Stony Brook University

Using Growth Mixture Modeling to identify loci associated with the progression of disease

A Dissertation Presented

by

TONG SHEN

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

August 2011

Stony Brook University

The Graduate School

TONG SHEN

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Stephen Finch, Professor
Stony Brook University, Department of Applied Math and Statistics

Nancy Mendell, Professor
Stony Brook University, Department of Applied Math and Statistics

Wei Zhu, Deputy Chair, Professor
Stony Brook University, Department of Applied Math and Statistics

Derek Gordon, Associate Professor
Rutgers University, Department of Genetics

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Using Growth Mixture Modeling to identify loci associated with the progression of disease

by

TONG SHEN

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2011

In a genome-wide association study (GWAS) for a longitudinal quantitative trait, the trait is measured at multiple time points. GWAS is the examination of marker loci to identify loci associated with the progression of the quantitative trait.

I use two models, a single locus model and a multi locus model, to simulate a longitudinal quantitative trait. I use the growth mixture modeling (GMM) method to assign each member of a sample into one of a small number of trajectory groups. The clinically important trajectory group is the one with fastest progression. The Bayesian posterior probability (BPP) of being in the clinically important group is used as a quantitative trait. I test for association with marker loci. I also use the modal BPP in the association test and perform a case/control association analysis. Finally, I compare these methods with the contingency table method. I

evaluate the empirical type I error and empirical power using null simulations and power simulations.

The principal results are that: (1) Both the BPP method and modal BPP method maintain the correct type I error rate, but the empirical null rejection rate is increasing less than the nominal rate as the nominal type I error rate increases. (2) Both the BPP and modal BPP methods have very high power to detect the disease locus in the single locus model. (3) Both the BPP and modal BPP methods have significant power to detect the disease loci in the multi locus model. The powers of detecting a specific locus are proportional to minor allele frequency (MAF) of loci. (4) Both the BPP and modal BPP methods are better than the contingency table method with regard to the empirical power and the power of the BPP is essentially equal to the power of the modal BPP.

To Joey and my parents with all my love

Table of Contents

List of Figures viii

List of Tables ix

List of Abbreviations xi

Chapter 1 Introduction 1

1.1 Empirical issues 1

1.1.1 Adolescent Idiopathic Scoliosis (AIS) 1

1.1.2 Diagnosis of AIS 2

1.1.3 Etiology of AIS 3

1.1.4 Genetic studies of AIS 4

1.2 Longitudinal Genome-Wide Association studies 4

1.2.1 Genetic mapping 4

1.2.2 Genome-Wide Association Study 5

1.2.3 Longitudinal GWAS 5

1.3 Growth Mixture Modeling 7

1.4 Analysis software 11

1.4.1 SAS TRAJ procedure 11

1.4.2 M-plus software 12

1.4.3 Comparison of the two procedures 12

1.4.4 PLINK software 14

Chapter 2 Methodology 15

2.1 Sample used in the analysis 15

2.2 Genotypes used in the analysis 16

2.3 Simulated longitudinal phenotype 21

2.3.1 Generation of null simulation data 22

2.3.2 Generation of power simulation data 25

2.4 Group separation 31

2.5 Methods for testing the association of longitudinal phenotypes with genotype data 31

2.5.1 Method I: Using Bayesian Posterior Probability (BPP) as phenotype in the association test 31
2.5.2 Method II: Using modal BPP in the association test 32
2.5.3 Method III: Post hoc contingency table test 32
2.6 The factorial design 33
2.6.1 Disease loci 33
2.6.2 Genetic model 33
2.6.3 Separation of groups 34
2.6.4 Data transformation 34
2.7 Definitions of empirical type I error rate and empirical power 34
2.7.1 Empirical type I error rate 34
2.7.2 Empirical power 35
Chapter 3 Results 36
3.1 Null distribution 36
3.1.1 Null Simulation I 36
3.1.2 Null Simulation II 40
3.2 Simulated Power Results 52
3.2.1 Single locus model 52
3.2.2 Multi-locus model 65
Chapter 4 Conclusions and discussions 73
References 76
Appendix 79
I. IDs of 1599 unrelated participants 79
II. PLINK 85

List of Figures

Figure 1.1 Adolescent idiopathic scoliosis (AIS)…….....…….....……2

Figure 1.2 Definition of Cobb angle in AIS…..…….....…….……..3

Figure 3.1 Proportion of replicates for which the number of SNP markers within 10 markers of the disease SNP locus are in the top 5%, top 10% and top 25% of chromosome 13 markers by the number of significant SNP markers …….....……..57

Figure 3.2 Power of Procedure by MAF of Locus in Multigenic Model, Normally Distributed Data, for selected target levels of significance…….....…….…….....….68

Figure 3.3 Power of Procedure by MAF of Locus in Multigenic Model, Data Square of Normally Distributed Data, for selected target levels of significance …….....……69

List of Tables

Table 2.1 Distribution of the sample by generation……..……..…16

Table 2.2 SNP markers on chr 13 used as disease loci in single-locus association analysis….....…..…..17

Table 2.3 SNP markers on chr 13 used as disease loci in multi-locus association analysis….....18

Table 2.4 Normalized disequilibrium coefficient D' of ten disease loci in multi-locus model…….…….....19

Table 2.5 Correlation coefficient p-values of ten disease loci in multi-locus model……..…….....20

Table 2.6 Distribution of variable *score* ……..…….…….....…29

Table 2.7 Variance of genotype over the variance of *score* in multi-locus model…….....30

Table 3.1 Empirical type I error rate and its 95% confidence interval for chromosome 13 scoliosis data (null model I)…….…….....38

Table 3.2 ANOVA table for empirical type I error rate using null model I……..….....39

Table 3.3 Empirical type I error rate and its 95% confidence interval for chromosome 13 scoliosis data (null model II)…….....41

Table 3.4 ANOVA table of empirical type I error rate using null model II……..…….....44

Table 3.5 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the target SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (null model II) using BPP method…….…….....46

Table 3.6 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the target SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (null model II) using modal BPP method…….....…….....48

Table 3.7 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the target SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (null model II) using contingency table method……..…….....50

Table 3.8 Empirical power to detect association of disease SNP on chromosome 13 with scoliosis data (single-locus model)…….…….....……53

Table 3.9 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the disease SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (single-locus model) using BPP method.....59

Table 3.10 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the disease SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (single-locus model) using model BPP method.....61

Table 3.11 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the disease SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (single-locus model) using contingency table method.....63

Table 3.12 Power Simulation to detect the association of disease SNP on chromosome 13 with scoliosis data (multi-locus model) …….....…….....….70

Table 3.13 ANOVA table of empirical power using multi-locus model.....71

Table 3.14 P-values of Cochran's test and McNemar's test when comparing the three methods in multi locus model.....72

List of Abbreviations

AIC: Akaike information criterion;

AIS: adolescent idiopathic scoliosis;

ANOVA: analysis of variance;

BIC: Bayesian information criterion;

BMI: body mass index;

BPP: Bayesian posterior probability;

CT: contingency table;

CVD: cardiovascular disease;

EM: expectation-maximization;

FHS: Framingham heart study;

FIML: full information maximum likelihood imputation;

GAW: genetic analysis workshop;

GMM: growth mixture modeling;

GWAS: genome-wide association study;

HC: human chromosome;

HWE: Hardy-Weinberg equilibrium;

LCGA: latent class growth analysis;

LD: linkage disequilibrium;

MAF: minor allele frequency;

MLE: maximum likelihood estimates;

QTL: quantitative trait loci;

SNP: single nucleotide polymorphism;

TG: trajectory group;

TVC: time-varying covariates;

Acknowledgments

I would like to thank my advisor, Professor Stephen Finch, for the exciting topics suggested, for interesting discussions and for his guidance and continuous support.

I would like to thank my dissertation committee members, Professor Derek Gordon, Professor Nancy Mendell, and Professor Wei Zhu for sharing their valuable experience and insights. It is my great honor to have them on my dissertation committee.

I would like to thank all my friends in the department for all their support and help.

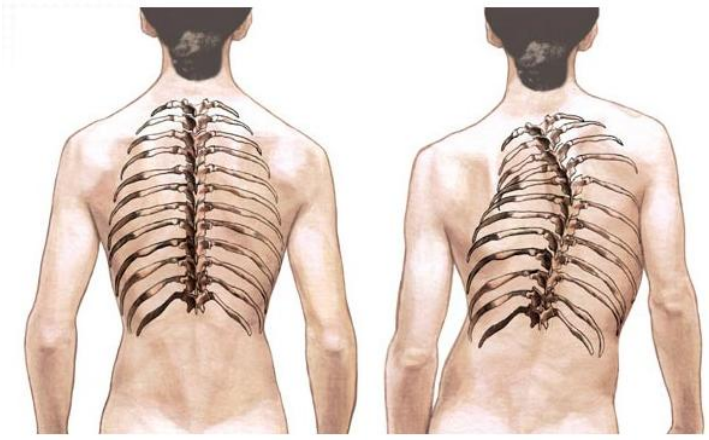
Chapter 1 Introduction

1.1 Empirical issues

1.1.1 Adolescent Idiopathic Scoliosis (AIS)

Adolescent Idiopathic Scoliosis (AIS) is the most common spinal deformity in children, affecting about 1-3% of children worldwide[1,2]. Patients with AIS may have one shoulder higher than the other, and their clothes may no longer fit correctly. Some severe cases of scoliosis can lead to diminished lung capacity, which can then put pressure on the heart and lead to restriction of physical activities[3]. Figure 1.1 is a schematic of the disease. I focus on methodological issues derived from research on AIS.

Figure 1.1 Adolescent Idiopathic Scoliosis (AIS)

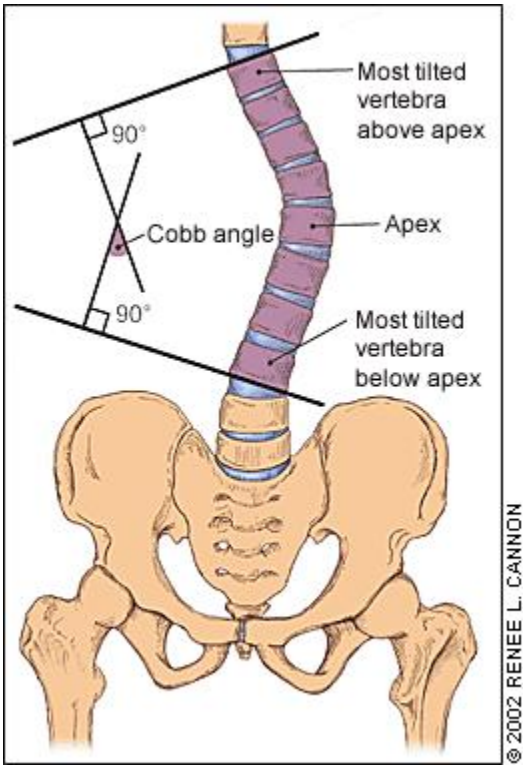


Normal spine (left) — Scoliosis (right)

1.1.2 Diagnosis of AIS

In practice, AIS is diagnosed using standing posteroanterior radiographs of the full spine to assess lateral curvature with the Cobb angle method[4,5]. The Cobb angle is the angle between two lines, drawn perpendicular to the upper endplate of the uppermost vertebrae involved and the lower endplate of the lowest vertebrae involved, as shown in Figure 1.2[6].

Figure 1.2 Definition of Cobb angle in AIS



1.1.3 Etiology of AIS

The etiology of AIS is still unknown, but it is believed to be multi-factorial, including complex genetic factors. Single Nucleotide Polymorphism (SNP) markers that are significantly associated with AIS have been identified from Genome-Wide Association Study (GWAS) research[7,8]. Although the genetic model of AIS is complicated, an autosomal dominant inheritance model is generally accepted[9]. Some studies also show an evidence of an X-linked susceptibility in AIS[10].

1.1.4 Genetic studies of AIS

Genetic studies of the progression of AIS are clinically important. After examining how the genetic factors affect the progression of disease and how an individual's genetic inheritance affects the body's response to drugs, physicians may be able to prescribe drugs tailor-made for individuals in the future. Compared with the traditional therapy, the individual customized therapy may enhance both the efficacy and safety of treatments[11]. Early prediction of maximal severity may allow earlier intervention, which may be more effective.

1.2 Longitudinal Genome-Wide Association studies

1.2.1 Genetic mapping

Many quantitative traits or human diseases are controlled by specific loci. When the trait is a quantitative measure such as body mass index (BMI), these loci are called quantitative trait loci (QTL). Genetic mapping can offer evidence that a disease transmitted from parent to child is linked to the QTL. Statistical methods for genetic mapping have been developed using two main approaches: linkage analysis and association analysis. Linkage is the tendency for loci and other genetic markers to be inherited together because of their location near one another on the same chromosome. In the process of meiosis, because there is some crossing over of DNA when the chromosomes segregate, alleles on the same chromosome can be separated and go to different daughter cells. Generally, in the same chromosome, the probability of recombination fraction

between two loci near each other is very low. Thus, a low recombination fraction means the two loci are near each other. Linkage analysis can help find the rough position of human disease loci relative to known genetic markers.

Association analysis, also known as association mapping or linkage disequilibrium (LD) mapping, is a method which is based on linkage disequilibrium to study the quantitative traits and genetic polymorphisms. LD is the association between two alleles located near each other on a chromosome, such that they are inherited together more frequently than expected by chance, which decays by recombination distance. So LD will be observed between two loci if they are in tight linkage. If we observe LD between candidate loci and markers, then we can claim that they are nearly located near each other.

1.2.2 Genome-Wide Association Study

In human genetics, a genome-wide association study (GWAS) is an examination of locus variations on a genome to discover loci that have associations with a disease. As of December 2010, over 1200 human GWASs have examined over 200 diseases and traits. Almost 4000 SNP associations have been found[12].

1.2.3 Longitudinal GWAS

Longitudinal studies use repeated observations of variables of interest over time. Fields such as psychology, sociology and medical research make extensive use of longitudinal studies. In genetics, there are a number of longitudinal GWAS[13]. In those studies, quantitative traits are measured at fixed time points. Then an association analysis or linkage analysis is conducted to detect the quantitative trait loci (QTL) in the genome. For example, researchers have found evidence for a disease locus influencing blood pressure on chromosome 17 using a genome scan[14]. GWAS analysis can also be used to detect interactions between the longitudinal traits and environment[15,16].

Wu and colleagues[17,18,19] propose a mapping strategy, call functional mapping, which integrates the mathematical aspects of biological processes into a statistical mapping framework for QTL mapping. The model is constructed within the traditional maximum-likelihood framework implemented with the expectation-maximization (EM) algorithm. A biologically meaningful growth curve, the logistic growth curve, is employed to model time-specific genetic values. An autoregressive model is used to structure the residual variance-covariance matrix among different time points. Because of a reduced number of parameters being estimated and the incorporation of biological principles, the functional mapping model displays increased statistical power to detect QTL. Later, Wu and colleagues generalize the functional mapping framework to more general models of time dependence of residuals. Wu's research group applies functional mapping model to QTL mapping of traits describing trees, as well as an HIV study[20].

1.3 Growth Mixture Modeling

Growth Mixture Modeling (GMM) is a method that can classify heterogeneous participants into discrete subgroups. GMM also describes the longitudinal pattern in each sub-population[21]. GMM applies mixture analysis methods to estimate the number of trajectory components and to estimate the probability that a trait variable (such as a genotype) affects the probability of trajectory component membership. The procedure allows for controlling for time-varying covariates (TVC) as well.

In 1999, Bengt Muthen and Kerby Shedden[22] propose a model that combines the features of conventional growth modeling and latent class growth modeling. Their research discusses a longitudinal study using a random coefficient model to assess the influence of latent growth trajectory class membership on the probability of a binary disease outcome. It is motivated by a study concerned with the longitudinal development of heavy drinking and its relation to alcohol dependence. In their paper, the EM algorithm is used for estimation. They analyze the influence of membership in different growth curve classes for heavy drinking from ages 18 to 25.

Later in 2000, Bengt Muthen and Linda K. Muthen[23] give a brief overview of new methods that integrate variable- and person-centered analyses. A variable-centered approach, such as regression analysis, factor analysis, and structural equation modeling, focuses on relationships among variables. A person-centered approach, such as cluster analysis, finite mixture analysis, latent class analysis, and latent transition analyses, focuses on relationships

among individuals. The goal is to group individuals into categories, each one of which contains individuals who are similar to each other and different from individuals in other categories. The methods that they discuss include latent class analysis, latent transition analysis, latent class growth analysis, growth mixture modeling, and general growth mixture modeling. Growth mixture modeling (GMM) is based on conventional growth modeling and combines the features of latent class growth analysis (LCGA). Conventional growth modeling estimates a mean growth curve under the assumption that all individuals in the sample come from a single population. Individual variation around the mean growth curve is captured by the estimation of the growth factor variances. LCGA estimates a mean growth curve for each class. No individual variation around the mean growth curves is allowed. As a result, the variation in the growth factors within each class is assumed to be zero. However, GMM estimates mean growth curves for each class and captures individual variation around these growth curves by the estimation of growth factor variances for each class. GMM can also be incorporated into a more general latent variable framework that allows combinations of the models mentioned above. This is referred to as general growth mixture modeling (GGMM). It is the statistical framework used in M-plus.

In 2002, Bengt Muthen[24] et al. present a novel application of growth mixture modeling to preventive intervention trials in which individuals are randomized into intervention and control groups and measured repeatedly before and after the start of the intervention. They apply four analyses, two of which are GMM. Comparison of models with different numbers of classes, however, is accomplished by a Bayesian information criterion (BIC). The larger the BIC value, the better the model. They conclude that the growth mixture modeling is a powerful analytic tool when applied to randomized trials as well as to non-experimental research.

Daniel S. Nagin[25] proposes a method to analyze developmental trajectories in 1999. It is a semi-parametric, group-based approach for identifying distinctive groups of individual trajectories within the population and for profiling the characteristics of group members. It can handle three data types—count, binary, and psychometric scale data. Four capabilities are demonstrated in their model: the capability to identify distinctive groups of trajectories; the capability to estimate the proportion of the population following each such trajectory group; the capability to relate group membership probability to individual characteristics and circumstances; the capability to use the group membership probabilities for other purposes such as creating profiles of group members. They also discuss two important issues in model selection: determination of the optimal number of groups in the mixture and the determination of the appropriate order of the polynomial used to model each group's trajectory. Here "order" refers to the degree of the polynomial used to model the group's trajectory.

In 2001, Daniel S. Nagin and Richard E. Tremblay[26] demonstrate a group-based method for joining developmental trajectories of distinct but theoretically related behaviors. This method will aid the analysis of comorbidity and heterotypic continuity. It is based on the method Nagin proposed in 1999. First, the statistical model underlying the estimation of a group-based trajectory model for a single behavior is summarized; then, the approach used to link two univariate models to form a joint model is described. They obtain three major outputs: the form of the trajectory of distinctive subpopulations for both measurement series; the probability of membership in each such trajectory group; the joint probability of membership in trajectory groups across behaviors. They apply the model to two examples: one is the data from research in physical aggression and hyperactivity in children; the other is the data from study of criminal behavior. Nagin et al. introduce a new SAS procedure that analyzes longitudinal data

(developmental trajectories) by fitting a mixture model. The TRAJ procedure will be discussed later.

GMM has been applied in alcohol use studies and smoking behavior studies. Li and colleagues[27] examine developmental trajectories in adolescent alcohol using piecewise GMM. Alcohol use typically begins in adolescence. The research about alcohol use suggests distinct developmental periods depending on age. It is assumed that onset and rates of change of adolescent alcohol use are not homogeneous but consist of subgroups that have different growth patterns and social, family and individual influence systems. Li et al. describe their model using the framework for GMM proposed by Muthen in 1999. They examine distinct trajectories from middle school to high school in the development of alcohol use. Two subgroups or trajectories are reported. They also analyze the influences of background variables, such as middle school entry and midpoint time-invariant predictors.

Colder et al.[28] apply GMM to identify trajectories of adolescent smoking. In their article, discrete patterns of smoking are identified on the basis of level of smoking. Analyses reveal considerable heterogeneity in how smoking unfolded over time during adolescence. They show that compared with the traditional growth models, GMM has more power to identify subpopulations on the basis of distinct growth trajectories.

There has been increasing interest in using GMM to identify the SNPs associated with a longitudinal quantitative trait. In Genetics Analysis Workshop 16, Chang et al.[29] examine the properties of GMM to find longitudinal QTL. She studies the trajectory model's Bayesian posterior probability and tests the association with 17 SNPs on Human Chromosome (HC) 22. Kerner and Muthen[30] apply GMM to longitudinal data of blood pressure in Framingham heart

study. They test SNPs on HC 8 for association with the class membership probabilities. Both studies find GMM to be a useful tool to detect subgroups in heterogeneous populations in GWAS.

1.4 Analysis software

There are two computer programs to perform GMM on longitudinal data. One is the SAS TRAJ procedure, and the other is M-plus.

1.4.1 SAS TRAJ procedure

The SAS Trajectory Procedure (Proc Traj)[31,32,33] fits a discrete mixture model to longitudinal data. The model groups data trajectories, with different parameter values for each component. Components may identify distinct subpopulations. Proc Traj estimates a longitudinal regression model for each component group within the population. The focus of this procedure is on group membership and identifying distinct subgroups within the population. SAS PROC TRAJ analysis reports the estimated frequency of each trajectory group, the t-statistic and the maximum likelihood estimates (MLEs) of the trajectory group parameters, the Bayesian posterior probability (BPP) that each subject is member of each trajectory component and Bayesian information criterion (BIC) for model selection. The model which has the largest BIC value and at least 10 subjects estimated to be in each trajectory component is often reported as the best model in actual studies.

1.4.2 M-plus software

M-plus is another statistical modeling program that provides researchers with a flexible tool to analyze longitudinal data. M-plus takes a multivariate approach to growth modeling. This approach allows flexible modeling of relationships between the outcomes such as correlated residuals over time and regressions among the outcomes over time. M-plus uses the principle of maximum likelihood estimation and employs the EM algorithm for maximization. M-plus program provides three measures of each model: Akaike information criterion (AIC); BIC; a sample-size adjusted BIC (ABIC). Researchers often use BIC to choose the best model. M-plus provides estimates of probabilities of class membership for each individual. For example, in a five-class solution, five probabilities are estimated for each individual in the data, where each estimates the likelihood that an individual is a member of one of the classes. For each individual, these probabilities sum to 1.0. Ideally, for each individual, one of these probabilities would be very high and the others very low, indicating little ambiguity about class membership.

1.4.3 Comparison of the two procedures

M-plus specifies latent GMM in the context of a general structural equation model. This allows considerable flexibility in model specification. However, M-plus uses full information maximum likelihood imputation (FIML) to deal with missing data. In contrast, Proc Traj in SAS allows for the inclusion of cases with missing data and specification of a variety of distributions

for the observed variables but does not permit specification of latent variables. In growth modeling, M-plus uses random effects to capture individual differences in development. In contrast with the M-plus model, there is no random effect capability within the Proc Traj model.

In each procedure, heterogeneous longitudinal data is classified into a few discrete growth trajectory groups such that there is an estimated probability that an individual belongs to a particular trajectory group. This probability is called the Bayesian posterior probability (BPP). Both programs report the estimated coefficients of the polynomial trajectory functions, their corresponding t-statistics and p-values, the MLE of the trajectory group parameters, the AIC and the BIC for a specified number of trajectory groups.

One major difference between the programs is the treatment of within-class variability. Since the SAS TRAJ procedure assumes no variation in growth parameters within each class, any individual deviations from the class mean trajectories are attributed to random error. M-plus, however, allows for within-class variation in individual trajectories; that is, the coefficients of the M-plus model are random. For example, in SAS TRAJ, one of the trajectory groups in my research is modeled as

$$Y_{i(\omega),t} = \beta_0 + \beta_1(t - 0.25) + \varepsilon_{i(\omega),t},$$

here, ω refers to the specific individual, i refers to the trajectory group and t refers to the time point. The intercept β_0 and slope β_1 are fixed, and the individual variation $\varepsilon_{i(\omega),t}$ in this trajectory group is attributed to random error. In M-plus, the parameters β_0 and β_1 are modeled as random. That is, they differ among individuals in the same trajectory group. In the fixed effects model (growth factor variances and covariance equal to zero), estimation of parameters is

easier, and time to convergence is faster. Muthen (2004) has suggested that use of both approaches may be useful. For example, he suggests using the simpler model group-based trajectory approach as a first step to identify the number of trajectory groups and cut points on the growth factors. Then researchers can use more complex variance/covariance constraints in a growth mixture model.

1.4.4 PLINK software

PLINK (Appendix II) is a whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner. I used PLINK to test the association between the quantitative trait and 1498 SNPs on chromosome 13.

Chapter 2 Methodology

2.1 Sample used in the analysis

The Framingham heart study (FHS)[34], which started in 1948, aims to identify the common factors or characteristics that contribute to cardiovascular disease (CVD). FHS recruits a large group of participants from the town of Framingham, Massachusetts. The Genetic Analysis Workshop (GAW 16) simulated dataset includes a total of 6,476 participants with actual genotype data from the FHS. The participants are in 936 pedigrees distributed among 3 generations with the 187 singleton subjects; that is, participants not related to any other participant.

My study focuses on the 1599 genetically unrelated participants of the FHS. I first include those who married into the pedigree and the 187 singletons. Next, I build a "family tree" for each pedigree and choose the first generation participants if any are genotyped. In this case, I exclude all of their children and grandchildren. If the first genotyped participants are in the

second generation, then I choose one at random and I exclude all of their children. If the first genotyped participants are in the third generation, then I choose one at random. The distribution of the 1599 unrelated participants extracted from FHS is shown in Table 2.1. Appendix I contains the participant IDs of the 1599 unrelated participants.

Table 2.1 Distribution of the sample by generation

Generation	Frequency	Percentage
First generation	260	16.6
Second generation	436	27.4
Third generation	160	10.0
Participants who married-in	556	34.9
Singletons	187	11.1

2.2 Genotypes used in the analysis

In my study, I use the genotype data from chromosome 13 for the 1599 genetically unrelated individuals in the FHS. There are 1498 SNP markers on Chromosome 13.

For the single-locus association analysis, I choose two SNP markers on chromosome 13 listed in Table 2.2, as the simulated disease loci. Table 2.2 shows the SNP ID, base pair, major allele, minor allele, MAF, Hardy-Weinberg Equilibrium (HWE) test p-value and missing rate. The null hypothesis of HWE test is that the SNP is in HWE. Both of the SNP markers are apparently in HWE from the table. The missing rate indicates the proportion of missing SNPs in 1599 unrelated participants. The SNP markers I studied have minor allele frequency (MAF) close to 0.5 and 0.15, respectively. I study the two loci separately, as well as other markers on chromosome 13. I use the normalized disequilibrium coefficient D' to measure Linkage Disequilibrium (LD). For the two loci in Table 2.2, $D' = 0.046$.

Table 2.2 SNP markers on chr13 used as disease loci in single-locus association analysis

SNP	Base-Pair	Major Allele	Minor Allele	MAF	HWE test p-value	Missing rate
rs4133063	39205119	C	T	0.49	0.68	0.0044
rs7990928	91893219	C	A	0.15	0.42	0.0037

For the multi-locus association analysis, I select ten SNP markers on chromosome 13 listed in Table 2.3, as the simulated disease loci. All the SNP markers are in HWE based on the chi-square goodness of fit test. They are rare variant loci, with MAF less than 0.05. The measurements of LD are shown in Table 2.4. The association test p-values are shown in Table

2.5. SNP rs17090361 is highly associated with SNP rs16943207; SNP rs7331979 is highly associated with SNP rs12863734.

Table 2.3 SNP markers on chr13 used as disease loci in multi-locus association analysis

SNP	Base-Pair	Major Allele	Minor Allele	MAF	HWE test p-value	Missing rate
rs9599854	71021185	T	C	0.026	1	0.0013
rs9542756	71309666	T	C	0.038	0.33	0.0038
rs9543107	72217237	C	T	0.016	1	0.0006
rs17090361	73186500	T	C	0.050	0.72	0.0013
rs9593132	75293621	C	T	0.048	1	0.0006
rs5352	77373231	C	T	0.012	1	0
rs7331979	78836214	T	G	0.033	0.22	0.0325
rs12863734	85268572	G	A	0.015	1	0.0025
rs9522610	89110831	C	T	0.026	0.44	0.0225
rs16943207	89144779	C	G	0.031	1	0.0088

Table 2.4 Normalized disequilibrium coefficient D' of ten disease loci in multi-locus model

D'	rs9599854	rs9542756	rs9543107	rs17090361	rs9593132	rs5352	rs7331979	rs12863734	rs9522610	rs16943207
rs9599854	NA	0.017	1**	0.715	0.132	0.078	1**	0.511	0.157	0.007
rs9542756	0.017	NA	1**	1**	0.065	0.082	0.998*	0.003	0.016	0.001
rs9543107	1**	1**	NA	0.581	0.481	1**	1**	0.007	0.023	0.787
rs17090361	0.715	1**	0.581	NA	0.008	0.029	0.008	0.059	0.014	0.037
rs9593132	0.132	0.065	0.481	0.008	NA	0.986*	0.017	0.064	0.622	0.457
rs5352	0.078	0.082	1**	0.029	0.986*	NA	1**	0.020	0	0.041
rs7331979	1**	0.998*	1**	0.008	0.017	1**	NA	0.007	0.688	0.868
rs12863734	0.511	0.003	0.007	0.059	0.064	0.020	0.007	NA	0.072	0.767
rs9522610	0.157	0.016	0.023	0.014	0.622	0	0.688	0.072	NA	0.579
rs16943207	0.007	0.001	0.787	0.037	0.457	0.041	0.868	0.767	0.579	NA

Notes: The normalized disequilibrium coefficients with * indicate the two SNPs are in linkage disequilibrium.

Table 2.5 Correlation coefficient p-values of ten disease loci in multi-locus model

association test p-value	rs9599854	rs9542756	rs9543107	rs17090361	rs9593132	rs5352	rs7331979	rs12863734	rs9522610	rs16943207
rs9599854	1	0.8392	0.3464	0.8702	0.8644	0.4624	0.491	0.5803	0.9164	0.9988
rs9542756	0.8392	1	0.129	0.5278	0.416	0.8855	0.2279	0.413	0.7341	0.9622
rs9543107	0.3464	0.129	1	0.3418	0.4524	0.8507	0.0974	0.5616	0.8485	0.5488
rs17090361	0.8702	0.5278	0.3418	1	0.4384	0.9092	0.8148	0.9566	0.4754	**0.007**
rs9593132	0.8644	0.416	0.4524	0.4384	1	0.9051	0.6388	0.4957	0.6481	0.6218
rs5352	0.4624	0.8855	0.8507	0.9092	0.9051	1	0.4103	0.6384	0.9416	0.7445
rs7331979	0.491	0.2279	0.0974	0.8148	0.6388	0.4103	1	**0.0291**	0.685	0.2734
rs12863734	0.5803	0.413	0.5616	0.9566	0.4957	0.6384	0.0291	1	0.441	0.7319
rs9522610	0.9164	0.7341	0.8485	0.4754	0.6481	0.9416	0.685	0.441	1	0.1923
rs16943207	0.9988	0.9622	0.5488	0.007	0.6218	0.7445	0.2734	0.7319	0.1923	1

2.3 Simulated longitudinal phenotype

The longitudinal phenotype is simulated based on the data describing the progression of AIS provided by Carol Wise, M.D. There are 334 Adolescent Idiopathic Scoliosis (AIS) study participants in the data used here. The quantitative longitudinal variable is the Cobb angle. A Proc Traj analysis identifies three linear trajectory groups for this data. In my simulation study, I specify a linear growth mixture model with the longitudinal trajectory functions for each individual w as follows:

$$f_{i(\omega),t} = \begin{cases} 50, i(\omega) = 1 \\ 50 + 28(t - 0.25), i(\omega) = 2 \\ 50 + 56(t - 0.25), i(\omega) = 3 \end{cases} \text{ where } t = 0.25,0.4,0.55,0.7,0.85,1;$$

When $i(w) = 1$, the individual w is in the constant trajectory group; $i(w) = 2$ indicates the individual w is in the intermediate increase trajectory group; $i(w) = 3$ indicates the individual w is in the fast increase trajectory group. The slope trajectory parameters are roughly equal to those estimated from Wise data. Since $f_{i(\omega),t}$ describes the Cobb angle of each individual, which should be a non-negative value, I set the intercept to be 50 so that $f_{i(\omega),t}$ is positive.

For each replicate, I randomly select 700 individuals without replacement from the sampling set of 1599 unrelated (independent) individuals. Then I divide the individuals into three trajectory groups; that is, the constant trajectory group, the intermediate increase trajectory group and the fast increase trajectory group. The allocation rules for both null simulations and power simulations are described below.

2.3.1 Generation of null simulation data

2.3.1.1 Null simulation I

The probability of each individual being in any of the trajectory groups is 1/3 for each group. Thus, for each of the 700 individuals in a replicate, I create U, a random $U(0,1)$ number. If $U \le \frac{1}{3}$, then the individual is in the constant trajectory group; if $\frac{1}{3} < U \le \frac{2}{3}$, then the individual is in the intermediate trajectory group; if $\frac{2}{3} < U \le 1$, then the individual is in the fast trajectory group.

2.3.1.2 Null simulation II

2.3.1.2.1 Penetrance matrix

In the null simulation II model and the single locus model (described below), I use the penetrance matrix X to define the allocation rule; that is, the relationship between disease SNP genotype and trajectory group membership. For a selected individual ω, let the penetrance matrix

$$X = \left(x_{j+1,i}\right) = \Pr\left(i(\omega) = i \mid j(\omega) = j\right), i = 1..., G, j = 0,1,2.$$

Here, the value j represents the number of copies of the minor allele at the disease locus, where the locus has two alleles. Thus, $j = 0,1,2$ refers to major homozygote, heterozygote and minor homozygote, respectively. The value i represents the trajectory group (TG), where G is the number of trajectory groups for the simulated data. In my study, I set $G = 3$. That is, all the individuals are divided into three trajectory groups. In matrix form, we have:

$$X = \left(x_{j+1,i}\right) = \begin{pmatrix} \Pr\left(i(\omega)=1 \mid j(\omega)=0\right) \text{L} & \Pr\left(i(\omega)=G \mid j(\omega)=0\right) \\ \Pr\left(i(\omega)=1 \mid j(\omega)=1\right) \text{L} & \Pr\left(i(\omega)=G \mid j(\omega)=1\right) \\ \Pr\left(i(\omega)=1 \mid j(\omega)=2\right) \text{L} & \Pr\left(i(\omega)=G \mid j(\omega)=2\right) \end{pmatrix}.$$

2.3.1.2.2 Allocation rule for null simulation II

The probability of each individual being in one of the trajectory groups is determined by HWE proportions and the penetrance matrix. If the MAF for a SNP is denoted by p, then for an arbitrary individual ω,

$$Y = \begin{pmatrix} \Pr\left(j(\omega)=0\right) \\ \Pr\left(j(\omega)=1\right) \\ \Pr\left(j(\omega)=2\right) \end{pmatrix} = \begin{pmatrix} q^2 \\ 2pq \\ p^2 \end{pmatrix},$$

where q is the major allele frequency and $q = 1 - p$.

The penetrance matrix is given by

$$X=\left(x_{j+1,i}\right)=\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}$$

$$=\begin{pmatrix} \Pr\left(i(\omega)=1\middle|j(\omega)=0\right) & \Pr\left(i(\omega)=2\middle|j(\omega)=0\right) & \Pr\left(i(\omega)=3\middle|j(\omega)=0\right) \\ \Pr\left(i(\omega)=1\middle|j(\omega)=1\right) & \Pr\left(i(\omega)=2\middle|j(\omega)=1\right) & \Pr\left(i(\omega)=3\middle|j(\omega)=1\right) \\ \Pr\left(i(\omega)=1\middle|j(\omega)=2\right) & \Pr\left(i(\omega)=2\middle|j(\omega)=2\right) & \Pr\left(i(\omega)=3\middle|j(\omega)=2\right) \end{pmatrix}$$

$$=\begin{pmatrix} 1-\rho & 3\rho/4 & \rho/4 \\ \rho/2 & 1-\rho & \rho/2 \\ \rho/4 & 3\rho/4 & 1-\rho \end{pmatrix}.$$

Thus, the probability of each individual belongs to a specific trajectory group is given by:

$$\begin{pmatrix} \Pr\left(i(\omega)=1\right) \\ \Pr\left(i(\omega)=2\right) \\ \Pr\left(i(\omega)=3\right) \end{pmatrix}=$$

$$\begin{pmatrix} \Pr\left(i(\omega)=1\middle|j(\omega)=0\right)\Pr\left(j(\omega)=0\right)+\Pr\left(i(\omega)=1\middle|j(\omega)=1\right)\Pr\left(j(\omega)=1\right)+\Pr\left(i(\omega)=1\middle|j(\omega)=2\right)\Pr\left(j(\omega)=2\right) \\ \Pr\left(i(\omega)=2\middle|j(\omega)=0\right)\Pr\left(j(\omega)=0\right)+\Pr\left(i(\omega)=2\middle|j(\omega)=1\right)\Pr\left(j(\omega)=1\right)+\Pr\left(i(\omega)=2\middle|j(\omega)=2\right)\Pr\left(j(\omega)=2\right) \\ \Pr\left(i(\omega)=3\middle|j(\omega)=0\right)\Pr\left(j(\omega)=0\right)+\Pr\left(i(\omega)=3\middle|j(\omega)=1\right)\Pr\left(j(\omega)=1\right)+\Pr\left(i(\omega)=3\middle|j(\omega)=2\right)\Pr\left(j(\omega)=2\right) \end{pmatrix}$$

$$=\begin{pmatrix} \Pr\left(i(\omega)=1\middle|j(\omega)=0\right) & \Pr\left(i(\omega)=1\middle|j(\omega)=1\right) & \Pr\left(i(\omega)=1\middle|j(\omega)=2\right) \\ \Pr\left(i(\omega)=2\middle|j(\omega)=0\right) & \Pr\left(i(\omega)=2\middle|j(\omega)=1\right) & \Pr\left(i(\omega)=2\middle|j(\omega)=2\right) \\ \Pr\left(i(\omega)=3\middle|j(\omega)=0\right) & \Pr\left(i(\omega)=3\middle|j(\omega)=1\right) & \Pr\left(i(\omega)=3\middle|j(\omega)=2\right) \end{pmatrix}\begin{pmatrix} \Pr\left(j(\omega)=0\right) \\ \Pr\left(j(\omega)=1\right) \\ \Pr\left(j(\omega)=2\right) \end{pmatrix}$$

$$=X^{T}Y=\begin{pmatrix} 1-\rho & \rho/2 & \rho/4 \\ 3\rho/4 & 1-\rho & 3\rho/4 \\ \rho/4 & \rho/2 & 1-\rho \end{pmatrix}\begin{pmatrix} q^2 \\ 2pq \\ p^2 \end{pmatrix}.$$

In this null simulation, I generate a random number from the uniform distribution on the interval (0,1) for every individual. If the random number is in

$$(0,\theta] = \left(0,(1-\rho)q^2 + \rho pq + \rho p^2/4\right],$$

then this individual is in group 1, the constant trajectory group. If the random number is in

$$(1-\psi,1) = \left(1-\left(\rho q^2/4 + \rho pq + (1-\rho)\,p^2\right),1\right),$$

then this individual is in group 3, the fast increase trajectory group. Otherwise, the individual is in the intermediate increase trajectory group.

2.3.2 Generation of power simulation data

The longitudinal data for power simulations are generated based on the selected individual's real genotype. I examine both single-locus model and multi-locus model.

2.3.2.1 Single-locus association

The penetrance matrix X is given by

$$X = \begin{pmatrix} 1-\rho & 3\rho/4 & \rho/4 \\ \rho/2 & 1-\rho & \rho/2 \\ \rho/4 & 3\rho/4 & 1-\rho \end{pmatrix},$$

where $\rho \in [0,1]$. In my study, I set $\rho = 0.1$ and 0.4. I call the model with $\rho = 0.1$ the 'high penetrance locus model' and the model with $\rho = 0.4$ the 'low penetrance locus model'.

2.3.2.1.1 High penetrance

The high penetrance matrix X is given by:

$$X = \begin{pmatrix} 0.9 & 0.075 & 0.025 \\ 0.05 & 0.9 & 0.05 \\ 0.025 & 0.075 & 0.9 \end{pmatrix}$$

For this model, in row one of the matrix, individuals with major homozygote genotype ($j(\omega) = 0$) are in the constant trajectory group with probability 0.9, in the intermediate increase group with probability 0.075 and in the fast increase group with probability 0.025. In row two of the matrix, individuals with heterozygote genotype ($j(\omega) = 1$) are in the intermediate increase group with probability 0.9, in the constant trajectory group with probability 0.05 and in the fast increase group with probability 0.05. In row three of the matrix, individuals with minor homozygote genotype ($j(\omega) = 2$) are in the fast increase group with probability 0.9, in the constant trajectory group with probability 0.025 and in the intermediate group with probability 0.075.

2.3.2.1.2 Low penetrance

The low penetrance matrix is given by:

$$X = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}.$$

This matrix indicates that in the first row, an individual with major homozygote genotype is in the constant trajectory group (trajectory group 1) with probability 0.60, in the intermediate group (trajectory group 2) with probability 0.30, and in the fast increase group (trajectory 3) with probability 0.10. For an individual with heterozygote genotype (the second row), an individual is in the intermediate trajectory group with probability 0.60, in the fast increase trajectory group with probability 0.20, and in the constant trajectory group with probability 0.20. For an individual with minor homozygote genotypes (the third row), an individual is in the fast increase trajectory group with probability 0.60, in the intermediate trajectory group with probability 0.30, and in the constant group with probability 0.10.

2.3.2.1.3 Other parameter settings

In each particular trajectory group, an individual's phenotype follows the normal distribution at each time point. I set the standard deviation σ at each time point to be 4 or 8. That is, the quantitative trait at each time point $Y_{i(\omega),t}$ for individual w is:

$$Y_{i(\omega),t} = f_{i(\omega),t} + N\left(0, \sigma^2\right), i = 1,2,3.$$

$$\text{Thus, } Y_{i(\omega),t} = \begin{cases} 50 + N\left(0,\sigma^2\right), i(\omega) = 1 \\ 50 + 28\left(t - 0.25\right) + N\left(0,\sigma^2\right), i(\omega) = 2 \\ 50 + 56\left(t - 0.25\right) + N\left(0,\sigma^2\right), i(\omega) = 3 \end{cases}.$$

The generated data above are linearly related to the time variable. The other model I examined is that the value seen is the square of the linearly generated value.

2.3.2.2 Multi-locus association

I also include a multi-locus model in my study. The model has ten disease loci, each of which has MAF<0.05. For each disease locus, I create the variable that is the count of the number of minor alleles; that is,

$$cnt = \begin{cases} 0, \text{ if the SNP has 0 minor allele;} \\ 1, \text{ if the SNP has 1 minor allele; .} \\ 2, \text{ if the SNP has 2 minor alleles.} \end{cases}$$

Thus, an individual ω has a vector:

$$\left(cnt_{1,\omega}, cnt_{2,\omega}, \text{L } , cnt_{10,\omega}\right).$$

For example, an individual who has minor homozygote on SNP5 and SNP7, has heterozygote on SNP 10, and has major homozygote on the other disease SNPs, has vector

$$\left(0,0,0,0,2,0,2,0,0,1\right).$$

The variable *score* is the sum of the ten counts:

$$score_{\omega} = \sum_{i=1}^{10} cnt_{i,w} .$$

For the individual above, *score* =5. The model is that if an individual's score is greater than or equal to 3, then the individual is in the fast increase trajectory group; if the score is 2, then this individual is in the intermediate increase trajectory group; if the score is 0 or 1, then the individual is in the constant trajectory group. The distribution of *score* of unrelated individuals is in Table 2.6. The mean of *score* is 0.588 and the standard deviation is 0.745. If the fast increase trajectory group is the clinically important group, then in this model, the prevalence of the disease is around 2%.

Table 2.6 Distribution of variable *score*

score	Frequency (%)
0	54.78
1	33.65
2	9.63
3	1.88
4	0.06

For the ten disease loci, I further calculate the variance of genotype over the variance of $score$: $Var(genotype) \big/ Var(score)$. They are displayed in Table 2.7.

Table 2.7 Variance of genotype over the variance of $score$ in multi-locus model

SNP	Base-Pair	MAF	$Var(genotype) \big/ Var(score)$
rs9599854	71021185	0.026	0.086
rs9542756	71309666	0.038	0.141
rs9543107	72217237	0.016	0.052
rs17090361	73186500	0.050	0.184
rs9593132	75293621	0.048	0.161
rs5352	77373231	0.012	0.044
rs7331979	78836214	0.033	0.098
rs12863734	85268572	0.015	0.063
rs9522610	89110831	0.026	0.084
rs16943207	89144779	0.031	0.108

2.4 Group separation

After generating the simulated longitudinal data, I used the SAS PROC TRAJ to classify the heterogeneous longitudinal data into three trajectory groups. The procedure can also report the BPP and BIC of the model. For some of the replicates, the trajectory curve failed to converge. In these replicates, no estimated parameters will be reported. Thus, I delete these replicates and do not include them when calculating the type I error rate and empirical power. The failure rate of each parameter settings will be reported in Chapter 3.

2.5 Methods for testing the association of longitudinal phenotypes with genotype data

I apply the SAS TRAJ procedure to the simulated longitudinal phenotype data. For each of the replicates, I set the number of trajectory groups to 3. The trajectory group with the largest slope is identified as the clinically important group. The SAS TRAJ procedure estimates the BPP that each subject belongs to each group. Specifically, the BPP that each subject belongs to the fast increase trajectory group (clinically important group) is used as a quantitative trait.

2.5.1 Method I: Using Bayesian Posterior Probability (BPP) as phenotype in the association test

I use the BPP of the clinically important group as a quantitative trait in PLINK. The association between each SNP on chromosome 13 and the quantitative trait is reported.

Additionally, PLINK reports SNP identifier, base-pair, Wald test statistic and the asymptotic p-value for the test.

2.5.2 Method II: Using modal BPP in the association test

I create a dummy variable, which is 1 if modal BPP is in the clinically important group, and is 0 if otherwise. Then I use this variable to perform a case/control association analysis in PLINK. The basic allelic test chi-square and its asymptotic p-value is reported in the PLINK output.

2.5.3 Method III: Post hoc contingency table test

I classify a subject as belonging to the clinically important trajectory group when the subject's modal BPP is the clinically important group. The row variable in the contingency table is whether or not the subject is classified into the clinically important group. The subject is simultaneously in one of three genotypes: major homozygote, heterozygote and minor homozygote for each SNP analyzed. That is, the genotypes are the column variables in the contingency table. The contingency table test of independence between genotype and membership in the clinically important group is used to test for association with a SNP. For each SNP, we have a chi-square statistic and the corresponding asymptotic p-value. For some replicates with small sample size (<5) in the fast increase group, I use Fisher's exact test instead.

2.6 The factorial design

2.6.1 Disease loci

For the single-locus association study, I use two loci to simulate the disease loci, one with MAF near 0.5 and the other with MAF near 0.15.

For the multi-locus association study, I use ten loci, each of which has MAF<0.05.

2.6.2 Genetic model

For the single-locus study, the genetic model is defined by the penetrance matrix X. There are two settings for this factor:

high penetrance ($X = \begin{pmatrix} 0.9 & 0.075 & 0.025 \\ 0.05 & 0.9 & 0.05 \\ 0.025 & 0.075 & 0.9 \end{pmatrix}$) and

low penetrance ($X = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.3 & 0.6 \end{pmatrix}$).

For the multi-locus study, the genetic model is defined by the variable *cnt* and *score* described above.

2.6.3 Separation of groups

The standard deviation of an individual time measurement σ is set to be 4 and 8. That is, for the low variance setting,

$$Y_{i(\omega),t} = \begin{cases} 50 + N(0,16), i(\omega) = 1 \\ 50 + 28(t - 0.25) + N(0,16), i(\omega) = 2 \\ 50 + 56(t - 0.25) + N(0,16), i(\omega) = 3 \end{cases}$$

For the high variance setting,

$$Y_{i(\omega),t} = \begin{cases} 50 + N(0,64), i(\omega) = 1 \\ 50 + 28(t - 0.25) + N(0,64), i(\omega) = 2 \\ 50 + 56(t - 0.25) + N(0,64), i(\omega) = 3 \end{cases}$$

2.6.4 Data transformation

There are two settings for this factor. The one setting is that the data are linearly related to the time variable. The other is that the data is the square of data linearly related to time. That is, the value seen is the square of the linearly generated value.

2.7 Definitions of empirical type I error rate and empirical power

2.7.1 Empirical type I error rate

Type I error rate is defined as

$$\alpha = \Pr\left(reject\ H_0 \middle| H_0 \text{ is true}\right)$$

In null simulations where trajectory group assignments are not correlated with an individual's genotype, the empirical type I error rate is defined as the proportion of replicates in which the target SNPs are significant. For example, if in 4 out of 1000 replicates, a target SNP association test p-value is significant among the top 0.5% SNP p-values on one chromosome, then the empirical type I error rate of this target SNP will be 0.004.

2.7.2 Empirical power

The power is defined as

$$power = \Pr\left(reject\ H_0 \middle| H_1 \text{ is true}\right)$$

In power simulations where the trajectory group assignments are dependent on an individual's genotype, empirical power is defined as the proportion of replicates in which the disease SNP is detected. For example, if in 800 out of 1000 replicates, a disease SNP p-value is in the top 0.5% of SNP p-values on one chromosome, then the power to detect this disease SNP will be 80%.

Chapter 3 Results

3.1 Null distribution

3.1.1 Null Simulation I

Table 3.1 showed the empirical type I error rate and its 95% confidence interval for target SNP detection using the model with three equi-probable trajectory components unrelated to any locus. A confidence interval in bold did not contain the target α. As α increased, the number of intervals not containing the target α increased. That is, true type I error rate was increasing below α as α increased. The failure rate of the TRAJ model was shown in Table 3.1.

Next, I reported an analysis of variance test of the empirical type I error rate as dependent variable with five independent variables: the indicator variable of whether or not the data was normal (denoted by *NORMAL*), the value of sigma (denoted by *SIGMA*), the MAF (denoted by

MAF), the target α for each method and the three methods. Here, the variable *NORMAL* had two levels: the value of 0 indicated the data was normally distributed at each time point and the value of 1 indicated the data was the square of normally distributed data. The variable *SIGMA* had two levels: 4 and 8. The variable MAF had three levels: 0.49, 0.15 and 0.05. Nominal α had four levels: 0.005, 0.01, 0.05 and 0.10. The variable *METHOD* had three levels: BPP method, modal BPP method and contingency table method. The analysis of variance table (ANOVA) was shown in Table 3.2. The variables *NORMAL*, *MAF* and α were significant. The regression model was that:

$$\begin{aligned}\text{empirical type I error} = {} & 0.0045(\pm 0.0014) + 0.0023(\pm 0.0009)\times(NORMAL = 0) \\ & - 0.0066(\pm 0.0011)\times(MAF = 0.05) + 0.0681(\pm 0.0013)\times(\alpha = 0.1) \\ & + 0.0037(\pm 0.0013)\times(\alpha = 0.01) + 0.0333(\pm 0.0013)\times(\alpha = 0.05)\end{aligned} \quad \text{with}$$

$R^2 = 0.96$.

There were no significant differences among those three methods with regard to the empirical type I error rate.

Table 3.1 Empirical type I error rate and its 95% confidence interval for chromosome 13 scoliosis data

Null Model 1: Three equiprobable trajectories

Significance Level			α=0.005			α=0.01			α=0.05			α=0.10			Failure rate (%)
Method			BPP	Modal	CT	BPP	Modal	CT	BPP	Modal	CT	BPP	Modal	CT	
Normal	Sigma	MAF													
No: Normal Squared	8	=0.49	0.005 ±0.004	0.004 ±0.004	0.006 ±0.006	0.012 ±0.007	0.012 ±0.007	0.01 ±0.008	0.045 ±0.013	0.044 ±0.013	0.05 ±0.02	**0.08** ±0.017	0.083 ±0.017	0.08 ±0.024	0
		=0.15	0.006 ±0.005	0.008 ±0.005	0.006 ±0.006	0.011 ±0.006	0.01 ±0.006	0.012 ±0.009	0.041 ±0.012	0.045 ±0.013	0.038 ±0.017	**0.078** ±0.016	0.087 ±0.017	**0.062** ±0.02	
		<0.05	0.003 ±0.003	0.003 ±0.003	0.003 ±0.003	0.006 ±0.005	0.006 ±0.005	0.006 ±0.006	**0.032** ±0.01	**0.033** ±0.01	**0.03** ±0.015	**0.069** ±0.016	**0.068** ±0.016	**0.06** ±0.02	
	4	=0.49	0.005 ±0.004	0.005 ±0.004	0.006 ±0.006	0.009 ±0.006	0.01 ±0.006	0.012 ±0.009	**0.034** ±0.011	**0.035** ±0.01	**0.03** ±0.015	**0.066** ±0.015	**0.072** ±0.016	**0.056** ±0.02	0.3
		=0.15	0.006 ±0.005	0.006 ±0.005	0.004 ±0.004	0.008 ±0.005	0.008 ±0.005	0.008 ±0.008	0.047 ±0.013	0.049 ±0.013	0.038 ±0.017	**0.087** ±0.017	0.09 ±0.018	0.084 ±0.024	
		<0.05	0.004 ±0.004	0.004 ±0.004	0.003 ±0.003	0.007 ±0.005	0.007 ±0.005	0.005 ±0.005	**0.032** ±0.011	**0.033** ±0.01	**0.03** ±0.015	**0.069** ±0.016	**0.069** ±0.016	**0.064** ±0.02	
Yes	8	=0.49	0.002 ±0.003	0.005 ±0.005	0.002 ±0.004	0.008 ±0.006	0.01 ±0.006	0.007 ±0.007	0.043 ±0.013	0.039 ±0.013	0.048 ±0.02	0.081 ±0.02	**0.08** ±0.018	0.082 ±0.026	12.2
		=0.15	0.002 ±0.003	0.003 ±0.003	0.002 ±0.004	0.007 ±0.005	**0.004** ±0.004	0.005 ±0.006	**0.023** ±0.01	**0.03** ±0.01	**0.03** ±0.016	**0.06** ±0.016	**0.06** ±0.016	**0.06** ±0.02	
		<0.05	0.003 ±0.004	0.003 ±0.003	0.003 ±0.005	0.006 ±0.005	0.006 ±0.005	0.007 ±0.007	**0.033** ±0.012	**0.032** ±0.01	**0.028** ±0.015	**0.061** ±0.016	**0.06** ±0.016	**0.062** ±0.02	
	4	=0.49	0.004 ±0.004	0.004 ±0.004	0.006 ±0.007	0.008 ±0.005	0.007 ±0.005	0.01 ±0.009	0.045 ±0.013	0.046 ±0.013	0.044 ±0.018	**0.074** ±0.016	**0.073** ±0.017	0.086 ±0.025	3
		=0.15	**0.001** ±0.002	0.002 ±0.003	0.002 ±0.004	**0.003** ±0.003	**0.003** ±0.003	0.004 ±0.006	0.038 ±0.012	0.04 ±0.012	0.042 ±0.018	**0.078** ±0.017	**0.077** ±0.017	0.08 ±0.024	
		<0.05	0.003 ±0.003	0.003 ±0.003	0.003 ±0.005	0.006 ±0.005	0.006 ±0.005	0.006 ±0.007	**0.032** ±0.011	**0.03** ±0.01	**0.031** ±0.016	**0.068** ±0.016	**0.067** ±0.016	**0.061** ±0.02	

Notes: The confidence intervals in bold indicate the parameter settings where α is not contained in the respective confidence interval.

Table 3.2 ANOVA table for empirical type I error rate using null model I

Variable source	DF	Mean square	F value	Pr>F
NORMAL	1	0.00019136	6.29	0.0133
SIGMA	1	0.00001344	0.44	0.5074
MAF	2	0.00056355	18.52	<.0001
α	3	0.03607492	1185.48	<.0001
METHOD	2	0.00002347	0.77	0.4645
Error	134	0.00003043		

3.1.2 Null Simulation II

Table 3.3 showed the empirical type I error rate and its 95% confidence interval using the second null model, which set the trajectory frequencies to Hardy Weinberg values and penetrance matrix values (the penetrance coefficient is denoted by $PENE$) as described in Chapter 2. Here, the variable $PENE$ had two levels: high penetrance and low penetrance. The failure rates of the TRAJ model using three methods were the same and they were shown in the first table of Table 3.3.

Table 3.4 showed the ANOVA table of empirical type I error. The variables $NORMAL$ and $SIGMA$ were significant. The variable α is highly significant. The regression model was

$$\begin{aligned}\text{empirical type I error} = {} & 0.0021(\pm 0.0012) + 0.0031(\pm 0.0007)\times(NORMAL = 0) \\ & + 0.0016(\pm 0.0007)\times(SIGMA = 4) + 0.0681(\pm 0.0010)\times(\alpha = 0.1) \quad \text{with} \\ & + 0.0036(\pm 0.0010)\times(\alpha = 0.01) + 0.0331(\pm 0.0010)\times(\alpha = 0.05)\end{aligned}$$

$R^2 = 0.968$.

The BPP and modal BPP had different empirical type I error rates ($p = 0.0004$). The empirical type I error rate for the modal BPP method was greater than the error rate using BPP method. Overall, the Modal BPP empirical type I error rate was closer to the target α. The empirical type I error rate of the contingency table method was essentially the same as that of that modal BPP method.

Table 3.3 Empirical Type I error rate and 95% confidence interval for chromosome 13 scoliosis data using BPP method

Null Model 2: Hardy Weinberg Distribution for Trajectory Groups

Normal	MAF	Penetrance Model	σ	$\alpha = 0.005$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$		Failure rate (%)
				OBS	95% CI	OBS	95% CI	OBS	95% CI	OBS	95% CI	
No—Normal Squared	MAF=0.49	High Penetrance	$\sigma = 8$	0.005	(0.001,0.009)	0.008	(0.002,0.013)	0.039	(0.027,0.051)	0.076	**(0.059,0.092)**	0
			$\sigma = 4$	0.003	(0,0.006)	0.006	(0.001,0.011)	0.040	(0.028,0.052)	0.078	**(0.061,0.095)**	0
		Low Penetrance	$\sigma = 8$	0.003	(0,0.006)	0.004	**(0,0.008)**	0.044	(0.031,0.057)	0.083	(0.066,0.100)	0
			$\sigma = 4$	0.005	(0.001,0.009)	0.007	(0.002,0.012)	0.038	(0.026,0.050)	0.075	**(0.059,0.091)**	0.1
	MAF=0.14	High Penetrance	$\sigma = 8$	0.007	(0.002,0.012)	0.010	(0.004,0.016)	0.037	**(0.025,0.049)**	0.061	**(0.046,0.076)**	0
			$\sigma = 4$	0.004	(0,0.008)	0.006	(0.001,0.011)	0.039	(0.027,0.051)	0.069	**(0.053,0.085)**	0
		Low Penetrance	$\sigma = 8$	0.004	(0,0.008)	0.006	(0.001,0.011)	0.036	(0.024,0.050)	0.076	**(0.059,0.092)**	0
			$\sigma = 4$	0.002	(0,0.005)	0.005	**(0.001,0.009)**	0.029	**(0.018,0.039)**	0.068	**(0.052,0.084)**	0
Yes	MAF=0.49	High Penetrance	$\sigma = 8$	0.003	(0,0.006)	0.007	(0.002,0.012)	0.027	**(0.017,0.037)**	0.053	**(0.039,0.067)**	27.8
			$\sigma = 4$	0.004	(0,0.008)	0.009	(0.003,0.015)	0.033	**(0.022,0.044)**	0.067	**(0.051,0.082)**	0.2
		Low Penetrance	$\sigma = 8$	0.001	**(0,0.003)**	0.003	**(0,0.006)**	0.025	**(0.015,0.035)**	0.052	**(0.038,0.066)**	27.4
			$\sigma = 4$	0.003	(0,0.006)	0.008	(0.002,0.013)	0.039	(0.027,0.051)	0.075	**(0.059,0.091)**	1.8
	MAF=0.14	High Penetrance	$\sigma = 8$	0.005	(0.001,0.009)	0.008	(0.002,0.013)	0.039	(0.027,0.051)	0.077	**(0.06,0.093)**	0
			$\sigma = 4$	0.004	(0,0.008)	0.006	(0.001,0.011)	0.038	(0.026,0.050)	0.070	**(0.054,0.086)**	0.1
		Low Penetrance	$\sigma = 8$	0.002	(0,0.005)	0.004	**(0,0.008)**	0.027	**(0.017,0.037)**	0.069	**(0.053,0.085)**	0
			$\sigma = 4$	0.003	(0,0.006)	0.007	(0.002,0.012)	0.031	**(0.020,0.042)**	0.072	**(0.056,0.088)**	0

Notes: The confidence intervals in bold indicate the parameter settings where α is not contained in the respective confidence interval.

Table 3.3 (continued) Empirical Type I error rate and 95% confidence interval for chromosome 13 scoliosis data using Modal BPP method

Null Model 2: Hardy Weinberg Distribution for Trajectory Groups

Normal	MAF	Penetrance Model	σ	$\alpha = 0.005$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
				OBS	95% CI	OBS	95% CI	OBS	95% CI	OBS	95% CI
No: Normal Squared	MAF=0.49	High Penetrance	$\sigma = 8$	0.005	(0.001,0.009)	0.010	(0.004,0.016)	0.049	(0.036,0.062)	0.079	**(0.062,0.096)**
			$\sigma = 4$	0.003	(0,0.006)	0.007	(0.002,0.012)	0.045	(0.032,0.058)	0.083	(0.066,0.1)
		Low Penetrance	$\sigma = 8$	0.003	(0,0.006)	0.007	(0.002,0.012)	0.046	(0.033,0.059)	0.085	(0.068,0.102)
			$\sigma = 4$	0.007	(0.002,0.012)	0.007	(0.002,0.012)	0.042	(0.029,0.054)	0.078	**(0.061,0.095)**
	MAF=0.14	High Penetrance	$\sigma = 8$	0.005	(0.001,0.009)	0.009	(0.003,0.015)	0.037	**(0.025,0.049)**	0.063	**(0.048,0.078)**
			$\sigma = 4$	0.004	(0,0.008)	0.008	(0.002,0.013)	0.042	(0.029,0.054)	0.074	**(0.058,0.09)**
		Low Penetrance	$\sigma = 8$	0.002	(0,0.005)	0.009	(0.003,0.015)	0.036	**(0.024,0.047)**	0.067	**(0.051,0.082)**
			$\sigma = 4$	0.002	(0,0.005)	0.005	**(0.001,0.009)**	0.034	**(0.023,0.045)**	0.070	**(0.054,0.086)**
Yes	MAF=0.49	High Penetrance	$\sigma = 8$	0.004	(0,0.008)	0.007	(0.002,0.012)	0.026	**(0.016,0.036)**	0.046	**(0.033,0.059)**
			$\sigma = 4$	0.005	(0.001,0.009)	0.008	(0.002,0.013)	0.035	**(0.024,0.046)**	0.070	**(0.054,0.086)**
		Low Penetrance	$\sigma = 8$	0.001	**(0,0.003)**	0.003	**(0,0.006)**	0.030	**(0.019,0.041)**	0.056	**(0.042,0.07)**
			$\sigma = 4$	0.004	(0,0.008)	0.008	(0.002,0.013)	0.040	(0.028,0.052)	0.077	**(0.06,0.093)**
	MAF=0.14	High Penetrance	$\sigma = 8$	0.004	(0,0.008)	0.006	(0.001,0.011)	0.043	(0.03,0.055)	0.083	(0.066,0.1)
			$\sigma = 4$	0.004	(0,0.008)	0.007	(0.002,0.012)	0.042	(0.029,0.054)	0.073	**(0.057,0.089)**
		Low Penetrance	$\sigma = 8$	0.002	(0,0.005)	0.006	(0.001,0.011)	0.027	**(0.017,0.037)**	0.069	**(0.053,0.085)**
			$\sigma = 4$	0.004	(0,0.008)	0.007	(0.002,0.012)	0.035	**(0.024,0.046)**	0.073	**(0.057,0.089)**

Notes: The confidence intervals in bold indicate the parameter settings where α is not contained in the respective confidence interval.

Table 3.3(continued) Empirical Type I error rate and 95% confidence interval for chromosome 13 scoliosis data using contingency table method

Null Model 2: Hardy Weinberg Distribution for Trajectory Groups

Normal	MAF	Penetrance Model	σ	$\alpha = 0.005$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
				OBS	95% CI	OBS	95% CI	OBS	95% CI	OBS	95% CI
No—Normal Squared	MAF=0.49	High Penetrance	$\sigma = 8$	0.003	(0,0.006)	0.010	(0.004,0.016)	0.043	(0.030,0.056)	0.072	**(0.056,0.088)**
			$\sigma = 4$	0.002	(0,0.005)	0.006	(0.001,0.010)	0.036	**(0.024,0.048)**	0.084	(0.067,0.101)
		Low Penetrance	$\sigma = 8$	0.002	(0,0.005)	0.008	(0.002,0.014)	0.040	(0.028,0.052)	0.080	**(0.063,0.097)**
			$\sigma = 4$	0.003	(0,0.006)	0.009	(0.003,0.015)	0.042	(0.029,0.054)	0.085	(0.068,0.102)
	MAF=0.14	High Penetrance	$\sigma = 8$	0.003	(0,0.006)	0.008	(0.002,0.014)	0.034	**(0.023,0.045)**	0.072	**(0.056,0.088)**
			$\sigma = 4$	0.002	(0,0.005)	0.008	(0.002,0.014)	0.043	(0.030,0.056)	0.076	**(0.060,0.092)**
		Low Penetrance	$\sigma = 8$	0.004	(0,0.008)	0.007	(0.002,0.012)	0.039	(0.027,0.051)	0.075	**(0.059,0.091)**
			$\sigma = 4$	0.001	**(0,0.003)**	0.005	**(0.001,0.009)**	0.032	**(0.021,0.043)**	0.059	**(0.044,0.074)**
Yes	MAF=0.49	High Penetrance	$\sigma = 8$	0.004	(0,0.008)	0.008	(0.002,0.014)	0.030	**(0.019,0.040)**	0.061	**(0.046,0.076)**
			$\sigma = 4$	0.005	(0,0.009)	0.009	(0.003,0.015)	0.034	**(0.023,0.045)**	0.072	**(0.056,0.088)**
		Low Penetrance	$\sigma = 8$	0.002	(0,0.005)	0.005	**(0.001,0.009)**	0.031	**(0.020,0.042)**	0.057	**(0.042,0.071)**
			$\sigma = 4$	0.002	(0,0.005)	0.009	(0.003,0.015)	0.041	(0.029,0.053)	0.076	**(0.060,0.092)**
	MAF=0.14	High Penetrance	$\sigma = 8$	0.004	(0,0.008)	0.007	(0.002,0.012)	0.038	(0.026,0.050)	0.078	**(0.061,0.095)**
			$\sigma = 4$	0.003	(0,0.006)	0.007	(0.002,0.012)	0.035	**(0.024,0.046)**	0.069	**(0.053,0.085)**
		Low Penetrance	$\sigma = 8$	0.002	(0,0.005)	0.005	**(0.001,0.009)**	0.035	**(0.024,0.046)**	0.071	**(0.055,0.087)**
			$\sigma = 4$	0.003	(0,0.006)	0.008	(0.002,0.014)	0.030	**(0.019,0.041)**	0.075	**(0.059,0.091)**

Notes: The confidence intervals in bold indicate the parameter settings where α is not contained in the respective confidence interval.

Table 3.4 ANOVA table of empirical type I error rate using null model II

Variable source	DF	Mean square	F Value	Pr>F
NORMAL	1	0.00045942	17.47	<.0001
SIGMA	1	0.00012192	4.64	0.0326
MAF	1	0.00001813	0.69	0.4074
α	3	0.04793349	1823.23	<.0001
PENE	1	0.00005105	1.94	0.1652
METHOD	2	0.00003032	1.15	0.3179
Error	182	0.00002629		

Table 3.5, Table 3.6 and Table 3.7 showed the proportion of replicates (1000 replicates per setting) for which the number of SNP markers within 10 markers of the target SNP locus were in the top 5%, top 10% or top 25% of markers on chromosome 13, using BPP method, modal BPP method and contingency table method respectively, using the second null model (Hardy Weinberg model). For example, on Table 3.5, there were 46.6% of the replicates for which no SNP markers within 10 markers of the target SNP locus were in the top 5% of the markers on the whole chromosome, under the parameter setting $NORMAL = 0$, $MAF = 0.49$, $PENE = 0$, and $SIGMA = 8$ and using the BPP method. Similarly, 47.5% of the replicates for which only 1 or 2 out of 10 SNP markers were in the top 10% of all the markers on chromosome 13 under the same parameter setting and using BPP method. The mean was calculated as:

$$MEAN = \sum_{i=0}^{20} i \times \Pr(i),$$

Here, i =Number of SNP markers within 10 markers of disease SNP. Among the independent variables: $NORMAL$, MAF, $PENE$, and $SIGMA$, $MEAN$ was correlated with $NORMAL(p < .05)$ and $MAF(p < .05)$. The independent variables $PENE$, $SIGMA$ were not associated with the dependent variable. The results obtained from the three methods were essentially the same.

Table 3.5 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the target SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (null model II) using BPP method

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
No: Normal Squared	0.49	High	$\sigma = 8$	5	46.6	47.5	5.6	0.3	0	0	0	0	0	0.82	0.90
				10	24.3	52	21	2.5	0.2	0	0	0	0	1.57	1.81
				25	3.1	25.2	38.5	23.9	7.7	1.4	0.2	0	0	3.77	4.02
			$\sigma = 4$	5	45.9	49.4	4.6	0.1	0	0	0	0	0	0.79	0.81
				10	21.9	57.4	18.1	2.5	0.1	0	0	0	0	1.54	1.60
				25	2.7	27.2	38.2	22.9	8.2	0.8	0	0	0	3.68	3.66
		Low	$\sigma = 8$	5	47.3	47.5	4.9	0.2	0.1	0	0	0	0	0.79	0.89
				10	22.2	56.2	18.8	2.4	0.4	0	0	0	0	1.58	1.75
				25	1.7	24.9	39.8	24.5	7.8	1.1	0.2	0	0	3.81	3.70
			$\sigma = 4$	5	49.85	45.74	4	0.4	0	0	0	0	0	0.73	0.82
				10	25.03	55.05	17.72	1.9	0.2	0.1	0	0	0	1.49	1.73
				25	3.6	26.63	38.54	23.53	6.4	1.2	0.1	0	0	3.64	3.76
	0.14	High	$\sigma = 8$	5	48	47.7	4.1	0.2	0	0	0	0	0	0.76	0.78
				10	22.6	57.4	18.2	1.7	0.1	0	0	0	0	1.51	1.54
				25	2.7	22.9	40.1	25.3	8.3	0.7	0	0	0	3.81	3.63
			$\sigma = 4$	5	47.9	47.3	4.8	0	0	0	0	0	0	0.76	0.78
				10	24.6	56.7	16.6	1.9	0.2	0	0	0	0	1.46	1.55
				25	4.1	26.6	39.7	23.4	5.2	0.8	0.2	0	0	3.57	3.55
		Low	$\sigma = 8$	5	49.8	46.1	4.1	0	0	0	0	0	0	0.71	0.74
				10	23.7	56.6	17	2.5	0.2	0	0	0	0	1.49	1.63
				25	3	26.2	38.1	23.3	8.3	1	0.1	0	0	3.75	3.89
			$\sigma = 4$	5	52.4	44.3	3.3	0	0	0	0	0	0	0.65	0.65
				10	24	57.2	18	0.8	0	0	0	0	0	1.44	1.38
				25	3	24.5	40.3	25.7	5.8	0.7	0	0	0	3.71	3.39

Table 3.5 (continued)

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
Yes	0.49	High	$\sigma = 8$	5	66.8	32	1.1	0.1	0	0	0	0	0	0.43	0.70
				10	51.7	41.3	6.9	0.1	0	0	0	0	0	0.81	1.01
				25	32.7	30.2	27	8.4	1.7	0	0	0	0	1.99	1.85
			$\sigma = 4$	5	51.5	43.69	4.71	0.1	0	0	0	0	0	0.72	0.89
				10	23.65	58.62	15.83	1.7	0.2	0	0	0	0	1.45	1.21
				25	2.3	27.16	40.28	23.45	5.51	1.2	0.1	0	0	3.61	1.86
		Low	$\sigma = 8$	5	73.7	24.4	1.9	0	0	0	0	0	0	0.36	0.69
				10	53.5	39.7	6.2	0.6	0	0	0	0	0	0.77	1.04
				25	32.7	33.1	24.9	8.3	1	0	0	0	0	1.86	1.78
			$\sigma = 4$	5	51.02	44.4	4.38	0.2	0	0	0	0	0	0.69	0.87
				10	27.39	53.26	17.32	2.03	0	0	0	0	0	1.44	1.26
				25	3.97	27.7	39.72	22.5	4.79	1.32	0	0	0	3.51	1.91
	0.14	High	$\sigma = 8$	5	50.1	44.5	5.2	0.2	0	0	0	0	0	0.75	0.91
				10	25.2	53	19.1	2.5	0.2	0	0	0	0	1.49	1.31
				25	2.2	24.5	38.4	26.2	7.3	1.4	0	0	0	3.83	1.93
			$\sigma = 4$	5	48.85	46.44	4.7	0	0	0	0	0	0	0.74	0.88
				10	24.42	54.85	18.72	2	0	0	0	0	0	1.51	1.27
				25	2.6	25.23	40.44	24.63	6.6	0.4	0.1	0	0	3.68	1.82
		Low	$\sigma = 8$	5	51.9	43.9	3.9	0.3	0	0	0	0	0	0.69	0.89
				10	24	56.2	17.9	1.7	0.2	0	0	0	0	1.47	1.26
				25	2.6	26.2	39.3	24.2	6.4	1.3	0	0	0	3.69	1.89
			$\sigma = 4$	5	51.4	44.4	4.2	0	0	0	0	0	0	0.69	0.86
				10	25.5	55.5	17.1	1.7	0.2	0	0	0	0	1.44	1.24
				25	2.6	26.5	38.7	24.3	6.2	1.7	0	0	0	3.74	1.94

Table 3.6 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the target SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (null model II) using Modal BPP method

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
No	0.49	High	$\sigma = 8$	5	49.9	43.9	5.9	0.3	0	0	0	0	0	0.76	0.95
				10	22.9	54.4	20.6	1.9	0.2	0	0	0	0	1.56	1.31
				25	3.6	24.8	39.5	24.3	6.2	1.4	0.2	0	0	3.71	1.98
			$\sigma = 4$	5	47.4	48.7	3.8	0.1	0	0	0	0	0	0.76	0.88
				10	23.1	58.3	16.5	2.1	0	0	0	0	0	1.48	1.23
				25	3.5	28.3	38.2	21.7	7.7	0.6	0	0	0	3.56	1.91
		Low	$\sigma = 8$	5	50	45.8	4	0.2	0	0	0	0	0	0.73	0.88
				10	24.6	56.6	16.8	1.8	0.2	0	0	0	0	1.47	1.25
				25	1.6	25.5	40.6	24.3	6.5	1.4	0.1	0	0	3.76	1.89
			$\sigma = 4$	5	50.65	45.54	3.4	0.4	0	0	0	0	0	0.69	0.87
				10	26.83	54.55	16.41	1.9	0.3	0	0	0	0	1.42	1.28
				25	3.5	29.03	38.74	21.82	5.8	1	0.1	0	0	3.53	1.91
	0.14	High	$\sigma = 8$	5	45.4	49.1	5.3	0.2	0	0	0	0	0	0.81	0.92
				10	21.2	57.9	18.7	2.2	0	0	0	0	0	1.56	1.25
				25	2.1	22.4	38.1	28.2	8.1	1.1	0	0	0	3.91	1.88
			$\sigma = 4$	5	45.6	49	5.4	0	0	0	0	0	0	0.81	0.9
				10	23.2	56.7	17.9	2.1	0.1	0	0	0	0	1.52	1.25
				25	3.8	25.1	39.5	24.1	6.5	0.9	0.1	0	0	3.67	1.9
		Low	$\sigma = 8$	5	48	47	4.8	0.2	0	0	0	0	0	0.76	0.9
				10	22.1	57.8	17.1	3	0	0	0	0	0	1.56	1.3
				25	2.5	24.4	36.2	26.9	8.2	1.7	0.1	0	0	3.87	2.01
			$\sigma = 4$	5	48.8	47.3	3.8	0.1	0	0	0	0	0	0.72	1.11
				10	22.4	57.4	18.9	1.3	0	0	0	0	0	1.52	1.21
				25	2.3	22.9	39.8	28	6.5	0.5	0	0	0	3.82	1.84

Table 3.6 (continued)

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
Yes	0.49	High	$\sigma = 8$	5	68.2	30.8	1	0	0	0	0	0	0	0.41	0.68
				10	51.7	40.8	7.3	0.2	0	0	0	0	0	0.79	1.00
				25	32.8	31	27.1	7.9	1.2	0	0	0	0	1.92	1.79
			$\sigma = 4$	5	53.71	42.48	3.71	0.1	0	0	0	0	0	0.67	0.86
				10	25.65	58.12	14.53	1.5	0.2	0	0	0	0	1.38	1.19
				25	2.3	27.65	41.69	22.35	5.21	0.8	0	0	0	3.52	1.81
		Low	$\sigma = 8$	5	76.2	22.3	1.5	0	0	0	0	0	0	0.32	0.65
				10	54.5	38.6	6.5	0.4	0	0	0	0	0	0.74	1.00
				25	33.4	34.1	24.7	6.3	1.5	0	0	0	0	1.81	1.76
			$\sigma = 4$	5	52.95	42.97	4.07	0	0	0	0	0	0	0.66	0.84
				10	29.12	53.56	15.58	1.73	0	0	0	0	0	1.36	1.24
				25	4.28	28.62	41.14	20.16	4.68	1.12	0	0	0	3.42	1.87
	0.14	High	$\sigma = 8$	5	48.7	45.7	5.5	0.1	0	0	0	0	0	0.79	0.93
				10	20.6	57.3	19.6	2.4	0.1	0	0	0	0	1.56	1.28
				25	1.7	22.7	39.3	26.4	8.2	1.6	0.1	0	0	3.90	1.92
			$\sigma = 4$	5	47.15	47.94	4.9	0	0	0	0	0	0	0.77	0.90
				10	22.22	55.95	19.22	2.5	0.1	0	0	0	0	1.58	1.29
				25	2.8	23.53	39.74	25.63	7.61	0.6	0.1	0	0	3.79	1.87
		Low	$\sigma = 8$	5	49.6	46.1	3.9	0.4	0	0	0	0	0	0.73	0.90
				10	23.2	55.5	19.3	1.5	0.5	0	0	0	0	1.54	1.31
				25	2.5	23.1	40.9	24.2	7.8	1.3	0.2	0	0	3.83	1.92
			$\sigma = 4$	5	48.7	46.4	4.9	0	0	0	0	0	0	0.74	0.87
				10	23.3	55.8	18.9	1.8	0.2	0	0	0	0	1.51	1.26
				25	2.4	24.9	38.1	26.2	6.3	2.1	0	0	0	3.83	1.95

Table 3.7 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the target SNP locus are in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (null simulation II) using contingency table method

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
No: Normal Squared	0.49	High	$\sigma = 8$	5	45.9	48.2	5.9	0	0	0	0	0	0	0.93	0.94
				10	21	58	18.7	2	0.2	0.1	0	0	0	1.66	1.65
				25	2.2	23.5	41	25.3	6.2	1.5	0.3	0	0	3.82	3.84
			$\sigma = 4$	5	44.24	52.15	3.5	0.1	0	0	0	0	0	0.91	0.80
				10	20.22	59.06	18.81	1.8	0.1	0	0	0	0	1.65	1.51
				25	1.9	24.83	42.34	23.93	6.2	0.6	0.2	0	0	3.71	3.47
		Low	$\sigma = 8$	5	45.55	49.75	4.6	0.1	0	0	0	0	0	0.91	0.88
				10	21.82	58.05	18.72	1.3	0.1	0	0	0	0	1.60	1.47
				25	1.4	24.23	39.54	25.83	8.21	0.8	0	0	0	3.86	3.64
			$\sigma = 4$	5	51.35	44.94	3.7	0	0	0	0	0	0	0.80	0.82
				10	25.63	54.25	17.42	2.4	0.3	0	0	0	0	1.58	1.76
				25	1.4	30.13	42.54	20.33	4.6	1	0.1	0	0	3.51	3.34
	0.14	High	$\sigma = 8$	5	46.75	48.55	4.3	0.4	0	0	0	0	0	0.90	0.93
				10	22.12	58.86	17.32	1.5	0.2	0	0	0	0	1.59	1.49
				25	2.3	24.03	41.64	24.22	6.6	1.2	0	0	0	3.76	3.63
			$\sigma = 4$	5	49.39	47.16	3.45	0	0	0	0	0	0	0.83	0.80
				10	24.04	58.21	16.43	1.21	0.1	0	0	0	0	1.52	1.43
				25	2.94	28.29	43.51	19.58	4.77	0.91	0	0	0	3.47	3.36
		Low	$\sigma = 8$	5	51.9	44.1	4	0	0	0	0	0	0	0.80	0.84
				10	25.9	55.5	16.2	2.2	0.2	0	0	0	0	1.53	1.65
				25	2.5	28.2	41.2	21.1	6.3	0.7	0	0	0	3.56	3.53
			$\sigma = 4$	5	50	46.4	3.6	0	0	0	0	0	0	0.82	0.81
				10	24.6	58	17	0.4	0	0	0	0	0	1.49	1.30
				25	2.6	27.7	42.1	23.4	3.8	0.4	0	0	0	3.50	3.11

Table 3.7 (continued)

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
Yes	0.49	High	$\sigma = 8$	5	66.6	32.3	0.8	0.3	0	0	0	0	0	0.53	0.64
				10	52.1	41.6	6.2	0.1	0	0	0	0	0	0.85	1.01
				25	34	31.7	25	7.9	1.4	0	0	0	0	1.89	3.38
			$\sigma = 4$	5	50.7	46.29	3.01	0	0	0	0	0	0	0.80	0.77
				10	25.03	60.14	13.33	1.4	0.1	0	0	0	0	1.45	1.35
				25	2.4	28.98	39.71	22.62	5.19	1	0.1	0	0	3.56	3.61
		Low	$\sigma = 8$	5	69.9	28	2.1	0	0	0	0	0	0	0.49	0.64
				10	52.8	41.2	5.7	0.3	0	0	0	0	0	0.83	1.02
				25	31.5	33.2	25.5	8.6	1.2	0	0	0	0	1.95	3.33
			$\sigma = 4$	5	53.48	42.56	3.96	0	0	0	0	0	0	0.78	0.84
				10	24.85	55.15	18.62	1.38	0	0	0	0	0	1.55	1.52
				25	4	30.26	40.69	20.64	3.23	1.18	0	0	0	3.37	3.45
	0.14	High	$\sigma = 8$	5	48.2	46.7	4.9	0.2	0	0	0	0	0	0.88	0.93
				10	26.5	54.5	17.1	1.8	0.1	0	0	0	0	1.52	1.60
				25	2.5	24.3	38.6	26.1	7.2	1.3	0	0	0	3.81	3.84
			$\sigma = 4$	5	47.16	48.09	4.74	0	0	0	0	0	0	0.89	0.87
				10	24.04	56.06	18.19	1.7	0	0	0	0	0	1.57	1.54
				25	2.3	27.4	39.5	23.2	6.2	1.4	0	0	0	3.67	3.77
		Low	$\sigma = 8$	5	52.9	45.3	1.8	0	0	0	0	0	0	0.74	0.69
				10	25.2	57	16.6	1.2	0	0	0	0	0	1.50	1.42
				25	2.9	26.1	40.2	24.1	5.6	1.1	0	0	0	3.65	3.63
			$\sigma = 4$	5	53.5	42.6	3.9	0	0	0	0	0	0	0.77	0.83
				10	25.8	57.3	16.2	0.7	0	0	0	0	0	1.46	1.34
				25	2.8	27.7	37.7	24	6.4	1.4	0	0	0	3.67	3.91

3.2 Simulated Power Results

3.2.1 Single locus model

Table 3.8 showed the simulated power to detect the disease locus associated with the progression of disease in a single-locus model. All the three methods had very high power ($power \approx 100\%$) to detect the disease locus for each setting of MAF, $SIGMA$, $PENE$ and $NORMAL$. The failure rates of the TRAJ procedure using the three methods are the same and they are shown in the first table of Table 3.8.

Table 3.8 Empirical power to detect association of disease SNP on chromosome 13 with scoliosis data (single locus model) using BPP method

Normal	MAF	Penetrance Model	σ	$\alpha = 0.005$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$		Failure rate (%)
				OBS	95% CI	OBS	95% CI	OBS	95% CI	OBS	95% CI	
No: Normal Squared	MAF=0.49	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
		Low Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
	MAF=0.14	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1.6
		Low Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
			$\sigma = 4$	0.999	(0.994,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
Yes	MAF=0.49	High Penetrance	$\sigma = 8$	0.999	(0.994,1)	0.999	(0.994,1)	0.999	(0.994,1)	1	(0.995,1)	1.0
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0.2
		Low Penetrance	$\sigma = 8$	0.993	**(0.98,0.997)**	0.993	**(0.98,0.997)**	0.993	**(0.98,0.997)**	0.994	**(0.985,0.998)**	18.4
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0.2
	MAF=0.14	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0.8
		Low Penetrance	$\sigma = 8$	0.997	**(0.99,0.999)**	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	0

Notes: The confidence intervals in bold indicate the parameter settings where α is not contained in the respective confidence interval.

Table 3.8 (continued) Empirical power to detect association of disease SNP on chromosome 13 with scoliosis data (single-locus model) using Modal BPP method

Normal	MAF	Penetrance Model	σ	$\alpha = 0.005$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
				OBS	95% CI	OBS	95% CI	OBS	95% CI	OBS	95% CI
No: Normal Squared	MAF=0.49	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
		Low Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
	MAF=0.14	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
		Low Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
Yes	MAF=0.49	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
		Low Penetrance	$\sigma = 8$	0.994	**(0.985,0.998)**	0.994	**(0.985,0.998)**	0.994	**(0.985,0.998)**	0.994	**(0.985,0.998)**
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
	MAF=0.14	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
		Low Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)

Notes: The confidence intervals in bold indicate the parameter settings where α is not contained in the respective confidence interval.

Table 3.8 (continued) Empirical power to detect association of disease SNP on chromosome 13 with scoliosis data (single-locus model) using contingency table method

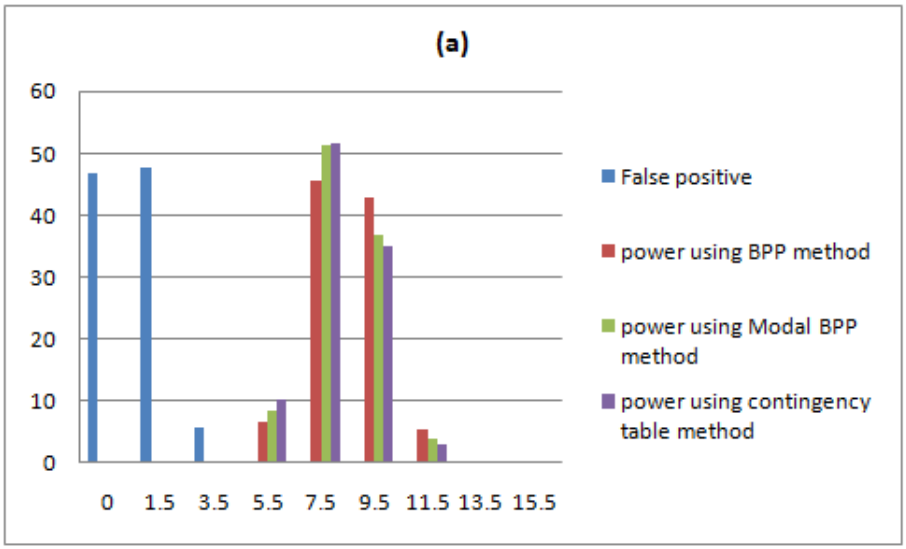
Normal	MAF	Penetrance Model	σ	$\alpha = 0.005$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
				OBS	95% CI	OBS	95% CI	OBS	95% CI	OBS	95% CI
No: Normal Squared	MAF=0.49	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
		Low Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
	MAF=0.14	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
		Low Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
Yes	MAF=0.49	High Penetrance	$\sigma = 8$	0.998	(0.992,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
		Low Penetrance	$\sigma = 8$	0.994	**(0.985,0.998)**	0.994	**(0.985,0.998)**	0.996	**(0.988,0.999)**	0.997	**(0.989,0.999)**
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
	MAF=0.14	High Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
		Low Penetrance	$\sigma = 8$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)
			$\sigma = 4$	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)	1	(0.995,1)

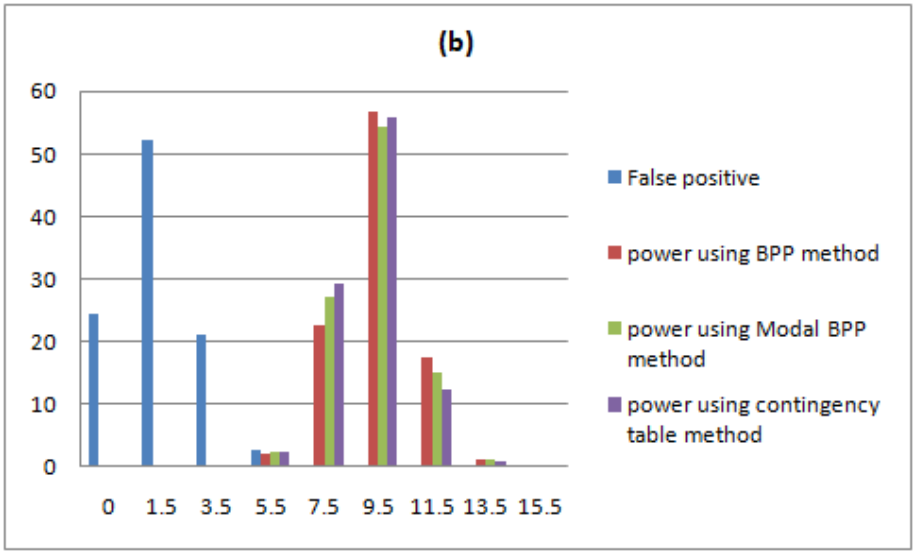
Notes: The confidence intervals in bold indicate the parameter settings where α is not contained in the respective confidence interval.

Tables 3.9, 3.10 and 3.11 showed the proportion of replicates for which the number of SNP markers within 10 markers of the disease SNP locus were in the top 5%, top 10% or top 25% of markers on chromosome 13 using BPP method, modal BPP method and contingency table method, respectively using the power simulation model. The proportions were 'drifting right' in tables 3.9, 3.10 and 3.11 (compared to tables 3.5, 3.6 and 3.7), indicating the clustering phenomena of the markers around the disease locus. That is, the markers around the disease locus had much higher probability to rank in the top 5%, 10% and 25% among all the loci on chromosome 13.

Figure 3.1 showed the proportion of replicates by the number of significant SNPs. That is, the x-axis was the number of the SNPs which were in the top 5% (figure (a)), top 10% (figure (b)) and top 25% (figure (c)) of the markers on chromosome 13. The y-axis was the proportion of replicates in which the number of significant markers was as given. For the null simulations (false positive), the number of significant markers was most likely 0, 1 or 2, while for the single locus power simulations, the number of significant markers was mostly 7 and above. This figure showed the clustering phenomenon of significant SNP markers around the disease SNP. For both disease SNPs at low MAF level and high MAF level, a few significant SNPs were clustering around the disease SNP. My study also showed that for high MAF level disease SNPs, there were more significant SNP markers around the loci.

Figure 3.1 Proportion of replicates for which the number of SNP markers within 10 markers of the disease SNP locus were in the top 5% (figure(a)), top 10% (figure(b)) and top 25% (figure(c)) of chromosome 13 by the number of significant SNP markers (parameter setting: no normal square transformation, MAF=0.49, high penetrance model and $\sigma = 8$)





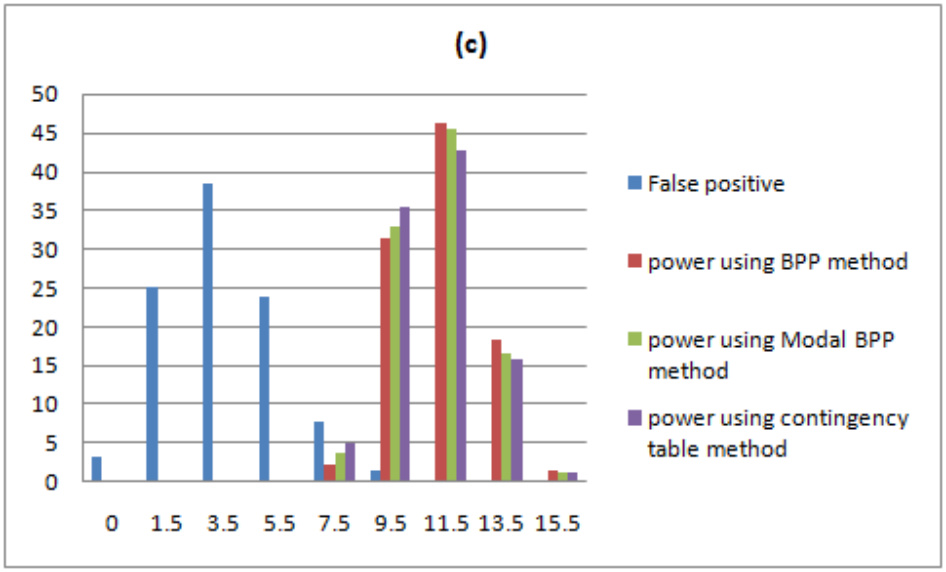

(c)
50
45
40
35
30
25
20
15
10
5
0
0
1.5
3.5
5.5
7.5
9.5
11.5
13.5
15.5
False positive
power using BPP method
power using Modal BPP method
power using contingency table method

Table 3.9 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the disease SNP locus were in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (single-locus model) using BPP method

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
No	0.49	High	$\sigma = 8$	5	0	0	0.1	6.4	45.5	42.8	5.1	0.1	0	8.39	1.31
				10	0	0	0	2.0	22.7	56.6	17.6	1.1	0	9.36	1.36
				25	0	0	0	0	2.3	31.5	46.2	18.4	1.6	11.21	1.52
			$\sigma = 4$	5	0	0	0	6.2	41.8	44.8	7.0	0.2	0	8.56	1.34
				10	0	0	0	1.4	19.5	54.5	22.5	2.1	0	9.58	1.41
				25	0	0	0	0.1	2.1	24.2	51.1	19.4	3.1	11.43	1.51
		Low	$\sigma = 8$	5	0	0	13.4	47.5	31.1	7.5	0.5	0	0	6.18	1.57
				10	0	0	4.2	29.8	41.7	19.5	4.6	0.1	0.1	7.34	1.77
				25	0	0	0.5	5.1	25.2	36.8	24.2	6.8	1.4	9.61	2.06
			$\sigma = 4$	5	0	0	13.5	46.9	32.2	7.1	0.3	0	0	6.19	1.53
				10	0	0	3.3	29.0	41.6	21.6	4.1	0.4	0	7.40	1.74
				25	0	0	0.2	4.0	22.9	37.3	26.7	8.4	0.5	9.76	1.96
	0.14	High	$\sigma = 8$	5	0	65.0	30.7	3.8	0.5	0	0	0	0	2.23	1.16
				10	0	38.0	45.7	14.0	1.9	0.4	0	0	0	3.12	1.47
				25	0	5.9	27.8	38.2	20.4	7.1	0.5	0.1	0	5.46	1.97
			$\sigma = 4$	5	0	2.84	54.07	39.33	3.66	0.1	0	0	0	4.37	1.15
				10	0	1.42	25.3	54.88	16.57	1.83	0	0	0	5.34	1.39
				25	0	0.2	3.56	24.7	39.64	27.03	4.27	0.51	0.1	7.58	1.83
		Low	$\sigma = 8$	5	0	74.9	23.8	1.3	0	0	0	0	0	1.96	0.97
				10	0	45.7	42.3	11.4	0.5	0.1	0	0	0	2.83	1.33
				25	0	5.8	31.0	38.9	19.3	4.4	0.6	0	0	5.24	1.88
			$\sigma = 4$	5	0	74.2	25.1	0.7	0	0	0	0	0	1.95	0.93
				10	0	45.8	44.7	8.8	0.6	0.1	0	0	0	2.79	1.28
				25	0	5.8	33.5	41.6	14.8	3.8	0.4	0.1	0	5.07	1.81

Table 3.9(continued)

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
Yes	0.49	High	$\sigma = 8$	5	0	0.1	0	9.6	45.15	40.7	4.35	0.1	0	8.29	1.39
				10	0	0	0.1	1.31	23.94	55.56	17.57	1.51	0	9.38	1.41
				25	0	0	0.1	0.1	2.72	29.8	46.57	18.49	2.23	11.23	1.61
			$\sigma = 4$	5	0	0	0.1	5.1	43.8	46.3	4.5	0	0.2	8.52	1.38
				10	0	0	0	1.1	21.6	57.9	17.9	1.3	0.2	9.47	1.42
				25	0	0	0	0	2.2	29.6	47.8	18.1	2.3	11.28	1.61
		Low	$\sigma = 8$	5	0	0.12	27.12	55.36	14.3	3.08	0	0	0	5.36	1.41
				10	0	0	10.85	46.24	32.55	8.63	1.73	0	0	6.39	1.64
				25	0	0	0.99	14.92	38.1	31.69	12.58	1.48	0.25	8.40	1.91
			$\sigma = 4$	5	0	0	22	52.4	21.9	3.2	0.3	0	0.2	5.68	1.61
				10	0	0	8.1	39.2	37.4	13.1	2	0	0.2	6.77	1.8
				25	0	0	0.8	9.3	31.9	34.9	18.3	4	0.8	9.01	2.11
	0.14	High	$\sigma = 8$	5	0	78.6	20.4	1	0	0	0	0	0	1.89	0.93
				10	0	48.1	42.4	8.9	0.6	0	0	0	0	2.72	1.26
				25	0	7.5	35.2	39.1	15	3.2	0	0	0	4.93	1.78
			$\sigma = 4$	5	0	67.3	26.5	5.1	0.3	0	0	0	0.8	2.35	2.05
				10	0	41.9	41.9	12.9	2.2	0.3	0	0	0.8	3.15	2.19
				25	0	5.5	27.8	39.2	20.2	5.4	1.1	0	0.8	5.52	2.4
		Low	$\sigma = 8$	5	0	79.6	19.3	1.1	0	0	0	0	0	1.83	0.89
				10	0	50.2	42	7.2	0.6	0	0	0	0	2.64	1.25
				25	0	7.8	38	35.6	15.2	3.2	0.2	0	0	4.86	1.83
			$\sigma = 4$	5	0	79.5	19.1	1.4	0	0	0	0	0	1.87	0.92
				10	0	48.7	41.9	8.7	0.7	0	0	0	0	2.71	1.27
				25	0	7	35.8	40	13.5	3.3	0.4	0	0	4.92	1.78

Table 3.10 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the disease SNP locus were in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (single-locus model) using Modal BPP method

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
No	0.49	High	$\sigma = 8$	5	0	0	0.1	8.4	51.1	36.7	3.7	0	0	8.17	1.3
				10	0	0	0	2.3	27.2	54.4	15.0	1.1	0	9.19	1.39
				25	0	0	0	0	3.8	32.8	45.6	16.5	1.3	11.10	1.52
			$\sigma = 4$	5	0	0	0	7.7	44.8	41.5	5.8	0.2	0	8.43	1.34
				10	0	0	0	1.6	22.2	54.7	20.2	1.3	0	9.45	1.39
				25	0	0	0	0.1	2.4	26.1	50.8	17.9	2.7	11.33	1.52
		Low	$\sigma = 8$	5	0	0	16.9	48.7	28.5	5.5	0.4	0	0	5.99	1.52
				10	0	0	3.9	33.4	41.1	18.4	2.8	0.3	0.1	7.16	1.74
				25	0	0	0.3	6.4	26.3	37.5	22.9	5.6	1.0	9.45	2.05
			$\sigma = 4$	5	0	0	14.5	50	29.8	5.5	0.2	0	0	6.05	1.48
				10	0	0	4.0	32	40.9	19.3	3.4	0.4	0	7.24	1.71
				25	0	0	0.2	4.6	23.8	38.8	24.7	7.4	0.5	9.65	1.96
	0.14	High	$\sigma = 8$	5	0	66.8	28.6	4.2	0.4	0	0	0	0	2.21	1.17
				10	0	37.2	44.8	15.6	2.2	0.2	0	0	0	3.13	1.48
				25	0	5.1	28.8	37.5	20.9	6.2	1.4	0.1	0	5.47	2.0
			$\sigma = 4$	5	0	2.74	47.36	45.53	4.26	0.1	0	0	0	4.54	1.16
				10	0	1.32	21.35	57.11	18.4	1.73	0.1	0	0	5.46	1.38
				25	0	0.2	3.05	23.68	41.05	26.93	4.57	0.41	0.1	7.63	1.8
		Low	$\sigma = 8$	5	0	74.7	23.8	1.4	0.1	0	0	0	0	1.99	0.96
				10	0	43.9	43.6	11.2	1.2	0.1	0	0	0	2.89	1.36
				25	0	4.9	32.1	40.2	17.5	4.7	0.6	0	0	5.25	1.85
			$\sigma = 4$	5	0	72.9	26	1.1	0	0	0	0	0	1.99	0.95
				10	0	42.9	46.6	10	0.5	0	0	0	0	2.87	1.28
				25	0	5.7	31.7	41.5	16.7	4.1	0.3	0	0	5.18	1.82

Table 3.10 (Continued)

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
Yes	0.49	High	$\sigma = 8$	5	0	0	0	12.74	50.25	33.67	3.24	0.1	0	8.05	1.35
				10	0	0	0	1.92	30.33	51.46	14.96	1.31	0	9.17	1.43
				25	0	0	0	0.1	3.44	32.86	44.39	16.88	2.33	11.13	1.61
			$\sigma = 4$	5	0	0	0.1	6.6	46.6	43.2	3.3	0	0.2	8.36	1.36
				10	0	0	0	1.3	24.4	57.1	16	1	0.2	9.35	1.42
				25	0	0	0	0	2.5	30.8	47.2	17.3	2.2	11.21	1.62
		Low	$\sigma = 8$	5	0	0.12	29.59	55.73	12.58	1.97	0	0	0	5.22	1.33
				10	0	0.12	12.83	48.21	30.21	7.27	1.36	0	0	6.23	1.60
				25	0	0	1.61	16.53	38.85	31.57	9.25	2.09	0.12	8.26	1.94
			$\sigma = 4$	5	0	0	24	53.2	20	2.3	0.3	0	0.2	5.57	1.55
				10	0	0	8.7	40.3	37.6	11.7	1.5	0	0.2	6.67	1.76
				25	0	0	0.9	10.6	34.4	32.5	17.4	3.4	0.8	8.87	2.10
	0.14	High	$\sigma = 8$	5	0	76.4	22.8	0.8	0	0	0	0	0	1.93	0.92
				10	0	47.6	41.6	10.3	0.5	0	0	0	0	2.77	1.28
				25	0	6.7	32.8	40.2	15.8	4.3	0.2	0	0	5.07	1.81
			$\sigma = 4$	5	0	66	27.6	5.3	0.3	0	0	0	0.8	2.38	2.06
				10	0	38.9	44	13.3	2.7	0.3	0	0	0.8	3.23	2.21
				25	0	5.5	26.2	39.6	20.7	6.1	1.1	0	0.8	5.62	2.40
		Low	$\sigma = 8$	5	0	77.3	21.6	1.1	0	0	0	0	0	1.85	0.90
				10	0	48.2	44.1	7.1	0.6	0	0	0	0	2.68	1.26
				25	0	8.2	34.2	38.3	16.7	2.3	0.3	0	0	4.95	1.78
			$\sigma = 4$	5	0	77.1	20.8	2.1	0	0	0	0	0	1.93	0.96
				10	0	46.5	43.4	9.2	0.9	0	0	0	0	2.77	1.29
				25	0	6.4	35.1	39.1	15.2	3.6	0.6	0	0	5.03	1.80

Table 3.11 Proportion of replicates (%) for which the number of SNP markers within 10 markers of the disease SNP locus were in the top 5%, top 10% or top 25% of markers on chromosome 13 for scoliosis data (single-locus model) using contingency table method

Normal	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
No	0.49	High	$\sigma = 8$	5	0	0	0.4	10.2	51.6	34.9	2.9	0	0	8.09	1.98
				10	0	0	0.1	2.1	29	55.9	12.1	0.8	0	9.10	1.99
				25	0	0	0	0.1	4.9	35.4	42.8	15.7	1.1	10.95	2.73
			$\sigma = 4$	5	0	0	0	7.2	46.9	43.3	2.6	0	0	8.33	1.75
				10	0	0	0.1	1.8	24	53.8	19.1	1.2	0	9.37	2.22
				25	0	0	0	0.1	3.3	29	48.8	16.4	2.4	11.21	2.68
		Low	$\sigma = 8$	5	0	0	17.4	50.3	26.8	5.2	0.3	0	0	5.91	2.54
				10	0	0	3.6	35.2	41.2	17.5	2.2	0.3	0	7.11	2.99
				25	0	0	0.2	6.2	27.4	39.8	20.1	5.4	0.9	9.36	4.13
			$\sigma = 4$	5	0	0	14.7	49.4	30.9	4.9	0.1	0	0	6.03	2.37
				10	0	0	3.9	32.2	43	17.4	3.3	0.2	0	7.19	3.11
				25	0	0	0.1	4.5	26.1	37.5	24.3	7.1	0.4	9.59	4.04
	0.14	High	$\sigma = 8$	5	0	64.9	30.7	4.1	0.3	0	0	0	0	2.29	1.36
				10	0	36.6	46.7	14.7	1.9	0.1	0	0	0	3.14	2.26
				25	0	3.8	28.7	40.2	20.8	5.5	0.9	0.1	0	5.47	3.85
			$\sigma = 4$	5	0	1.9	48.27	46.12	3.69	0.01	0	0	0	4.53	1.45
				10	0	1.11	22.13	56.75	18.21	1.71	0.1	0	0	5.45	2.09
				25	0	0.15	3.74	24.75	40.27	26.17	4.48	0.35	0.07	7.57	3.58
		Low	$\sigma = 8$	5	0	76.3	22.9	0.6	0.2	0	0	0	0	1.99	0.84
				10	0	44.8	42.7	10.6	1.6	0.3	0	0	0	2.89	2.22
				25	0	5.7	31.6	40.3	17.3	4.6	0.5	0	0	5.20	3.69
			$\sigma = 4$	5	0	71.4	27.6	1	0	0	0	0	0	2.09	0.91
				10	0	42.1	47.5	10.1	0.3	0	0	0	0	2.87	1.74
				25	0	5.5	31.6	42.6	16.2	3.9	0.2	0	0	5.14	3.36

Table 3.11 (continued)

Box-Cox	MAF	Penetrance Model	σ	Top (%)	Number of SNP markers within 10 markers of disease SNP									Mean	SD
					0	1-2	3-4	5-6	7-8	9-10	11-12	13-14	15-20		
Yes	0.49	High	$\sigma = 8$	5	0	0	0	12.06	51.76	33.28	2.83	0.07	0	8.04	1.99
				10	0	0	0	1.54	31.29	51.57	14.41	1.17	0	9.14	2.15
				25	0	0	0	0.06	3.12	34.09	45.3	15.3	2.13	11.08	2.66
			$\sigma = 4$	5	0	0	0	6.3	47.2	43.7	2.4	0.3	0.1	8.37	1.79
				10	0	0	0	1.2	25.6	56.8	15.6	0.7	0.1	9.29	1.94
				25	0	0	0	0	1.6	31.3	47.1	17.6	2.4	11.26	2.54
		Low	$\sigma = 8$	5	0	0.21	29.76	56.72	11.63	1.67	0	0	0	5.19	1.87
				10	0	0.15	12.99	48.15	30.06	7.3	1.35	0	0	6.21	2.90
				25	0	0	1.64	18.09	40.11	30.71	7.67	1.79	0.01	8.10	3.72
			$\sigma = 4$	5	0	0	21.9	52.9	22.2	2.7	0.1	0	0.2	5.64	2.41
				10	0	0	8.9	42	36.7	11.1	1.1	0	0.2	6.59	3.02
				25	0	0	1.3	11.2	35.9	31.3	16.5	3.2	0.6	8.75	4.52
	0.14	High	$\sigma = 8$	5	0	74.6	23.4	2	0	0	0	0	0	2.05	0.95
				10	0	48.4	41.9	9.4	0.3	0	0	0	0	2.73	1.77
				25	0	7.4	33.9	40.4	14.5	3.6	0.2	0	0	4.97	3.49
			$\sigma = 4$	5	0	67.3	26.7	5.1	0.2	0	0	0	0.7	2.35	2.61
				10	0	39.6	45.2	12	2.3	0.2	0	0	0.7	3.14	3.38
				25	0	5.6	27	39.8	19.7	6.1	1	0	0.8	5.51	4.90
		Low	$\sigma = 8$	5	0	76.5	22.2	1.3	0	0	0	0	0	1.99	0.85
				10	0	47.9	45.2	6.5	0.4	0	0	0	0	2.69	1.58
				25	0	7.9	36.7	36.4	16.5	2.2	0.3	0	0	4.88	3.47
			$\sigma = 4$	5	0	77.6	19.7	2.7	0	0	0	0	0	2.00	0.97
				10	0	46.8	44.5	8.4	0.3	0	0	0	0	2.74	1.68
				25	0	6.5	36.2	38.3	14.9	3.5	0.6	0	0	4.99	3.59

3.2.2 Multi-locus model

Table 3.12 showed the simulated power to detect each locus associated with the progression of disease in a multi-locus model. The failure rates using the TRAJ procedure are 0 in all settings. I fit a generalized linear model to the variables. Here, the simulated power was the dependent variable, and $NORMAL, MAF$, target α level and $METHOD$ were the independent variables. In Table 3.13, the variables MAF and α level were significant ($p < 0.0001$) in the model, but not $NORMAL$ variable ($p = 0.3552$) or $METHOD$ variable ($p = 0.4688$). The empirical power of the three methods was essentially the same. The regression model was

$$\begin{aligned}\text{empirical power} = &-1.9673(\pm 3.5579) + 1254.92(\pm 79.4324) \times MAF + 46.2867(\pm 2.8546) \times (\alpha = 0.1) \\ &+ 23.9600(\pm 2.8546) \times (\alpha = 0.01) + 42.2400(\pm 2.8546) \times (\alpha = 0.05)\end{aligned}$$

with $R^2 = 0.71$.

I further examined the power controlling for whether or not the data was normally distributed ($NORMAL$ =0 or 1). Figure 3.2 displayed three charts, each of which showed the power to detect each locus by method when the data was normally distributed ($NORMAL$ =1). The horizontal axis was the MAF of a causal SNP. The vertical axis was the power. Each graph contained four curves, one for each of the four target alpha levels. The three charts in Figure 3.2 were quite similar: the powers were approximately proportional to MAF. That is, the power usually increased as MAF increased, except for one SNP with $MAF = 0.016$. The power strictly increased as α increased. Figure 3.3 contained the results when the data was the square of normally distributed values. The patterns were the same as in Figure 3.2.

There was a power drop at MAF=0.016 in Figure 3.2, at the SNP rs9543107. In Table 2.4 in Chapter 2, the SNP rs9543107 was in high linkage disequilibrium with four other SNPs. This could be an explanation of the power drop. Another power drop in Figure 3.2 happened at MAF=0.033, at the SNP rs7331979. In Table 2.4, the SNP rs7331979 was also in high linkage disequilibrium with other three SNPs. In Figure 3.2, for the other SNPs which did not have more than two high linkage disequilibrium with other SNPs, the power increased as the MAF increased.

To understand the differences among the three methods, I compared any two paired methods using t-test. From the results, BPP method and Modal BPP method were both significantly different from contingency table method with regard to the power comparison ($p < 0.001$), for both normally distributed data and data that was the square of normally distributed data. The power using the BPP method was not significantly different from the power using modal assignment method when the data was normally distributed ($p = 0.40$). When the data was the square of normally distributed data, the power of the modal BPP was somewhat higher than the power using the BPP ($p = 0.0165$).

I also used the Cochran's test (or Cochran's Q test) and McNemar's test to compare the distributions of the detection rate using three methods, which was shown in Table 3.14. For each SNP, I set a binary variable

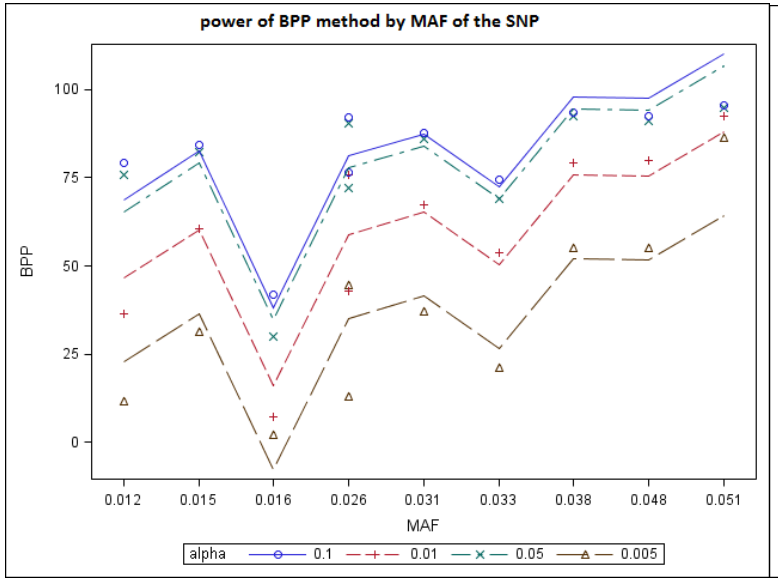
$$z_i = \begin{cases} 1, \text{SNP i is detected using one of the methods} \\ 0, \text{otherwise} \end{cases}.$$

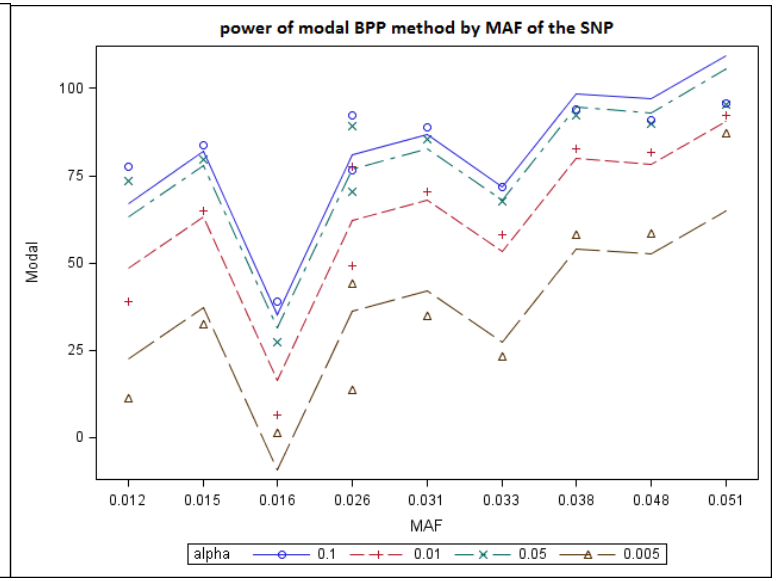
Then z_i was measured using three methods. The hypothesis of Cochran's test was:

H_0: the marginal probability of a positive response was unchanged across the three methods.

For 70% of the SNPs, the p-values of the Cochran test were small ($p<0.05$), which indicated that the probability of detecting each of the loci using three methods were different. Further, I used the McNemar's test to compare any pair of the methods. The BPP method and the modal BPP method were essentially the same with regard to the detection rates. However, for some of the loci, comparing the BPP method and modal BPP method with the contingency table method, the detection rates were different.

Figure 3.2 Power of Procedure by MAF of Locus in Multi-locus Model, Normally Distributed Data, for selected target levels of significance





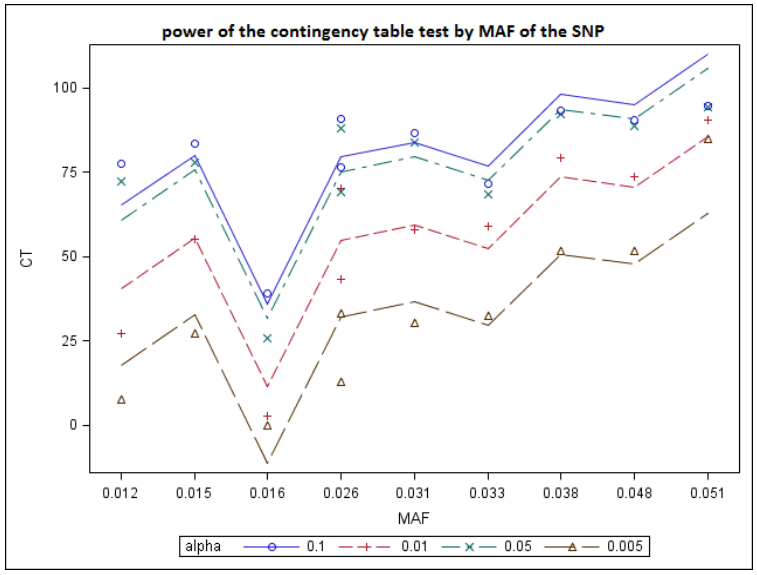


Figure 3.3 Power of Procedure by MAF of Locus in Multi-locus Model, Data Square of Normally Distributed Data, for selected target levels of significance

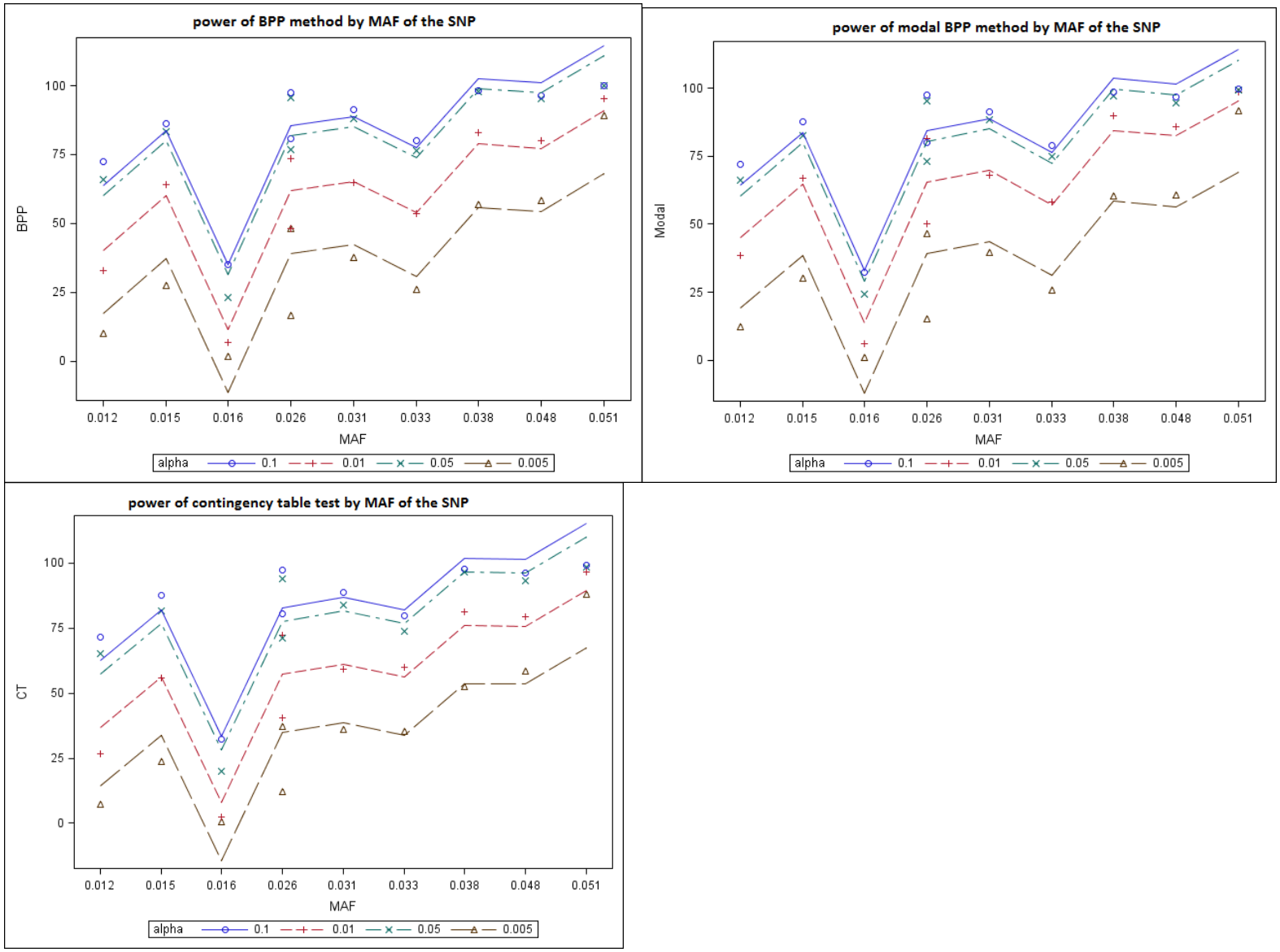


Table 3.12 Power Simulation to detect the association of disease SNP on chromosome 13 with scoliosis data (multi-locus model)

Normal	SNP	BP	MAF	α=0.005			α=0.01			α=0.05			α=0.10		
				BPP	Modal	CT	BPP	Modal	CT	BPP	Modal	CT	BPP	Modal	CT
No: Normal Squared	rs9599854	71021185	0.026	12.8	13.6	12.8	42.8	49.2	43.2	72	70.4	69.2	76.4	76.4	76.4
	rs9542756	71309666	0.038	55.2	58	51.6	79.2	82.8	79.2	92.4	92.4	92.4	93.6	94	93.2
	rs9543107	72217237	0.016	2	1.2	0	7.2	6.4	2.8	30	27.2	26	42	38.8	39.2
	rs17090361	73186500	0.051	86.4	87.2	84.8	92.4	92.4	90.4	94.8	95.2	94.4	95.6	95.6	94.8
	rs9593132	75293621	0.048	55.2	58.4	51.6	80	81.6	73.6	91.2	90	88.8	92.4	90.8	90.4
	rs5352	77373231	0.012	11.6	11.2	7.6	36.4	38.8	27.2	76	73.6	72.4	79.2	77.6	77.6
	rs7331979	78836214	0.033	21.2	23.2	32.4	53.6	58	59.2	69.2	67.6	68.4	74.4	71.6	71.6
	rs12863734	85268572	0.015	31.2	32.4	27.2	60.4	64.8	55.2	82.4	79.6	78	84.4	83.6	83.6
	rs9522610	89110831	0.026	44.4	44	33.2	76	77.6	70.4	90.4	89.2	88	92	92.4	90.8
	rs16943207	89144779	0.031	37.2	34.8	30.4	67.2	70.4	58	86	85.6	84	87.6	88.8	86.8
Yes	rs9599854	71021185	0.026	16.4	15.2	12	48	50	40.4	76.8	73.2	71.2	80.8	80	80.4
	rs9542756	71309666	0.038	56.8	60.4	52.4	82.8	90	81.2	98	97.2	96.4	98.4	98.4	97.6
	rs9543107	72217237	0.016	1.6	0.8	0.4	6.8	6	2.4	23.2	24.4	20	35.2	32.4	32.4
	rs17090361	73186500	0.051	89.2	91.6	88	95.2	98.4	96.4	100	99.2	98.4	100	99.6	99.2
	rs9593132	75293621	0.048	58.4	60.8	58.4	80	86	79.2	95.2	94.4	93.2	96.4	96.8	96
	rs5352	77373231	0.012	10	12.4	7.2	32.8	38.4	26.8	66	66	65.2	72.4	72	71.6
	rs7331979	78836214	0.033	26	25.6	35.2	53.6	58	60	76.4	74.8	73.6	80	78.8	79.6
	rs12863734	85268572	0.015	27.6	30	23.6	64	66.8	56	83.2	82.4	81.6	86.4	87.6	87.6
	rs9522610	89110831	0.026	48	46.4	37.2	73.6	81.6	72.4	95.6	95.2	94	97.6	97.6	97.2
	rs16943207	89144779	0.031	37.6	39.6	36	64.8	68	59.2	88	88.4	84	91.2	91.2	88.8

Table 3.13 ANOVA table of empirical power using multi-locus model

Variable source	DF	Mean square	F Value	Pr>F
NORMAL	1	209.81400	0.86	0.3552
MAF	1	61017.53801	249.60	<.0001
α	3	26748.84333	109.42	<.0001
METHOD	2	185.79467	0.76	0.4688
Error	232	56715.9680		

Table 3.14 P-values of Cochran's test and McNemar's test when comparing the three methods in multi-locus model

Normal	SNP	BP	MAF		McNemar's test		
				Cochran test	BPP vs. Modal	BPP vs. CT	Modal vs. CT
No Normal Squared	rs9599854	71021185	0.026	0.81	0.77	0.99	0.75
	rs9542756	71309666	0.038	**0.01**	0.26	0.14	**0.009**
	rs9543107	72217237	0.016	0.07	0.69	0.99	NA
	rs17090361	73186500	0.051	0.21	0.73	0.45	0.15
	rs9593132	75293621	0.048	**0.008**	0.15	0.16	**0.004**
	rs5352	77373231	0.012	**0.02**	0.78	**0.02**	**0.03**
	rs7331979	78836214	0.033	**<0.0001**	0.38	**<0.0001**	**<0.0001**
	rs12863734	85268572	0.015	**0.05**	0.71	0.13	**0.02**
	rs9522610	89110831	0.026	**<0.0001**	0.86	**<0.0001**	**<0.0001**
	rs16943207	89144779	0.031	**0.008**	0.40	**0.003**	0.05
Normal Squared	rs9599854	71021185	0.026	**0.04**	0.70	**0.01**	0.09
	rs9542756	71309666	0.038	**0.003**	0.17	0.09	**<0.0001**
	rs9543107	72217237	0.016	0.10	0.50	0.25	0.99
	rs17090361	73186500	0.051	**0.05**	0.21	0.63	**0.01**
	rs9593132	75293621	0.048	0.46	0.34	0.99	0.38
	rs5352	77373231	0.012	**0.003**	0.26	0.06	**<0.0001**
	rs7331979	78836214	0.033	**<0.0001**	0.82	**<0.0001**	**<0.0001**
	rs12863734	85268572	0.015	**0.006**	0.33	0.09	**<0.0001**
	rs9522610	89110831	0.026	**<0.0001**	0.58	**<0.0001**	**<0.0001**
	rs16943207	89144779	0.031	0.17	0.40	0.52	0.11

Notes: p-values in bold indicate the significance under confidence level $\alpha = 0.05$.

Chapter 4 Conclusions and discussions

In this dissertation, I examined three methods: using the BPP as the quantitative trait, using the indicator variable that modal BPP was in the clinically important group as the trait, and the contingency table method to test the association with the SNPs on chromosome 13. I simulated two genetic models, the single-locus model and the multi-locus model. In the single locus model, I assumed that the disease is caused by a single locus, and I studied two disease SNPs, with MAF at 0.15 and 0.5 respectively. In the multi-locus model, I assumed that the disease is caused equally by ten rare variant SNPs, each with MAF smaller than 0.05. I conducted the null simulation and the power simulation and reported the empirical type I error rate and empirical power to detect the disease SNPs using the three methods.

In the null simulations, my study suggested that the empirical type I error rate generally held the nominal α rate when α was small. However, when α increased (α near 0.05), there was a decrease of the empirical type I error rate below the nominal rate as the nominal rate

increased. There were no significant differences among the three methods. Null model I and null model II had the similar results. The failure rate of the TRAJ procedure was higher in the squared data model than the normal data model. Among the squared data models, those with high within group variance had much higher failure rates.

In power simulations of single-locus model, all the three methods had very high power (>99%) to detect the disease SNPs. I also examined ten markers around the disease SNP. All methods showed significant power to detect the markers around the locus. This finding might be important because instead of locating a specific SNP, we could locate a region on chromosome, in which the disease SNP may occur.

In power simulations of the multi-locus model, the power to detect the disease SNPs was generally proportional to the MAF; that is, as the MAF increased, the power usually increased. However, if a SNP was in high linkage disequilibrium with many other SNPs, the power to detect this SNP would drop substantially. Both the BPP method and modal BPP method were significantly better than contingency table method with regard to power. The difference in power between BPP method and modal BPP method was not significant.

In this dissertation, I only examined the genetic factors. For future work, diverse factors, like environment factors or other non-genetic covariates, could be considered. Additionally, in my study, I set the trajectory group to be three when I run SAS PROC TRAJ. This was because generally, the three trajectory group model had the best BIC value. However, it could be problematic because in a few cases, a two trajectory group model or a four trajectory group model had better BIC value. For future work, one could consider the best model that BIC picked and examined how it would affect the empirical type I error rate and the power. Also in my

study, I dropped the replicates in which TRAJ model failed to converge. In future study, one could examine those replicates.

In my study, I examined 1498 SNPs on HC13. In future research, one could choose another chromosome and examine the SNPs on it. In the multi-locus model, I simulated the prevalence of disease to be around 2%. In future research, one could use more disease locus and set a different prevalence of disease.

References

1. Weinstein SL, Dolan LA, Cheng JC, Danielsson A, and Morcuende JA., 2008, Adolescent Idiopathic Scoliosis. *Lancet* 371(9623): 1527-37

2. Salehi LB, Mangino M, De Serio S, De Cicco D, Capon F, Semprini S, Pizzuti A, Novelli G, and Dallapiccola B., 2002, Assignment of a locus for autosomal dominant Idiopathic Scoliosis (AIS) to human chromosome 17p11. *Human Genetics* 111: 401-404

3. http://www.scoliosisassociates.com/subject.php?pn=idiopathic-scoliosis-009

4. Morrissy RT, Goldsmith GS, Hall EC, Kehl D, and Cowie GH., 1990, Measurement of the Cobb angle on radiographs of patients who have scoliosis. Evaluatoin of intrinsic error. *Journal of Bone and Joint Surgery*. 72: 320-327

5. Wise C, Barnes R, Gillum J, Herring J, Bowcock A, and Lovett M., 2000, Localization of susceptibility to familial idiopathic scoliosis. *Spine* 25-18: 2372-2380

6. Greiner KA, 2002, Adolescent Idiopathic Scoliosis: Radiologic Decision-making. *American Family Physician* 65(9): 1817-1823, http://www.aafp.org/afp/2002/0501/p1817.html

7. Qiu XS, Tang NL, Yeung HY, Lee KM, Hung VW, Ng BK, Ma SL, Kwok RH, Qin L, Qiu Y, and Cheng JC, 2007, Melatonin receptor 1B (MTNR1B) locus polymorphism is associated with the occurrence of adolescent idiopathic scoliosis. *Spine* 32-16: 1748-1753

8. Gao X, Gordon D, Zhang D, Browne R, Helms C, Gillum J, Weber S, Devroy S, Swaney S, Dobbs M, Morcuende J, Sheffield V, Lovett M, Bowcock A, Herring J, and Wise C, 2007, CHD7 locus polymorphisms are associated with susceptibility to idiopathic scoliosis. *American Journal of Human Genetics* 80(5): 957-965

9. Ward K, Ogilvie J, Argyle VA, Nelson L, Meade M, Braun J, and Chettier R, 2010, Polygenic Inheritance of Adolescent Idiopathic Scoliosis: A study of extended families in Utah. *American Journal of Medical Genetics*: 152A-5: 1178-1188

10. Justics CM, Miller NH, Marosy B, Zhang J, and Wilson AF, 2003, Familial idiopathic scoliosis: evidence of an X-linked susceptibility locus. *Spine* 28(6): 589-94

11. http://www.ornl.gov/sci/techresources/Human_Genome/medicine/pharma.shtml

12. http://en.wikipedia.org/wiki/Genome-wide_association_study

13. Hirschhorn JN, Daly MJ, 2005, Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* 6:95-108

14. Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples A and Myers RH, 2000, Evidence for a locus influencing blood pressure on chromosome 17. *Hypertension* 36: 477-483

15. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M, Sovio U, Ruokonen A, Laitinen J, Jakkula E, Coin L, Hoggart C, Collins A, Turunen H, Gabriel S, Elliot P, McCarthy MI, Daly MJ, Jarelin MR, Freimer NB and Peltonen L, 2009, Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics* 41(1): 35-46

16. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, and Froguel P, 2007, A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445: 881-885

17. Ma CX, Casella G, and Wu R, 2002, Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* 161: 1751-1762

18. Wu R, and Lin M, 2006, Functional mapping-how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics* 7:229-237

19. Wu R, Ma CX, Lin M, and Casella G, 2004, A general framework for analyzing the genetic architecture of developmental characteristics. *Genetics* 166: 1541-1551

20. Wang Z and Wu R, 2004, A statistical model for high-resolution mapping of quantitative trait loci determining HIV dynamics. *Statistics In Medicine*, 23, 3033-3051

21. Bauer DJ, and Curran PJ, 2003, Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods* 8(3): 338-363

22. Muthén, B., and Shedden, K., 1999, Finite mixture modeling with mixture outcomes using the EM algorithm, *Biometrics*, 55, 463-469.

23. Muthén, B., and Muthén, L., 2000, Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, 24, 882-891.

24. Muthén, B., Brown, C.H., Masyn, K., Jo, B., Khoo, S.T., Yang, C.C., Wang, C.P., Kellam, S., Carlin, J., and Liao, J., 2002, General growth mixture modeling for randomized preventive interventions, *Biostatistics*, 3, 459-475.

25. Nagin , D., 1999, Analyzing Developmental Trajectories: A Semi-parametric, Group-based Approach, *Psychological Methods*, 4, 139-177.

26. Nagin , D. and Tremblay, R E., 2001, Analyzing Developmental Trajectories of Distinct but Related Behaviors: A Group-Based Method, *Psychological Methods*, 6, 18-34.

27. Li F., Duncan, T. E., and Hops, H., 2001, Examining Developmental Trajectories in Adolescent Alcohol Use Using Piecewise Growth Mixture Modeling Analysis, *Journal of Studies on Alcohol and Drugs*, 62 (2), 2001.

28. Colder C. R., Mehta P., Balanda K., Campbell R. T., Mayhew K., Stanton W. R., Pentz M. A. and Flay B. R., 2001, Identifying Trajectories of Adolescent Smoking: An Application of Latent Growth Mixture Modeling, *Health Psychology*, 20 (2), 127–135.

29. Chang SW, Choi SH, Li K, Fleur RS, Huang C, Shen T, Ahn K, Gordon D, Kim W, Wu R, Mendell NR and Finch SJ, 2009, Growth mixture modeling as an exploratory analysis tool in longitudinal quantitative trait loci analysis, *BMC Proceedings 3* (Suppl 7): S112

30. Kerner B, and Muthen BO, 2009, Growth mixture modeling in families of the Framingham Heart Study, *BMC Proceedings 3* (Suppl 7): S114

31. Jones BL, Nagin DS, and Roeder K, 2001, A SAS procedure based on mixture models for estimating developmental trajectories, *Sociological Methods and Research* 29(3): 374-393

32. Jones BL, and Nagin DS, 2007, Advances in group-based trajectory modeling and an SAS procedure for estimating them, *Sociological Methods and Research* 35(4): 542-571

33. Nagin DS, Tremblay RE, 1999, Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency, *Child Development* 70(5): 1181-1196

34. http://www.framinghamheartstudy.org/

Appendix

I. IDs of 1599 unrelated participants

9	36	64	65	66	74	92	97	108	128	175	183
186	190	210	225	279	291	304	306	309	326	339	374
392	412	418	434	499	510	521	534	571	606	618	620
626	640	653	660	671	708	718	726	800	801	841	877
914	972	977	1001	1016	1054	1055	1114	1148	1156	1173	1195
1196	1239	1241	1248	1252	1278	1281	1285	1293	1295	1296	1302
1303	1305	1313	1317	1318	1322	1323	1355	1373	1375	1378	1437
1438	1441	1459	1486	1517	1528	1539	1542	1561	1588	1601	1620
1639	1655	1660	1673	1713	1738	1743	1754	1770	1780	1803	1804
1808	1849	1857	1859	1868	1879	1887	1920	1921	1925	1928	1938
1939	1998	2038	2040	2052	2060	2098	2116	2155	2160	2170	2207
2212	2232	2242	2253	2257	2277	2294	2301	2313	2334	2337	2340
2367	2395	2408	2426	2469	2470	2471	2490	2498	2499	2529	2540
2593	2637	2657	2658	2696	2700	2737	2776	2785	2805	2806	2816
2843	2862	2864	2905	2920	2945	2947	2970	2975	2981	3017	3058
3077	3119	3125	3162	3181	3205	3237	3238	3239	3253	3254	3274
3279	3303	3356	3368	3373	3376	3398	3399	3410	3413	3422	3486
3492	3516	3527	3550	3567	3610	3623	3634	3637	3655	3676	3715
3726	3739	3746	3751	3767	3779	3782	3794	3796	3798	3802	3822

3827	3852	3858	3871	3877	3880	3889	3900	3928	3946	3982	4027
4032	4050	4080	4089	4094	4097	4098	4105	4137	4146	4173	4174
4175	4183	4187	4217	4274	4276	4280	4284	4286	4321	4343	4351
4370	4379	4384	4393	4404	4416	4420	4423	4433	4440	4442	4464
4475	4517	4522	4526	4530	4541	4549	4551	4573	4581	4602	4606
4615	4635	4651	4652	4670	4706	4718	4756	4767	4775	4777	4788
4790	4810	4831	4837	4839	4841	4847	4854	4859	4875	4903	4926
4993	5006	5023	5077	5082	5122	5133	5153	5156	5190	5191	5201
5216	5220	5233	5253	5256	5285	5292	5297	5299	5302	5317	5331
5335	5342	5352	5355	5357	5397	5400	5403	5449	5493	5513	5533
5556	5558	5571	5596	5615	5628	5630	5635	5685	5694	5703	5733
5777	5782	5790	5869	5883	5904	5911	5923	5925	5955	5969	5978
5986	5990	6001	6002	6017	6019	6052	6056	6060	6102	6126	6127
6131	6143	6147	6191	6195	6204	6227	6234	6250	6263	6331	6332
6351	6370	6388	6406	6410	6416	6422	6458	6518	6535	6548	6560
6561	6566	6586	6616	6636	6654	6731	6742	6757	6766	6785	6788
6789	6795	6797	6829	6841	6844	6862	6871	6916	6938	6945	6953
6973	7009	7018	7047	7050	7056	7059	7076	7079	7114	7123	7175
7197	7229	7249	7270	7322	7326	7345	7360	7368	7379	7395	7402
7404	7422	7431	7433	7444	7468	7470	7483	7500	7502	7509	7520
7526	7531	7544	7545	7605	7622	7630	7644	7659	7671	7714	7746
7747	7775	7781	7802	7827	7853	7890	7913	7930	7939	7961	7969
8043	8055	8070	8073	8077	8089	8124	8140	8160	8161	8163	8165

8177 8182 8194 8211 8216 8221 8228 8230 8231 8235 8260 8283

8304 8319 8382 8383 8394 8413 8429 8455 8456 8511 8514 8588

8590 8610 8612 8629 8660 8668 8671 8674 8676 8699 8723 8731

8743 8760 8774 8796 8828 8870 8888 8903 8909 8912 8958 8963

8971 8984 9005 9025 9034 9041 9044 9055 9106 9131 9143 9166

9179 9199 9203 9261 9278 9297 9333 9389 9400 9423 9444 9462

9467 9506 9518 9524 9544 9547 9555 9558 9585 9589 9609 9620

9643 9644 9672 9736 9742 9746 9748 9762 9773 9782 9784 9790

9802 9805 9810 9850 9859 9893 9906 9913 9929 9933 9955 9960

9976 9980 9992 10010 10011 10014 10060 10066 10111 10137 10163 10167

10168 10181 10190 10198 10205 10227 10232 10255 10311 10334 10336 10340

10375 10376 10390 10401 10431 10442 10458 10466 10469 10478 10480 10510

10513 10537 10538 10543 10552 10557 10599 10613 10614 10617 10652 10655

10657 10680 10687 10703 10712 10719 10753 10771 10785 10800 10815 10835

10845 10852 10854 10865 10880 10890 10895 10950 10972 10978 10986 10995

11017 11040 11041 11078 11081 11100 11107 11119 11159 11211 11216 11243

11251 11280 11284 11297 11330 11331 11345 11352 11353 11359 11368 11369

11410 11445 11452 11459 11465 11486 11507 11543 11592 11593 11612 11642

11649 11663 11682 11693 11696 11735 11745 11761 11778 11786 11802 11806

11815 11827 11834 11847 11857 11886 11898 11905 11907 11914 11954 11957

11960 11993 12012 12049 12084 12098 12106 12182 12202 12241 12261 12265

12274 12303 12335 12347 12365 12383 12394 12416 12417 12428 12431 12439

12489 12500 12511 12552 12563 12568 12582 12585 12595 12597 12604 12626

12627 12642 12675 12722 12735 12773 12775 12776 12800 12802 12806 12819

12854 12859 12869 12939 12952 12968 12984 13020 13022 13039 13082 13104

13124 13128 13129 13132 13147 13161 13168 13169 13170 13178 13203 13217

13219 13303 13310 13315 13323 13324 13331 13336 13349 13354 13389 13394

13407 13415 13444 13445 13461 13471 13486 13494 13508 13511 13521 13578

13585 13593 13625 13638 13654 13658 13663 13678 13708 13709 13732 13824

13864 13890 13894 13920 13926 13959 13982 13983 14012 14014 14030 14037

14038 14070 14090 14102 14110 14147 14157 14161 14186 14193 14197 14215

14220 14221 14250 14260 14268 14286 14290 14305 14309 14320 14339 14353

14396 14402 14428 14462 14492 14506 14522 14564 14632 14663 14675 14702

14715 14716 14745 14761 14793 14794 14800 14824 14873 14877 14906 14921

14940 15001 15030 15033 15039 15078 15086 15091 15093 15099 15100 15129

15149 15154 15160 15179 15193 15229 15239 15241 15246 15275 15289 15304

15306 15351 15366 15368 15375 15379 15393 15406 15408 15433 15448 15463

15490 15503 15504 15558 15574 15585 15591 15600 15628 15634 15672 15687

15767 15781 15810 15811 15844 15851 15858 15862 15881 15886 15910 15931

15941 15955 15956 15974 16046 16054 16057 16099 16114 16122 16188 16251

16260 16274 16299 16327 16352 16367 16429 16430 16433 16450 16462 16473

16490 16514 16515 16518 16526 16529 16532 16575 16604 16619 16623 16630

16671 16725 16734 16758 16790 16805 16853 16902 16905 16923 16927 16939

16951 16994 16995 17012 17027 17030 17031 17043 17054 17057 17077 17079

17100 17113 17117 17142 17147 17182 17240 17251 17254 17259 17274 17275

17281 17311 17313 17320 17323 17327 17338 17373 17408 17412 17421 17431

17441 17448 17458 17474 17493 17508 17516 17529 17532 17542 17583 17600

17603 17609 17621 17641 17673 17718 17720 17731 17734 17767 17774 17784

17786 17791 17806 17810 17836 17885 17897 17921 17938 17940 17943 17949

17951 17960 17973 17986 17995 18005 18008 18038 18055 18080 18085 18107

18111 18259 18263 18265 18270 18285 18298 18310 18385 18388 18393 18408

18412 18436 18440 18471 18485 18511 18545 18563 18570 18626 18630 18679

18687 18701 18714 18719 18737 18751 18753 18771 18786 18825 18903 18906

18911 18921 18941 18947 18954 18961 18986 18987 19002 19063 19068 19076

19086 19142 19159 19179 19185 19188 19211 19221 19236 19252 19280 19360

19374 19378 19385 19387 19391 19393 19407 19443 19452 19455 19470 19477

19495 19530 19585 19592 19613 19618 19627 19640 19653 19659 19676 19677

19688 19703 19711 19751 19757 19758 19770 19778 19779 19797 19800 19815

19832 19840 19876 19902 19930 19946 19965 19974 20023 20063 20089 20095

20105 20135 20175 20179 20189 20204 20220 20236 20246 20276 20323 20347

20359 20384 20420 20423 20427 20432 20472 20475 20489 20499 20501 20530

20562 20581 20601 20602 20609 20617 20644 20678 20697 20746 20774 20780

20801 20816 20820 20845 20855 20856 20861 20873 20887 20908 20925 20930

20936 21002 21029 21064 21101 21107 21130 21141 21151 21202 21207 21209

21213 21216 21224 21237 21254 21291 21335 21340 21408 21438 21444 21478

21530 21557 21597 21617 21625 21639 21647 21651 21653 21696 21715 21766

21819 21839 21843 21864 21885 21891 21906 21911 21928 21931 21937 21957

21984 21994 21999 22005 22006 22012 22017 22041 22046 22068 22070 22116

22117 22126 22129 22144 22148 22151 22156 22173 22174 22177 22182 22185

22249 22257 22268 22295 22312 22319 22320 22349 22360 22415 22427 22455
22459 22478 22485 22486 22507 22517 22549 22554 22606 22608 22617 22639
22645 22662 22714 22717 22741 22769 22789 22805 22808 22813 22817 22828
22847 22896 22902 22915 22948 22953 22958 22970 22994 23007 23030 23058
23102 23112 23126 23141 23145 23185 23189 23192 23208 23220 23273 23305
23306 23323 23329 23368 23374 23382 23383 23398 23402 23439 23449 23451
23472 23474 23546 23549 23562 23576 23579 23583 23616 23640 23662 23669
23678 23688 23706 23721 23752 23765 23800 23808 23810 23841 23842 23864
23892 23906 23913 23920 23926 23956 23979 23983 24029 24033 24050 24069
24086 24096 24098 24125 24136 24148 24159 24163 24164 24204 24215 24230
24234 24235 24240 24272 24294 24300 24318 24325 24336 24387 24417 24425
24432 24469 24481 24503 24523 24529 24547 24553 24589 24607 24616 24624
24666 24730 24788 24790 24797 24811 24816 24845 24847 24857 24873 24890
24928 24933 24971 24999 25057 25059 25063 25068 25071 25079 25088 25095
25107 25112 25121 25122 25151 25194 25217 25238 25242 25247 25259 25267
25284 25290 25297 25335 25358 25391 25417 25419 25443 25480 25484 25491
25492 25532 25539 25546 25572 25582 25605 25609 25623 25625 25635 25642
25663 25671 25748 25754 25776 25782 25794 25808 25812 25823 25845 25856
25861 25881 25932 25952 25961 25980 26017 26021 26030 26047 26071 26110
26193 26224 26247 26251 26271 26320 26337 26340 26344 26347 26377 26387
26443 26446 26471 26485 26487 26488 26489 26491 26492 26505 26515 26527
26533 26580 26582 26657 26675 26684 26693 26705 26708 26725 26766 26782
26789 26792 26799

II. PLINK

1. Basics

PLINK is a command line program written in C/C++. All commands involve typing "plink" at the command prompt, followed by a number of options (all starting with "--option") to specify the data files/methods to be used. A complete list of all options and output file types is given in this link:

http://pngu.mgh.harvard.edu/~purcell/plink/reference.shtml

To run PLINK, one should start from typing "plink --file mydata" (there is a space before the dashes). The data is in two files: in this case, mydata.ped and mydata.map. If the PED and MAP files have different names, they can be specified separately, with the command: "plink --ped mydata.ped --map autosomal.map".

The PED file contains the demographic and phenotypic information about the subjects. It is a white-space (that is, space or tab) delimited file with the columns:

"Family ID, Individual ID, Paternal ID, Maternal ID, Sex(1=male; 2=female; other=unknown), Phenotype". The PED file can have one and only one phenotype, which is given in the sixth column. The phenotype can be either a quantitative trait or an indicator (0 or 1) variable.

If the PED file has some missing fields, one can use a command to indicate which columns, if any, are missing. For instance, "--no-fid" indicates there is no Family ID column (the first column); "--no-parents" indicates there are no paternal and maternal ID columns (third and fourth columns); "--no-sex"indicates there is no sex field (fifth column) and all individuals set to have a missing sex code; "--no-pheno" indicates there is no phenotype field (sixth column).

The MAP file contains the genotype location information. By default, each line of the MAP file describes a single marker and must contain exactly 4 columns:

"chromosome # (1-22, X, Y or 0 if unplaced)", rs # or SNP identifier, Genetic distance (morgans), Base-pair position (bp units)"

If "Genetic distance" is missing in MAP file, one can add a flag: "--map3", that is: "plink --file mydata --map3" In this case, the three columns in MAP are expected to be "chromosome, rs# and Base-Pair".

2. Summary statistics calculated in PLINK:

(1) Hardy-Weinberg Equilibrium:

To test HWE for each SNP, use the option: "plink --file data --hardy". PLINK then creates the file: plink.hwe, which has the following format:

"SNP: SNP identifier; TEST: code indicating sample; A1: minor allele code; A2: major allele code; GENO: genotype counts:A1A1/A1A2/A2A2; O(HET): observed heterozygosity; E(HET): expected heterozygosity; P: HW p-value."

Thus, if HW p-value is significant, then we'll conclude that this SNP is not in HWE.

(2) Minor Allele Frequency (MAF):

To generate a list of MAF for each SNP, one can use the command: "plink --file data --freq", which will create a file: plink.frq with five columns: "CHR: chromosome; SNP: SNP identifier; A1: allele 1 code (minor allele); A2: allele 2 code (major allele); MAF: minor allele frequency; NCHROBS: non-missing allele count".

3. Association analysis

(1) Basic case/control association test:

To perform a standard case/control association test, one can use the option: "plink --file mydata --assoc". PLINK then will generate a file "plink.assoc", which contains the fields: "CHR: chromosome; SNP:SNP ID; BP: base-pair; A1: minor allele name; F_A: frequency of this allele in cases; F_U: frequency of this allele in controls; A2: major allele name; CHISQ: basic allelic test chi-square (1df); P: asymptotic p-value for this test; OR: estimated odds ratio."

If the p-value is significant, we conclude that this SNP is associated with the disease. In addition, if the option "--ci 0.95" is included, then "L95: lower bound of 95% CI for odds ratio" and "U95: upper bound of 95% CI for odds ratio" will be appended to the output.

In my study, I used the Modal BPP (1 as in the fast trajectory group and 0 otherwise) as the phenotype, and test the association with the SNPs on chromosome 13.

(2) Quantitative trait association:

If the phenotype (column 6 of the PED file) is quantitative, the PLINK will automatically treat the analysis as a quantitative trait analysis. One can use the same command as for case/control association: "plink --file mydata --assoc", which will generate the file "plink.assoc". The file has the following fields:

"CHR: chromosome; SNP: SNP ID; BP: base-pair; NMISS: # of non-missing genotypes; BETA: regression coefficient; SE: standard error; R2: regression r-squared; T: Wald test t-statistic; P: Wald test asymptotic p-value." If the p-value is significant, then we conclude that the SNP is highly associated with the disease.

In my study, I used BPP as the quantitative trait, and then tested the association with the SNPs on chromosome 13.