variable Y from U (0, 1).

$$\begin{cases} sib \quad genotype \quad = \quad AA, \quad if \quad 0 < y < \frac{1}{4} \\ sib \quad genotype \quad = \quad AB, \quad if \quad \frac{1}{4} < y < \frac{3}{4} \\ sib \quad genotype \quad = \quad BB, \quad if \quad \frac{3}{4} < y < 1 \end{cases} \qquad (3.2.11)$$

Similarly, we can get the probability for sibling genotype given mating type AA*AB, AB*AB, AB*BB and BB*BB.

Figure 3.3 shows the simulation results based on model (p= p=0.01, q=0.99, f_{AA}=0.5, f_{AB}=0.25, f_{BB}=0, z_0=0, z_1=2, z_2=4).

Table 3.3 shows the simulation results and population results on the probability of sibling's genotype.

genotype(p=.01, q=.99)	Population results	Simulation results
AA	0.01%	0.83%
AB	1.98%	50.89%
BB	98.01%	48.28%

From the simulation, we got 0.83% sibling's genotype AA, 50.89% sibling's genotype is AB and 48.28% sibling's genotype is BB. Also we got 13.33% sibling are affected with disease from simulation. See the chart below (of course, the probability of affected proband is 100%).

In Step 3a: Simulate Sibling's Quantitative Trait Values Based on his PL Genotype.

In this step, we are using the similar method to simulate the sibling's quantitative trait values based on sibling's genotype instead of proband genotype. Since we have 3 types of genotype for sibling, here we generate another random number from N (0, 1) for genotype BB. Below shows the schematic plots on PT based on genotype in this model, from the simulation results, we found that the quantitative trait values fit normal distribution with mean 3.94 and standard deviation 0.92 (standard error 0.10 skewness -0.14 and kurtosis -0.75) for genotype AA, The 95% confidence interval (CI) is $3.94 \pm 1.96.\sigma/\sqrt{n} = 3.94 \pm 0.21$ = [3.87, 4.01]. Also we got the quantitative trait values fit normal distribution with mean 2.018 and standard deviation 0.99 (standard error 0.01 skewness 0.002 and kurtosis 0.04) for genotype AB. The 95% confidence interval (CI) is $2.018 \pm 1.96.\sigma/\sqrt{n} = 2.018 \pm 0.03$ = [1.97, 2.01]. Similarly, for genotype BB, we got the trait values with mean -0.028 and standard deviation 1.003 (standard error 0.01 skewness 0.05 and kurtosis 0.02). The 95% CI is $\sqrt{n}\ 0.028 \pm 1.96.\sigma/\sqrt{n} = -0.028 \pm 0.03$ = [-0.058, 0.002]. Figure 3.3 shows the histogram of quantitative trait distribution conditional on the sib's genotype.

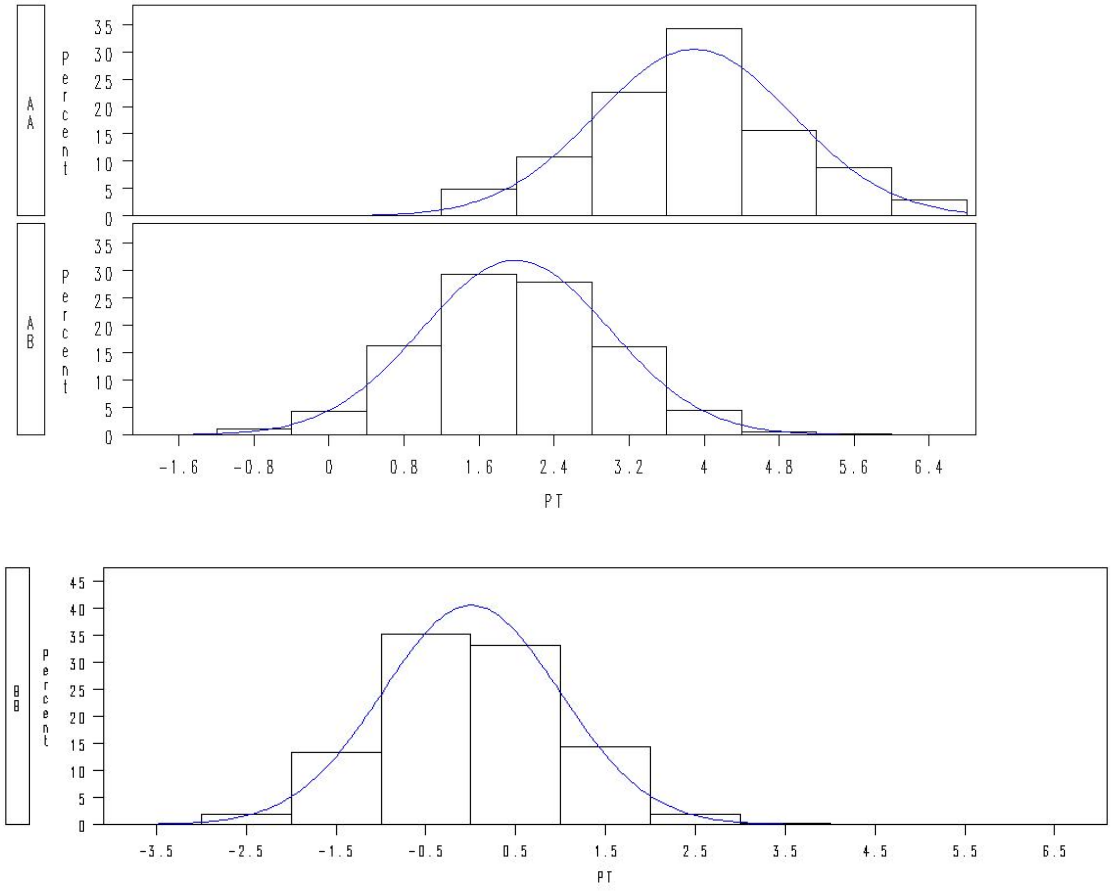


Figure 3.3 Quantitative trait distributions given sib's genotype

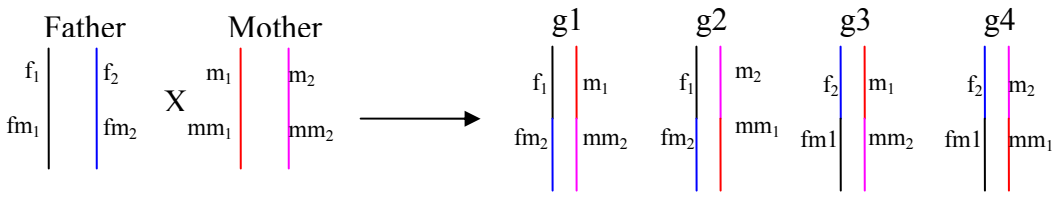
In Step 4: Simulate Sibling's and Proband's Marker Locus Genotypes Based on PL Genotypes and Recombination Fraction θ

We already knew both pleiotropic trait locus and marker genotypes of the parents. In this step, we assign disease and marker genotype to the offspring. Since there are 4 combinations of haplotypes inherited from father and mother, there are 4 possible haplotypes for the offspring. (1). Paternal haplotype is recombinant and maternal haplotype is recombinant; (2). Paternal haplotype is recombinant and maternal haplotype is non-recombinant; (3). Paternal haplotype is non-recombinant and maternal haplotype is recombinant; (4). Paternal haplotype is non-recombinant and maternal haplotype is non-recombinant.

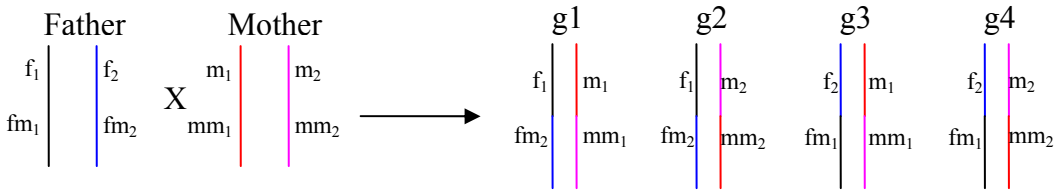
Let r_f and r_m denote recombinant status on the haplotype inherited from father and mother respectively. We assign r_f, r_m=1 if the haplotype is a recombinant haplotype and let r_f, r_m=0 if the haplotype is a non-recombinant.

For each case, there are 4 possible trait and marker genotypes which can be passed on to the offspring; each has equal possibility (25%). Let f_1 , f_2 denote the two alleles at father's pleiotropic locus, and let m_1 , m_2 denote two alleles at mother's pleiotropic locus Let fm_1, fm_2 denote the two alleles at father's marker locus, and mm_1, mm_2 denote the two alleles at mother's marker locus. Let g1, g2, g3 and g4 denote the four possible trait and marker genotypes. Below are 4 possible cases and the 4 possible haplotype passed on to the offspring in each case.

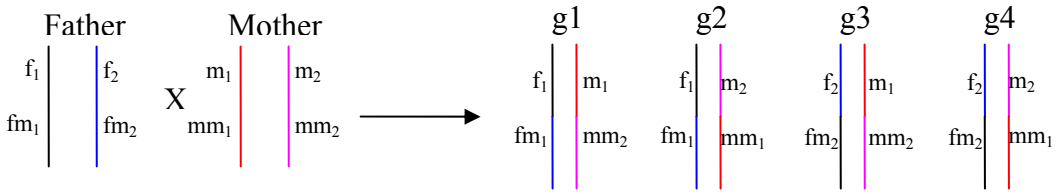
Case I: r_a=1 and r_m=1 (both parents have recombinant)



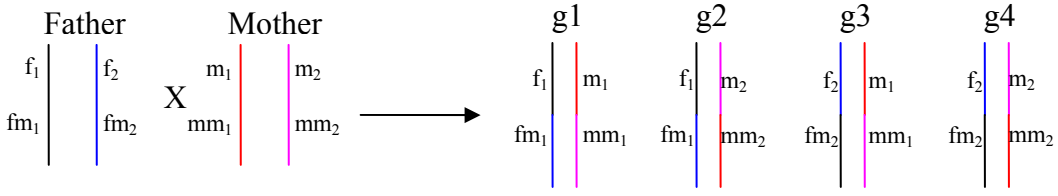
Case II: r_a=1 and r_m=0 (only paternal haplotype has recombinant)



Case III: r_a=0 and r_m=1 (only maternal haplotype has recombinant)



Case IV: r_a=0 and r_m=0 (neither has recombinant)



Now we generate independent random variables X, Y from U (0, 1),

$$\begin{cases} r_f = 1 & if \quad 0 < y < \theta \\ r_f = 0 & otherwise \end{cases}$$

$$\begin{cases} r_m = 1 & if \quad 0 < x < \theta \\ r_m = 0 & otherwise \end{cases}.$$

We will choose our simulation case is which one from these 4 cases according to the values of f_a and f_m. For example, If r_f=1 and r_m=1 then we have case 1.After we decided case 1 will be used, we will generate another random variable Z from U (0, 1). From the Z value, we will assign the trait and marker genotype to the sibling.

$$\begin{cases} offspring \quad haplotype = g1 \quad if \quad 0 < z < 0.25 \\ offspring \quad haplotype = g2 \quad if \quad 0.25 < z < 0.50 \\ offspring \quad haplotype = g3 \quad if \quad 0.50 < z < 0.75 \\ offspring \quad haplotype = g4 \quad if \quad 0.75 < z < 1 \end{cases}$$

We use same procedure to assign genotype to sibling conditional on the recombination status (case 1, 2, 3, and 4).

Marker genotypes of proband and offspring of the proband are assigned from the parents' marker genotype. To simplify, parents are given marker genotype 1, 2, 3, 4. When we assign the parents' marker genotype, We first assign the marker alleles to father, whose first allele can random choose from 1,2,3,4, then his second allele can choose from the rest 3 numbers. After assigned the marker genotype for the father, then we assign the marker genotype for the mother. The mother's first allele is random chosen from the rest 2 numbers, and then assign mother's second allele from the last number. Now we have the parents' marker genotypes and we can differ from each other. Since we assume parents' markers are fully informative. According to the recombination parameter, the sibling is assigned both disease and marker genotypes.

As for the proband, it's a bit different since we already assigned the disease genotype at the beginning of the simulation, so I let the computer keep assigning the disease and marker genotypes to the proband until at one iteration that the assigned disease proband is in agreement with the known disease genotype. Appendix Table A2.1 shows the whole

simulated pedigree data based on a particular simulation model (p=0.01, q=0.99, f_{AA}=0.5, f_{AB}=0.25, f_{BB}=0, z_0=0, z_1=2, z_2=4).

Chapter 4

Power study

4.1 Study Design: Generating Different Models Based on Heritability.

For the given generating model, there are a lot of options for the parameter values. We first focused on varying trait heritability values. For given trait heritability ranging from 0.1-0.5, we considered several disease parameters settings and then get the trait values bases on different trait models (dominant, additive and recessive). The design of the study of the parameter values is shown in Table 4.1. The heritability values and dominance values and gene frequency determine the quantitative trait means.

Aside from the value of the allele frequency, we consider three parameters for the quantitative trait values and 3 parameters for the disease models if the P (D+|BB) =0 and 4 parameters for the disease models if the P (D+|BB) =0.001.

Table 4.1: Values of the generating model parameters for sib pair simulations:

Disease/Trait Allele frequency	p=0.01 q=0.99	p=0.05 q=0.95	
Trait Heritability	0.1, 0.2, 0.3, 0.4, 0.5		
Trait Dominance (d)	-1, 0, 1		
Disease Penetrance Values	$p(D+\|AA)=f_{AA}=0.5$ $p(D+\|AA)=f_{AA}=0.3$ $p(D+\|AA)=0.83$	$p(D+\|AB)=f_{AB}=f_{AA}$ $p(D+\|AB)=\frac{f_{AA}+f_{BB}}{2}$ $p(D+\|AB)=f_{BB}$ $p(D+\|AB)=\sqrt{f_{AA}\cdot f_{BB}}$	$p(D+\|BB)=f_{BB}=0$ $p(D+\|BB)=f_{BB}=0.001$

Based on the different trait and disease parameter values, we have 90 dominant, 90 additive, 90 recessive parameter settings at $f_{BB}=f_{0D}=0$. Also we have 120 dominant, 120 additive, 120 recessive and 120 log additive values for the case where $f_{BB}=f_{0D}=0.001$.

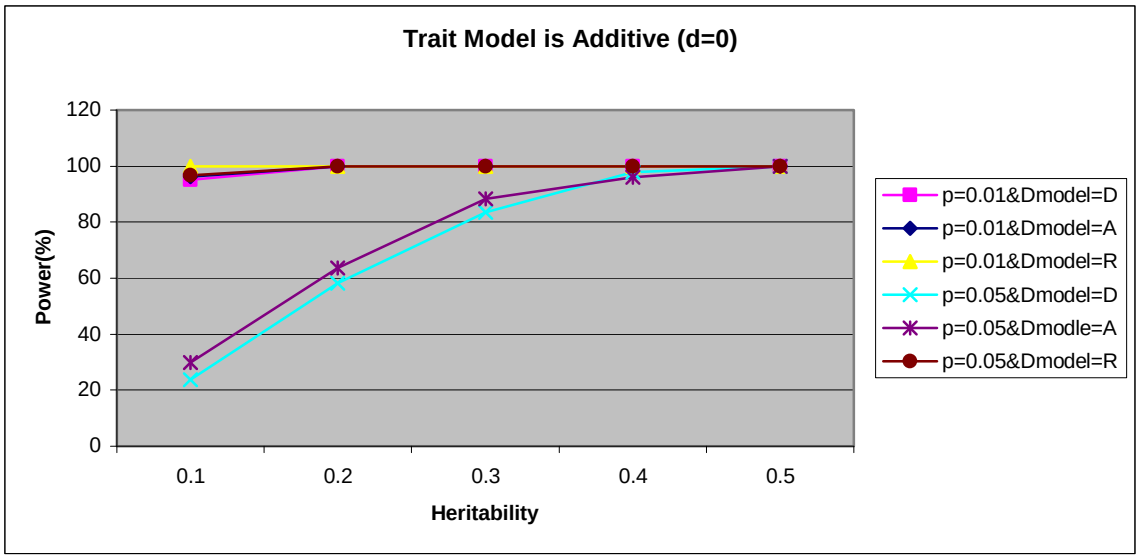
In each dataset, there are 100 nuclear families with 4 per family, 2 parents and 2 offspring (one disease affected proband and the one random chosen sibling). The results are based on simulating N=1000 simulations of n=100 families.

Once the pedigree datasets are simulated, for a given generating parameter value, we had the information on the marker genotype, pleiotropic locus genotype and quantitative trait values for each individual. We then use the sib pair information on the

marker genotype, pleiotropic genotype and quantitative trait values to get the squared difference of the quantitative trait values and the proportion of shared IBD number at the marker locus for each sib pair. We obtained the t-value for the regression coefficient after applying the Haseman-Elston regression based method. The power is calculated with significance level equal to α =0.05, i.e. t≤ -1.65. Appendix Tables A4.1, A4.2, A4.3 and A4.4 shows some simulation results based on different genetic models. Figure 4.1 to Figure 4.3 that follow show the power values for different p values given f_{0D}=0, Figure 4.4 to Figure 4.6 shows the power comparison for different f_{2D} values given f_{0D}=0. Figure 4.6 to Figure 4.9 shows the power comparison for different p values given f_{0D}=0.001, Figure 4.10 to Figure 4.12 shows the power values for different f_{2D} values given f_{0D}=0.001.

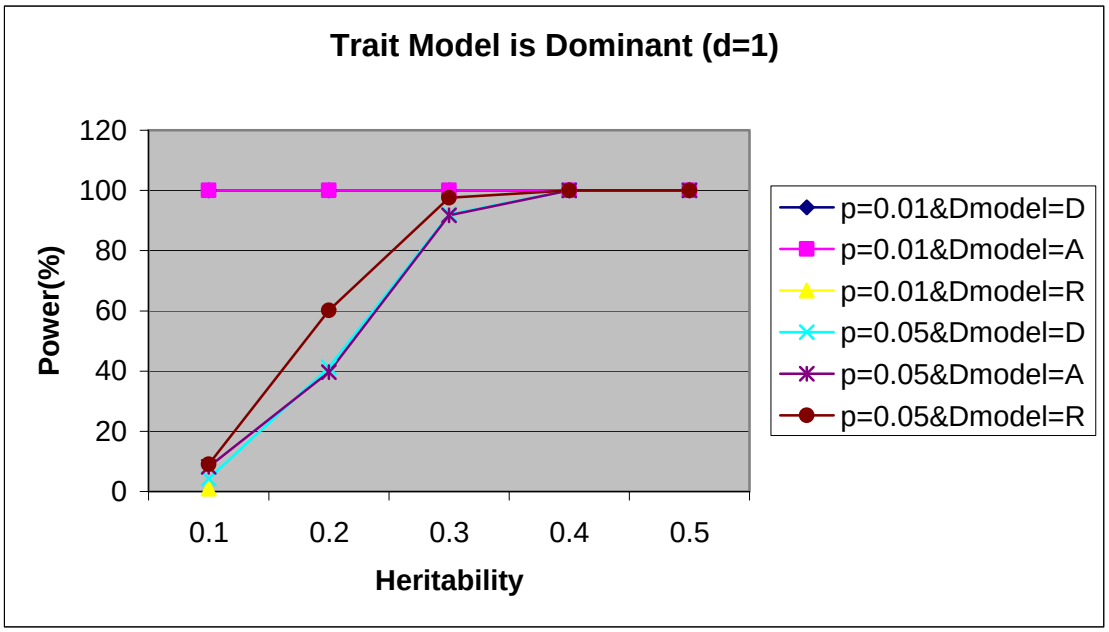
4.2 Power Values for f_{0D}=0

Figure 4.1 Power comparison for p=0.01 and p=0.05 given trait model is additive, f_{2D}=0.5, f_{0D}=0, θ=0.01 and n=100, N=1000



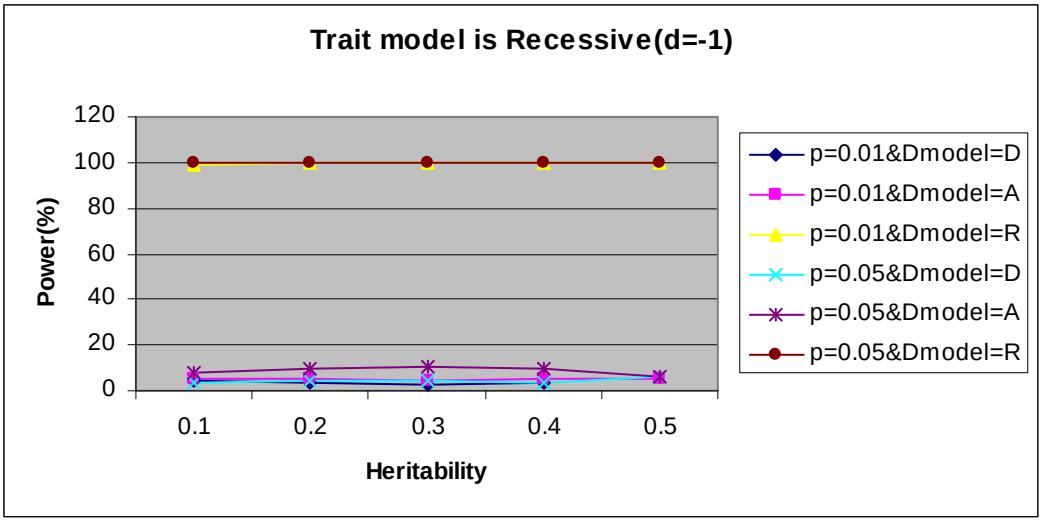
Note that p=endophenotype/allel frequency, Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model and R denotes recessive model.

Figure 4.2 Power comparison for p=0.01 and p=0.05 given trait model is dominant, f_{2D}=0.5, f_{0D}=0, θ=0.01 and n=100, N=1000



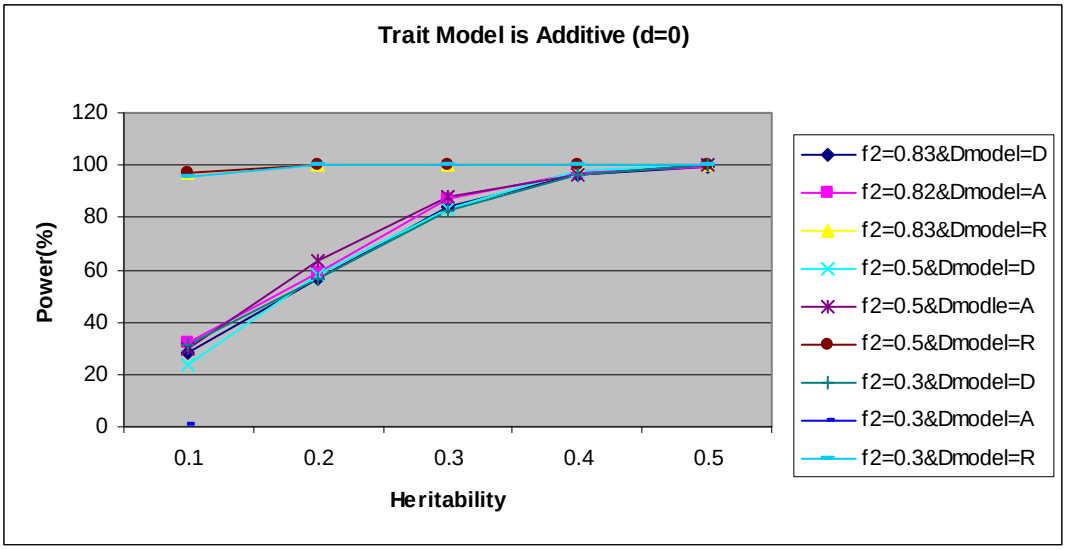
Note that p=endophenotype/allel frequency, Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model and R denotes recessive model.

Figure 4.3 Power comparison for p=0.01 and p=0.05 given trait model is recessive, f_{2D}=0.5, f_{0D}=0, θ=0.01 and n=100, N=1000



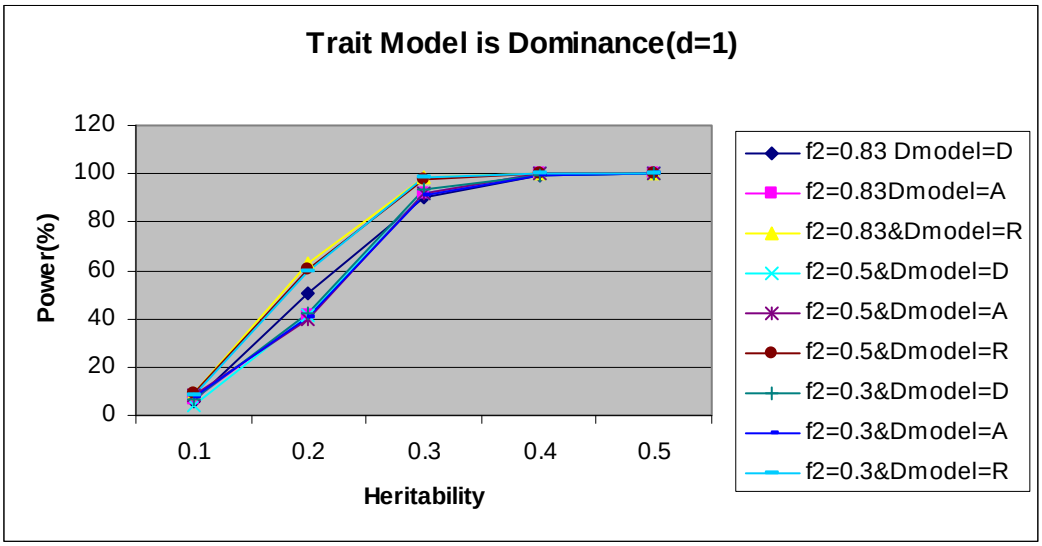
Note that p=endophenotype/allele frequency, Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model and R denotes recessive model.

Figure 4.4 Power comparison for f_{2D}=0.83, f_{2D}=0.5 and f_{2D}=0.3 given trait model is additive, p=0.05, f_{0D}=0, θ=0.01 and n=100, N=1000



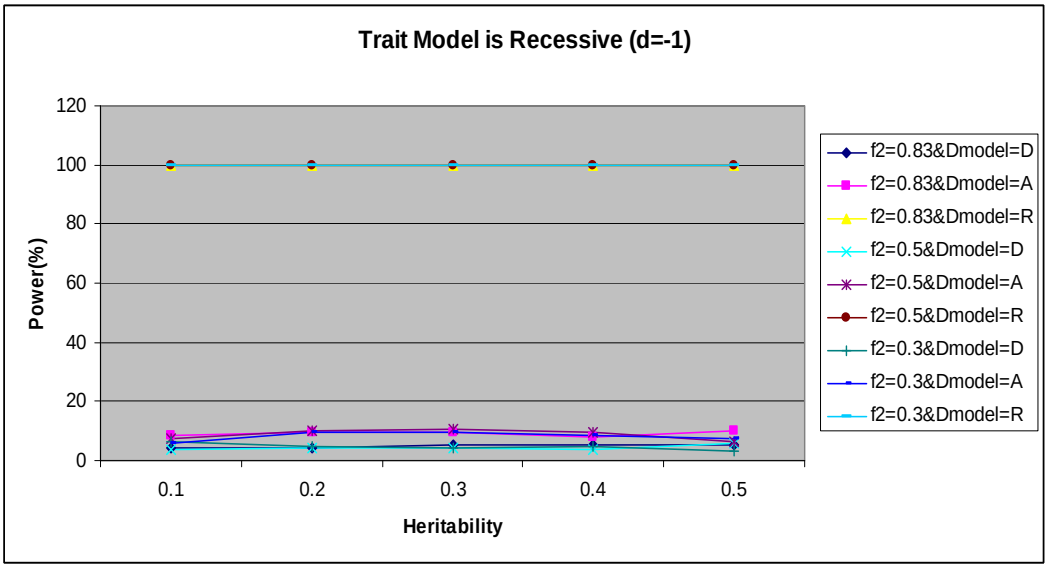
Note that f_2 is p(D+|AA), Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model and R denotes recessive model.

Figure 4.5 Power comparison for f_{2D}=0.83, f_{2D}=0.5 and f_{2D}=0.3 given trait model is dominant, p=0.05, f_{0D}=0, θ=0.01 and n=100, N=1000



Note that f_2 is p(D+|AA), Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model and R denotes recessive model.

Figure 4.6 Power comparison for f_{2D}=0.83, f_{2D}=0.5 and f_{2D}=0.3 given trait model is recessive, p=0.05, f_{0D}=0, θ=0.01 and n=100, N=1000



Note that f_2 is p(D+|AA), Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model and R denotes recessive model.

Results and discussion:

1). From Figure 4.1 to 4.6, we found that the power increases as the heritability increases, i.e. the trait value is an important factor, higher trait values will result in higher power if all other generate parameter values are equal. For example, for the model p=0.05, f_{2D}=0.05, trait model additive and disease model dominant, the power goes from 23.8%to 100% as the heritability goes from 0.1 to 0.5.

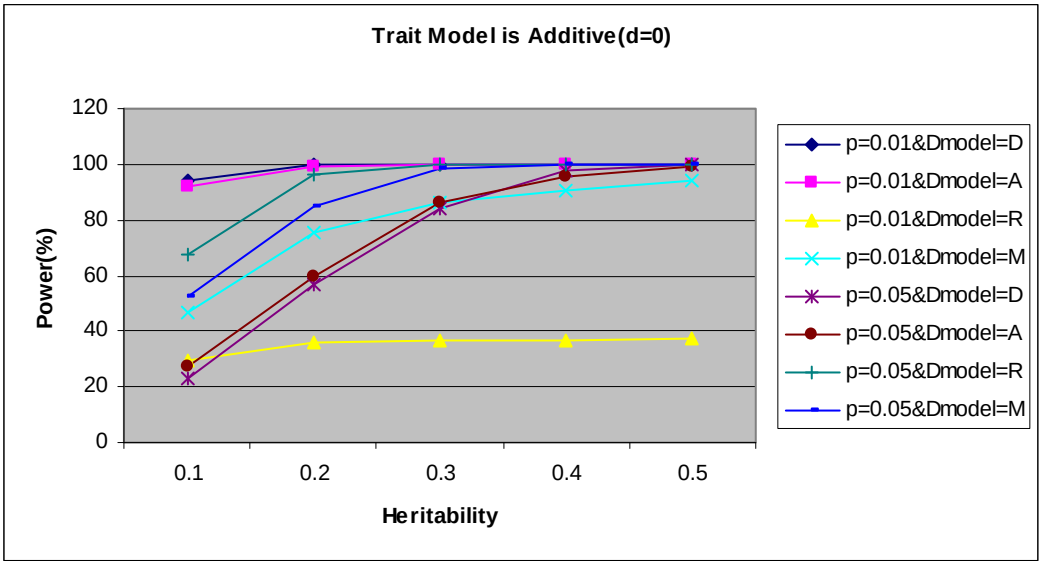
2). From Figure 4.4 to 4.6, we found that for given trait models (Additive, Dominant and Recessive) and disease models (Additive, Dominant, Recessive), we can see that for different f_{2D} (0.5, 0.83 and 0.3), the power for higher penetrance is only slightly higher in most cases if other generating parameters are same, but there is no significant difference among them. For example, given trait model additive, disease model dominant, p=0.05, the power for heritability from 0.1 to 0.5 and f_{2D}=0.5 are 23.8, 58.2, 83.6, 97.8, 100 and for f_{2D}=0.3 are 31.4, 56.2, 82.2, 96 and 100.

3). If trait model is dominant, that is, $\mu_1 = \mu_2$ and μ_0 =0 and f_{2D}=0.5, we found that the power for disease allele frequency 0.01 is higher than for 0.05. We found the similar results if the trait model additive, i.e., $\mu_1 = \frac{\mu_2 + \mu_0}{2}$. The power values for p=0.01 are all high (greater than 95 %) when the trait model is additive or dominant, you can see this from Figure 4.2 in which the line for the power is horizontal. But the Figure of the power for p=0.05 is not horizontal; it is dramatically increasing as the heritability goes up. It starts around 10% and goes up to 95% when the heritability goes up to 0.3 and the power get to 100% at when heritability is 0.5.

4) If the trait model is recessive, i.e., $\mu_1 = \mu_0$, the power is high only if the disease model is recessive for both disease allele frequencies (p=0.01 and p=.05) and different f_{2D} (0.83, 0.5 and 0.3). But the power is low (less than 20%) for other disease models (dominant and additive). From the Figure 4.3, we can see that the power for p=0.01 and p=0.05 is almost the same.

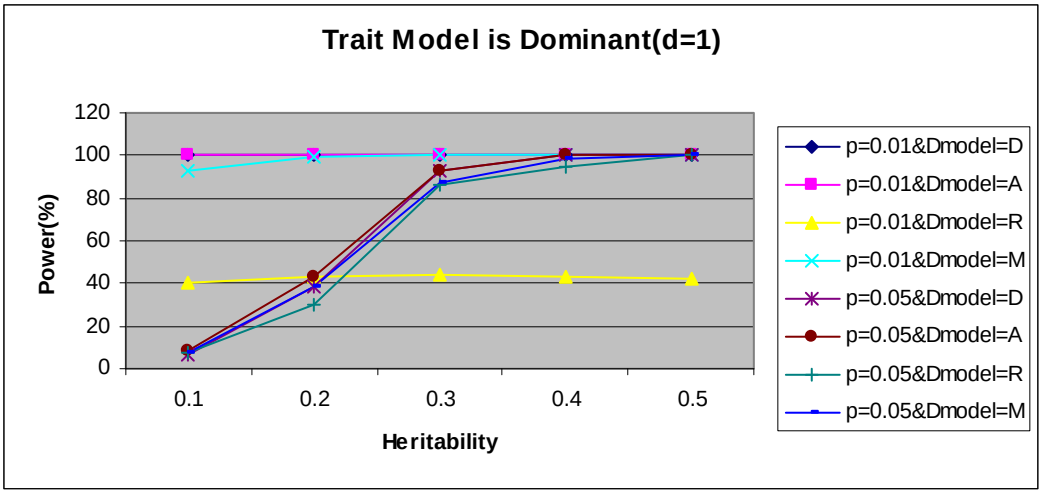
4.3 Power Comparison for f_{0D}=0.001

Figure4.7 Power comparison for p=0.01 and p=0.05 given trait model is additive, f_{2D}=0.5, f_{0D}=0.001, θ=0.01 and n=100, N=1000



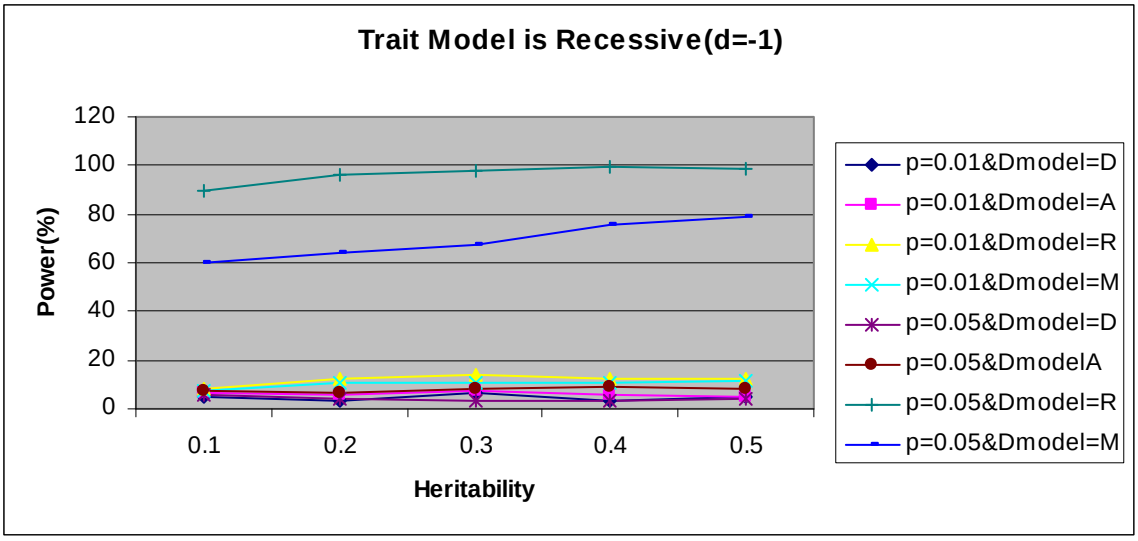
Note that p=endophenotype/allele frequency, Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model and M denotes multiplicative model.

Figure 4.8 Power comparison for p=0.01 and p=0.05 given trait model is dominant, f_{2D}=0.5, f_{0D}=0.001, θ=0.01 and n=100, N=1000



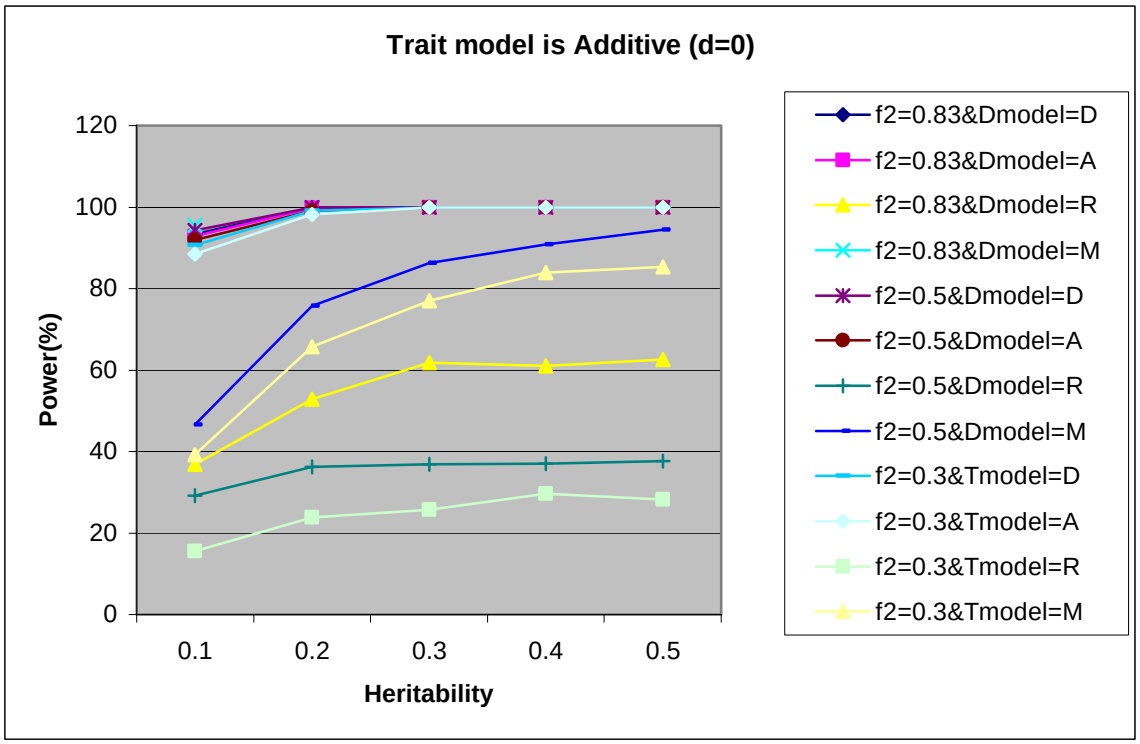
Note that p=endophenotype/allele frequency, Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model and M denotes multiplicative model.

Figure 4.9 Power comparison for p=0.01 and p=0.05 given trait model is recessive, f_{2D}=0.5, f_{0D}=0.001, θ=0.01 and n=100, N=1000



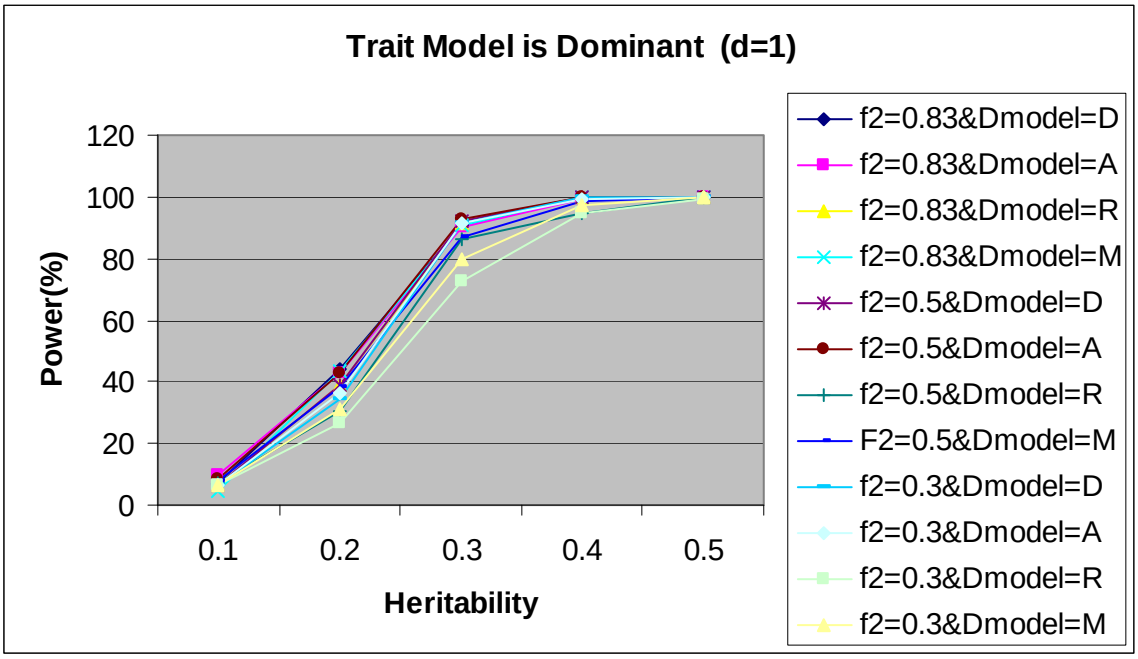
Note date p=endophenotype/allele frequency, Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model and M denotes multiplicative model.

Figure 4.10 Power comparison for f_{2D}=0.83, f_{2D}=0.5 and f_{2D}=0.3 given trait model is additive, p=0.05, f_{0D}=0.001, θ=0.01 and n=100, N=1000



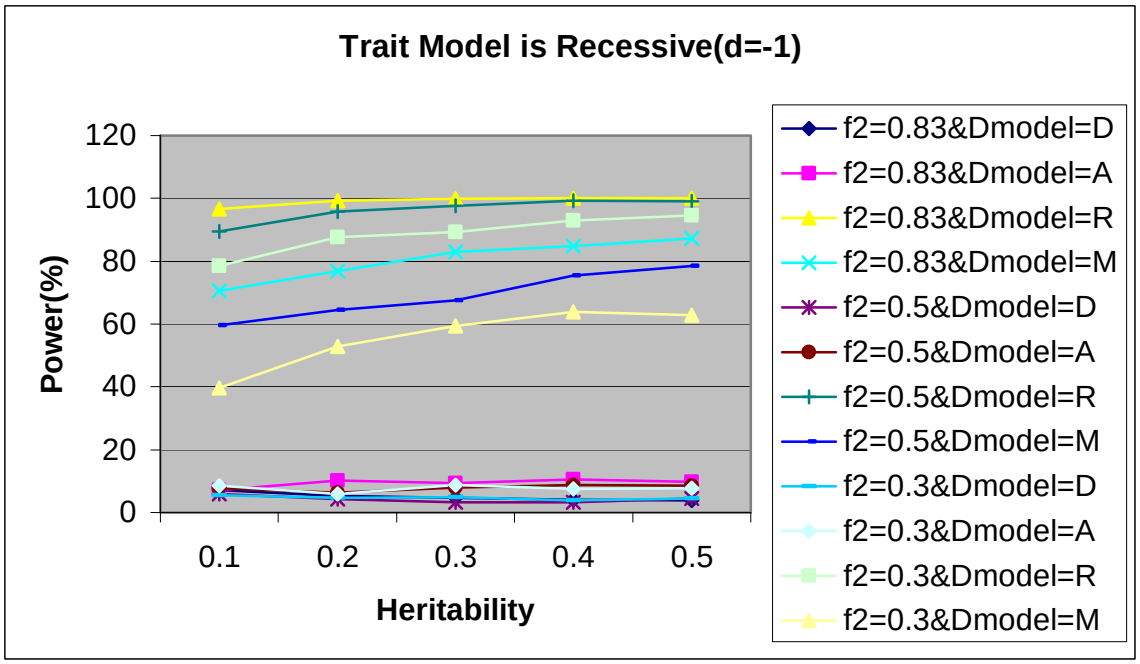
Note that f2=p(D+|AA), Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model and M denotes multiplicative model.

Figure 4.11 Power comparison for f_{2D}=0.83, f_{2D}=0.5 and f_{2D}=0.3 given trait model is dominant, p=0.05, f_{0D}=0.001, θ=0.01 and n=100, N=1000



Note that f2=p(D+|AA), Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model and M denotes multiplicative model.

Figure 4.12 Power comparison for f_{2D}=0.83, f_{2D}=0.5 and f_{2D}=0.3 given trait model is recessive, p=0.05, f_{0D}=0.001, θ=0.01 and n=100, N=1000



Note that f2=p(D+|AA), Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model and M denotes multiplicative model.

Results and Discussions

The results from considering a situation where there are disease phenocopies, i.e. $f_{0D}>0$ showed essentially the same trends as those where we have a 0.0 phenocopy rate for the disease/endophenotype locus. Our major observations from Figure 4.7 -4.12 are as follows:

1) The power goes up as the quantitative trait heritability goes up, which means that the trait value is an important factor. For example, given that the trait model is additive , p=0.01 and disease model is log-additive, we can see that power increases from 46.6 to 75.8, 86.2, 90.8, and 94.4 as the quantitative trait heritability goes up from 0.1 to 0.2,0.3,0.4,0.5.

2) The value of f_{2D} has little or no effect on the power. This is the especially if the trait model is dominant, as we can see in Figure 4.11.

3) The power for p=0.01 is higher than for p=0.05 in most cases. If the trait model is additive, the disease model is dominant or additive and p=0.01 then the power is uniformly high regardless of the (greater than 90%). However in these situations if p=0.05the power increases from around 30% to 100% as the heritability goes up from 0.1 to 0.5. If the disease model is log-additive then the decrease in disease allele frequency from 0.05 to 0.01 has little effect on the power. In both cases, power increases from around 50 % to 100%. But for disease model is recessive, the power for p=0.01 does not change a lot, it's all around 30 % while for p=0.05 the power increase from 50% to 100%.

4) If trait model is dominant, and the disease model is additive we can see from the Figure 4.11, that the power for p=0.01 and disease model is high regardless of the heritability values considered. The power for p=0.01 and disease model is recessive is

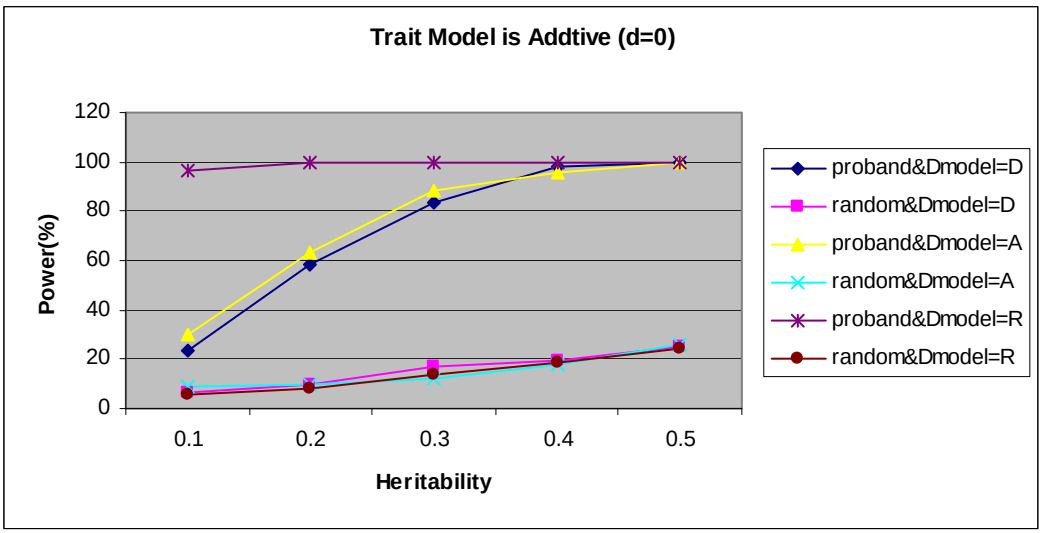
horizontal too and round 40 %. The power for other models (p=0.05 and disease model is dominant) increase as the quantitative trait heritability goes up.

5). If trait model is recessive, we found the power for all disease models (dominant, additive, recessive and log-additive) are low (less than 15 %) for disease allele frequency p=0.01. But for disease allele frequency p=0.05, the power in the case where the disease model is dominant and additive is low as well, less than 15%. However with a recessive trait model and a recessive disease or a log-additive disease model, the power is high and with the power for a recessive disease model being a bit higher than log-additive model. For example, for the case f_{2D}=0.5, the power values if both the trait and the disease models are recessive are 89.4, 95.8, 97.6, 99.2, and 99 as the quantitative trait heritability goes up from 0.1 to 0.5 and for this same situation, the power values for the log-additive model are 59.6, 64.4, 67.6, 75.4 and 78.6.

4.4 Power Comparison for Study Design: Random vs. Selected

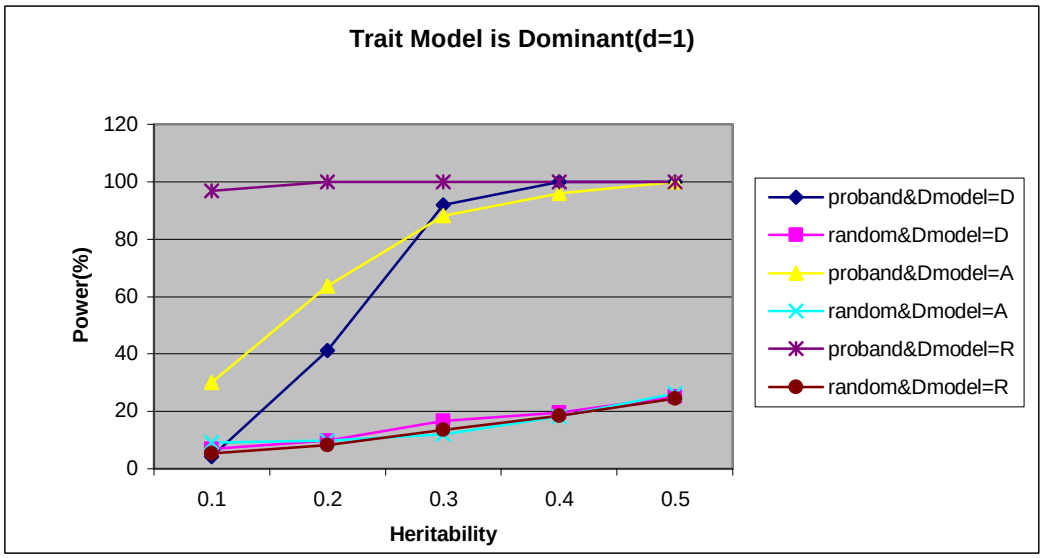
This study focuses on a selected sample of sib pairs with at least one affected proband. The traditional power studies have focused on random sib pairs and other selected study designs. Here we compare the power for two different study designs. Figure 4.13 to Figure 4.15 shows a power comparison for the linkage analysis of a quantitative endophenotype when based on a random sample of sib pairs as compared to a sample in which sib pairs are selected so as to have at least one disease affected.

Figure 4.13 Power comparison for random sib-pairs and selected sib-pairs given trait model is additive, p=0.05, f_{2D}=0.5&f_{0D}=0, θ=0.01 and n=100, N=1000



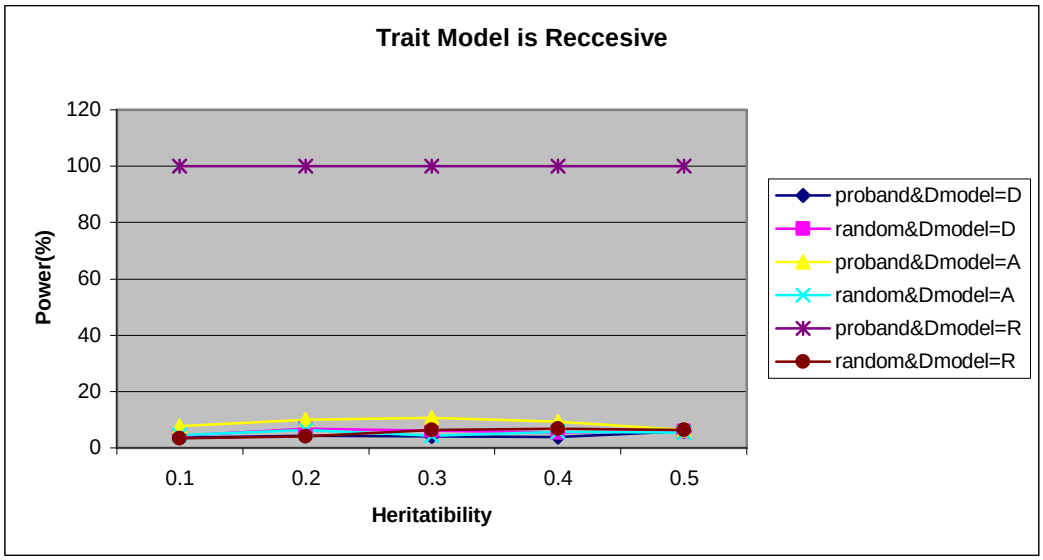
Note that proand indicates the each of the 100 sib pairs analyzed consisted of one disease affected proband and a randomly chosen sib. Random indicates that the 100 sib pairs analyzed consisted of 100 randomly chosen sib pairs. Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model.

Figure 4.14 Power comparison for random sib-pairs and selected sib-pairs given trait model is dominant, p=0.05, f_{2D}=0.5&f_{0D}=0, θ=0.01 and n=100, N=1000



Note that proand indicates the each of the 100 sib pairs analyzed consisted of one disease affected proband and a randomly chosen sib. Random indicates that the 100 sib pairs analyzed consisted of 100 randomly chosen sib pairs. Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model.

Figure 4.15 Power comparison for random sib-pairs and selected sib-pairs given trait model is recessive, p=0.05, f_{2D}=0.5&f_{0D}=0, θ=0.01 and n=100, N=1000



Note that proand indicates the each of the 100 sib pairs analyzed consisted of one disease affected proband and a randomly chosen sib. Random indicates that the 100 sib pairs analyzed consisted of 100 randomly chosen sib pairs. Dmodel denotes disease model, D denotes dominant disease model, A denoted additive model, R denotes recessive model.

.

Results and Discussions

The results from comparing the power for the proband sib pairs are greater than that for the random sib pairs. Our major observations from Figure 4.13 -4.15 are as follows:

1). From Figure 4.13 and Figure 4.14, we can see that for given trait model is additive or dominant, the power for the proband sib-pair increases dramatically as the quantitative trait values goes up. The power excess 80% when the heritability is 0.3. For the random sib-pair sample, the power increases as the heritability goes up, but very slow, the power for each the disease model is around 20% when the quantitative trait heritability is 0.5.

2). From Figure 4.15, for given the trait model is recessive, we can the power for the disease model is recessive, the power is a straight line and on the top for the proband sib-pairs, for other models either for random sib-pair or proband sib-pair, the power variants around 10%, it almost is a straight line and at the bottom.

3). From the simulation results above, we can see that the power for the sib-pairs including proband has a much higher power than random sib-pairs in most cases. To confirm this, we used ttest and found there is a significant difference (p_value=0.001) between them.

Chapter 5

Regression Analysis

5.1 Regression Analysis

In this chapter, we carried out regression analyses to see which factors and interactions have significant effects on the power. Since there are 3 disease dominance values if the penetrance for the BB genotype when f_{0D} equals 0 and there are 4 values of penetrance for the heterozygote wherever $f_{0D} \geq 0$, we ran the analyses separately. We consider the mean of the T-value instead of the observed power because it can be estimated with greater precision. For example, the power for the model (trait model=Additive, disease model is dominant, p=0.01 and f_{2D} =0.5 and f_{0D} =0) is 95.2, 100, 100, 100, 100 as the quantitative trait heritability goes from 0.1 to 0.5, yet from the average of the Z values, are -3.25, -4.78,-5.86, -6.53 and -7.01 respectively. We can calculate the approximate power at type I error level α using:

$$Z_\beta = \frac{E(Z) - Z_\alpha}{\sigma} = E(Z) - Z_\alpha$$

Power=1- β . Thus the average value of Z is directly related to power.

Below is the result of the regression Z values on different genetic generating parameters values.

5.2 Regression Analysis for $f_{0D}=0$

The analysis was first done using all main effects and all two, three and four way interactions. The variables considered are trait model donated as 'tmodel', disease model, denoted as 'dmodel', allele frequency, p, heritability of the quantitative trait, $h^{2,}$ and penetrance of the disease in individuals with genotype AA, f_2. For tmodel and dmodel we generated two dummy variables as follows: TA=1 if the trait model is additive, that is if tmodel='A', TR=1 if tmodel='R' i.e., the trait model is recessive. Similarly DA=1 if the disease model is additive, i.e. if dmodel='A' and DR=1 if dmodel='R'. Appendix Table A5.1 shows the ANOVA table and the estimates of the coefficients of these variables and the interaction terms. In this model, the sum of square of error is 17.5, the mean square of error is 0.16 and the adjusted R^2=0.99. The overall F statistic is significant (F=232.23, p < 0.0001). The five-way interactions, four-way Interactions and three-way interactions are not significant (p > 0.05). The two-way interactions of pxTA and pxTR are significant. The main effect of p, TA and TR are significant. The main effects of f2, h2, DA and DR are not significant. So we used the backward deletion method to determine the significant factors and interactions and develop a final model. The results are shown in Table 5.1.

Table 5.1 shows the ANOVA table and parameter estimates.

Dependent Variable: Z					
Analysis of Variance					
Source	DF	Sum of Square	Mean Square	F Value	Pr > F
Model	15	2625.80	175.05	422.00	<.0001
Error	164	68.03	0.41		
Corrected Total	179	2693.83			

Root MSE	0.64	R-Square	0.9747
Dependent Mean	-4.79	Adj R-Sq	0.9724
Coeff Var	-13.43		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.70817	0.2742	-24.46	<.0001
P	1	129.3222	4.49054	28.8	<.0001
h2	1	-8.125	0.72009	-11.28	<.0001
TA	1	2.81012	0.38103	7.37	<.0001
TR	1	6.92812	0.38103	18.18	<.0001
DR	1	-2.3915	0.44096	-5.42	<.0001
TAxDR	1	-0.66287	0.585	-1.13	0.2588
TRxDR	1	-2.10087	0.585	-3.59	0.0004
h2xTA	1	0.04875	1.01836	0.05	0.9619
h2xTR	1	7.80125	1.01836	7.66	<.0001
h2xDR	1	-3.7775	1.24723	-3.03	0.0029
pxTA	1	-52.4083	5.8795	-8.91	<.0001
pxTR	1	-144.433	5.8795	-24.57	<.0001
pxDR	1	44.60833	5.09179	8.76	<.0001
h2xTAxDR	1	-8.59375	1.76385	-4.87	<.0001
h2xTRxDR	1	-8.07875	1.76385	-4.58	<.0001

Note that p is endophenotype/allele frequency, h2 is the trait heritability, TA denotes the trait model is additive, TR denotes the trait model is recessive, DA denoted the disease model is additive and DR denotes the disease model is recessive.

The overall F statistic is still significant (F=422.00, p<0.0001) and the adjusted R^2=0.97. The fitted model is

$$Z=-6.71+129.32*p-8.13*h2$$
$$+2.81*TA+6.93*TR- 2.39*DR$$
$$+0.05*h2*TA+7.8* h2xTR -3.78* h2xDR$$
$$-52.41*pxTA -144.43*pxTR +44.61*pxDR$$
$$-8.59*h2xTAxDR-8.08*h2xTRxDR-0.66*TAxDR-2.10*TRxDR \quad (5.1)$$

Where p, h2, TA, TR, DA and DR are defined above.

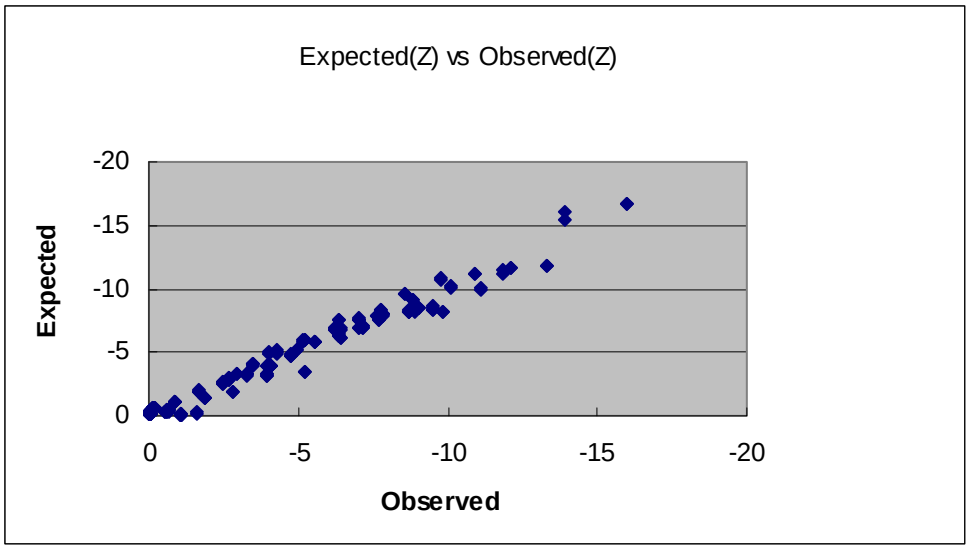
From the fitted model, we can get the different regression equations for the various combinations of trait model and disease model. Table 5.2 shows the regression coefficients for p and h2 for predicting the mean Z value for different models and Figure 5.1 shows the comparison of the observed average Z value to the predicted "expected" mean Z value. From Figure 5.1, we can see that these regression models fit fairly since all points are close to the line y=x.

Table 5.2 The estimated regression coefficients for obtaining the mean value of Haseman-Elston statistic for different models.

Tmodel	Dmodel	Intercept	px10	h2
D	D	-6.71	12.93	-8.13
D	A	-6.71	12.93	-8.13
D	R	-9.1	17.39	-11.91
A	D	-3.9	7.69	-8.08
A	A	-3.9	7.69	-8.08
A	R	-6.95	12.15	-20.45
R	D	0.22	-1.51	-0.33
R	A	0.22	-0.51	-0.33
R	R	-4.75	2.95	-12.19

Note that p is the disease allele frequency, h2 is the trait heritability, Tmodel='A' denotes the trait model is additive, Tmodel='D' denotes the trait model is dominant, Tmodel='R' denotes the trait model is recessive, Dmodel='A' denotes the disease model is additive, Dmodel='D' denotes the disease model is dominant and Dmodel='R' denotes the disease model is recessive.

Figure 5.1 Mean Z value (Haseman-Elston statistic): Comparison of predicted expected value and observed (simulated) average value.



5.3 Regression Analysis for f_{0D}=0.001

The statements begin the analysis including the entire main factors and all the interactions. Appendix Table A5.2 shows the ANOVA table and parameter estimates. In this model, we created one more dummy variable for disease model, DM=1 if dmodel='M' that is when the disease model is Multiplicative (Log-additive). From the results, we can see that sum of square of error is 4.29, the mean square error is 0.03 and the adjust R^2=0.99.The overall F statistic is significant (F=324.11, $p < 0.0001$). All the five-way interactions and four-way Interactions are not significant ($p > 0.05$). The three-way interactions only pxh2xTA and pxh2xTR are significant. The two-way interactions between disease model and trait model TDxDR, TRxDM, pxh2, pxTA, pxTR, pxDR, pxDM are significant. The main effects of p, TA, TR, DR, DM are significant. So we refitted the model, the results is shown in Table 5.3.

Table5.3 shows ANOVA table and parameter estimates for mean of Z

Dependent Variable: Z					
Analysis of Variance					
Source	DF	Sum of Square	Mean Square	F Value	Pr > F
Model	17	868.10	51.06	210.09	<.0001
Error	222	53.96	0.24		
Corrected Total	239	922.06			

Root MSE	0.49	R-Square	0.9415
Dependent Mean	-2.48	Adj R-Sq	0.9370
Coeff Var	-19.87		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-7.01544	0.24105	-29.1	<.0001
P	1	159.0667	6.65637	23.9	<.0001
h2	1	1.16375	0.70264	1.66	0.0991
TA	1	3.01056	0.32957	9.13	<.0001
TR	1	6.63406	0.33641	19.72	<.0001
DR	1	5.02425	0.15095	33.28	<.0001
DM	1	3.322	0.15095	22.01	<.0001
TRxDR	1	-3.8085	0.16536	-23.03	<.0001
TRxDM	1	-2.42525	0.16536	-14.67	<.0001
PxDR	1	-101.075	3.89757	-25.93	<.0001
PxDM	1	-64.1917	3.89757	-16.47	<.0001
pxh2	1	-252	19.48786	-12.93	<.0001
PxTA	1	-95.9063	9.14062	-10.49	<.0001
PxTR	1	-148.65	9.14062	-16.26	<.0001
h2xTA	1	-5.79125	0.99369	-5.83	<.0001
h2xTR	1	-1.61938	0.99369	-1.63	0.1046
pxh2xTA	1	202.625	27.56	7.35	<.0001
pxh2xTR	1	239.0625	27.56	8.67	<.0001

Note that p is endophenotype/allele frequency, h2 is the trait heritability, TA denotes the trait model is additive, TR denotes the trait model is recessive, DM denotes the disease model is multiplicative and DR denotes the disease model is recessive.

The overall F statistic is still significant (F=210.09, p<0.0001) and the adjusted R^2=0.94. The fitted final model is

$$\begin{aligned} Z= & -7.02+159.07*p+1.16*h2-252*pxh2 \\ & +3.01*TA+6.63*TR+5.02*DR+3.32*DM \\ & -3.81*TRxDR-2.43*TRxDM-101.08*pxDR-64.19*pxDM \\ & -95.91*pxTA-148.65*pxTR-5.79*h2xTA-1.62*h2xTR \\ & +202.63*pxh2xTA+239.06*pxh2xTR \end{aligned} \quad (5.2)$$

From the fitted model, we can get the different regression equations for different generating models. Table 5.4 shows the regression coefficient for the predicting the Z value for different models and Figure 5.2 shows the comparison of the predicted mean Z

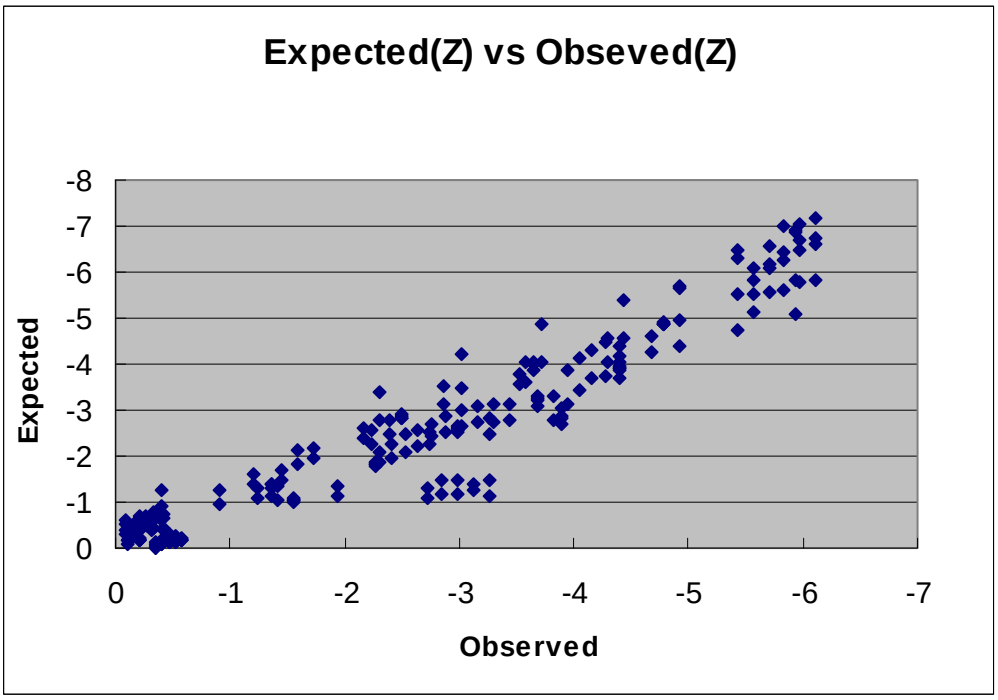
value expected and the observed average Z value in the simulations. From Figure 5.2, we can see that it's a good model and most points are close to the line y=x.

Table 5.4: Regression coefficients for predicting the mean value of the Haseman-Elston statistic based on allele frequency and trait heritability for different pleiotropic models.

Tmodel	Dmodel	intercept	px10	h2	pxh2x10
D	D	-7.02	15.91	1.16	-25.2
D	A	-7.02	15.91	1.16	-25.2
D	R	-2	-5.8	1.16	-25.2
D	M	-3.7	9.49	1.16	-25.2
A	D	-4.01	6.32	-4.63	-4.94
A	A	-4.01	6.32	-4.63	-4.94
A	R	1.01	-3.79	-4.63	-4.94
A	M	-0.69	-0.1	-4.63	-4.94
R	D	-0.39	1.04	-0.46	-1.3
R	A	-0.39	1.04	-0.46	-1.3
R	R	0.82	-9.07	-0.46	-1.3
R	M	-0.02	-5.38	-0.46	-1.3

Note that p is the disease allele frequency, h2 is the trait heritability, Tmodel='A' denotes the trait model is additive, Tmodel='D' denotes the trait model is dominant, Tmodel='R' denotes the trait model is recessive, Dmodel='A' denotes the disease model is additive, Dmodel='D' denotes the disease model is dominant, Dmodel='R' denotes the disease model is recessive and Dmodel='M' denoted the disease model is multiplicative.

Figure 5.2 Z value comparison: expected value vs. average simulated value



Chapter 6

Discussion and Future Work

6.1 Sample Selection

From our simulation results, we can see that the power of linkage analysis of quantitative trait endophenotype can be increased by using a selected sample of sib-pairs. That is, the power for random samples is much lower than the sample which includes at least one affected proband. We have investigated only one approach to linkage analysis specifically the Haseman-Elston analysis. Other statistical approaches to linkage analysis of endophenotypes such as variance components analysis and association analysis are likely to generate greater power using the selection approach we proposed. These can be studied as well using our simulation software.

References:

Alarcon M, Cantor RM, Liu J, Gillian TC, Geschwind DH. (2002): Evidence for a language trait locus on chromosome 7q I multiplex autism families. American Journal of Human Genetics. 70(1): 60-71.

Almasy L, Dyer TD, Blangero J. (1997): Bivariate quantitative trait linkage analysis: pleiotropic versus co-incident linkages. Genet Epidemiology. 14(6): 953-8.

Amos CI. , Elston RC., Bonney GE., Keats BJ. Berenson GS. (1990): A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. American Journal of Human Genetics. 47: 247-254.

Amos CI. (1994): Robust variance-components approach for assessing genetic linkage in pedigrees. American Journal of Human Genetics. 54(3): 535-543.

Arya R, Duggirala R, Almasy L, Rainwater DL, Mahaney MC, Cole S, Dyer TD, Williams K, Leach RJ, Hixson JE, MacCluer JW, O'Connell P, Stern MP, Blangero J. (2002): Linkage of high-density lipoprotein-cholesterol concentrations to a locus on chromosome 9p in Mexican Americans. Nature Genetics. 30: 102–105.

Blackwelder WC. (1977): Statistical methods for detecting genetic linkage from sibship data. Institute of Statistics Mimeo Series 1114.

Blangero J. (1995): Genetic analysis of a common oligogenic trait with quantitative correlates: Summary of GAW9 results. Genet Epidemiology. 12: 689-706.

Cuenco KT, Szatkiewicz JP, Feingold E. (2003): Recent advances in human quantitative-trait-locus mapping: comparison of methods fro selected sibling pairs. American Journal of Human Genetics. 73: 863-873.

Elston RC, Buxbaurn S, Jacobs KB, Olson JM. (2000): Haseman and Elston revisted. Genetic Epidemiology. 19: 1-17.

Goldberg TE, Green MF. (2002): Neurocognitive functioning in patients with schizophrenia: an overview, in Neuropsychopharmacology: The Fifth Generation of Progress. Edited by Davis KL, Charney DS, Coyle JT, Nemeroff C. Philadelphia, Lippincott Williams & Wilkins. pp 657-669.

Gottesman I, Gould T. (2003): The endophenotype concept in Psychiatry: Etymology and Strategic Intentions.

Haseman JK, Elston RC. (1972): The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 3:3-19.

Holzman PS, Kinney DK, Jacobsen B, Jansson L, Faber B, Hildebrand W, Kasell E, Zimbalist ME. (1997): Thought disorder in schizophrenic and control adoptees and their relatives. Archives of General Psychiatry. 54(5): 475-9.

Jayakar SD. (1970): On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. Biometrics. 26(3): 451-64.

Johannsen W. (1911): The genotype conception of heredity. American Naturalist 45(531): 129-159.

Kraft P, Bauman L, Cantor RM, Horvath S, Yuan JY. (2002): Multivariate quantitative trait linkage analysis of longitudinal blood pressure measurements. GAW 13 Participant Contributions: 293-297.

Levy DL, Holzman PS, Matthysse S, Mendell NR. (1993): Eye tracking dysfunction and schizophrenia: a critical perspective. Schizophr Bull. 19(4): 685.

Morton NE. (1955) Sequential tests for the detection of linkage. American Journal of Human Genetics 7**.** pp. 277-318.

Olswold C, de Andrade M. (2003): Comparison of longitudinal variance components and regression-based approached for linkage detection on chromosome 17 for systoic blood pressure. BMC Genetics. 4: S17.

Ott J. (1997): Analysis of human genetic linkage, third edition, Baltimore and London, The Johns Hopkins University Press.

Penrose, LS. (1938): Genetic Linkage in graded Human Characters. Ann. Engen. 8: 233-237.

Pankratz MJ, Zinke I, Schutz CS, Katzenberger JD, Bauer M (2002): Nutrient control of gene expression in Drosophila: microarray analysis of starvation and sugar-dependent response. EMBO journal. 21: 6162-6173.

Rao S, Li X. (2000): Strategies for genetic mapping of categorical traits. Genetica. 109: 183-197.

Sham PC, Purcell S. (2001): Equivalence between Haseman-Elston and Variance-Components linkage analyses for sib pairs. American Journal of Human Genetics. 68: 1527-1532.

Sing CF, Boerwinkle EA. (1987): Genetic architecture of inter-individual variability in apolipoprotein, lipoprotein and lipid phenotypes. Ciba Found Symp. 130:99-127.

Sung H. (2005): Power Calculation for Linkage Analysis of a Disease Related Trait. Ph.D. Dissertation.

Ulgen A. (2001): Statistical Properties of 3 Model Based Tests for Linkage. Ph.D. Dissertation.

Williams GC. (1957): Pleiotropy, natural selection, and the evolution of senescence. Evolution 11: 398-411.

Wright FA. (1997): The phenotypic difference discards sibpair QTL linkage information. American Journal of Human Genetics. 60: 740-742.

Xu X, Weiss S, Xu X, Wei LJ. (2000): A unified Haseman-Elston method for testing linkage with quantitative traits. American Journal of Human Genetics. 67: 1025-1028.

Appendix:

Appendix Table A1.1 T-value based on a particular simulation model (p=0.01, q=0.99, f_{AA}=0.5, f_{AB}=0.25, f_{BB}=0, z_0=0, z_1=2, z_2=4).

-2.56713	-3.21316	-3.17492	-2.87639	-3.35963	-1.65019	-2.09628
-2.80922	-3.24851	-3.69736	-3.9387	-3.47578	-3.93237	-2.50926
-3.77627	-4.21305	-2.06752	-1.85226	-2.90998	-2.14643	-2.46431
-4.14368	-2.60941	-2.99032	-2.72398	-1.73651	-1.89192	-3.46352
-3.47074	-3.2976	-1.88292	-3.09127	-1.55	-3.24751	-1.99104
-1.23574	-3.29563	-2.94044	-2.13032	-2.87098	-2.50369	-3.21095
-1.94212	-4.30661	-3.00031	-4.15056	-2.45279	-1.78194	-2.0228
-2.02564	-3.13499	-4.41064	-1.63671	-3.4178	-2.60192	-2.06675
-2.92414	-4.07143	-2.06494	-3.31607	-1.9794	-2.58062	-2.34036
-2.46406	-1.28652	-2.13611	-3.05888	-2.21638	-2.8425	-2.60882
-3.04074	-2.78188	-3.42644	-3.28118	-4.05328	-2.89153	-4.30098
-2.03794	-1.53771	-2.68968	-2.39289	-2.2802	-3.54955	-2.49937
-3.36501	-3.07525	-1.62436	-2.39179	-2.7288	-1.9018	-2.8883
-2.25108	-2.71529	-3.3466	-1.98527	-1.84702	-2.38616	-1.99121
-2.20849	-2.78427					

Appendix Table A2.1 simulated pedigree data based on a particular simulation model ($p=0.01$, $q=0.99$, $f_{AA}=0.5$, $f_{AB}=0.25$, $f_{BB}=0$, $z_0=0$, $z_1=2$, $z_2=4$).

familyID	individualID	fatherID	motherID	sex	GD[0]	GD[1]	GM[0]	GM[1]	PT	GD[f2]	GM[f2]
1	3	1	2	2	12	22	43	12	3.353604	12	41
1	4	1	2	1	12	22	43	12	2.839293	12	42
2	3	1	2	1	12	22	23	41	3.451288	12	21
2	4	1	2	2	12	22	23	41	1.130917	12	24
3	3	1	2	1	12	22	32	41	1.923809	12	31
3	4	1	2	2	12	22	32	41	5.033534	12	31
4	3	1	2	2	12	22	13	24	1.664957	12	14
4	4	1	2	1	12	22	13	24	1.596932	12	12
5	3	1	2	2	12	22	34	12	2.420756	12	31
5	4	1	2	2	12	22	34	12	-0.89957	22	42
6	3	1	2	1	12	22	31	24	5.204206	12	32
6	4	1	2	1	12	22	31	24	-0.22507	22	12
7	3	1	2	1	12	22	42	31	2.613266	12	43
7	4	1	2	1	12	22	42	31	3.082184	12	41
8	3	1	2	1	12	22	24	31	2.325471	12	23
8	4	1	2	2	12	22	24	31	4.077521	12	23
9	3	1	2	2	12	22	24	13	3.776028	12	23
9	4	1	2	2	12	22	24	13	0.957283	22	41
10	3	1	2	2	12	22	24	31	2.838258	12	21
10	4	1	2	1	12	22	24	31	2.191085	12	21

Note that GD[0] is father's PL genotype, GD[1] is mother's PL genotype, GM[0] is father's marker genotype, GM[1] is mother's marker genotype, GD[f2] is offsping's PL genotype, GM[f2] is offspring's marker genotype, IndividualID=3 is proband, individualID=4 is sibling.

Appendix Table A4.1 shows the simulation results of the power for gene frequency p=0.01, f0=0 and θ=0.01, α=0.05, n=100, N=1000

p	f2	f0	h2	Trait model	Disease model	Power
0.01	0.5	0	0.1	A	D	95.2
0.01	0.5	0	0.1	A	A	96.2
0.01	0.5	0	0.1	A	R	100
0.01	0.5	0	0.2	A	D	100
0.01	0.5	0	0.2	A	A	100
0.01	0.5	0	0.2	A	R	100
0.01	0.5	0	0.3	A	D	100
0.01	0.5	0	0.3	A	A	100
0.01	0.5	0	0.3	A	R	100
0.01	0.5	0	0.4	A	D	100
0.01	0.5	0	0.4	A	A	100
0.01	0.5	0	0.4	A	R	100
0.01	0.5	0	0.5	A	D	100
0.01	0.5	0	0.5	A	A	100
0.01	0.5	0	0.5	A	R	100

Note that p is disease allele frequency, f2=p(D+|AA), f0=p(D+|BB), h2 is trait heritability, Trait model='A' denoted the trait model is additive and Disease model ='A' denotes the disease model is additive, Disease model='D' denotes the disease model is dominant, Disease model ='R' denotes the disease model is recessive.

Appendix Table A4.2 shows the simulation results of the power for p=0.01, f0=0.001, θ=0.01 and n=100, N=1000

p	f2	f0	h2	Trait model	Disease model	Power
0.01	0.5	0.001	0.1	A	D	94.2
0.01	0.5	0.001	0.1	A	A	92
0.01	0.5	0.001	0.1	A	R	29.2
0.01	0.5	0.001	0.1	A	M	46.6
0.01	0.5	0.001	0.2	A	D	100
0.01	0.5	0.001	0.2	A	A	99.2
0.01	0.5	0.001	0.2	A	R	36.2
0.01	0.5	0.001	0.2	A	M	75.8
0.01	0.5	0.001	0.3	A	D	100
0.01	0.5	0.001	0.3	A	A	100
0.01	0.5	0.001	0.3	A	R	36.8
0.01	0.5	0.001	0.3	A	M	86.2
0.01	0.5	0.001	0.4	A	D	100
0.01	0.5	0.001	0.4	A	A	100
0.01	0.5	0.001	0.4	A	R	37
0.01	0.5	0.001	0.4	A	M	90.8
0.01	0.5	0.001	0.5	A	D	100
0.01	0.5	0.001	0.5	A	A	100
0.01	0.5	0.001	0.5	A	R	37.6
0.01	0.5	0.001	0.5	A	M	94.4

Note that p is disease allele frequency, f2=p(D+|AA), f0=p(D+|BB), h2 is trait heritability, Trait model='A' denoted the trait model is additive and Disease model ='A' denotes the disease model is additive, Disease model='D' denotes the disease model is dominant, Disease model ='R' denotes the disease model is recessive.

Appendix Table A4.3 shows the simulation results of the power for p=0.05, f0=0, θ=0.01 and n=100, N=1000

p	f2	f0	h2	Trait model	Disease model	Power
0.05	0.5	0	0.1	A	D	23.8
0.05	0.5	0	0.1	A	A	30
0.05	0.5	0	0.1	A	R	96.8
0.05	0.5	0	0.2	A	D	58.2
0.05	0.5	0	0.2	A	A	63.6
0.05	0.5	0	0.2	A	R	100
0.05	0.5	0	0.3	A	D	83.6
0.05	0.5	0	0.3	A	A	88.2
0.05	0.5	0	0.3	A	R	100
0.05	0.5	0	0.4	A	D	97.8
0.05	0.5	0	0.4	A	A	96
0.05	0.5	0	0.4	A	R	100
0.05	0.5	0	0.5	A	D	100
0.05	0.5	0	0.5	A	A	100
0.05	0.5	0	0.5	A	R	100

Note that p is disease allele frequency, f2=p(D+|AA), f0=p(D+|BB), h2 is trait heritability, Trait model='A' denoted the trait model is additive and Disease model ='A' denotes the disease model is additive, Disease model='D' denotes the disease model is dominant, Disease model ='R' denotes the disease model is recessive.

Appendix Table A4.4 shows the simulation results of the power for p=0.05, f0=0.001, θ=0.01 and n=100, N=1000

p	f2	f0	h2	Trait model	Disease model	Power
0.05	0.5	0.001	0.1	A	D	22.8
0.05	0.5	0.001	0.1	A	A	27.4
0.05	0.5	0.001	0.1	A	R	67.4
0.05	0.5	0.001	0.1	A	M	52.4
0.05	0.5	0.001	0.2	A	D	56.8
0.05	0.5	0.001	0.2	A	A	59.4
0.05	0.5	0.001	0.2	A	R	96.4
0.05	0.5	0.001	0.2	A	M	84.6
0.05	0.5	0.001	0.3	A	D	84.4
0.05	0.5	0.001	0.3	A	A	86
0.05	0.5	0.001	0.3	A	R	99.8
0.05	0.5	0.001	0.3	A	M	98.2
0.05	0.5	0.001	0.4	A	D	98
0.05	0.5	0.001	0.4	A	A	95.4
0.05	0.5	0.001	0.4	A	R	100
0.05	0.5	0.001	0.4	A	M	99.6
0.05	0.5	0.001	0.5	A	D	100
0.05	0.5	0.001	0.5	A	A	99.2
0.05	0.5	0.001	0.5	A	R	100
0.05	0.5	0.001	0.5	A	M	100

Note that p is disease allele frequency, f2=p(D+|AA), f0=p(D+|BB), h2 is trait heritability, Trait model='A' denoted the trait model is additive and Disease model ='A' denotes the disease model is additive, Disease model='D' denotes the disease model is dominant, Disease model ='R' denotes the disease model is recessive and Disease model ='M' denotes the disease model is multiplicative.

Appendix Table A5.1 ANOVA table and parameter estimates

The Reg Procedure					
Dependent Variable: Z					
Analysis of Variance					
Source	DF	Sum of Square	Mean Square	F Value	Pr > F
Model	71	2676.30	37.69	232.23	<.0001
Error	108	17.53	0.16		
Corrected Total	179	2693.83			

Root MSE	0.40	R-Square	0.9935
Dependent Mean	-4.79	Adj R-Sq	0.9892
Coeff Var	-8.40		

PTAxDTRxDAmeter Estimates					
VTAxDRiable	DF	PTAxDTRxDAmeter Estimate	StandTAxDRd ETRxDRor	t Value	Pr > |t|
Intercept	1	-8.33112	1.57039	-5.31	<.0001
p	1	183.2125	43.55479	4.21	<.0001
f2	1	-0.79875	3.80876	-0.21	0.8343
h2	1	-1.98375	4.7349	-0.42	0.6761
TA	1	5.24712	2.22087	2.36	0.0199
TR	1	8.44225	2.22087	3.8	0.0002
DA	1	-0.08575	2.22087	-0.04	0.9693
DR	1	-3.809	2.22087	-1.72	0.0892
pxTA	1	-134.813	61.59578	-2.19	0.0308
pxTR	1	-197.975	61.59578	-3.21	0.0017
pxDA	1	8.875	61.59578	0.14	0.8857
pxDR	1	93.5	61.59578	1.52	0.1319
h2xTA	1	-7.65875	6.69617	-1.14	0.2553
h2xTR	1	1.45	6.69617	0.22	0.829
h2xDA	1	0.2825	6.69617	0.04	0.9664
h2xDR	1	0.14	6.69617	0.02	0.9834
f2xTA	1	0.36625	5.38639	0.07	0.9459
f2xTR	1	0.525	5.38639	0.1	0.9225
f2xDA	1	0.5175	5.38639	0.1	0.9236
f2xDR	1	0.4475	5.38639	0.08	0.9339
pxh2xTA	1	269.875	185.7183	1.45	0.1491
pxh2xTR	1	220.5	185.7183	1.19	0.2377
pxh2xDA	1	-30.25	185.7183	-0.16	0.8709
pxh2xDR	1	-140	185.7183	-0.75	0.4526
pxf2xTA	1	-3.625	149.3917	-0.02	0.9807
pxf2xTR	1	-5	149.3917	-0.03	0.9734
pxf2xDA	1	-29.75	149.3917	-0.2	0.8425
pxf2xDR	1	-17.75	149.3917	-0.12	0.9056

h2xf2xTA	1	-1.6875	16.24059	-0.1	0.9174
h2xf2xTR	1	-0.525	16.24059	-0.03	0.9743
h2xf2xDA	1	-2.175	16.24059	-0.13	0.8937
h2xf2xDR	1	-0.975	16.24059	-0.06	0.9522
TAXDA	1	0.04188	3.14078	0.01	0.9894
TAXDR	1	1.22413	3.14078	0.39	0.6975
TRXDA	1	0.00963	3.14078	0	0.9976
TRXDR	1	1.22288	3.14078	0.39	0.6978
pxf2	1	25.875	105.6359	0.24	0.807
pxh2	1	-197.625	131.3226	-1.5	0.1353
f2xh2	1	1.4875	11.48383	0.13	0.8972
pxh2xf2	1	-58.75	318.5042	-0.18	0.854
pxTAXDA	1	-2.2875	87.10958	-0.03	0.9791
pxTAXDR	1	-66.4625	87.10958	-0.76	0.4471
pxTRXDA	1	-2.8125	87.10958	-0.03	0.9743
pxTRXDR	1	-113.288	87.10958	-1.3	0.1962
f2xTAXDA	1	-0.52375	7.61751	-0.07	0.9453
f2xTAXDR	1	0.31875	7.61751	0.04	0.9667
f2xTRXDA	1	-0.12125	7.61751	-0.02	0.9873
f2xTRXDR	1	0.78875	7.61751	0.1	0.9177
h2xTAXDA	1	0.46375	9.46981	0.05	0.961
h2xTAXDR	1	-15.3513	9.46981	-1.62	0.1079
h2xTRXDA	1	-0.32375	9.46981	-0.03	0.9728
h2xTRXDR	1	-16.5338	9.46981	-1.75	0.0837
pxh2xf2xTA	1	8.75	450.4329	0.02	0.9845
pxh2xf2xTR	1	-2.5	450.4329	-0.01	0.9956
pxh2xf2xDA	1	107.5	450.4329	0.24	0.8118
pxh2xf2xDR	1	47.5	450.4329	0.11	0.9162
pxf2xTAXDA	1	8.375	211.2718	0.04	0.9685
pxf2xTAXDR	1	-7.375	211.2718	-0.03	0.9722
pxf2xTRXDA	1	0.625	211.2718	0	0.9976
pxf2xTRXDR	1	-24.875	211.2718	-0.12	0.9065
pxh2xTAXDA	1	-12.375	262.6453	-0.05	0.9625
pxh2xTAXDR	1	241.625	262.6453	0.92	0.3596
pxh2xTRXDA	1	6.875	262.6453	0.03	0.9792
pxh2xTRXDR	1	297.375	262.6453	1.13	0.26
f2xh2xTAXDA	1	0.7125	22.96766	0.03	0.9753
f2xh2xTAXDR	1	-2.0875	22.96766	-0.09	0.9277
f2xh2xTRXDA	1	1.0375	22.96766	0.05	0.9641
f2xh2xTRXDR	1	-2.9875	22.96766	-0.13	0.8967
pxh2xf2xTAXDA	1	-1.25	637.0083	0	0.9984
pxh2xf2xTAXDR	1	43.75	637.0083	0.07	0.9454
pxh2xf2xTRXDA	1	-28.75	637.0083	-0.05	0.9641
pxh2xf2xTRXDR	1	58.75	637.0083	0.09	0.9267

Note that p is the disease allele frequency, h2 is the trait heritability, Tmodel='A' denotes the trait model is additive, Tmodel='D' denotes the trait model is dominant, Tmodel='R'

denotes the trait model is recessive, Dmodel='A' denotes the disease model is additive, Dmodel='D' denotes the disease model is dominant, Dmodel='R' denotes the disease model is recessive.

Appendix Table A5.2 shows ANOVA table and pTAxDTRxDAmeter estimates

The Reg Procedure					
Dependent VTAxDRiable: Z					
Analysis of VTAxDRiance					
Source	DF	Sum of SquTAxDRe	Mean SquTAxDRe	F Value	Pr > F
Model	95	917.77	9.66	324.11	<.0001
ETRxDRor	144	4.29	0.03		
CoTRxDRected Total	239	922.06			

Root MSE	0.17	R-SquTAxDRe	0.9953
Dependent Mean	-2.48	Adj R-Sq	0.9923
Coeff VTAxDR	-6.96		

PTAxDTRxDTAxDMeter Estimates					
VTAxDRiable	DF	PTAxDTRxDTAxDMeter Estimate	StandTAxDRd ETRxDRor	t Value	Pr > |t|
Intercept	1	-6.81325	0.67296	-10.12	<.0001
p	1	161.725	18.66455	8.66	<.0001
f2	1	-1.8025	1.63217	-1.1	0.2713
h2	1	0.86	2.02905	0.42	0.6723
TA	1	3.61963	0.95171	3.8	0.0002
TR	1	6.80362	0.95171	7.15	<.0001
DA	1	0.83775	0.95171	0.88	0.3802
DR	1	5.66875	0.95171	5.96	<.0001
DM	1	4.034	0.95171	4.24	<.0001
TAXDA	1	-1.061	1.34592	-0.79	0.4318
TAXDR	1	-2.08712	1.34592	-1.55	0.1232
TRXDA	1	-0.62125	1.34592	-0.46	0.6451
TRXDR	1	-4.52087	1.34592	-3.36	0.001
TAXDM	1	-1.83213	1.34592	-1.36	0.1756
TRXDM	1	-3.54487	1.34592	-2.63	0.0094
pxf2	1	22.25	45.26818	0.49	0.6238
pxh2	1	-267	56.27573	-4.74	<.0001
f2xh2	1	-2.35	4.92117	-0.48	0.6337
pxh2xf2	1	65	136.4887	0.48	0.6346
pxTA	1	-110.013	26.39566	-4.17	<.0001
pxTR	1	-169.763	26.39566	-6.43	<.0001
pxDA	1	-22.225	26.39566	-0.84	0.4012

pxDR	1	-123.325	26.39566	-4.67	<.0001
pxDM	1	-84.45	26.39566	-3.2	0.0017
f2xTA	1	2.52375	2.30823	1.09	0.2761
f2xTR	1	1.94375	2.30823	0.84	0.4011
f2xDA	1	-0.27	2.30823	-0.12	0.907
f2xDR	1	0.195	2.30823	0.08	0.9328
f2xDM	1	-0.2225	2.30823	-0.1	0.9233
h2xTA	1	-3.73125	2.86951	-1.3	0.1956
h2xTR	1	-1.00125	2.86951	-0.35	0.7277
h2xDA	1	1.6825	2.86951	0.59	0.5586
h2xDR	1	1.395	2.86951	0.49	0.6276
h2xDM	1	0.9875	2.86951	0.34	0.7312
pxf2xTA	1	-24.875	64.01888	-0.39	0.6982
pxf2xTR	1	-20.375	64.01888	-0.32	0.7507
pxf2xDA	1	16.5	64.01888	0.26	0.797
pxf2xDR	1	20	64.01888	0.31	0.7552
pxf2xDM	1	9.75	64.01888	0.15	0.8792
pxh2xTA	1	204.625	79.58591	2.57	0.0112
pxh2xTR	1	270.125	79.58591	3.39	0.0009
pxh2xDA	1	-18.75	79.58591	-0.24	0.8141
pxh2xDR	1	39	79.58591	0.49	0.6249
pxh2xDM	1	28.75	79.58591	0.36	0.7184
h2xf2xTA	1	-12.8875	6.95959	-1.85	0.0661
h2xf2xTR	1	1.8625	6.95959	0.27	0.7894
h2xf2xDA	1	-3.2	6.95959	-0.46	0.6464
h2xf2xDR	1	2.675	6.95959	0.38	0.7013
h2xf2xDM	1	2.8	6.95959	0.4	0.688
pxh2xf2xTA	1	183.75	193.0242	0.95	0.3427
pxh2xf2xTR	1	-66.25	193.0242	-0.34	0.7319
pxh2xf2xDA	1	35	193.0242	0.18	0.8564
pxh2xf2xDR	1	-142.5	193.0242	-0.74	0.4616
pxh2xf2xDM	1	-125	193.0242	-0.65	0.5183
pxTAXDA	1	33.7	37.3291	0.9	0.3681
pxTAXDR	1	45.5625	37.3291	1.22	0.2242
pxTRXDA	1	14.725	37.3291	0.39	0.6938
pxTRXDR	1	28.5875	37.3291	0.77	0.445
pxTAXDM	1	33.9625	37.3291	0.91	0.3644
pxTRXDM	1	35.2375	37.3291	0.94	0.3468
h2xTAXDA	1	3.925	4.0581	0.97	0.3351
h2xTAXDR	1	0.17625	4.0581	0.04	0.9654
h2xTRXDA	1	-2.0975	4.0581	-0.52	0.606
h2xTRXDR	1	-1.90875	4.0581	-0.47	0.6388
h2xTAXDM	1	0.08875	4.0581	0.02	0.9826
h2xTRXDM	1	-0.68625	4.0581	-0.17	0.866
f2xTAXDA	1	1.3625	3.26434	0.42	0.677
f2xTAXDR	1	-3.41125	3.26434	-1.05	0.2978
f2xTRXDA	1	-0.3625	3.26434	-0.11	0.9117

f2xTRXDR	1	-2.26375	3.26434	-0.69	0.4891
f2xTAXDM	1	-1.68625	3.26434	-0.52	0.6063
f2xTRXDM	1	-0.71125	3.26434	-0.22	0.8278
pxf2xTAXDA	1	-57.25	90.53636	-0.63	0.5282
pxf2xTAXDR	1	5.625	90.53636	0.06	0.9505
pxf2xTRXDA	1	-2.75	90.53636	-0.03	0.9758
pxf2xTRXDR	1	71.375	90.53636	0.79	0.4318
pxf2xTAXDM	1	9.125	90.53636	0.1	0.9199
pxf2xTRXDM	1	27.625	90.53636	0.31	0.7607
pxh2xTAXDA	1	-120.5	112.5515	-1.07	0.2861
pxh2xTAXDR	1	-46.125	112.5515	-0.41	0.6826
pxh2xTRXDA	1	17.75	112.5515	0.16	0.8749
pxh2xTRXDR	1	-97.125	112.5515	-0.86	0.3896
pxh2xTAXDM	1	-28.375	112.5515	-0.25	0.8013
pxh2xTRXDM	1	-64.875	112.5515	-0.58	0.5652
f2xh2xTAXDA	1	-9.075	9.84234	-0.92	0.3581
f2xh2xTAXDR	1	17.2625	9.84234	1.75	0.0816
f2xh2xTRXDA	1	4.425	9.84234	0.45	0.6537
f2xh2xTRXDR	1	-1.9875	9.84234	-0.2	0.8403
f2xh2xTAXDM	1	12.2875	9.84234	1.25	0.2139
f2xh2xTRXDM	1	-4.3375	9.84234	-0.44	0.6601
pxh2xf2xTAXDA	1	277.5	272.9774	1.02	0.3111
pxh2xf2xTAXDR	1	-331.25	272.9774	-1.21	0.2269
pxh2xf2xTRXDA	1	-52.5	272.9774	-0.19	0.8478
pxh2xf2xTRXDR	1	178.75	272.9774	0.65	0.5136
pxh2xf2xTAXDM	1	-213.75	272.9774	-0.78	0.4349
pxh2xf2xTRXDM	1	188.75	272.9774	0.69	0.4904

Note that p is the disease allele frequency, h2 is the trait heritability, Tmodel='A' denotes the trait model is additive, Tmodel='D' denotes the trait model is dominant, Tmodel='R' denotes the trait model is recessive, Dmodel='A' denotes the disease model is additive, Dmodel='D' denotes the disease model is dominant, Dmodel='R' denotes the disease model is recessive and Dmodel='M' denoted the disease model is multiplicative.