Stony Brook University

Growth Mixture Modeling as an Exploratory Analysis Tool in a Longitudinal Quantitative Trait Locus Analysis

A Dissertation Presented

by

Su-Wei Chang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

May 2009

Stony Brook University

The Graduate School

Su-Wei Chang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Stephen J. Finch
Professor
Department of Applied Mathematics and Statistics

Nancy R. Mendell
Professor
Department of Applied Mathematics and Statistics

Wei Zhu
Professor
Department of Applied Mathematics and Statistics

Derek Gordon
Associate Professor
Department of Genetics, Rutgers University

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Growth Mixture Modeling as an Exploratory Analysis Tool in a Longitudinal Quantitative Trait Locus Analysis

by

Su-Wei Chang

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2009

I examined the properties of growth mixture modeling (GMM) in finding longitudinal quantitative trait loci. Two software packages are commonly used in GMM analyses: Mplus and the SAS TRAJ procedure. I analyzed the 200 replicates of the simulated data from the Genetic Analysis Workshop 16 with these programs using three tests: the likelihood ratio test statistic (LRTS), a direct test of genetic model coefficients, and the chi-square test classifying subjects based on the trajectory model's posterior Bayesian probability. The Mplus program was not effective in this application due to its computational demands. The distributions of these tests applied to genes not related to the trait were sensitive to departures from Hardy-Weinberg equilibrium (HWE). Genotyping error might be partially responsible for this departure. It may not be valid to apply GMM procedures to single-nucleotide polymorphisms (SNPs) that are apparently not in HWE. The LRTS was not usable in this application as its distribution was far from the expected asymptotic distributions when applied to markers with no genetic relation to the quantitative trait. The other two tests were satisfactory. Power was still substantial when

markers near the gene rather than the gene itself were used. That is, GMM may be useful in genome wide association studies. The direct test of the coefficients and the posterior Bayesian probability chi-squared test had essentially the same power when analyzing genes in the disease mechanisms. When analyzing data from markers near the true gene, there was somewhat greater power for the direct test of the coefficients and less power for the posterior Bayesian probability chi-squared test.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my greatest gratitude to my advisor Dr. Stephen J. Finch for his constant guidance and support throughout the last four and half years. His advice and encouragement aided my move into genetic statistics and made this dissertation possible.

Thanks go to my committee members Dr. Nancy R. Mendell, Dr. Wei Zhu, and Dr. Derek Gordon for their constructive suggestions and inspiring feedback.

Thanks to my fellow students in the statistical genetics research group. The interactive and brainstorming discussions among group members helped fulfill this study.

I am thankful to my friends and family in Taiwan for their unconditional love and care. With their support, I have stayed on the right track and concentrated on what I wanted to do without worry and fear.

A special thank goes to my beloved YW. Although distant from me over years, he has been my greatest cheer leader and comforter. With his love and encouragement, I have had strength to keep pushing myself and came a long way to achieve my goal eventually.

Chapter 1 Introduction

1.1 Background

Growth mixture modeling (GMM) is an important tool for analyzing longitudinal data (Muthén & Shedden, 1999; Li, Duncan, and Hops, 2001; Colder et al., 2001). GMM is a combination of the conventional growth curve model and finite mixture modeling (Muthén, 2004). Use of GMM not only provides researchers the opportunity to study growth curves of a single or multiple measurable characteristics, such as phenotypes or traits, but also provides the chance to improve the accuracy for detection of genetic or environmental factors that influence growth change. The conventional growth curve model treats the data as inferred from a homogeneous population where population members follow a common developmental process of growth or decline. In contrast, GMM hypothesizes that there is a fixed but unknown number of components of distinctive trajectory patterns observed within the population. GMM applies mixture analysis methods to estimate the number of trajectory components and the probability that a trait variable (such as a genotype) affects the trajectory component membership. The modeling technique generalizes multilevel random effects growth modeling to model a combination of continuous and categorical latent variables. The continuous latent variables denote the growth parameters, such as intercept or slope, and determine the trajectory shapes, while the latent categorical variables represent the latent trajectory components underlying the

latent growth variables. Under the assumption of multinormally distributed random effects, GMM allows researchers to test for the departure of an individual's latent growth parameters from the population mean growth parameters, which can be modeled as functions of risk factors (time-invariant covariates) or time-varying covariates (TVCs). Further, the model has estimates of the probability that the risk factors affect the trajectory component membership. The posterior probability of membership of an individual in each latent component is used to assign latent class membership. Such latent trajectory class membership can further be used as a covariate in a post-hoc cluster analysis.

There are two software packages for GMM. One is the SAS TRAJ procedure developed by Nagin and colleagues (Nagin & Land, 1993; Nagin, 1999; Nagin & Tremblay, 2001; Jones, Nagin, and Roeder, 2001). The other is Mplus, a widely used structural equation modeling software package created by Bengt Muthén, Linda Muthén, and colleagues (Muthén & Muthén, 2000; Muthén et al., 2002; Kreuter & Muthén, 2008). The main difference between the two GMM analytic tools is that the variance and covariance matrix of growth parameters are held to be zero in the SAS TRAJ procedure, while the Mplus GMM program allows for the variation of these parameters. That is, all individuals are assumed to behave identically within a trajectory component using the SAS TRAJ procedure. The term "growth mixture modeling" originally was used by Muthén and his colleagues. They regarded the approach used by SAS PROC TRAJ as a simplified version of GMM and called it "latent class growth analysis" (LCGA) in Mplus to signify the difference. The principle advantage of Mplus GMM compared with the SAS TRAJ procedure is that fewer number of trajectory components may be required to

identify a satisfactory model by allowing variations about the group mean (Muthén, 2004; Nagin & Tremblay, 2005a). To accommodate such variation, Mplus uses a far more complex set of parameters to model trajectory components. This increases the computational complexity and instability of the analysis. Muthén and Muthén (1998-2007) suggested that, before conducting GMM, one should use Mplus LCGA as a preliminary analytic tool, since model convergence is generally easier and faster to achieve with that subroutine.

1.2 Literature Review

1.2.1 Application of Growth Mixture Modeling

In recent years, the prevalence of GMM modeling has been increasing in social and psychological studies as well as other scientific disciplines. The application of GMM can be traced back to 1990s. It was first used to study criminal behaviors longitudinally. Nagin and Land (1993) introduced the fundamental framework of the trajectory-based approach and used it to study the life course of individual offending patterns. The popularity of this approach among criminologists and sociologists and its advantages in the study of the outcome of change over time or at different ages drew attention from many researchers.

A great number of papers applying GMM have been focused on the relations between behavioral problems of children, such as antisocial acts, physical aggression,

opposition, and physical violence, and adolescent delinquency (e.g. Nagin and Tremblay, 1999; Broidy et al., 2003; Schaeffer et al., 2003; Wiesner & Capaldi, 2003, etc.). Different developmental trajectories of problem behavior in childhood may lead to different types of juvenile delinquency. For example, trajectory component members with chronic oppositional tendency and with constant low-level physical aggression and hyperactivity at age 6 through 15 were more likely to commit a covert crime such as theft, while trajectory component members with physical aggression behaviors and with minor opposition and hyperactivity were more likely to commit an overt crime and serious delinquent acts (Nagin and Tremblay, 1999). Using data from multiple sites in three countries, Broidy et al. (2003) found that for males, constant physical aggression during the elementary school years was associated with increased risk of continued physical violence as well as other nonviolent forms of delinquency during adolescence. Schaeffer et al. (2003) reported that boys with trajectories defined by chronically high and increasing ratings of aggression, evaluated longitudinally from the 1st to the 7th grade by school teachers, appeared to be at increased risk for antisocial personality disorder, conduct disorder, and juvenile and adult arrest

There are a considerable number of studies applying GMM in cigarette smoking (Colder et al., 2000; White, Pandina, and Chen, 2002), alcohol drinking (Li, Duncan, and Hops, 2000; Chassin, Pitts, and Prost, 2002), drug or substance use (Ellickson, Martino, and Collin, 2004; Hix-Small et al., 2004; Tucker et al., 2005). GMM was used to identify developmental trajectory components and potential predictors or risk factors underlying them. For instance, Chassin et al. (2002) showed that among the three drinking trajectory

components identified, the trajectory component members who started drinking early (at about age 13) and were heavy drinkers were characterized by parental alcoholism and antisocial tendency, peer drinking, drug use, and (for boys) high levels of externalizing behavior, but low depression. The infrequent drinking trajectory component members distinguished themselves by having parental alcoholism and (for girls) adolescent depression, while the trajectory component members who started drinking late (at about age 16) and were modest drinkers showed the most favorable adolescent psychosocial status.

Researchers also used GMM to study the life course of mental illness (Tremblay et al, 2004; Aneshensel et al, 2004; Romano et al, 2006; Xie, Drake, and McHugo, 2006; Odgers et al., 2007), and patterns of medication or therapy visits (Mojtabai et al., 2009). Aneshensel et al. (2004) identified four trajectories of depressive symptoms over time among caregivers following bereavement. They reported that caregivers were not identical in their emotional responses to bereavement. The caregivers followed distinct trajectory patterns connected with their previous experiences as care-givers, in particular exposure to stress and access to resources. Mojtabai et al. (2009) employed GMM to identify four trajectory patterns of mental health service use for a community sample of schizophrenia patients during the four year period after their first admission.

In genetic studies, finite mixture modeling approaches have been applied to microarray gene expression data to cluster genes with distinctive gene-expression levels in organisms (Yeung et al., 2001; Pan, Lin, and Le, 2002; McLachlan, Bean, and Peel,

2002; Allison et al., 2002; McLachlan, Do, and Ambroise 2004). Rodriguez-Zas et al. (2006) used GMM methods to characterize gene expression trajectories across time. To date, I have not found any research articles applying GMM for the identification of a longitudinal quantitative trait locus.

1.2.2 Current Studies about Identification of Quantitative Trait Loci

For genetic studies, there is currently considerable interest in quantitative traits such as blood pressure, body mass index, and cholesterol levels. A quantitative trait locus (QTL) is a region of a chromosome that has been shown through genetic mapping to contain one or more of the genes that contribute to quantitative phenotypic differences.

A wide variety of QTL mapping techniques have been developed to allow the dissection of quantitative traits in a certain populations (Haseman & Elston, 1972; Goldgar 1990; Zeng, 1993 & 1994; Lynch & Walsh, 1998; George et al., 2000). Most of these studies have focused on traits measured at a single time point. The genetic mechanism of some traits may be better understood by collecting and analyzing them longitudinally. Macgregor et al. (2005) proposed a flexible random regression model to analyze longitudinal QTL data based on the covariance function (CF) structure. They showed that the change in the genetic effects over time can be well characterized by this approach and that including parameters to model the change in effect with age can result in a substantial increase in power to detect QTL compared with repeated measure or univariate techniques.

A third technique to identify longitudinal QTL is the functional mapping approach developed by Rongling Wu and his colleagues (Ma, Casella, and Wu, 2002; Wu et al., 2004). They claimed that this mapping technique can characterize the QTLs and nucleotides (QTNs) that underlie a complex dynamic trait in a single analysis, showing a substantial improvement on the method proposed by Weiren Wu and his group (Wu et al., 2002). Functional mapping estimates parameters that describe the developmental mechanisms of traits and expression for each QTL or QTN. The modeling approach also allows for assessing the interplay between gene actions or interactions between developmental changes.

The value of functional mapping has been affirmed in mapping longitudinal QTL (Zhao et al., 2004a, 2004b; Wu & Lin, 2006). However, the construction of functional mapping within the context of simple interval mapping makes it unsuitable for analyzing multiple linked QTLs that jointly affect developmental patterns. Zeng (1993, 1994) and Jansen and Stam (1994) proposed composite interval mapping to simultaneously model two flanking markers and to test for the existence of a QTL by interval mapping and the markers outside the interval by a partial regression analysis. Incorporating the strengths of functional mapping and composite interval mapping, Yang et al. (2006) presented a so-called "composite functional mapping" framework, which allowed for modeling the time-varying genetic effects of a QTL tested within a marker interval, and aimed at increasing the resolution of multiple QTL on the same region of a chromosome.

1.3 Research Problems and Specific Aims

There are no precedents in which GMM has been used in the discovery of longitudinal QTL in genome-wide association searches. There has not been previous work evaluating the statistical properties of GMM applied to longitudinal quantitative genetic traits when the underlying mechanism of the data is known a priori. My goal is to evaluate the strength and limitations of methods using GMM through a simulation study. I will analyze the 200 replicates of the Genetic Analysis Workshop (GAW) 16 simulated datasets with the SAS TRAJ procedure and the Mplus GMM programs using three tests: the likelihood ratio test statistic (LRTS), a direct test of genetic model coefficients, and the chi-square test classifying subjects based on the trajectory model's posterior Bayesian probability.

There are several research questions that I would like to answer in this study. First, using 200 replicates of the GAW 16 simulated data on the coronary artery calcification (CAC) measurements taken at the three visits, I would like to assess whether genotypes appear to be associated with trajectory component membership and hence identify longitudinal quantitative trait loci (QTL) employing GMM techniques. I will also evaluate the applicability of the two GMM software packages to this kind of study.

Second, to estimate the empirical power for each test, it is necessary to estimate its empirical null distribution. I would like to explore the properties of the empirical null

distributions of three proposed measures of association for genes not in the genetic mechanism for CAC and compare them with the conjectured null distributions.

Third, using posterior probability for the assignment of trajectory component membership and using such latent component membership as a predictor of other outcomes of interest has been commonly used in a variety of research articles (Nagin, 1999; White, Bates, and Buyske, 2001; Tremblay et al., 2004; Nagin & Tremblay, 2005b). Since the statistical properties of such an analytic approach have never been discussed or studied, I would like to compare the power of GMM analyses that explicitly incorporate genotype measurements of the genes in the genetic model for CAC into the mixture modeling to GMM analyses that assess genetic association with post hoc tests.

Fourth, I would like to investigate the change in power using markers close to the true gene rather than the gene itself and assess whether GMM might be useful in genome wide association studies.

Fifth and finally, the evaluation of the effects of genotyping errors is crucial, since their consequences might be devastating. Existence of genotyping errors may influence the empirical null distributions, increase the critical value, and thus reduce the power of the study. Therefore, I will evaluate the effect that genotyping errors have on the three proposed procedures.

Chapter 2 Growth Mixture Modeling

Growth Mixture Modeling (GMM) extends the conventional mixed effects model and finite mixture analysis and models a mixture of continuous and categorical latent variables. The continuous latent variables define growth within classes with factors for baseline level and trend, and the latent categorical variable defines the unobserved developmental trajectory components.

GMM permits estimation of trajectory shapes (eg, linear, quadratic, cubic), trajectory classification probabilities for each participant (posterior probabilities), class-specific growth parameter variance, and regression of the latent trajectory class variable on covariates for trajectory characterization. With multinomial logistic regression methods, the characterization allows for identification of the most likely members of a given trajectory in relation to a comparison trajectory, which is generally the most common trajectory component or the trajectory with mean values closest to zero. Adding a binary variable (a distal outcome) or another growth process in the Mplus GMM model will make it a generalized growth mixture modeling (GGMM), which is a special case of GMM where the distal outcome is regressed on the latent trajectory variable and covariates can be added to improve model specification.

2.1 The Growth Mixture Model Structure

The GMM model I use throughout the study is based on the group-based trajectory model proposed by Nagin (2005). It has been seen as a special case of GMM since the variance and covariance of growth parameters are held to be zero, and the model assumes that there is no variation among individuals within the same trajectory component. Let $Y_i = \{y_{i_1}, y_{i_2}, ..., y_{i_t}\}$ denote the longitudinal sequence of independent observations for individual *i* over *t* time periods. The simple heterogeneity model assumes that the population sampled is heterogeneous and consists of a mixture of *K* underlying sub-populations. The probability density function for the data *Y* is given by

$$f(y_i) = \sum_{k=1}^{K} P(C_i = k) P(Y_i = y \mid C_i = k) = \sum_{k=1}^{K} p_k(\lambda_k) f(y_i, \mu_k), \tag{2.1}$$

where $p_k(\lambda_k)$ represents the probability of membership *C* in component *k* given λ_k. The corresponding parameters λ_k are time-invariant covariates (time-stable covariates or risk factors), and μ_k's are time-varying covariates (TVCs) that do not affect the probability of individual *i* belonging to a component *k*.

Since risk factors influence only the probability of belonging to a trajectory component, it is assumed that no more information can be acquired from the data *Y* through the risk factor *Z* given component membership *C*. Therefore, suppose for individual *i*, there are *R* risk factors $Z_i = \{Z_{i1}, Z_{i2}, ..., Z_{iR}\}$ and a sequence of time-varying covariates $W_i = \{W_{i1}, W_{i2}, ..., W_{it}\}$ over *t* time periods. Given that there are *K* trajectory

components, the conditional distribution of the observed data Y_i in (2.1) can be rewritten as

$$f(y_i \mid z_i, w_i) = \sum_{k=1}^{K} P(C_i = k \mid Z_i = z_i) P(Y_i = y_i \mid C_i = k, W_i = w_i), \tag{2.2}$$

The effect of the risk factor Z on component membership C is modeled with a multinomial logistic regression function as follows:

$$P(C_i = k \mid Z_i = z_i) = \frac{\exp(\theta_k + \lambda_k' z_i)}{\sum_{k=1}^{K} \exp(\theta_k + \lambda_k' z_i)}. \tag{2.3}$$

where $\theta_k = (\theta_1, ..., \theta_K)$ is a vector of K scalar, and $\lambda_k' = (\lambda_{1k}, ..., \lambda_{Rk})$ is a vector of length R , with θ_1 and λ_1 set to be zero.

2.2 Modeling for Trajectories

There are three options for the conditional distributions of observed data in the SAS PROC TRAJ program. The censored normal model is useful for modeling continuous outcome or interval scale data. The zero-inflated Poisson model is used to analyze count data when there are more zeros than would be expected under the Poisson assumption (Lambert 1992, Jones et al., 2001). The binary logit model is suitable for the analysis when the outcome at each measurement point is binary. In Mplus, for mixture modeling with longitudinal data, observed outcome variables can be continuous, censored,

binary, ordered categorical (ordinal), counts, or combinations of these variable types (Muthén & Muthén, 1998-2007).

2.2.1 Model Specification for the Censored Normal Distribution in the SAS TRAJ Procedure

Since the outcome of interest CAC is the longitudinal quantity in my research and is continuous, I will apply GMM using the censored normal distribution. The censored normal model is applicable to estimate trajectory models when the observed outcome, such as a psychometric scale, tends to cluster at the scale maximum or minimum. For the censored normal model, the linkage between observed outcome and age (or time) when the outcome is measured is established via a variable y_{it}^{*}. Up to a fifth-order polynomial relationship is assumed between y_{it}^{*} and age (or time) such that

$$y_{it}^{*} = \beta_{0k} + \beta_{1k} Age_{it} + \beta_{2k} Age_{ti}^{2} + \beta_{3k} Age_{it}^{3} + \beta_{4k} Age_{it}^{4} + \beta_{5k} Age_{it}^{5} + \varepsilon_{it}. \qquad (2.4)$$

where Age_{it}, Age_{it}^{2},..., Age_{it}^{5} are the age, age squared,..., and age to the fifth power for each individual i in trajectory component k, and ε_{it} is a disturbance assumed to be normally distributed with a zero mean and a standard deviation σ. The parameters $\beta_{ok}, \beta_{1k}, ..., \beta_{5k}$ determine the shape of the trajectory, which is allowed to vary freely across different trajectory components.

Let S_{min} and S_{max} denote the minimum and maximum possible score of the measured outcome, respectively. If the variable y_{it}^{*} is less than S_{min}, then the measured

outcome Y_i is set to be equal to S_{min}. If the variable y_{it}^* is greater than S_{max}, then the measured outcome Y_i is set to be equal to S_{max}. Only if y_{it}^* ranges between S_{min} and S_{max} does Y_i equal to y_{it}^*. The censored normal model can also be used for uncensored data by setting the scale minimum S_{min} less than all data values and setting the scale maximum S_{max} greater than all data values.

Let $\beta_k X_{it}$ denote $\beta_{0k} + \beta_{1k} Age_{it} + \beta_{2k} Age_{it}^2 + \beta_{3k} Age_{it}^3 + \beta_{4k} Age_{it}^4 + \beta_{5k} Age_{it}^5$ for notational convenience. Then Equation (2.4) can be written as $y_{it}^* = \beta_k X_{it} + \varepsilon_i$, where y_{it}^* is normally distributed with mean $\beta_k X_{it}$ and standard deviation σ. Hence, the probability of observing the trajectory for individual *i*, given membership in component *k*, is

$$P(Y_i = y_i \mid C_i = k) =$$

$$\prod_{y_i = S_{\min}} \Phi\left(\frac{S_{\min} - \beta_k X_{it}}{\sigma}\right) \prod_{S_{\min} < y_i < S_{\max}} \frac{1}{\sigma}\phi\left(\frac{y_i - \beta_k X_{it}}{\sigma}\right) \prod_{y_i = S_{\max}} \left(1 - \Phi\left(\frac{S_{\max} - \beta_k X_{it}}{\sigma}\right)\right), \qquad (2.5)$$

where $y_{it}^* = \beta_{0k} + \beta_{1k} Age_{it} + \beta_{2k} Age_{ti}^2 + \beta_{3k} Age_{it}^3 + \beta_{4k} Age_{it}^4 + \beta_{5k} Age_{it}^5 + \varepsilon_{it}$. Note that $Y_i = S_{min}$ if $y_{it}^* \leq S_{min}$, $Y_i = y_{it}^*$ if $S_{min} < y_{it}^* < S_{max}$, and $Y_i = S_{max}$ if $y_{it}^* \geq S_{max}$.

When adding L time-varying covariates $W_{it} = \{W_{i1t}, W_{i2t}, ..., W_{iLt}\}$ into the model, the specification of y_{it}^* for individual *i* at time *t* is restated by including them in Equation (2.4). Hence, the likelihood of observing the data trajectory for individual *i* at time t, given component membership *k* is

$$P(Y_i = y_{it} \mid C_i = k, W_i = w_{it}) =$$

$$\prod_{y_{it}=S_{\min}} \Phi\left(\frac{S_{\min} - \beta_k X_{it}}{\sigma}\right) \prod_{S_{\min}<y_{it}<S_{\max}} \frac{1}{\sigma}\phi\left(\frac{y_i - \beta_k X_{it}}{\sigma}\right) \prod_{y_{it}=S_{\max}} \left(1 - \Phi\left(\frac{S_{\max} - \beta_k X_{it}}{\sigma}\right)\right), \qquad (2.6)$$

where $y_{it}^{*} = \beta_{0k} + \beta_{1k} Age_{it} + ... + \beta_{5k} Age_{it}^{5} + \alpha_{1k} w_{i1t} + \alpha_{2k} w_{i2t} + ... + \alpha_{Lk} w_{iLt} + \varepsilon_{it}$.

2.2.2 Model Specification for the Uncensored Normal Distribution in the Mplus GMM

The Mplus GMM program allows the continuous outcome variable to be censored or uncensored. As noted in Chapter 1, the Mplus GMM model adds random effects to the growth parameter $\beta_{ok}, \beta_{1k}, ..., \beta_{5k}$, which define a component's mean trajectory such that

$$\beta_{mk} = \delta_{m0} + \delta_{m1} z_i + \gamma_{mi}, \qquad (2.7)$$

where m = 0, 1, ..., 5 denoting the polynomial order; β_{mk} are random growth parameters varying across individuals i = 1,..., n in a trajectory component. The residuals γ_{mi} are assumed to be normally distributed with zero means and uncorrelated with age or time, ε_{it} and other covariates.

2.3 Model Estimation

All analyses to be discussed can be carried out using maximum-likelihood estimation in GMM programs (Jones et al., 2001; Muthen & Muthen, 1998-2007). As of May 2009, the default program in Mplus 5.2 first generates 10 sets of random starting

values, runs through 10 iterations with each set, and then takes the set with the highest log-likelihood value and continues to iterate with that specific set until convergence criteria are satisfied. In the initial iterations, the Mplus program uses an expectation-maximization (EM) algorithm to improve the stability of estimation. It then switches to a Newton-Raphson, quasi-Newton, or Fisher scoring algorithm to increase the speed of convergence. The SAS PROC TRAJ macro uses a quasi-Newton algorithm and currently has no provision for automatically varying starting values, though one can manually input sets of starting values. The variance –covariance matrix for the parameter estimates is obtained from the inverse observed information matrix with the likelihood of the parameter estimates maximized (Nagin, 2005). With regard to missing data, Mplus GMM handles missing data using the "missing at random" (MAR) approach, while the SAS TRAJ procedure applies "missing completely at random" (MCAR) method (Rubin, 1976).

2.4 Testing for the Number of Trajectory Components and Selection Criteria

There is continuing debate about which criterion is best to decide on the optimal number of trajectory components in a growth mixture model is a complicated issue that is as unsettled. In general, researcher use a combination of different criterion, including Akaike Information Criterion (AIC; Akaike, 1987), Bayesian Information Criterion (BIC; Schwartz, 1978), Adjusted BIC (Sclove, 1987) and entropy. Additionally, to determine the number of trajectory components, the meaningfulness for conceptual interpretation of

the trajectories is also considered. There are a number of studies that show that the AIC overestimates the correct number of components in finite mixture models (Soromenho, 1993; Celeux & Soromenho, 1996), while the BIC has been reported to performed well (Roeder & Wasserman, 1997; Magidson & Vermunt, 2004). In this study, I follow the recommendation of D'Unger et al. (1998) and Nagin (1999) and use the BIC as the primary basis for the selection of the optimal model. For a given model in the SAS TRAJ procedure, BIC is defined as

$$BIC = \log(L) - 0.5 \cdot \log(n) \cdot (k), \tag{2.8}$$

where L represents the value of model's maximized likelihood, r is the number of parameters in the model, and n denotes sample size. Note that the value of BIC calculated in the SAS TRAJ procedure multiplied by -2 is the value of BIC calculated in Mplus GMM models.

A widely accepted rule to decide on the number of components is to model with increasing number of trajectories as long as the BIC continues to increase, with the restrictions that each trajectory component has at least ten subjects and each trajectory is interpretable and substantively meaningful.

2.5 Limitations and Important Issues

One limitation of GMM is that there is no guarantee of model convergence or existence of an optimal solution. Model failure often occurs due to excessive number of

model parameters or over-extraction of the trajectory components. Even when convergence is achieved, different sets of starting values may results in multiple solutions of the likelihood function. That is, the model may converge at a local rather than a global maximum. The failure to identify the global maximum of the likelihood function may result in serious consequences. Specifically, one may select an incorrect number of trajectory components. To find the global maximum of log-likelihood, Dolan, Jansen & van der Maas (2004) reported the use of as many as 5,000 randomized sets of starting values. However, the recommendation to vary the number of starting values provided by Mplus User's Guide or SAS TRAJ procedure is vague, and it is still unclear how many random starting values are necessary to get the optimal log-likelihood solution of a GMM model.

Bauer and Curran (2003, 2004) brought up several important issues about the implementation of GMM. First, when the data are drawn from non-normal distributions, incorrect estimation for the number of latent trajectory components may be likely, with fit indices, such as the AIC and the BIC, selecting a higher number of components than are present. Secondly, in GMM, the incorporation of covariates is used to assess their effects on the probability of belonging to certain trajectory components. If a covariate has differential effects on the growth parameters, that is, has positive effects on the intercept and negative effects on the slope, the capacity to detect the effects of covariates may be reduced and thus may lead to spurious estimation of the number of trajectory components. In addition, the results from the GMM analysis may not reflect actual population

heterogeneity by correctly estimating the number of trajectory components. Rather, it may over-simplify the description of a complex population distribution.

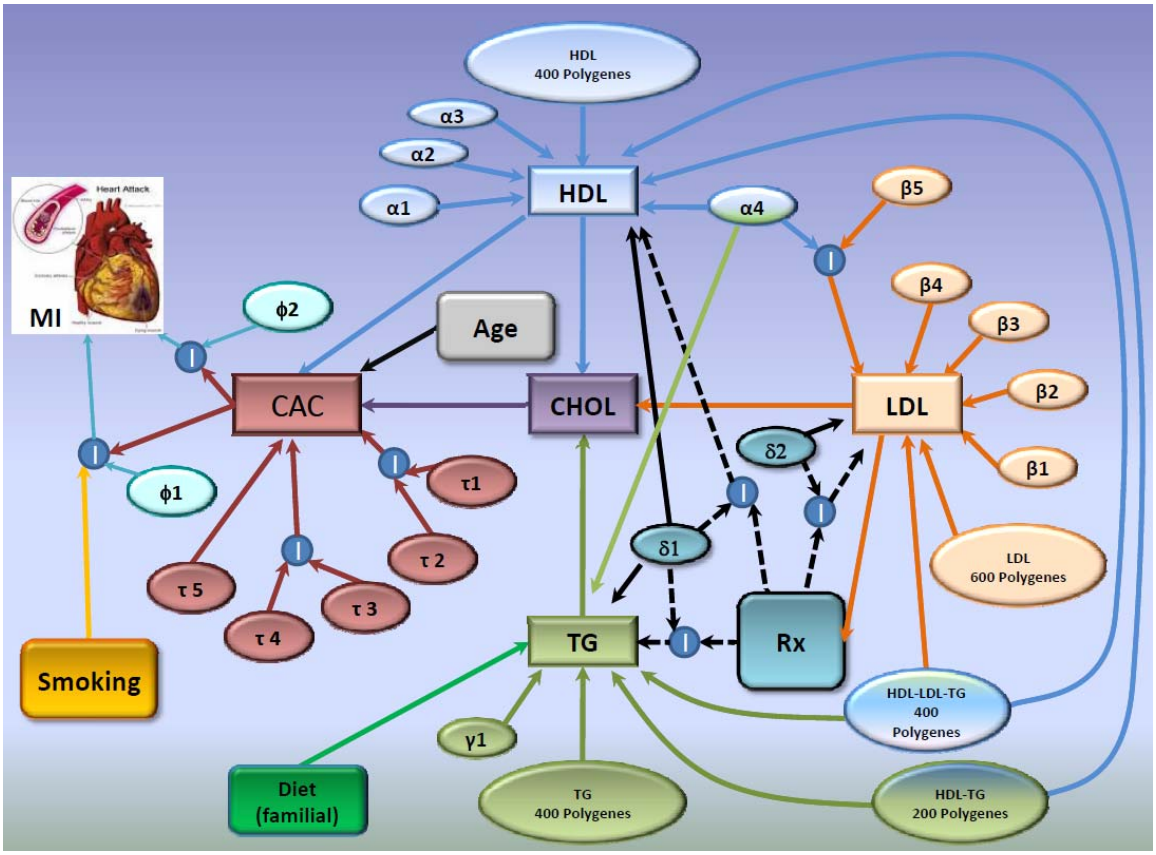
Although GMM is a common and important tool to evaluate population heterogeneity and to study the pattern and determinants of such heterogeneity in an outcome of interest over time, the interpretation of the modeling results may be difficult. Researchers should be aware of this complexity and apply GMM with cautions.

Chapter 3 Method

3.1 Genetic Models Known

The 200 replicates of data used in this research were generated from the Framingham Heart Study (FHS) using simulation with known genetic mechanisms and were given as GAW 16 Problem 3 (Kraja et al., 2008). Each replicate of the data includes a total of 6,476 participants with simulated phenotype and true genotype information. Specifically, each replicate contains 188 singletons (participants with no other relatives) and 942 pedigrees ranging across 3 generations. The measured genotypes include a total of approximately 550,000 SNPs (GeneChip® Human Mapping 500K Array Set and the 50K Human Gene Focused Panel). These are the actual genotypes from the FHS for both the genome-wide scan and additional candidate gene SNPs. Because the three generations of the family members in the FHS attended various examinations and were observed at different time points, Kraja et al. simulated the FHS pedigrees, calculated the family member's ages at a selected exam, and then assigned a simulated age at two subsequent time points, 10 and 20 years later. The details of the simulations for each phenotype generated can be found in Kraja et al. (2008). The simulated etiologic pathways of genes and risk factors determining quantitative traits are shown in Figure 3.1.

Figure 3.1 Simulated genetic mechanisms for GAW 16 data set



Source: Kraja et al., 2008.

3.2 Longitudinal Quantitative Trait CAC

In this research, the dependent variable I use is a simulated quantity called "coronary artery calcification" (CAC), given at 3 time points, with 10 year intervals between measurements in 6,476 individuals. Kraja et al. (2008) modeled the longitudinal CAC in two stages. First, they modeled an age independent CAC (CAC_{AI}) as a function of two lipid variables CHOL and HDL, and 5 genes $\tau_1,\ldots,\tau_5$ which had direct effects on

its development. The locations of $\tau_1,\ldots,\tau_5$ are given in Table 3.1. Note that a RefSNP (RS) is a reference SNP. A "RS" number is a RefSNP accession ID used to identify and cluster SNPs that are mapped to the same location on the genome.

Table 3.1 The identities of genes contributing to CAC and MI event

Trait	Factor	RS number	Chromosome
CAC	τ_1	rs6743961	2
	τ_2	rs17714718	19
	τ_3	rs1894638	6
	τ_4	rs1919811	7
	τ_5	rs213952	7
MI event	φ_1	rs12565497	1
	φ_2	rs11927551	3

Source: Kraja et al., 2008.

The values of CAC_{AI} were simulated using the following model:

$$CAC_{AI} = 500 + 20(\text{Total CHOL} - 200) - 25(\text{HDL} - 53) + \text{ME} + \text{PE} + \text{Het} + \varepsilon, \qquad (3.1)$$

where $\varepsilon \sim N(0,300)$. Since CAC cannot be negative, CAC_{AI} is set as 0 if the generated value is not positive. In the model, τ_1 and τ_2 has a joint 2-locus genetic effect on ME; however, the effect τ_1 displays is only minimal compared with a considerable additive

main effect from τ_2. From Kraja et al. (2008), the interactions between the more common/less common homozygous genotype of τ_1 (CC and TT, respectively) and the more common homozygous genotype of τ_2 (CC) decrease the mean effect of ME on CAC_{AI} by 250 points. The interaction between the heterozygous genotype of τ_1 (CT) and the less common homozygous genotype of τ_2 (TT) decreases the mean effect of ME on CAC_{AI} by 150 points; the interaction between the heterozygous genotype of τ_1 (CT) and the common homozygous genotype of τ_2 (CC) increases the mean effect of ME on CAC_{AI} by 150 points; the interactions between the homozygous genotypes of τ_1 (CC and TT) and the less common homozygote of τ_2 (TT) increase the mean effect of ME on CAC_{AI} by 250 points. The interactions between the genotypes of τ_1 (CC, CT and TT) and the heterozygous genotype of τ_2 (CT) do not have any effects.

The pair of genes, τ_3 and τ_4, have a joint 2-locus, purely epistatic effect on PE in Equation (3.1). The interactions between the heterozygous genotype of τ_3 (CT) and the more common/less common homozygous genotypes of τ_4 (AA and CC, respectively) and the interactions between the homozygous genotypes of τ_3 (CC and TT) and the heterozygous genotype of τ_4 (AC) both decrease the mean effect of PE on CAC_{AI} by 200 points. Other combinations increase the mean effect of PE on CAC_{AI} by 200 points.

The gene τ_5 has an over-dominant allele for high CAC_{AI} and determines the Het effect. The heterozygous genotype of τ_5 (AG) decreases CAC_{AI} by 100 points on average,

the more common homozygote AA increases CAC_{AI} by 25 points, and the less common homozygote GG increases CAC_{AI} by 400 points.

The residual value ε is drawn from $N(0,1)$ and then multiplied by 300. It represents the sum of deviations from the mean of normally distributed modeled genetic effects and "noise" from other environmental and genetic effects not explained by the factors described in Equation (3.1).

The simulated CAC is derived from CAC_{AI} using a piecewise linear function adjusted by age. Participants under age 20 have not developed measurable levels of CAC; for participants from age 20 to 60, the CAC progresses linearly; for participants older than 60, CAC is equal to CAC_{AI}.

As shown in Figure 3.1, CAC influences the chance of having a myocardial infarction (MI) event before each visit. In addition, smoking and two genetic loci ϕ_1 and ϕ_2 interact with CAC to determine the risk of an MI event. The MI data were not analyzed in this paper. The two SNPs ϕ_1 and ϕ_2 are not associated with CAC levels but are associated with the MI event. They will be used later in the study as candidate "null" genes, with the expectation that they are not CAC risk factors. The positions of ϕ_1 and ϕ_2 are listed in Table 3.1.

3.2.1 Genes Used in the Analysis

A total of 27 SNPs are studied in this analysis: 5 SNPs ($\tau_1, \tau_2, \tau_3, \tau_4, \tau_5$) that have effects on the simulated CAC, 2 SNPs (ϕ_1 and ϕ_2) which determine MI but not the CAC level, and 20 "null" SNPs ($u_1, u_2, \ldots, u_{10}$ and $v_1, v_2, \ldots, v_{10}$), randomly selected from human chromosome (HC) 5 and HC 22, respectively, that were not in the genetic mechanism determining the simulated CAC and myocardial infarction (MI) events. The minor allele frequency (MAF) for each of the 4 SNPs $\tau_1, \ldots, \tau_4$ is approximately 0.5; τ_5 has MAF equal to 0.2; all of the other 20 SNPs have MAF greater than 0.15.

For each gene considered, I create two indicator variables: whether the participant's genotype is the more common homozygote and whether the participant's genotype is the less common homozygote. These indicator variables are used as trait variables (also called "time-invariant covariates" in Mplus or "risk factors" in the SAS PROC TRAJ programs) in the GMM models. The results for $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$ are one basis of the empirical null distribution of the test statistics. The results for ϕ_1 and ϕ_2 should be similar to the results for $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$. I also report results for four randomly chosen SNPs near τ_5 and τ_2, respectively, that have MAF greater than 0.1 and have genotype frequencies that are in Hardy-Weinberg equilibrium (HWE) to demonstrate the possible applicability of the proposed procedures for genome wide association studies (GWAS).

3.2.2 Measures of Association with Genes

I use the SAS TRAJ procedure (Jones, Nagin, and Roeder, 2001) and the Mplus program (Muthén, 2004) to perform GMM and to assess whether genotypes appear to be associated with trajectory component membership and hence suggest longitudinal QTL. Each SAS TRAJ analysis reports the maximized log likelihood, the maximum likelihood estimates (MLEs) of the trajectory component parameters, the t-statistics of the trajectory component parameters, the estimated frequency of each trajectory component, the Bayesian posterior probability (BPP) that each subject is a member of each trajectory component and the Bayesian Information Criterion (BIC) statistic which is used to assess the number of trajectory components. Mplus also reports these statistics.

Two sets of analyses applied to the 200 replicates are considered. Each replicate of data consists of 6,476 participants with genotypes and simulated phenotypes. For each of the 27 candidate SNPs, the first set uses the longitudinal CAC measures with the two genetic indicator variables used as traits but without the TVCs CHOL and HDL. The second is the longitudinal CAC with the TVCs and with the two genetic indicator variables as traits. I use a quadratic trend function and set the number of components to 2 and 3. I treat each participant as an independent observation. That is, I ignore the relationships within a pedigree.

For each set of analyses, I analyze the 200 replicates of the simulated data using three tests and assess their power: the likelihood ratio test statistic (LRTS), a direct test of

genetic model coefficients, and the chi-squared test classifying subjects based on the trajectory model's posterior Bayesian probability.

3.2.3 Direct Coefficient Test

In an analysis that identifies c trajectory components, there are $2(c-1)$ indicator variables associated with gene $i, i \in \{\tau_1,\ldots,\tau_5,\phi_1,\phi_2,u_1,...,u_{10},\nu_1,\ldots,\nu_{10}\}$. For example, for the τ_5 gene (which has homozygous genotypes AA and GG), there are estimated coefficients for the two homozygous indicators in components 2 through c. Component 1 is a reference group with coefficients of trait variables set to 1 identically in the SAS TRAJ procedure. With τ_5, I calculate $S_{\tau_5} = \sum_{j=2}^{c}(T^2_{AA,j} + T^2_{GG,j})$ and approximate its null distribution with the empirical distribution for $u_1,u_2,...,u_{10},\nu_1,\nu_2,\ldots,\nu_{10}$. I call this the "direct coefficient test" (DCT) and use the empirical critical value corresponding to a level of significance equal to 0.05 from the distribution for $u_1,u_2,...,u_{10},\nu_1,\nu_2,\ldots,\nu_{10}$. I conjecture that a chi-squared random variable with $2(c-1)$ degrees of freedom may be a good approximation for this null distribution.

3.2.4 Bayesian Posterior Probability Chi-Squared Test

The second procedure is the Bayesian posterior probability (BPP) chi-squared test on the 3 genotype rows by c trajectory component column contingency table. I use the results of the GMM model and classify each subject into the trajectory component that has the largest BPP. A significant value of the chi-squared test for independence

($p < 0.05$ based on the empirical distribution of the chi-squared test for $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$) indicates association with the gene. I conjecture that the empirical distribution will be approximately a central chi-squared distribution with $2(c-1)$ degrees of freedom.

3.2.5 Likelihood Ratio Test Statistic

The third procedure is the LRTS. I take the difference of the likelihood function with the two genetic indicator variables and the likelihood function without the two genetic indicator variables. I perform this test without TVC and with TVC respectively. A significant value of the LRTS ($p < 0.05$ based on the distribution for $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$) indicates association with the gene. I conjecture that the distribution of the LRTS for $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$ is bounded by a central chi-squared distribution with $2(c-1)$ degrees of freedom.

3.3 Gene-Gene Interaction Analysis

Two pairs of the genes, τ_1 with τ_2 and τ_3 with τ_4, have epistatic associations with CAC. To evaluate the power of GMM to detect the interactions between τ_1 and τ_2, for each of the 6,476 participants, I create four mutually exclusive indicator variables ME_I, ME_{II}, ME_{III}, ME_{IV} based on each individual's level of the mean effect of ME on CAC_{AI}

induced by the epistasis of τ_1 and τ_2: whether CAC_{AI} increases by 250 points (ME_I = 1, and 0 otherwise), whether CAC_{AI} increases by 150 points (ME_{II} = 1, and 0 otherwise), whether CAC_{AI} decreases by 150 points (ME_{III} = 1, and 0 otherwise), and whether CAC_{AI} decreases by 250 points (ME_{IV} = 1, and 0 otherwise). For example, if a participant has the genotype CC for τ_1 and the genotype TT for τ_2, the interaction increases the mean effect of ME on CAC_{AI} by 250 points. Thus, for this participant, ME_I = 1, ME_{II} = 0, ME_{III} = 0, and ME_{IV} = 0. In a GMM analysis that identifies *c* trajectory components, there are estimated coefficients for the four indicators in components 2 through *c*, as well as t statistics which hypothesize that the parameter equals 0 and their corresponding p-values. I use the DCT procedure and calculate the sum of the t-squared statistics $S_{ME} = \sum_{j=2}^{c} (T^2_{ME_I,j} + T^2_{ME_{II},j} + T^2_{ME_{III},j} + T^2_{ME_{IV},j})$. I approximate its null distribution with a chi-squared random variable with $4(c-1)$ degrees of freedom using level of significance 0.05.

Similarly, since an individual has one of only two possible combinations of the mean effect level of PE on CAC_{AI} caused by the epistasis of τ_3 and τ_4, I create one indicator variable PE_{II} = 1 when CAC_{AI} decreases by 200 points, and PE_{II} = 0 when CAC_{AI} increases by 200 points. In a GMM analysis that identifies 2 through *c* trajectory components, I perform the DCT and calculate $S_{PE} = \sum_{j=2}^{c} T^2_{PE_{II},j}$ and approximate its null distribution with a central chi-squared distribution with $(c-1)$ degrees of freedom using level of significance 0.05. I do not study the BPP and LRTS for detecting these epistatic relations.

3.4 Tests for Hardy-Weinberg Equilibrium

To evaluate deviations from HWE for each SNP studied in this research, I use Pearson's chi-squared test. Suppose there is a single locus with two alleles *A* and *a*, with frequencies denoted by *p* and *q*, respectively. We have *P(A)* = *p*, *P(a)* = *q*, and *p* + *q* = 1. If HWE holds for the genotype distribution in the population, we will have $P(AA) = p^2$ for the homozygote *AA*, $P(aa) = q^2$ for the homozygote *aa*, and $P(Aa) = 2pq$ for the heterozygote *Aa*. Suppose the observed genotype frequencies for *AA*, *Aa*, and *aa* for a total of *N* individuals with complete genotype information are $obs(AA), obs(Aa)$, and $obs(aa)$. The allele frequencies can be estimated as: $\hat{p} = \frac{2 \times obs(AA) + obs(Aa)}{2N}$ and $\hat{q} = 1 - \hat{p}$. Under the hypothesis of HWE, the expected number of subjects for each genotype can be expressed as: $Exp(AA) = \hat{p}^2 N, Exp(Aa) = 2\hat{p}\hat{q}N$, and $Exp(aa) = \hat{q}^2 N$. Therefore, the Pearson's chi-square test statistic can be calculated as:

$$\chi_1^2 = \sum \frac{(O-E)^2}{E} = \frac{(obs(AA) - Exp(AA))^2}{Exp(AA)} + \frac{(obs(Aa) - Exp(Aa))^2}{Exp(Aa)} + \frac{(obs(aa) - Exp(aa))^2}{Exp(aa)}.$$

with one degree of freedom, since the degree of freedom equals the number of phenotypes minus the number of alleles. The 1% level of significance for χ_1^2 = 6.64 is used. If the chi-square statistic is larger than this value, the null hypothesis that the population is in HWE will be rejected.

Example: The gene τ_5 has the observed genotype frequencies for *AA*, *AG*, and *GG* for a total of 6,474 (genotypes for 2 participants were missing) individuals as 4,176, 2,014, and 284. The sample frequency of the less common allele *G* is $\frac{2\times 284 + 2,014}{2\times 6,474} = 0.1994$, which is also the estimated MAF for τ_5. The sample frequency for allele *A* is therefore equal to 1 – MAF = 0.8006. The expected number of subjects for the genotypes *AA*, *AG*, and *GG* are therefore $(0.8006)^2(6,474) = 4,149.6$, $2(0.8006)(0.1994)(6,474) = 2,067.0$, and $(0.1994)^2(6,474) = 257.4$, respectively. The chi-square test statistic with degree of freedom 1 is obtained as follows:

$$\chi_1^2 = \frac{(4,176 - 4,149.6)^2}{4,149.6} + \frac{(2,014 - 2,067.0)^2}{2,067.0} + \frac{(284 - 257.4)^2}{257.4} = 4.27 .$$

Since 4.27 is less than the critical value 6.64, we do not reject the null hypothesis and report that τ_5 appears to be in HWE.

3.5 Linkage Disequilibrium Measures and the Chi-Squared Test

In addition to the true genes, I study the power of the three procedures for nearby SNPs and evaluate the association between the linkage disequilibrium (LD) and change of power. LD between disease locus alleles and alleles at nearby markers can be used to refine the location of the disease locus. In general, LD is expected to be related to the

distance between two loci, but there are many factors that may affect disequilibrium, including recombination, migration, selection, mutation, and population admixture and stratification. There may even be disequilibrium between alleles at loci located on different chromosomes.

LD involves haplotype frequencies and refers to the association between tightly linked SNPs. Two markers are said to be in LD if their alleles are in statistical association. For example, if P_{AB} is the probability that allele A_1 at genetic locus A occurs together with allele B_1 at locus B on the same chromosome, LD occurs when $P_{A_1B_1} \neq P_{A_1} P_{B_1}$. Thus the A_1B_1 haplotype occurs either more or less frequently than would be expected on the assumption of statistical independence. Table 3.2 shows the observed haplotype frequencies between alleles at loci A and B.

Table 3.2 Haplotype frequencies between alleles at loci A and B

		Locus B		
		B_1	B_2	Total
Locus A	A_1	$P_{A_1B_1}$	$P_{A_1B_2}$	P_{A_1}
	A_2	$P_{A_2B_1}$	$P_{A_2B_2}$	P_{A_2}
	Total	P_{B_1}	P_{B_2}	1

There are a variety of LD measures. I will focus on three of the most common measures: the disequilibrium coefficient *D* (also called LD coefficient), Lewontin's *D'* (Lewontin, 1964), which is a normalized disequilibrium coefficient, and the squared correlation coefficient r^2. There are more measures discussed in Devlin and Risch (1995). Using the information in Table 3.2, the disequilibrium coefficient *D*, the most basic measure of LD, can be easily calculated (Lewontin and Kojima, 1960):

$$D = P_{A_1B_1}P_{A_2B_2} - P_{A_1B_2}P_{A_2B_1} = P_{A_1B_1} - P_{A_1}P_{B_1} = P_{A_2B_2} - P_{A_2}P_{B_2} = P_{A_1}P_{B_2} - P_{A_1B_2} = P_{A_2}P_{B_1} - P_{A_2B_1}.$$

The calculation of *D* depends only on observed frequencies. The value of *D* ranges from -0.25 to 0.25. If both haplotype frequencies are 0.5, *D* will be maximal. Although *D* captures the intuitive concept of disequilibrium, its numerical value is difficult to use for measuring and comparing the strength of LD.

Lewontin (1964) proposed a normalized *D* by diving *D* by the absolute maximum *D* which could be achieved from the observed haplotype frequencies. Lewontin's *D'* is defined as

$$D' = \begin{cases} \dfrac{D}{\min(P_{A_A}P_{B_2}, P_{A_2}P_{B_1})}, D \geq 0 \\ \dfrac{D}{\min(P_{A_1}P_{B_1}, P_{A_2}P_{B_2})}, D < 0 \end{cases}$$

The value of Lewontin's *D'* is between - 1 and 1. When $|D'| = 1$, the LD is said to be complete. However, $|D'| = 1$ may indicate that at least one haplotype is missing. Since Lewontin's *D'* is derived from population genetic considerations, there is no implication that $D' = 1$ should imply that the two markers carry the same information. The squared

correlation coefficient r^2 as used by Hill and Roberson (1968) and Franklin and Lewontin (1970) has this property. The measure is defined as

$$r^2 = \frac{D^2}{P_{A_1} P_{A_2} P_{B_1} P_{B_2}}.$$

The value of r^2 is between 0 and 1 with $r^2 = 1$ indicating perfect LD. That is, observations at one marker provide complete information about the other marker, making the second redundant. The value of r^2 can be small even when $|D'|$ is 1.

To evaluate the significance of LD, one can use the chi-square statistic to test whether the LD coefficient D between two markers is different from zero as follows (Weir, 1979 & 1990):

$$\chi^2_{df} = \frac{2nD^2}{P_{A_1} P_{A_2} P_{B_1} P_{B_2}},$$

where $df = (k-1)(l-1)$ for the pair of markers with k and l alleles, respectively; n is the number of individuals in the population. Here, the degree of freedom parameter equals 1. If the test statistic is larger than the critical value $\chi^2_1 = 3.84$ with 5% level of significance, D is apparently different from zero, and the population under study appears to be in LD.

3.6 Evaluation of the Two Software Packages

I ran the Mplus software on replicates 1 through 11 with two and three trajectory components specified with participants' age as individually-varying times of observations

for the outcome CAC. The software either failed to converge or failed to identify the solution due to excessive numbers of local maxima. I used at least 500 sets of starting values in the initial stage and 100 optimizations in the second stage. Mplus computation times were between 67 and 75 hours for each replicate to fit the 2-component models without any time-invariant or time-varying covariates. The Mplus software was not considered any further.

As for the SAS TRAJ procedure, the GMM modeling for each replicate took less than one minute to identify two trajectory components without adding any covariates, nearly one minute to identify two trajectory components with genetic indicator variables, and about one minute to identify two trajectory components with genetic indicator variables and time-varying covariates. It took about three minutes for the SAS TRAJ procedure to identify three trajectory components with genetic indicator variables and time-varying covariates for each replicate.

Chapter 4 Obtaining Empirical Values for the Null Distribution of Test Statistics

4.1 Null Distribution Based on Two Human Chromosomes Not in the Disease Mechanism

The 20 candidate SNPs $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$ for the empirical null distribution were chosen from HC 5 and HC 22 which were not in the simulated genetic model determining CAC or any of the CAC related traits (eg. CHOL and HDL). The MAF ranged from 0.16 to 0.49 for these SNPs. Half were in HWE, and half were not. The chi-squared test statistics for HWE and the corresponding p-values are given in Table 4.1.

Table 4.1 Summary characteristics of the twenty candidate SNPs used for the empirical null distribution

Chromo-some	SNP Label	RS number	Physical Position (cM)	MAF	χ^2 for HWE	*P*-value	In HWE
5	U1	rs819910	15.2689	0.30	187.38	<0.001	No
	U2	rs1754389	15.4451	0.39	2.31	0.1284	Yes
	U3	rs53610702	15.5514	0.31	0.00	0.9800	Yes
	U4	rs11542515	16.8158	0.28	2.04	0.1536	Yes
	U5	rs15648724	23.3948	0.28	0.38	0.5367	Yes
	U6	rs76666545	25.7556	0.45	0.07	0.7846	Yes
	U7	rs77537313	31.0333	0.41	25.60	<0.001	No
	U8	rs12098179	43.9496	0.26	58.04	<0.001	No
	U9	rs41622162	44.9991	0.38	16.53	<0.001	No
	U10	rs14007463	49.4872	0.29	24.27	<0.001	No
22	V1	rs15268900	0.8199	0.25	579.21	<0.001	No
	V2	rs15445079	1.7544	0.25	1.11	0.2906	Yes
	V3	rs15551377	53.6107	0.27	101.84	<0.001	No
	V4	rs16815794	11.5425	0.25	3.07	0.0799	Yes
	V5	rs23394809	156.4872	0.25	496.26	<0.001	No
	V6	rs25755592	76.6665	0.34	0.55	0.457	Yes
	V7	rs31033292	77.5373	0.40	31.94	<0.001	No
	V8	rs43949633	120.9818	0.49	177.68	<0.001	No
	V9	rs44999080	41.6222	0.16	0.51	0.4748	Yes
	V10	rs49487182	14.0075	0.29	0.39	0.5341	Yes

I ran the SAS TRAJ procedure for $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$ with two and three trajectory components, with and without TVCs. The distribution of the results from the three tests using the SAS TRAJ procedure had greater means and standard deviations for the ten SNPs that were not in HWE ($u_1, u_7 - u_{10}, v_1, v_3, v_5, v_7, v_8$) than for the ten in HWE ($u_2 - u_6, v_2, v_4, v_6, v_9, v_{10}$) as shown in Table 4.2. Out of the 200 replicates, the rates of

model failure for 2-component models without TVCs and 2-component models with TVCs were both 0%. The rates of model failure for 3-component models without TVCs CHOL and HDL and 3-component models with TVCs were 10% and 16% on average, respectively. The means and standard deviations of the DCT and BPP tests but not the LRTS for the 2 trajectory component models appeared to be relatively close to the value 2, which holds for a χ_2^2 distribution when the SNP was in HWE. The LRTS was well beyond the expected asymptotic distributions, particularly for the group of SNPs not in HWE. The use of TVC appeared to increase the mean and standard deviation observed for all the test statistics. I used the 95^{th} percentile for the ten markers in HWE as the critical value for subsequent tests.

Table 4.2 Summary statistics of three tests for the 20 null SNPs from HC5 and HC22, 200 replicates

Test, components, TVC	In HWE Mean (Std)	Not in HWE Mean (Std)	In HWE 95^{th} Empirical Percentile
LRTS, 2, no TVC	1680.85 (2890.89)	19159.28 (12361.63)	10083.25
LRTS, 2, TVC	15213.58 (2747.61)	31852.81 (11753.07)	23171.14
LRTS, 3, no TVC	1678.50 (2883.70)	19096.85 (12314.53)	10059.77
LRTS, 3, TVC	15654.12 (2752.35)	32056.14 (11639.50)	23583.94
DCT, 2, no TVC	1.77 (1.74)	2.28 (2.52)	5.22
DCT, 2, TVC	1.83 (1.84)	3.15 (3.35)	5.67
DCT, 3, no TVC	3.88 (3.13)	3.93 (3.15)	9.90
DCT, 3, TVC	4.29 (5.09)	4.57 (3.90)	11.16
BPP, 2, no TVC	2.96 (2.92)	4.06 (4.65)	8.79
BPP, 2, TVC	3.19 (3.03)	6.80 (11.51)	9.27
BPP, 3, no TVC	7.27 (5.43)	8.16 (6.37)	17.60
BPP, 3, TVC	9.99 (16.57)	14.06 (16.58)	22.98

4.2 Distribution of the DCT and BPP Tests

I compared the distributions of DCT and BPP for the groups of SNPs in HWE and for the group of SNPs not in HWE using Kolmogorov-Smirnov (K-S) tests. All analyses indicated that the distributions of the two groups differed significantly except for the DCT based on the 3-component models without TVCs. The comparison results are reported in Table 4.3.

Table 4.3 The asymptotic K-S statistics (ksa) and p-values: comparisons of the distributions for DCT and BPP from the 10 null SNPs in HWE and the distributions for DCT and BPP from the 10 null SNPs not in HWE

Ksa, p-value	2 components		3 components	
	No TVCs	TVCs	No TVCs	TVCs
DCT	2.17, $p < 0.001$	5.89, $p < 0.001$	0.67, $p = 0.766$	2.29, $p < 0.001$
BPP	2.75, $p < 0.001$	5.49, $p < 0.001$	1.88, $p = 0.002$	4.11, $p < 0.001$

As shown in Figure 4.1, the empirical distribution of the DCT applied to the ten SNPs in apparent HWE were significantly different from the distribution of DCT applied to the other ten SNPs not in HWE for the 2-component models without TVCs. The distribution of DCT of the 2-component models without adding TVCs for the ten SNPs in HWE has both the mean and the standard deviation close to 2, consistent with a χ_2^2 distribution. The DCT of the 2 trajectory component models without adding TVCs for the ten SNPs not in HWE has larger mean and standard deviation than the DCT for the ten SNPs in HWE. The results for the distributions of BPP applied to the SNPs in HWE and to the SNPs not in HWE are given in Figure 4.2. Like DCT, the mean and standard deviation for the BPP of the 2 trajectory component models without adding TVCs from the ten SNPs not in HWE are much larger than those from the ten SNPs in HWE. Similar results hold for the models with TVCs and for the 3-component models (also see Table 4.2).

Figure 4.1 Histograms of the empirical distributions of DCT applied to the 10 null SNPs in HWE and to the 10 null SNPs not in HWE for 2-component models without TVCs, 200 replicates

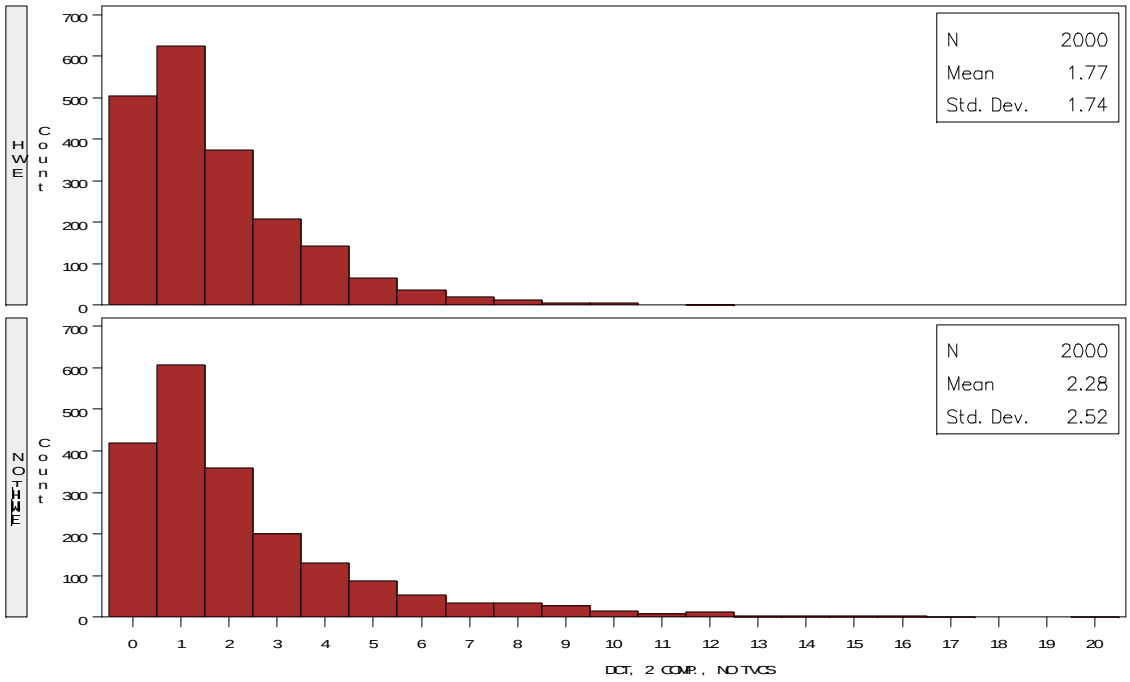
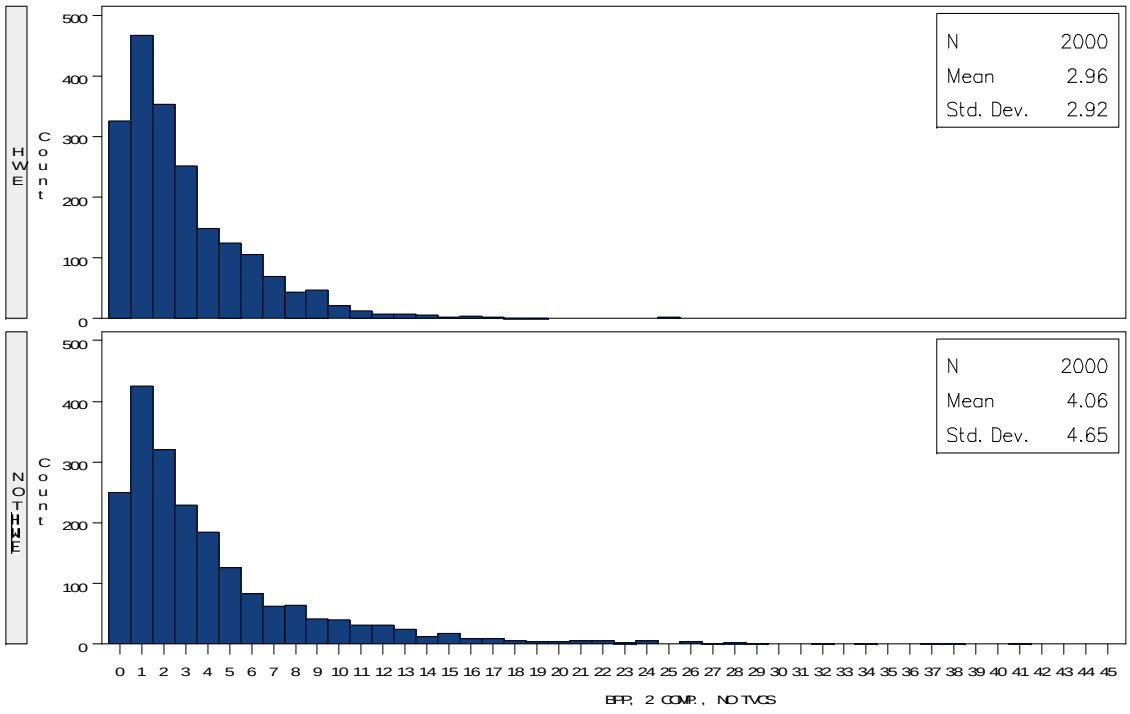


Figure 4.2 Histograms of the empirical distributions of BPP applied to the 10 null SNPs in HWE and to the 10 null SNPs not in HWE for 2-component models without TVCs, 200 replicates



The distributions of DCT and BPP tests for each of the ten SNPs in HWE are shown in Figures 4.3 (A) - (D). The descriptive characteristics for the distributions of each test using 2-component models are shown in Table 4.5 – Table 4.7. Among these ten markers, the distributions for U5 of DCT and BPP with 2-component models, with and without TVCs, appeared to be different from the distributions for all other SNPs. The U5 SNP had the highest means and standard deviations for the distributions of most of the tests. The U6 SNP also had high means and standard deviations and showed great variability in the distributions.

Figure 4.3 (A) Empirical distribution function plot for the distributions of DCT of 2-component models without TVCs for the 10 null SNPs in HWE

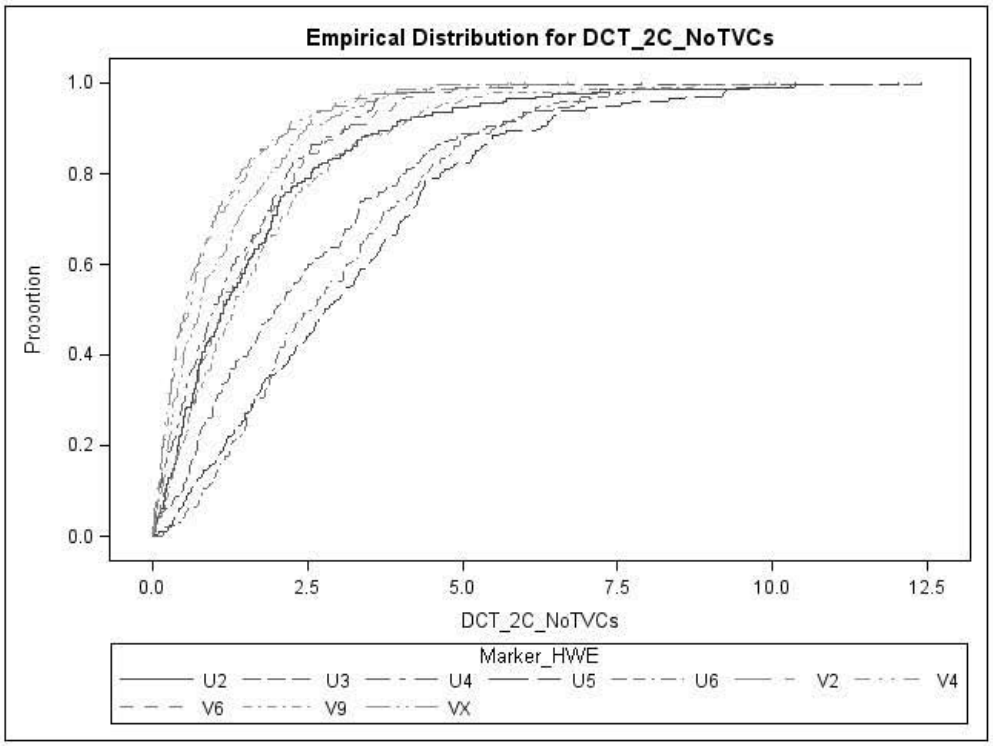


Figure 4.3 (B) Empirical distribution function plot for the distributions of DCT of 2-component models with TVCs for the 10 null SNPs in HWE

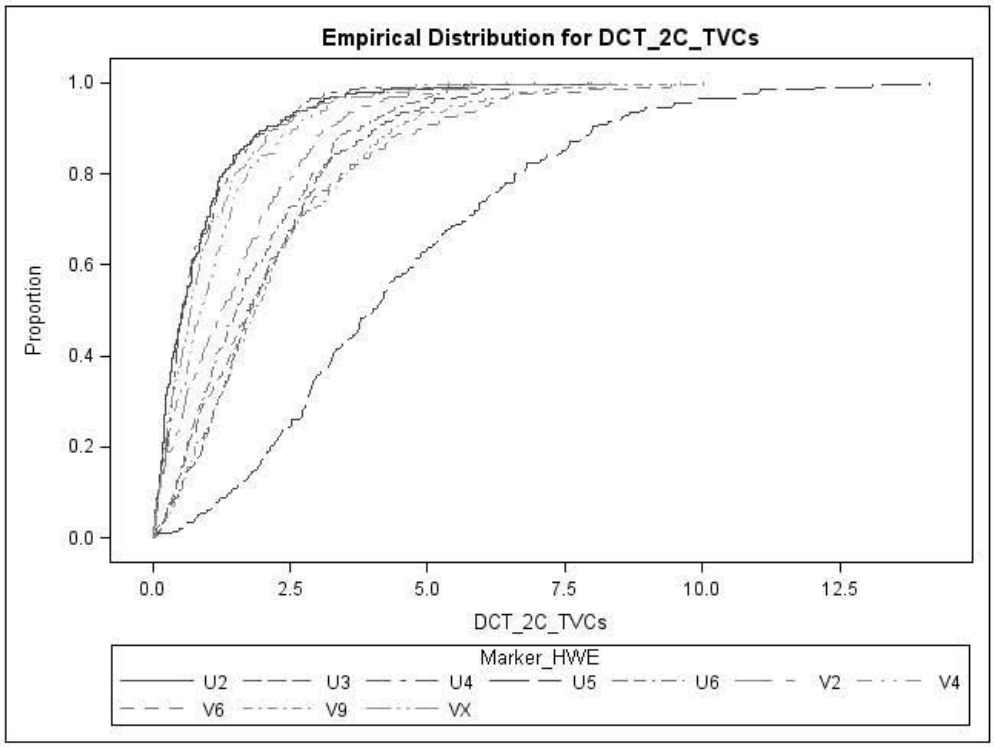


Figure 4.3 (C) Empirical distribution function plot for the distributions of BPP of 2-component models without TVCs for the 10 null SNPs in HWE

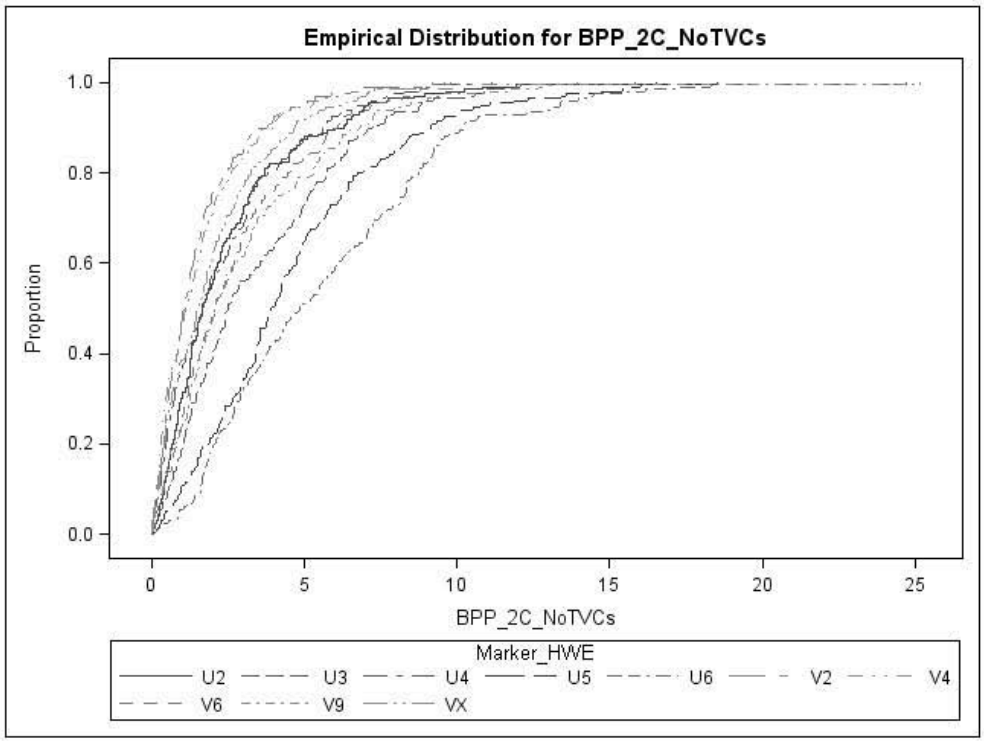


Figure 4.3 (D) Empirical distribution function plot for the distributions of BPP of 2-component models with TVCs for the 10 null SNPs in HWE

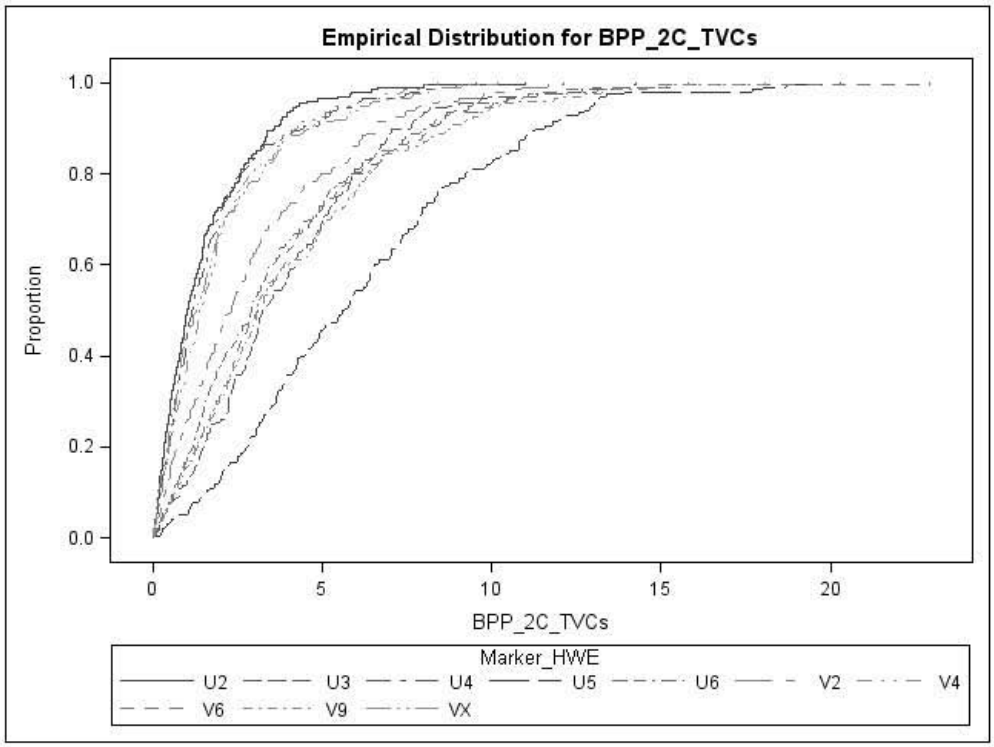


Table 4.4 Characteristics of the DCT test statistic values obtained without TVCs for the 10 null SNPs in HWE: 2-component trajectory models

SNP Label	Mean (Std)	95 percentile	99 percentile
U2	1.69 (1.77)	5.28	10.28
U3	2.49 (2.04)	6.87	8.14
U4	1.34 (1.22)	3.58	6.04
U5	3.16 (2.22)	7.56	10.35
U6	2.93 (1.87)	6.43	8.42
V2	0.91 (1.05)	3.17	5.58
V4	0.92 (1.03)	2.95	5.58
V6	1.46 (1.18)	3.92	5.45
V9	1.73 (1.59)	4.66	8.73
V10	1.11 (1.06)	3.42	4.56

Table 4.5 Characteristics of the DCT test statistic values obtained with TVCs for the 10 null SNPs in HWE: 2-component trajectory models

SNP Label	Mean (Std)	95 percentile	99 percentile
U2	0.87 (1.06)	3.01	5.67
U3	2.08 (1.45)	5.06	6.52
U4	0.89 (0.96)	2.74	5.38
U5	4.48 (2.69)	9.49	13.08
U6	1.84 (1.35)	4.59	5.73
V2	1.53 (1.34)	3.99	6.13
V4	1.13 (1.03)	3.21	4.81
V6	2.23 (1.88)	6.14	9.42
V9	2.25 (1.68)	5.68	8.26
V10	0.99 (1.02)	3.02	5.13

Table 4.6 Characteristics of the BPP test statistic values obtained without TVCs for the 10 null SNPs in HWE: 2-component trajectory models

SNP Label	Mean (Std)	95 percentile	99 percentile
U2	2.49 (2.46)	7.07	11.90
U3	3.45 (2.84)	8.82	12.08
U4	2.33 (2.20)	7.10	9.26
U5	4.66 (3.37)	11.07	16.02
U6	5.65 (3.94)	13.39	18.31
V2	1.55 (1.61)	5.07	9.25
V4	1.67 (1.89)	5.27	10.05
V6	2.70 (2.23)	7.33	9.14
V9	3.04 (3.05)	8.46	13.04
V10	2.03 (1.91)	6.39	9.16

Table 4.7 Characteristics of the BPP test statistic values obtained with TVCs for the 10 null SNPs in HWE: 2-component trajectory models

SNP Label	Mean (Std)	95 percentile	99 percentile
U2	1.54 (1.59)	4.32	8.01
U3	3.93 (2.79)	9.01	14.18
U4	1.78 (1.85)	5.96	8.94
U5	6.15 (3.90)	13.03	18.69
U6	3.70 (3.10)	9.45	15.32
V2	0.91 (1.05)	2.98	2.59
V4	0.92 (1.03)	1.90	1.76
V6	1.46 (1.18)	3.86	3.20
V9	1.73 (1.59)	10.05	14.80
V10	1.11 (1.06)	6.42	9.11

The empirical distributions of the DCT and BPP for genes not associated with CAC values appeared to depend on whether the gene was apparently in HWE. Since violation of HWE is often used as a test for large genotyping error rates (Leal, 2005), a question to be considered is the robustness of these procedures to genotyping error. I have evaluated the effect of genotyping errors on the empirical null distributions and of other genes and will present the results in Chapter 7.

Chapter 5 Genes in the Genetic Models Known

5.1 The Seven Genes in the Genetic Mechanisms

I ran the SAS TRAJ procedure for the 200 replicates and studied two and three trajectory components, with and without TVCs for the seven genes $\phi_1, \phi_2, \tau_1, \tau_2, \tau_3, \tau_4$, and τ_5 in the genetic model. I used the 95th percentile for the ten markers in HWE from HC 5 and HC 22 as the critical value in my power study (see Table 4.2). That is, the fraction of replicates that yield a value of the statistic greater than the critical value of a corresponding test is the estimated power of the test or the "rejection rate". Table 5.1 contains the rejection rates by gene for the analysis results of the three procedures using the 2 and 3 trajectory component models, either including or excluding TVCs.

5.2 Results for the Two Genes Associated with MI but Not CAC

For ϕ_1 and ϕ_2, which were genes associated with MI but not CAC, the DCT and BPP rejection rates were low and consistent with 5% level of significance as shown in Table 5.1. The LRTS rejection rates were all 0, suggesting that the test might not be well defined for this application.

Table 5.1 Rejection rates of each test by gene, 200 replicates

Gene, Test		2 components, no TVC	2 components, TVC	3 components, no TVC	3 components, TVC
ϕ_1	LRTS	0	0	0	0
	DCT	1	5	2	7.5
	BPP	6	7	3.5	9
ϕ_2	LRTS	0	0	0	0
	DCT	1	1	2	2
	BPP	0.5	1	1	1.5
τ_5	LRTS	0	0	0	0
	DCT	100	100	90	85
	BPP	100	100	90	85
τ_2	LRTS	0	0	0	0
	DCT	85	99.5	62	85.5
	BPP	90.5	100	78.5	85.5
τ_1	LRTS	0	0	0	0
	DCT	5.5	2	1.5	28
	BPP	4	0.5	1.5	15
τ_3	LRTS	0	0	0	0
	DCT	2	1	3	3
	BPP	1	1.5	1.5	3
τ_4	LRTS	0	0	0	0
	DCT	0.5	3	1	2.5
	BPP	0.5	2	0.5	0

5.3 Results for the Five Genes in the Genetic Mechanisms Determining CAC

For τ_5 , the rejection rate was 100% for both DCT and BPP using the 2 trajectory component model with and without TVCs. The rejection rate for τ_2 is 85% for DCT and 91% for BPP with the 2 trajectory component model without TVCs. When the TVCs were included, the rejection rate for both DCT and BPP increased for τ_2. For τ_1, τ_3, and τ_4, the rejection rates for DCT and BPP are mostly below 5%, the level of significance.

For the five genes, the rejection rate of DCT and BPP tests was nearly the same on average. Use of TVCs did not increase the power since approximately 17% of the replicates did not have a solution when three components were specified with TVCs. However, when solutions with TVCs existed for all the GMM analyses that identified 2 trajectory components, there was an apparent increase in the power of DCT and BPP. That is, using three trajectory components rather than two did not appear to increase power due to failure of solutions. Compared to a 0% failure rate for both 2 trajectory component models without TVCs and 2 trajectory component models with TVCs, the rate of model failure was 10% for all the 3 trajectory component models without TVCs and 16% - 20% for the 3 trajectory component models with TVCs.

5.4 Results for the Gene-Gene Interactions

I ran the SAS TRAJ procedure for the 200 replicates and studied the epistasis effect of the two pairs of genes, τ_1 with τ_2 and τ_3 with τ_4 on CAC, using two and three trajectory component models and with and without TVCs.

In the GMM analyses that identified 2 trajectory components, excluding or including TVCs, the τ_1 and τ_2 interaction had 100% rejection rate for the DCT, since the sum of t-squared statistics of all the 200 replications was larger than the critical value $\chi_4^2 = 9.488$ with 5% level of significance. The rejection rate was also 100% for the DCT of the τ_1 and τ_2 interaction in the analyses that identified 3 trajectory component models, with and without TVCs. The critical value $\chi_8^2 = 15.507$ with 5% level of significance was used. The DCT for the 2 trajectory component had much higher means and standard deviations than the DCT for the 3 trajectory component. Use of TVCs increased the mean and the standard deviation of the DCT for the interaction of τ_1 and τ_2 substantially. The analysis results for the DCT are given in Table 5.2.

Similarly, the interaction of τ_3 and τ_4 had 100% rejection rate for DCT in the analyses that identified 2 and 3 trajectory components, whether excluding or excluding TVCs (see Table 5.2). The means and standard deviations for DCT with the 2 trajectory component were much higher than those for DCT with the 3 trajectory component.

Inclusion of TVCs did not necessarily increase the mean and the standard deviation of the DCT associated with the interaction of τ_3 and τ_4.

Table 5.2 Descriptive characteristics of the DCT for the epistasis from τ_1 with τ_2 and from τ_3 with τ_4, 200 replicates

Epistasis	Model	Mean (Std)	Minimum	Maximum	Rate of model failure (%)
$\tau_1\tau_2$	2 components, no TVC	162.15 (13.64)	124.50	201.02	0
	2 components, TVC	246.07 (16.97)	208.07	301.82	0
	3 components, no TVC	134.70 (12.03)	101.41	173.56	10
	3 components, TVC	190.71 (44.03)	132.80	625.08	16.5
$\tau_3\tau_4$	2 components, no TVC	410.92 (20.13)	362.40	464.62	0
	2 components, TVC	699.42 (27.58)	619.59	795.99	0
	3 components, no TVC	355.00 (25.11)	290.47	431.24	10
	3 components, TVC	289.21 (60.09)	40.69	574.22	15.5

Chapter 6 Markers near the Actual Gene

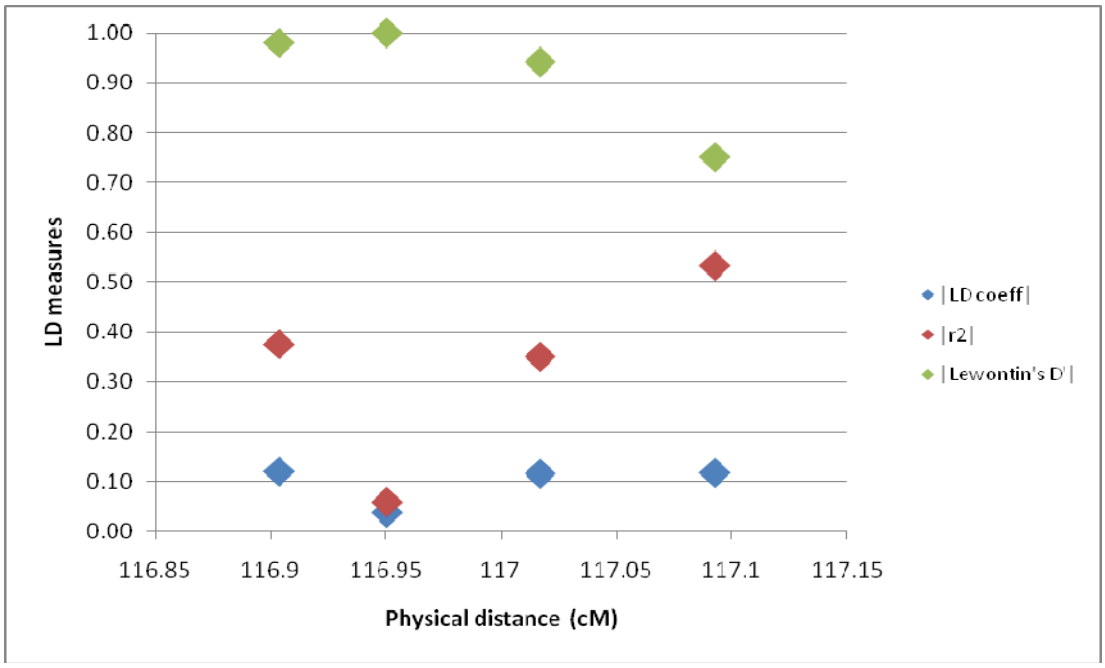
6.1 Markers near the Two Genes τ_5 and τ_2

I selected and analyzed four SNPs near τ_5 and τ_2 that had MAF greater than 0.1 and were in HWE respectively to demonstrate the possible applicability of the three procedures for genome-wide association studies (GWAS). The SAS ALLELE procedure (Czika et al., 2005) and the chi-squared test were used to test for linkage disequilibrium (LD) and to calculate the LD measures of τ_5 and τ_2 with their nearby markers. As shown in Table 6.1, the chi-squared test indicated that the LD coefficient of each of the four SNPs near τ_5 appeared to be significantly different from 0 ($p < 0.001$). That is, there was apparent LD between τ_5 and each of the four flanking markers. The position, LD measures, chi-squared statistics and the corresponding p-values of the SNPs near τ_5 are given in Table 6.1. The LD measures for the four SNPs near τ_5 by physical position on HC7 are shown in Figure 6.1. There was no apparent association between the LD measures and the distance between τ_5 and the four SNPs near τ_5. The Pearson correlation (0.30, 0.58, -0.89 for |LD coefficient|, r^2 , and |Lewontin's D'|, respectively, $p > 0.1$) confirmed that there was no significant association between the LD measures and the physical distance.

Table 6.1 Physical location and the LD measures for τ_5 and the 4 nearby SNPs

SNP Label	Position (cM)	LD measures			Chi-squared test for LD with τ_5	
		|LD coeff|	r^2	|Lewontin's D'|	χ^2	P-value
FM51	116.9034	0.12	0.38	0.98	2427.92	$p < 0.001$
FM52	116.9502	0.04	0.06	1	374.67	$p < 0.001$
τ_5	116.9907	.	.	.	.	.
FM53	117.0168	0.12	0.35	0.94	2258.14	$p < 0.001$
FM54	117.0926	0.12	0.53	0.75	3303.80	$p < 0.001$

Figure 6.1 LD measures for the 4 SNPs near τ_5 by physical position



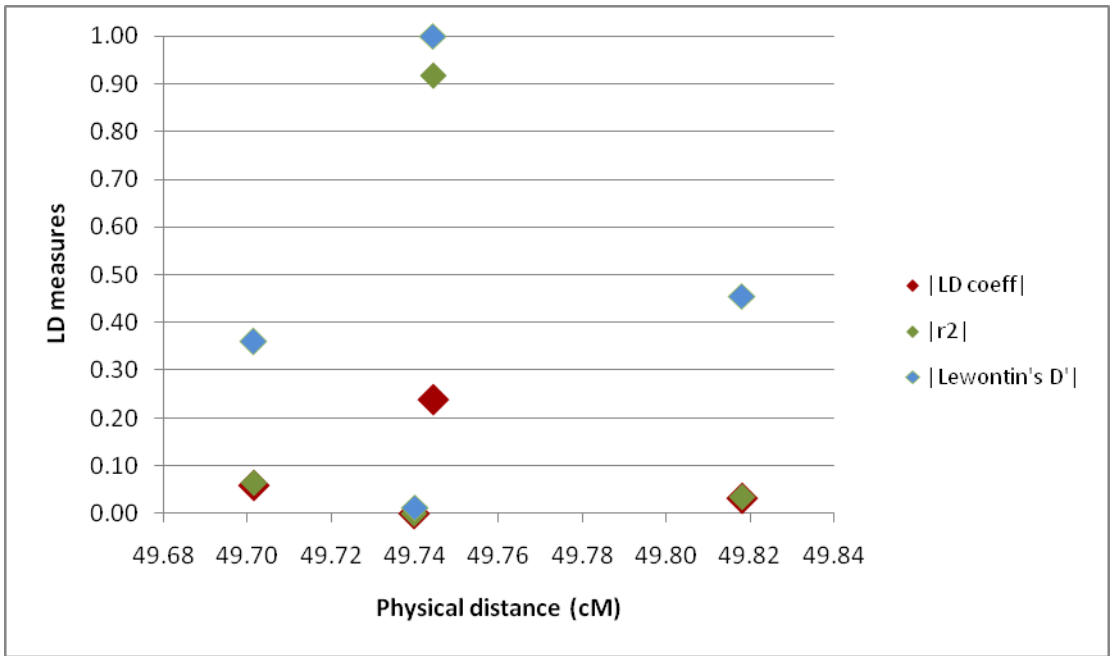
As shown in Table 6.2, the results of the chi-squared tests for LD indicated that three of the four SNPs near τ_2 were in LD with τ_2 ($p < 0.001$). The SNP FM22 had very

low values for all of the LD measures, and was not in LD with τ_2 ($p = 0.4816$). Figure 6.2 depicts the relationship between the three LD measures and the physical position on HC19 for the 4 SNPs near τ_2. There appeared to be no apparent association between the LD measures and the physical position (Pearson correlation = -0.15, -0.11, 0.1 for |LD coefficient|, r^2, and |Lewontin's D'|, respectively, $p > 0.8$).

Table 6.2 Physical location and the LD measures for τ_2 and the 4 nearby SNPs

SNP Label	Position (cM)	LD measures			Chi-squared test for LD with τ_2	
		|LD coeff|	r^2	|Lewontin's D'|	χ^2	P-value
FM21	49.7015	0.06	0.07	0.36	841.90	$p < 0.001$
FM22	49.7400	0.00	0.00	0.01	0.50	0.4816
τ_2	49.7426	.	.	.	.	.
FM23	49.7444	0.24	0.92	1	11863.58	$p < 0.001$
FM24	49.8182	0.03	0.03	0.46	450.38	$p < 0.001$

Figure 6.2 LD measures for the 4 SNPs near τ_2 by physical position

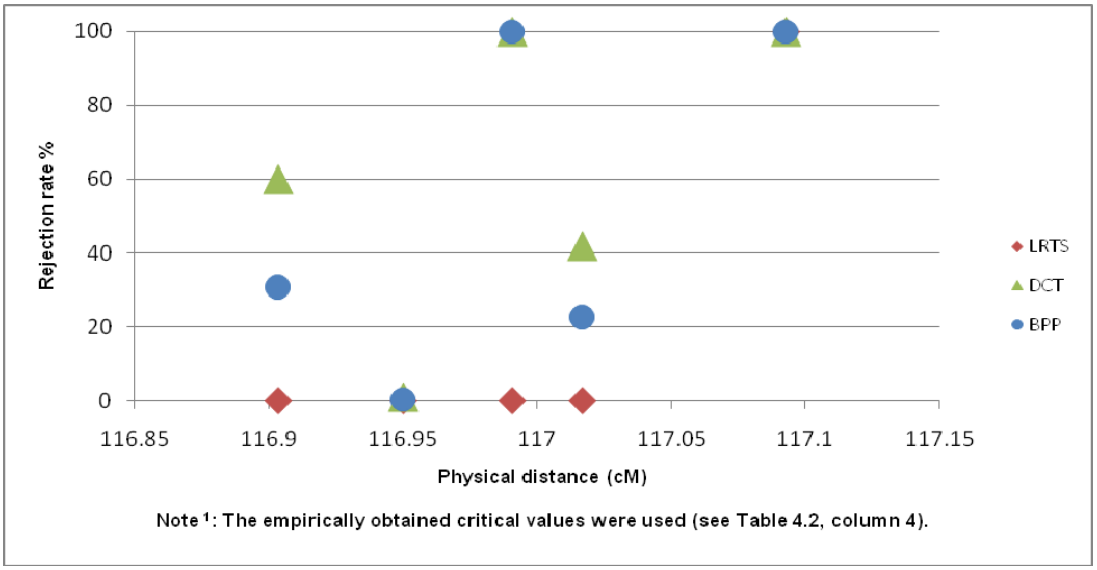


6.2 Markers flanking τ_5

I ran the SAS TRAJ procedure for the 200 replicates and studied the four SNPs near τ_5, using two and three trajectory component models and with and without TVCs. Figure 6.3 shows the rejection rate of the three procedures for τ_5 (116.9907 cM) and the four SNPs near τ_5 using two trajectory components without TVCs. The rejection rate for the nearby SNP FM54 (117.0926 cM) was 100% for all the three tests. The rejection rate was greater than 40% for DCT for two nearby SNPs (FM51 and FM53). The rejection rate for BPP was about half the rejection rate for DCT for the SNPs FM51, FM52 and

FM53. The rejection rate for both DCT and BPP for the SNP FM52 was at or below 1%. The LRTS had 0% rejection rate for three of the four markers near τ_5 except FM54.

Figure 6.3 Rejection rate[1] of tests for τ_5 and SNPs near τ_5 by physical position



I also examined three LD measures for the four SNPs near τ_5 and looked at their changes by the test as shown in Figure 6.4. The DCT rejection rate increased as the r^2 increased (Pearson correlation = 0.97, $p < 0.05$), while the LD coefficient showed a similar but weaker association (Pearson correlation = 0.81). The Lewontin's D' measure showed a negative association (Pearson correlation = -0.85). Similarly, the rejection rate of the BPP test appeared to be associated with the LD coefficient and the r^2 measures as shown in Figure 6.5. The Lewontin's D' measure showed a strong but negative association with the BPP rejection rate (Pearson correlation = -0.97, $p < 0.05$). The BPP

rejection rate and the $r^2$was also highly correlated (Pearson correlation = 0.87). The correlation between the BPP rejection rate and the LD coefficient was 0.59.

Figure 6.4 LD measures by DCT rejection rate for the 4 SNPs near τ_5

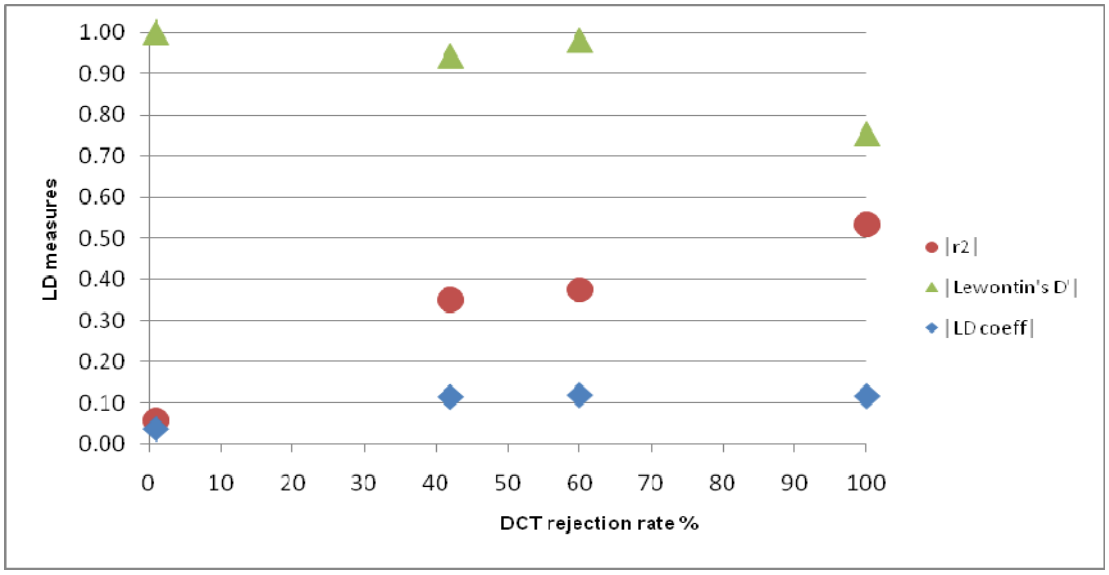
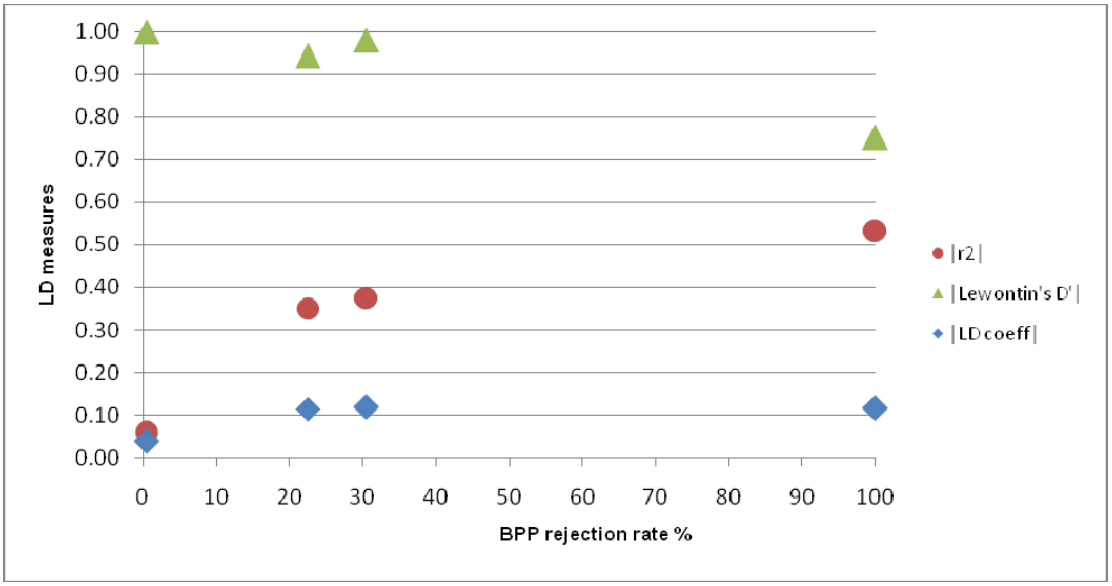


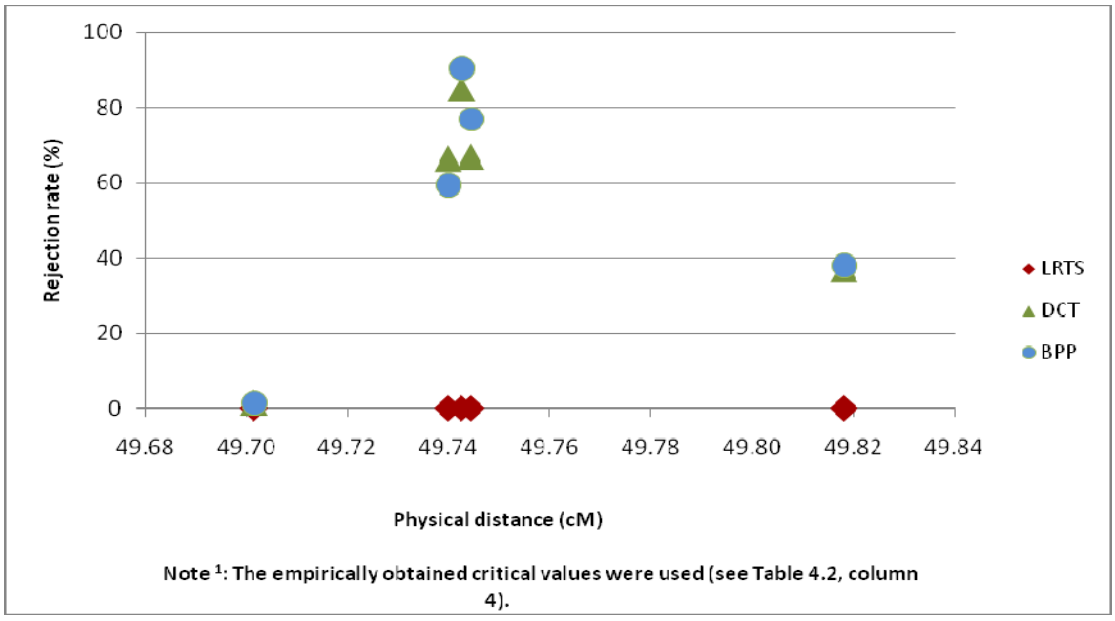
Figure 6.5 LD measures by BPP rejection rate for the 4 SNPs near τ_5



6.3 Markers flanking τ_2

I obtained analogous PROC TRAJ results for the rejection rate of tests for the four SNPs near τ_2. Figure 6.6 shows the rejection rate of the LRTS, DCT, and BPP procedures for τ_2 (49.8182 cM) and the four SNPs near using two trajectory component models without TVCs. The rejection rate for two of the nearby SNPs FM22 (49.7400 cM) and FM23 (49.7444 cM) was at or above 60% for both DCT and BPP. The rejection rate was about 38% for DCT and BPP for the nearby SNP FM24. The rejection rate for the remaining SNP FM21 was 1.5% for DCT and BPP. The rejection rate for BPP was nearly the same as the rejection rate for DCT. The LRTS rejection rate was 0 for all the four nearby SNPs.

Figure 6.6 Rejection rate of tests for τ_2 and the 4 SNPs near τ_2 by physical position



The SNP FM23, which is closest to the τ_2 gene had the highest value of the r^2 and Lewontin's *D'* as shown in Figure 6.7. All the three LD measures were very low for the SNP FM22, which is not in LD with τ_2. The power of the DCT and BPP tests appeared to have stronger association with the Lewontin's *D'* measure than with the LD coefficient or with the r^2 for the three SNPs near τ_2 other than FM22 as shown in Figure 6.7 and Figure 6.8. The Pearson correlation for the Lewontin's *D'* measure with the rejection rate of DCT was 0.91. The LD coefficient and the r^2 also appeared to be association with the DCT rejection rate (Pearson correlation = 0.77 and 0.82, respectively). The Pearson correlation with the BPP rejection rate for the Lewontin's *D'*, the LD coefficient and the r^2 were 0.93, 0.81, and 0.86, respectively.

Figure 6.7 LD measures by DCT rejection rate for the 4 SNPs near τ_2

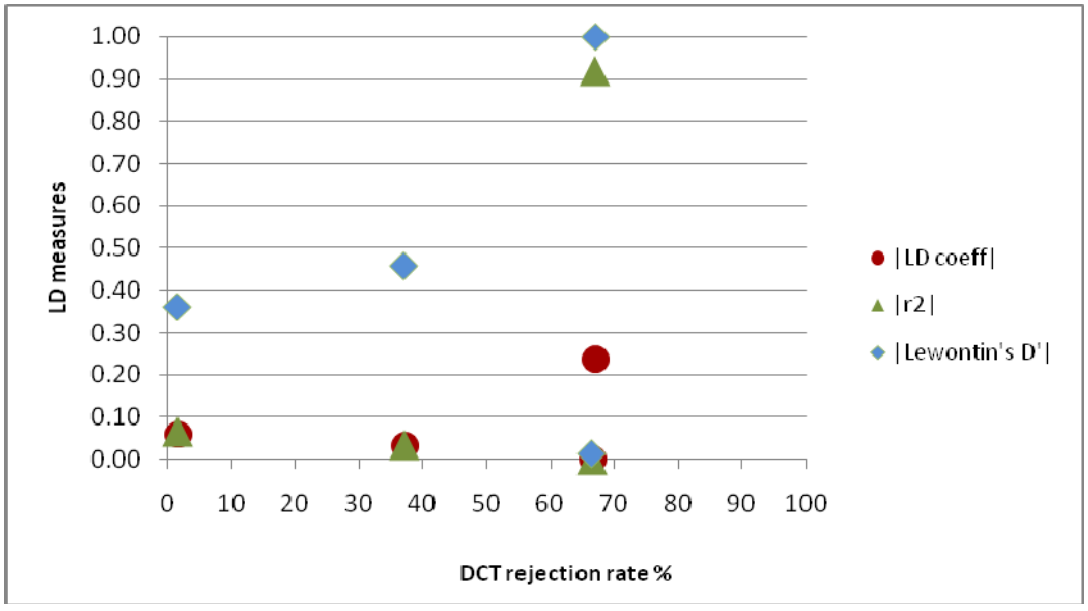
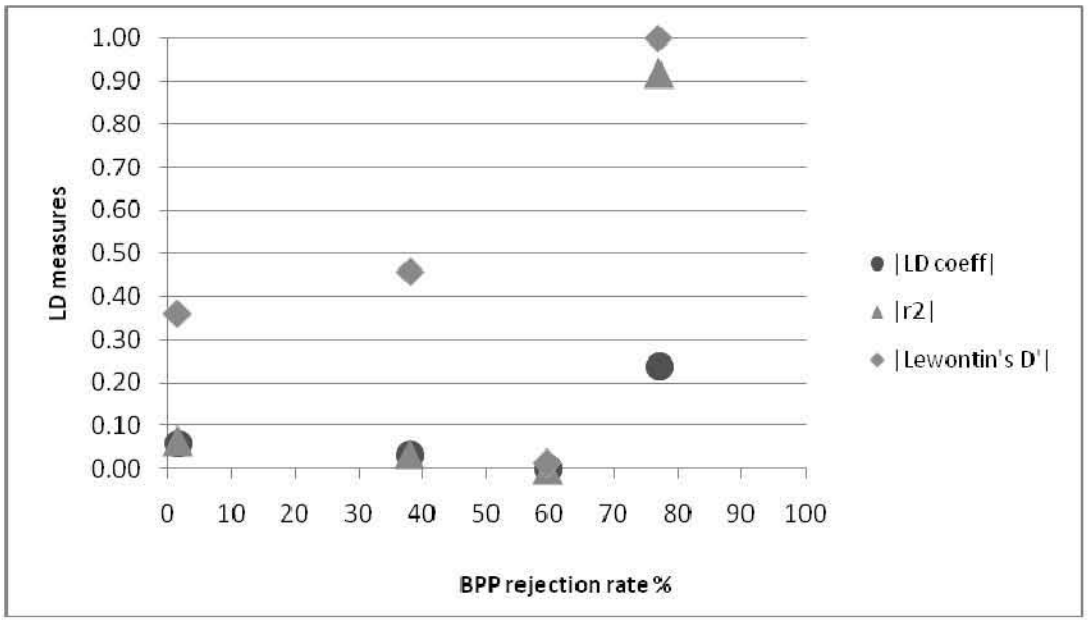


Figure 6.8 LD measures by BPP rejection rate for the 4 SNPs near τ_2



Chapter 7 Genotyping Error Study

7.1 Effects of Genotyping Errors

Genotyping errors can cause deviations from HWE and reduce the power of a genetic study (Leal, 2005). The error is particularly costly when misclassifying the more common homozygote as the less common homozygote, and the more common homozygote as the heterozygote, with the minimum sample size necessary to maintain constant asymptotic power that becomes infinitely increasing as the minor SNP allele frequency approaches zero (Kang et al., 2004, Ahn et al., 2006). From Chapter 4, the empirical distributions of DCT and BPP for genes not associated with CAC appeared to be sensitive to departures from HWE. Without further testing and examination, it is impossible to assess the extent that genotyping errors are responsible for such departures. In this chapter, I simulate different rates of genotyping errors and evaluate the inflation of the level of significance when the genotyping error is present.

7.2 Simulations of Genotyping Errors

I used a simple but realistic error model to simulate genotyping errors. This error model is a modified version from a general model for di-allelic marker loci as used by Mote and Anderson (1965), Kang et al. (2004), and Ahn et al. (2006). Table 7.1 represents the conditional probability of three observed genotypes given the true genotypes: the more common homozygote *AA*, the heterozygote *AB*, and the less common homozygote *BB*. Tintle et al. (2005) showed that the error rate of classifying a homozygote as the other homozygote is extremely rare (in only 0.00011% of the classifications), and the rates of misclassifying a homozygote as a heterozygote and misclassifying a heterozygote as a homozygote are roughly the same (about 0.2%). Therefore, I set the error rate of recoding a more common homozygote *AA* as a less common homozygote *BB* equal to 0, and vice versa. I set the error rates of the other four inconsistently identified classifications as identically ε, and the error rates of the three consistently identified classifications as 1 - ε or 1 - 2ε, as shown in Table 7.1. In my simulations, ε was set to 4% to imply a higher level of error rate, and 0.5% to imply a lower level of error rate.

Table 7.1 Conditional probability of the simulated genotypes given the true genotypes

	Simulated Genotype		
True Genotype	*AA*	*AB*	*BB*
AA	1 - ε	ε	0
AB	ε	1 - 2ε	ε
BB	0	ε	1 - ε

Based on Table 7.1, I created two sets of simulated genotypes for the 20 "null" SNPs $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$ that were not associated with CAC values or any of the CAC related traits for the 6,476 individuals in the sample using the error rates 4% and 0.5% respectively.

7.3 Effects of Genotyping Errors on the Empirical Null Distribution

For the 200 replicates, I ran the SAS TRAJ procedure for $u_1, u_2, ..., u_{10}, v_1, v_2, \ldots, v_{10}$ using the two sets of simulated genotypes with the number of trajectory components fixed to 2 and 3, and excluding or including TVCs. The results for the DCT and BPP tests are reported in Tables 7.2 (A) and (B), and Tables 7.3 (A) and (B).

Tables 7.2 (A) and (B) summarize the test statistics not including TVCs for the ten null SNPs in HWE and the ten null SNPs not in HWE, respectively. For the ten SNPs

in HWE, the mean and standard deviation for the DCT that identified 2 and 3 components and for the BPP that identified 2 components without TVCs appeared to be systematically increasing when the rate of genotyping error was increasing ($p < 0.05$, see Table 7.2 (A)). In contrast, the ten SNPs not in HWE, the mean and standard deviation of the DCT and the BPP that identified 2 components without TVCs tend to be decreasing with the rate of genotyping error increasing for ($p < 0.05$, see Table 7.2 (B)).

When the TVCs were included, for the ten SNPs in HWE, the mean and standard deviation for the DCT and BPP tests that identified 2 components appeared to be systematically increasing when the rate of genotyping error was increasing ($p < 0.001$), while the mean and standard deviation for the DCT and BPP tests that identified 3 components had a less clear tendency (see Table 7.3 (A)). With regard to the ten SNPs not in HWE, the mean and standard deviation of the BPP that identified 2 and 3 components with TVCs tend to be decreasing when the rate of genotyping error was increasing ($p < 0.01$), but the mean and standard deviation of the DCT did not appear to be systematically changing with the error rate as shown in Table 7.3 (B).

Table 7.2 (A) Mean and standard deviation of DCT and BPP tests for the 10 null SNPs from HC5 and HC22: in HWE, no TVCs

Test, components, TVC	In HWE Error rate			
	0% (Original)	0.5%	4%	*F* (*P*-value)
DCT, 2, no TVC	1.77 (1.74)	1.81 (1.77)	2.01 (2.06)	9.05, $p < 0.001$
DCT, 3, no TVC	3.88 (3.13)	3.90 (3.06)	4.19 (3.30)	5.60, $p = 0.004$
BPP, 2, no TVC	2.96 (2.92)	3.00 (2.98)	3.19 (3.23)	3.37, $p = 0.034$
BPP, 3, no TVC	7.27 (5.43)	7.41 (5.50)	7.68 (5.48)	2.67, $p = 0.069$

Table 7.2 (B) Mean and standard deviation of DCT and BPP tests for the 10 null SNPs from HC5 and HC22: not in HWE, no TVCs

Test, components, TVC	Not in HWE Error rate			
	0% (Original)	0.5%	4%	*F* (*P*-value)
DCT, 2, no TVC	2.28 (2.52)	2.11 (2.22)	2.10 (2.02)	3.88, $p = 0.021$
DCT, 3, no TVC	3.93 (3.15)	3.78 (3.03)	3.89 (3.09)	1.01, $p = 0.364$
BPP, 2, no TVC	4.06 (4.65)	3.75 (4.09)	3.61 (3.44)	6.41, $p = 0.002$
BPP, 3, no TVC	8.16 (6.37)	7.81 (5.93)	7.70 (5.50)	2.95, $p = 0.053$

Table 7.3 (A) Mean and standard deviation of DCT and BPP tests for the 10 null SNPs from HC5 and HC22: in HWE, with TVCs

Test, components, TVC	In HWE Error rate			
	0% (Original)	0.5%	4%	*F* (*P*-value)
DCT, 2, TVC	1.83 (1.84)	1.87 (1.90)	2.17 (2.27)	16.56, $p < 0.001$
DCT, 3, TVC	4.29 (5.09)	4.10 (3.49)	4.63 (5.09)	5.60, $p = 0.004$
BPP, 2, TVC	3.19 (3.03)	3.28 (3.15)	3.65 (3.58)	11.14, $p < 0.001$
BPP, 3, TVC	9.99 (16.57)	9.69 (11.21)	10.46 (11.78)	1.40, $p = 0.247$

Table 7.3 (B) Mean and standard deviation of DCT and BPP tests for the 10 null SNPs from HC5 and HC22: not in HWE, with TVCs

Test, components, TVC	Not in HWE Error rate			
	0% (Original)	0.5%	4%	*F* (*P*-value)
DCT, 2, TVC	3.15 (3.35)	2.95 (3.22)	3.09 (3.04)	2.10, $p = 0.122$
DCT, 3, TVC	4.57 (3.90)	4.66 (5.86)	4.65 (4.04)	0.20, $p = 0.816$
BPP, 2, TVC	6.80 (11.51)	6.07 (12.42)	5.29 (5.24)	10.87, $p < 0.001$
BPP, 3, TVC	14.06 (16.58)	13.65 (17.01)	12.19 (15.90)	5.94, $p = 0.003$

For the ten SNPs in HWE, the effect of genotyping error is noticeable. Without adding TVCs, the 95th empirical percentile of the tests was about 2% higher with the presence of 0.5% genotyping error rate, and about 10% higher on average with 4% genotyping error rate as shown in Table 7.4 (A). When TVCs were included, the 95th empirical percentile of the tests was about 3% higher when the genotyping error rate was

0.5%, and about 13% higher on average when 4% genotyping error rate was present, as shown in Table 7.4 (B).

Table 7.4 (A) The 95th empirical percentile of DCT and BPP tests for the 10 null SNPs from HC5 and HC22: in HWE, no TVCs

Test, components, TVC	In HWE 95th empirical percentile		
	0% (Original)	0.5%	4%
DCT, 2, no TVC	5.22	5.33	6.28
DCT, 3, no TVC	9.90	9.77	10.38
BPP, 2, no TVC	8.79	9.02	9.63
BPP, 3, no TVC	17.60	17.89	18.27

Table 7.4 (B) The 95th empirical percentile of DCT and BPP tests for the 10 null SNPs from HC5 and HC22: in HWE, with TVCs

Test, components, TVC	In HWE 95th empirical percentile		
	0% (Original)	0.5%	4%
DCT, 2, TVC	5.67	5.82	6.70
DCT, 3, TVC	11.16	11.04	12.21
BPP, 2, TVC	9.27	9.65	10.88
BPP, 3, TVC	22.98	23.45	24.49

7.4 Summary

For the ten "null" SNPs in HWE, the effect of genotyping error on the null distribution of the DCT and BPP is detectable. The empirical null distribution of DCT and BPP is thus sensitive to high genotyping error rates. The incorporation of the simulated genotyping errors in the analysis appeared to increase the mean and the standard deviation of the DCT and BPP statistics in the null case. In addition, the 95^{th} empirical percentile increased about 2% with presence of a 0.5% error rate, and could increase as high as 20% with the presence of a 4% error rate.

For the ten SNPs not in HWE, the incorporation of the simulated genotyping errors either increased or decreased the mean and the standard deviation of the DCT and BPP tests. As expected, in the event that the failure of HWE was due to genotyping errors, the addition of the simulated genotyping errors would be expected to have minor effect. Additionally, there seemed to be unknown causes responsible for the departure from HWE. More investigation in the potential genotyping errors of the data is needed. As of now, it appears not to be a wise choice to apply GMM procedures to SNPs that are apparently not in HWE.

Chapter 8 Conclusion and Discussion

The analyses based on the SAS TRAJ procedure have power to detect genes associated with CAC longitudinal QTLs. For CAC, the SAS TRAJ analysis of unadjusted CAC values with 2 trajectory components and no TVCs had excellent power (100% rejection rate for both DCT and BPP) to detect the τ_5 association and good power (85% rejection rate for DCT and 91% for BPP) to detect the τ_2 association. When TVCs were included and the solution existed, there was still 100% rejection rate for both DCT and BPP for τ_5, while for τ_2, there was a noticeable increase in the power of DCT (99.5% rejection rate) and BPP (100% rejection rate). The associations with τ_1, τ_3 and τ_4 were not detected with these procedures. There was 100% power to detect the epistasis between τ_1 and τ_2 and between τ_3 and τ_4 using DCT when the interaction mechanism was specified in the GMM model.

The LRTS was not usable, possibly due to the dependence of values taken from subjects within pedigrees and the non-normality of the distribution of CAC, especially values obtained from the first visit. In an actual genetic analysis, to reduce the chances that skewness of the data would result in an apparent genetic association, one should follow Maclean et al. (1976) and consider multiple transformations of the data.

Procedures to find the most effective transformation should be developed to enhance the applicability of GMM analysis.

The Mplus software was not effective in analyzing these simulated datasets due to computational instability and computer time needs. Computational instability also affected the SAS TRAJ Procedure. One effect of this instability was that using TVCs did not increase the overall power as had been expected. About 10% of the replicates did not have a PROC TRAJ solution in the analyses where three components were specified without TVCs. If TVCs were incorporated along with a 3-component model, the model convergence was even more difficult to achieve. In this case, about 17% of the replicates did not have a PROC TRAJ solution. However, since 100% of the 200 replicates had a solution for the GMM analyses that identified 2 trajectory components with TVCs and without TVCs, there was an apparent increase in the power of DCT and BPP with TVCs.

The empirical null distribution of the DCT and BPP for genes not associated with CAC values appeared to depend on whether the gene was apparently in HWE. The empirical null distribution of the tests was also sensitive to genotyping errors. With presence of a 4% error rate, the 95^{th} empirical percentile could increase by 20% for the ten null SNPs in apparent HWE. For the ten null SNPs not in HWE, there was no systematical change observed in the empirical distribution of these tests when genotyping error was present. In addition, there might be other underlying factors differentiating the SNPs in HWE and the SNPs not in HWE that influenced the empirical null distributions of the tests.

Another approach that might be used to calculate the p-value of the DCT or BPP statistics and to explore the properties of the empirical null distribution is use of permutation procedures (Fisher, 1922; Koehler, 1986; Mielke & Berry, 2001). Specifically, one can generate a large number of random permutations of the *n* vectors (here *n* = 6,476 participants) of CAC values. The fraction of permutations that yield a value of the statistic larger than the one observed is the permutation p-value.

For the markers flanking the actual genes, the rejection rates at a marker near τ_5 could be as large as the τ_5 rejection rate. The rejection rate for DCT was high for markers near τ_5 (42% - 100%), except as expected, for one marker with low LD measures. The BPP had approximately 50% lower rejection rates than the DCT. The rejection rates of DCT and BPP for the markers near τ_5 appeared to be associated with LD. Specifically, the markers with higher rejection rate for both DCT and BPP appeared to be with higher values of disequilibrium coefficient and r^2, with Pearson correlations above 0.5.

With regard to the markers near τ_2, the rejection rates at a marker near τ_2 were somewhat lower than the τ_2 rejection rate. The rejection rate for DCT was 37% - 67% for markers near τ_2, except for one marker with very low LD measures. One marker that was not in LD ($\chi^2 = 0.50$, $p = 0.48$) still had a modestly high rejection rate for both DCT (67%) and BPP (60%). On average, the BPP had essentially the same rejection rates as

DCT. As expected, the rejection rates of DCT and BPP for the markers near τ_2 appeared to be associated with LD. The markers with higher rejection rate of the two tests appeared to be correlated with higher values of Lewontin's *D*' measure (Pearson correlations > 0.9).

The BPP test seems to be as powerful as the DCT test for the identification of the genes directly affecting the CAC. However, when it comes to the detection of markers flanking the actual gene, in particular for τ_5, the BPP appeared to be less powerful than DCT. Overall, analyses that incorporate genotype measurements of the genes into the mixture modeling appeared to have somewhat greater power than GMM analyses that assess genetic association with post hoc tests.

The results of my study showed that the SAS TRAJ procedure was a useful tool in this application. The use of Mplus was computationally demanding, since the specification of model parameters was more complicated in the Mplus growth mixture modeling programs, and a large number of random staring points need to be specified to achieve convergence. DCT and BPP tests using the SAS TRAJ procedure had power to detect CAC longitudinal QTLs. In particular, the LRTS was not usable in this application in that its distribution was far from the expected asymptotic distributions when applied to markers with no genetic relation to the quantitative trait.

In this research, the procedures using growth mixture modeling has been applied on pedigree data. I have treated the participants from 942 pedigrees as independent

observations. The effect of dependence of observations taken from related individuals has not been evaluated. For future work, my fellow graduate students and I seek to expand our work and replicate our analyses on unrelated individuals. We will also consider the use of PC-POPCORN, a new principal-component-based method proposed by McPeek, Zhang, and Abney (2008) to correct for population stratification. Moreover, we will extend our work to the GMM analysis of a binary trait, myocardial infarction (MI) events using a binary logit regression model.

Bibliography

Ahn, K., Haynes, C., Kim, W., St. Fleur, R., Gordon, D., and Finch, S. J. (2006). The effects of SNP genotyping errors on the power of the cochran-armitage linear trend test for case/control association studies. *Annals of Human Genetics*, 71, 249–261.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika,* 52, 317-332.

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C. K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39, 1–20.

Aneshensel, C. S., Botticello, A. L., and Yamamoto-Mitani, N. (2004). When caregiving ends: The course of depressive symptoms after bereavement. *Journal of Health and Social Behavior*, 45 (4), 422-440.

Bauer, D. J. & Curran, P. J. (2003). Distributional assumptions of growth mixture models: Implications for over-extraction of latent trajectory classes. *Psychological Methods,* 8, 338–363.

Bauer, D. J. & Curran, P. J. (2004). The integration of continuous and discrete latent

variable models: Potential problems and promising opportunities. *Psychological Methods*, 9 (1), 3-29.

Broidy, L. M., Nagin, D. S., Tremblay, R. E., Bates, J. E., Brame, B., Dodge, K. A., Fergusson, D., Horwood, J. L., Loeber, R., Laird, R., Lynam, D. R., Moffitt, T. E., and Pettit,. G. S. (2003). Developmental trajectories of childhood disruptive behaviors and adolescent delinquency: A six-site, cross-national study. *Developmental Psychology*, 39 (2), 222-245.

Celeux, G. & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification,* 13, 195-212.

Chassin, L., Pitts, S. C., and Prost, J. (2002). Binge drinking trajectories from adolescence to emerging adulthood in a high-risk sample: Predictors and substance abuse outcomes. *Journal of Consulting and Clinical Psychology*, 70 (1), 67-78.

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.

Czika, W., Yu, X., Clark, V., and Pratt, R. (2005). *SAS/Genetics 9.1.3: user's guide*. SAS Publishing.

Devlin, B. & Risch, N. (1995). A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics*, 29 (2), 311-322.

Dolan, C. V., Jansen, B. R. J., and van der Maas, H. L. J. (2004). Constrained and unconstrained multivariate normal finite mixture modeling of Piagetian data. *Multivariate Behavioral Research,* 39 (1), 69–98.

D'Unger, A. V., Land, K. C., McCall, P. L., and Nagin, D. S. (1998). How many latent classes of delinquent/criminal careers? Results from mixed poisson regression analyses. *The American Journal of Sociology*, 103 (6), 1593-1630.

Ellickson, P. L., Martino, S. C., and Collin, R. L. (2004). Marijuana use from adolescence to young adulthood: Multiple developmental trajectories and their associated outcomes. *Health Psychology*, 23 (3), 299–307.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85 (1), 87-94.

Franklin, I. & Lewontin, R. C. (1970). Is the gene the unit of selection? *Genetics*, 65 (4), 707–734.

George, A. W., Visscher, P. M., and Haley, C. S. (2000). Mapping quantitative trait loci in complex pedigrees: A two-step variance component approach. *Genetics*, 156, 2081-

2092.

Goldgar, D. E. (1990). Multipoint analysis of human quantitative genetic variation. *American Journal of Human Genetics*, 47 (6), 957-967.

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics, 2 (1), 3-19.

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

Hix-Small, H., Duncan, T. E., Duncan, S. C., and Okut, H. (2004). A multivariate associative finite growth mixture modeling approach examining adolescent alcohol and marijuana use. *Journal of Psychopathology and Behavioral Assessment,* 26 (4), 255-270.

Jansen, R. C., & Stam, P. (1994). High resolution mapping of quantitative traits into multiple loci via interval mapping. *Genetics*, 136, 1447-1455.

Jones, B., Nagin, D., and Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. *Sociological Methods & Research*, 29, 374-393.

Kang, S. J., Gordon, D., and Finch, S. J. (2004). What SNP genotyping errors are most

costly for genetic association studies? *Genetic Epidemiology*, 26 (2), 132-141.

Koehler, K. J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association*, 81 (394), 483- 493.

Kraja, A. T., Culverhouse, R., Daw, E. W., Wu, J., Brunt, A. V., Province, M. A., and Borecki, I. B. (2008). *Genetics Analysis Workshop 16 Problem 3: FHS Simulated Data Set – The Answers*.

Kreuter, F. & Muthen, B. (2008). Longitudinal modeling of population heterogeneity: Methodological challenges to the analysis of empirically derived criminal trajectory profiles. In Hancock, G. R., & Samuelsen, K. M. (Eds.), *Advances in latent variable mixture models*, pp. 53-75. Charlotte, NC: Information Age Publishing, Inc.

Lambert D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34 (1), 1-14.

Leal, S. M. (2005). Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg Equilibrium. *Genetic Epidemiology*, 29, 204-214.

Lewontin, R. C. & Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14 (4), 458-472.

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

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.

Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA.

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

Macgregor, S., Knott, S. A., White, I., and Visscher, P. M. (2005). Quantitative trait locus analysis of longitudinal quantitative trait data in complex pedigrees. Genetics, 171 (3), 1365-1376.

Maclean, C. J., Morton, N. E., Elston, R. C., Yee, S. (1976). Skewness in commingled distributions. *Biometrics*, 32 (3), 695-699.

Magidson, J. & Vermunt, J. (2004). Latent Class Models. In D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences*. Newbury Park, CA: Sage Publications.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18 (3), 413-422.

McLachlan, G. J., Do, K., and Ambroise, C. (2004). Analyzing microarray gene expression data. Hoboken, NJ: Wiley.

McPeek, M. S., Zhang, J., and Abney, M. (2008). Association testing with principal-components-based correction for stratification: When and how does it work? *Genetic Epidemiology*, 32 (7), 707-707 (Meeting Abstract: 130).

Mielke, P. W. & Berry, K. J. (2001). *Permutation methods: A distance function approach*. Springer, New York.

Mojtabai, R., Fochtmann, L., Chang, S. W., Kotov, R., Craig, T. J., Bromet, E. J. (2009). Unmet need for care in schizophrenia. *Schizophrenia Bulletin* (in press).

Mote, V. L. & Anderson, R. L. (1965). An investigation of the effect of misclassification on the properties of χ^2-tests in the analysis of categorical data. Source: *Biometrika*, 52, 95-109.

Muthén, L. & Muthén, B. (1998-2007). Mplus user's guide. Fifth Edition. Los Angeles, CA: Muthén & Muthén.

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

Muthén, B. & 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.

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.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In: *Handbook of quantitative methodology for the social sciences*, pp. 345-368. Edited by Kaplan D. Thousand Oaks, CA: Sage.

Nagin, D. & Land, K. C. (1993). Age, Criminal careers, and population heterogeneity – specification and estimation of a nonparametric, mixed Poisson models. *Criminology*, 31 (3), 327-362.

Nagin, D. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods*, 4, 139-177.

Nagin, D. & Tremblay, R E. (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.

Nagin, D. & Tremblay, R E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods*, 6, 18-34.

Nagin, D. & Tremblay, R. E. (2005a). Developmental trajectory components: Fact or a useful statistical fiction? *Criminology*, 43, 873-904.

Nagin, D. & Tremblay, R E. (2005b). What has been learned from group-based trajectory modeling? Examples from physical aggression and other problem behaviors. Annals of the American Academy of Political and Social Science, 602, 82-117.

Nagin, D. S. (2005). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Odgers, C. L., Caspi, A., Broadbent, J. M., Dickson, N., Hancox, R. J., Harrington, H.L. Poulton, R., Sears, M. R., Thomson, W. M., and Moffitt T. E. (2007). Prediction of differential adult health burden by conduct problem subtypes in males. *Archives of general psychiatry*, 64 (4), 476-484.

Pan, W., Lin, J., and Le, C. T. (2002). Model-based cluster analysis of microarray gene-

expression data. *Genome Biology*, 3 (2): Research 0009.1–0009.8.

Rodriguez-Zas, S. L., Southey, B. R., Whitfield, C. W., and Robinson, G. E. (2006). Semiparametric approach to characterize unique gene expression trajectories across time. *BMC Genomics*, 7: 233.

Roeder, K. & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* 92, 894-902.

Romano, E., Tremblay, R. E., Farhat, A., and Cote, S. (2006). Development and prediction of hyperactive symptoms from 2 to 7 years in a population-based sample. *Pediatrics,* 117 (6), 2101-2110.

Rubin, D. B. (1976). Inferences and missing data. *Biometrika*, 63, 581-592.

Schaeffer, C. M., Petras, H., Ialongo, N., Poduska, J., and Kellam, S. (2003). Modeling growth in boys' aggressive behavior across elementary school: Links to later criminal involvement, conduct disorder, and antisocial personality disorder. *Developmental Psychology*, 39 (6), 1020-1035.

Schwartz, G. (1978). Estimating a dimension of a model. *The Annals of Statistics*, 6, 461-464.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika,* 52, 333-343.

Soromenho, G. (1993). Comparing approaches for testing the number of components in a finite mixture model. *Computational Statistics,* 9, 65–78.

Tintle, N. L., Ahn, K., Mendell, N. R., Gordon, D., and Finch, S. J. (2005). Characteristics of replicated single-nucleotide polymorphism genotypes from COGA: Affymetrix and center for inherited disease research. *BMC Genetic*, 6 (Suppl 1): S154.

Tremblay, R. E., Nagin, D. S., Séguin, J. R., Zoccolillo, M., Zelazo, P. D., Boivin, M., Pérusse, D., and Japel, C. (2004). Physical aggression during early childhood: Trajectories and predictors. *Pediatrics*, 114 (1), e43-e50.

Tucker, J. S., Ellickson, P. L., Orlando, M., Martino, S. C., and Klein, D. J. (2005). Substance use trajectories from adolescence to emerging adulthood: A comparison of smoking, binge drinking, and marijuana use. *Journal of Drug Issues*, 35 (2), 307-332.

Weir, B. S. (1979). Inferences about linkage disequilibrium. *Biometrics*, 35 (1), 235-254.

Weir, B. S. (1990). *Genetic data analysis*. Sunderland, Massachusetts' Slnauer.

White, H. R., Bates, M. E., and Buyske, S. (2001). Adolescence-limited versus persistent

delinquency: Extending Moffitt's hypothesis into adulthood. Journal of Abnormal Psychology, 110 (4), 600-609.

White, H. R., Pandina, R. J., and Chen, P. (2002). Developmental trajectories of cigarette use from early adolescence into young adulthood. *Drug and Alcohol Dependence*, 65 (2), 167-178.

Wiesner, M. & Capaldi, D. M. (2003). Relations of childhood and adolescent factors to offending trajectories of young men. *Journal of Research in Crime and Delinquency*, 40 (3), 231-262.

Wu, R., Ma, C., Lin, M., Wang, Z., and Casella, G. (2004). Functional mapping of quantitative trait loci underlying growth trajectories using a transform-both-sides logistic model. *Biometrics*, 60 (3), 729-738.

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

Wu, W., Zhou, Y., Li, W., Mao, D., and Chen, Q. (2002). Mapping of quantitative trait loci based on growth models. *Theoretical and Applied Genetics*, 105 (6-7), 1043-1049.

Xie, H., Drake, R., and McHugo, G. (2006). Are there distinctive trajectory components in substance abuse remission over 10 years? An application of the group-based modeling

approach. *Administration and Policy in Mental Health & Mental Health Services Research*, 33, 423–432.

Yang, R. Q., Tian, Q., and Xu, S. Z. (2006). Mapping quantitative trait loci for longitudinal traits in line crosses. Genetics, 173 (4), 2339-2356.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model based clustering and data transformations for gene expression data. *Bioinformatics*, 17 (10), 977-987.

Zeng, Z. B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, 90 (23), 10972-10976.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics*, 136 (4), 1457-1468.

Zhao, W., Wu, R. L., Ma, C. X., and Casella, G. (2004a). A fast algorithm for functional mapping of complex traits. *Genetics*, 167, 2133-2137.

Zhao, W., Zhu, J., Gallo-Meagher, M., and Wu, R. L. (2004b). A unified statistical model for functional mapping of genotype x environment interactions for ontogenetic development. *Genetics*, 168, 1751-1762.