ABSTRACT

LI, YEN-WEI. Implementation of Association Analysis Methods for Less Constrained Sampling Strategies. (Under the direction of Dr. Yi-Ju Li.)

Association tests have served as an important tool to detect genetic variants associated with qualitative or quantitative traits of interest. Association analysis aims to detect markers that may be in strong or perfect linkage disequilibrium with variants that affect the etiology or susceptibility of the disease, or modify the quantitative trait. In Chapter One, we review the analytical background for the field of gene mapping in human complex disease and the development of association methods. In order to gain sufficient statistical power, efficient utilization of all data available is imperative. This thesis presents our efforts in developing new association methods that can utilize different combinations of family or population datasets, with the goal of relaxing the recruitment criteria. In Chapter Two, we describe our work on the extension of the Monks and Kaplan (MK) method and the development of a novel program that performs both allele- and genotype-based association tests in general pedigree structures. The new genotype-based MK method (EMK method) provides a test statistic for each genotype and a global test for combined genotypes. In Chapter Three, we describe our work on developing a family-based association test for pedigree including half-sib data (PHAST) to fully utilize all possible information in the family data. This method will benefit understudied populations, especially for late-onset diseases that recruitment tends to be difficult. In Chapter Four, we develop a semiparametric additive mixed model (SAMM) for

quantitative traits association study which can cope with family and unrelated data simultaneously and control for population stratification through a smooth function. The extension and future plans are discussed in Chapter Five.

Implementation of Association Methods for Less Constrained Sampling Strategies

by
Yen-Wei Li

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2012

APPROVED BY:

_____	_____
Dr. Zhao-Bang Zeng Chair of Advisory Committee	Dr. Yi-Ju Li
_____	_____
Dr. Leonard Stefanski	Dr. Jung-Ying Tzeng

DEDICATION

This dissertation is dedicated to my parents for their support and love.

BIOGRAPHY

Yen-Wei Li was born on September 14, 1974, in Taiwan. He grew up and spent his earlier school years in Taipei. He received a bachelor degree in Mathematics from National Cheng Kung University (NCKU), Taiwan in June 1997. He enrolled in the Statistics Department in National Central University at the same year and completed his master's degree in 1999. He was employed as a research assistant at the Institute of Statistical Science, Academia Sinica after his two years military service. In August of 2003, he came to the United States to pursue his Ph.D. degree in the Statistics Department of North Carolina State University. He did his internship in Center for Human Genetics of Duke University and focused his PhD research on developing statistic methods for complex disease gene mapping under the direction of Dr. Yi-Ju Li.

ACKNOWLEDGMENTS

I would like to thank my advisor and mentor, Dr. Yi-Ju Li, for her guidance, advice, and enthusiasm throughout this research. Her encouragement and patience help me to overcome all the difficulties during the entire course of this work.

I would like to thank my committee members; co-chair Dr. Zhao-Bang Zeng, Dr. Leonard Stefanski, and Dr. Jung-Ying Tzeng, for their invaluable advice and comments throughout the period of my doctoral studies. Their support and constant inspiration was my energy to achieve my goal in this challenging journey.

I also appreciate all faculties and staffs at CHG for their suggestions and help to support my study. I also want to thank all my classmates and friends for their friendship.

Finally, I would like to express my deeply appreciation to my family. Thanks to Mom and Dad for their love, understanding, and tireless support throughout my life. Thanks to my sisters for their constant encouragement and support. Last but not least, thanks to my little cute rabbit.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Statistical Methods and Study Designs in Human Genetics

1.1. INTRODUCTION

Linkage and association studies are two primary methods for human disease gene mapping. Traditionally, the first step of gene mapping begins with a linkage study. Genetic linkage assumes that certain chromosomal regions were linked in inheritance through pedigree. This implies that some genes can be inherited together due to the lack of recombination events occurred in between. Accordingly, a genome wide linkage study is actually searching chromosomal regions that are linked (i.e. no recombination events) and correlated to the traits of interests. This approach does not require a priori knowledge about the location of potential susceptibility genes. Since the recombination rate is the target parameter to be tested, familial data are required for the linkage study. In particular, multiplex families (i.e. more than two affected siblings per families) are needed for a qualitative trait. Therefore, many of early phase recruitments for linkage studies are targeted for large extended pedigrees or nuclear families with affected sib pairs.

Linkage analysis can be grouped into two categories: parametric and non-parametric linkage analysis. Parametric (model-based) linkage tests require the assumption of a genetic model (e.g., dominant, additive, or recessive model) (Kruglyak, et al., 1996; Ott, 1999; Terwilliger and Ott, 1994), and is often evaluated by the logarithm of odds (LOD) score (Morton, 1955). LOD is the logarithm (base 10) of odds of likelihood with the recombination rate estimated from the observed data versus that with the null hypothesis recombination rate 0.5. A LOD score greater than 3.0 is usually considered as the evidence for linkage (Kruglyak, et al., 1996). The disadvantage of the parametric linkage

is that the analysis can be sensitive to misspecification of the genetic model (Clerget-Darpoux, et al., 1986) and may not have adequate power for complex diseases (Weeks and Lathrop, 1995).

When the knowledge of genetic models is opaque, non-parametric (model-free) linkage methods can do a better job. Non-parametric (parameter-free) linkage analysis compares identity-by descent (IBD) between affected relatives (usually affected sib-pairs) to see whether they are significantly different from Mendelian expectations (Kong and Cox, 1997; Kruglyak, et al., 1996; Sengul, et al., 2001; Whittemore and Halpern, 1994). If there is no linkage, the chances of an affected sib pair sharing 0, 1, and 2 alleles IBD are 0.25, 0.5, and 0.25, respectively.

Linkage can be accomplished by two-point analysis, considering one marker and the disease locus, or multipoint analysis, considering multiple markers with their relationship to a putative disease locus position (Kruglyak, et al., 1996). Although computationally intensive, multipoint analysis is more informative (Halpern and Whittemore, 2000). Multipoint and two-point analyses for either parametric or nonparametric approaches have been implemented in several software packages, such as, GENEHUNTER (Kruglyak, et al., 1996), VITESSE (O'Connell and Weeks, 1995; O'Connell, 2001), MENDLE (Lange, et al., 2001), SOLAR (Almasy and Blangero, 1998), and MERLIN (Abecasis, et al., 2002).

Linkage is a powerful tool for rare variants. By tradition, linkage analysis was the primary method for mapping simple Mendelian diseases such as Huntington's disease and cystic fibrosis. For human complex diseases, however, linkage analysis may not be

the most appropriate and promising method. Moreover, linkage analysis has the disadvantage of low mapping resolution (which can narrow down the search for causal variants to several megabases (Mb) in general) and could only determine roughly where the putative susceptibility gene is (Roberts, et al., 1999).

1.1.1. Characteristics of Complex Disease Gene Mapping

Complex diseases are common disorders. Several late-onset-diseases and congenital defects, including Alzheimer disease, Parkinson disease, scleroderma, asthma, diabetes, and many more, fall into this category (Hunter, 2005). A common disease-common variant (CD/CV) hypothesis (Collins, et al., 1997; Lander, 1996) assumes that "the genetic risk for common diseases will often be due to disease-producing alleles found at relatively high frequencies (>1%)". These disorders are popular because the influencing genes for these diseases are common. This CD/CV hypothesis also suggests that the effect of each disease gene is generally small or moderate (Becker, 2004). These modest genetic effects make it difficult to find the causative genes that underlie complex diseases.

The search for common variants affecting the complex diseases usually uses association studies. Rather than directly assessing allelic association between marker and disease alleles, the tactic is looking for associations between marker alleles and disease phenotype. However, linkage analysis still plays some roles for complex disease gene mapping. Especially nowadays, as the genome-wide association study (GWAS) produces

massive results, linkage evidences can help to prioritize them. Furthermore, some studies suffer insufficient sample sizes for a GWAS study. A linkage study can complement small scale association studies (Ott, et al., 2011; Wang, et al., 2005).

1.1.2. Genetic Association Studies

Historically, the main usage of association studies is to examine functional candidate genes (biologically associated to the trait of interest) or genes located in the linkage regions detected by the genome wide linkage scan studies. With recent advances in cost-effective high-throughput single nucleotide polymorphism (SNP) genotyping technologies, the field of association studies has progressed to fine-mapping of linkage regions and then GWAS using high density SNPs.

Association analysis aims to detect if there is an association between a trait and occurrence of a certain allele at some markers. The principle of the association study is looking for a significantly higher or lower frequency of a marker allele with a trait than that would happen randomly if there is no association. If association is present, a particular allele, genotype or haplotype of a polymorphism or polymorphism(s) will be seen more often than expected by chance in individuals carrying the trait.

The validity of association between genotype and phenotype depends on genotyping the direct biological functional polymorphisms (direct association) (e.g., the APOE-4 allele in Alzheimer's disease (AD)), or nearby genetic markers that are in linkage disequilibrium (LD) with functional polymorphism (indirect association). If the

association arises because of some underlying stratification or admixture of the population (confounded association) (Cordell and Clayton, 2005), this is called spurious association which should be avoided by employing appropriate study design and analysis methods.

Allelic association, also called linkage disequilibrium mapping, refers to the dependence and correlation between marker allele and a disease trait. Such studies utilize panels of SNPs (markers) and rely on the concept of linkage disequilibrium.

1.2. GENETIC CONCEPTS AND STUDY DESIGN

1.2.1. Linkage Disequilibrium

In population genetics, alleles at two loci in random association are said to be in a state of linkage equilibrium (LE). In contrast, linkage disequilibrium (LD) is the non-random association between alleles at two loci. In other words, LD reflects the phenomena that some combinations of alleles between two loci (i.e. haplotype) happen more or less frequently in a population than expected under random association.

There are several different measures of linkage disequilibrium. The basic measure is disequilibrium coefficient, *D*, which measures the deviation of the haplotype frequency from the expected value (the product of the marginal allele frequencies) (Lewontin and Kojima, 1960). The *D* is formulated as $D = P_{AB} - P_A P_B$, where assuming at loci A and B,

P_{AB} is the frequency of haplotype carrying the A and B alleles, and P_A and P_B are the marginal allele frequencies of allele A and B, respectively. If there is no allelic association, the expectation of *D* would be zero. The maximum value of disequilibrium coefficient depends on allele frequencies at the two loci. This undesirable property makes *D* hard to interpret because comparison between pairs of alleles with different allele frequencies is difficult.

Therefore, several alternative measures have been proposed that are less sensitive to marginal allele frequencies and standardized to lie between 0 and 1 (or -1 and 1). Two commonly used standardized measures are the Lewontin's D' and correlation coefficient (r^2). Lewontin's $D'(= D/D_{max})$ is simply the disequilibrium coefficient standardized by its maximum value D_{max}, where $D_{max} = \min\left(P_A P_B, (1-P_A)(1-P_B)\right)\ if\ D < 0$,and $\min(P_A(1-P_B), (1-P_A)P_B)\ if\ D > 0$ (Lewontin, 1964). If no allelic association, the expectations of Lewontin's D' (ranges from -1 to 1) would be zero. The case of D'equaling to 1 (complete LD) should imply that there is evidence for no recombination between the markers in the population. However, this conclusion is not always correct because D' is sensitive to very rare alleles and could be problematic if they exist (Ardlie, et al., 2002). Hill and Robertson (1968) proposed the correlation coefficient r^2 as follows:

$r^2 = D^2/P_A(1-P_A)P_B(1-P_B)$. The value of r^2 (ranges from 0 to 1) is a direct

measure of the correlation between two loci, and is more useful in most situations. r^2 equals 1 (perfect LD) only when the genotype at one locus perfectly predicts the genotypes at a second locus and equals 0 when they are not in perfect equilibrium (Lawrence, et al., 2005). These linkage disequilibrium measures can be used to measure allelic associations between loci, no matter linked or not. They are not measures of linkage.

Several mechanisms can cause LD, including genetic linkage, population substructure, recent admixture, founder effects, along with a number of other factors such as mutation, selection, random genetic drift, non-random mating, etc. After an original event like mutation creates a LD, this association will begin to decay over generations. The rate of decay largely depends on the recombination rate. Usually, the decay of association will be slower if the two loci are more tightly linked. In general population, LD is rarely found more than 1 Mb and often exists in much smaller distances. However, LD at unlinked loci is possible by random, even though it often declines rapidly in the evolutionary process. Therefore, in human disease gene mapping, there is an emphasis on association methods to test both association (LD between disease locus and marker) and linkage (physical association on the same chromosome) to the phenotype of interest (Spielman and Ewens, 1996).

1.2.2. Study Design

There are two fundamental designs applied to association studies: family-based design and population-based design. Family-based association study utilizes familial relationship to characterize the allelic association to the trait of interest. Population-based design uses related samples to make inference. With their respective advantages and limitations, both methods provide helpful tools for identifying genetic variants in complex disorders. More details and difference between these two study designs are presented in the following sections.

Family-Based Design

Family-based association studies use families to detect association between genetic markers and qualitative or quantitative traits. Early development of family-based association methods took family triads (a proband and both parents) as the sampling unit. Later on, many different nuclear families and extended pedigrees were also used for testing association. Due to the diversity of family structures, various statistical methods for family-based association have been developed (See Section 1.3.1).

Family-based association study was favorable for multiple reasons. First, unlike population-based designs, family-based methods are robust against the confounding

effects due to population admixture and stratification, environment interaction, or selection bias (Ewens and Spielman, 1995). Second, significant association can only be detected in the presence of linkage when applying family-based association tests (Ott, 1989; Whittaker and Morris, 2001). It means that significant results will infer both linkage and association at the same time. This property matches the intuitiveness that truly associated markers should be physically near the causal genetic variant. Third, family-based designs would be helpful for parent-of-origin studies (diagnosing imprinting effects) (Weinberg, 1999a). It is almost impossible for population-based designs to deal with such problem. Fourth, family-based designs could largely benefit model building, multiple-hypothesis testing (Van Steen, et al., 2005). In addition, many family data are already available for previous linkage studies. It is naturally following up the family-based association design in such situations for fine-mapping.

The main drawbacks for family-based designs are mostly the practical issue of recruitment and cost (Laird and Lange, 2006). Family-based association studies usually require more time and funding for recruitment of probands and their relatives. In addition, for the same sample size, family data is generally lower in statistical power than population data.

Population-based Design

Population samples are in general easier to recruit. Since many human disease gene mapping studies focus on binary phenotype such as disease status, the terminology of case-control study is often used to refer to population-based design. However, the population-based design also includes studies for quantitative phenotype (continuous variables). In this case, there is no “case-control” per se in the samples. Population association studies utilize unrelated individuals to identify risk alleles by studying patterns of polymorphisms that vary systematically between case and control individuals or increases the variation of quantitative phenotype. Because the genome is very enormous, patterns considered to be a causal polymorphism could happen only by chance. In order to distinguish causal from spurious signals, adequate statistical methods and high standard statistical significance are required (Balding, 2006).

Case-control study designs are a classical epidemiological tool and now widely used to study the genetic susceptibility in rare complex diseases. Basically, a case-control study involves recruiting a number of the diseased subjects (cases), and then sampling a comparable number of unaffected subjects (controls). To eliminate confounding effects, case-control studies usually match case and control groups to have equal number of samples and to keep homogenous on many other nuisance factors such as race, gender, and age (Greenland and Rothman, 1998). In a nutshell, the genetic case–control test

compares the observed allele or genotype frequencies in the cases and controls to the expected values under the null hypothesis of no association where the frequencies are the same in cases and controls (Risch, 2000). A significant difference in the frequency of an allele or genotype between cases and controls reflects the potential of the increase or decrease of the disease risk derived from the specific marker. The limitation of the case-control study is that the presence of confounding effects in the samples could cause association between unlinked loci (false positive) in the analysis (Devlin, et al., 2001; Ewens and Spielman, 1995; Risch, 2000). Confounding is caused by the population stratification (i.e., the existence of multiple population subgroups in a population) as well as other factors which are associated with affection risk or marker–allele frequency (Hennekens, et al., 1987). Therefore, it is essential to ensure that case and control populations have identical ethnicity, because genotype and haplotype frequencies may vary between ethnic or geographic populations.

Genome Wide Association

Candidate gene approaches and whole genome screens are two general methods to explore the molecular genetics of complex diseases. In contrast to candidate gene approaches which specifically examine functional related genes or genes located with a region of interest (i.e. linkage region), the whole genome screens investigate the entire

genome by sampling 500,000 or more SNPs from each subject. A genome-wide association study (GWAS) aims to unbiased identify associations between SNPs and a trait. Because of the advance of genotyping technology, GWAS has offered an unbiased way to search SNPs that are associated with the trait of interest.

In 2007, the Wellcome Trust Case-Control Consortium (WTCCC) implemented GWAS with 2,000 individuals for each of seven complex human diseases (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes, and type 2 diabetes) and 3000 shared controls. They had successfully identified 24 independent association signals at above diseases (Burton, et al., 2007). To June 2011, there have been 1,449 published GWAS getting significance results (p-value $\leq 5 \times 10^{-8}$) for 237 traits (http://www.genome.gov). In spite of many advantages, there are also some limitations in GWAS. For example, Maher (2008) indicated that the genes detected by GWAS were unable to explain most of the heritability of many common diseases. For some rare risk variants found by linkage, present GWAS has difficulty to detect such variants through LD with common SNPs. This has made some recurrent interest in using linkage methods to detect rare variants.

GWAS creates great demand for developing new approaches to identify genes that underlie phenotypic variation. While the data preprocessing and computation are more extensive in GWAS, the basic statistical concepts and analyzing methods described

previously are generally relevant to both candidate gene studies and GWAS. However, in order to get sufficient power in these large-genome scans, large sample size is required no matter whether population- or family-based design is used (Hirschhorn and Daly, 2005).

Most published GWAS nowadays used population-based designs. This is reasonable and inevitable because large samples for case–control studies are generally much easier and economical to obtain than large numbers of suitable families. If no population stratification, the case-control studies also have more power than the family-based design (Risch and Teng, 1998; Veal, et al., 2002). However, family-based designs still possess many irreplaceable functions which make these family studies competitive.

In theory, a combination of linkage and association methods could offer the most powerful tool to identify the disease susceptibility variants (Ott, et al., 2011). Family-based designs can offer a nature framework to this combined approach. Moreover, many familial samples were collected during the era of whole genome linkage scans, and these existing large collections of linkage data can be recycled and pave the way for the family-based GWAS. These family datasets provide good resources for association studies as either an initial investigation or a replication dataset for other association findings. In addition, family-based designs can offer better quality control. Family data are more robust not only for population stratification but also for detecting genotyping errors such as Mendelian errors than population-based samples. Family data have higher

accuracy in inferring haplotypes and confirming the role of rare variants to disease susceptibility. In some cases, family-based design can carry out genetic analyses that cannot be achieved using population-based design, such as testing imprinted genes (Kong, et al., 2009; Paterson, et al., 1999). Finally, as the next-generation sequencing is becoming an important approach to pinpoint the causal variant, familial samples have the advantage for the first pass of sequencing effort to discover both common and rare variants. Overall, familial-based designs will still remain important role in GWAS era.

1.3. ASSOCIATION METHODS

1.3.1. Family-Based Association Analysis

Qualitative Trait Association

TRANSMISSION/DISEQUILIBRIUM TEST

For qualitative (binary) traits, the transmission/disequilibrium test (TDT) (Spielman, et al., 1993) is a pioneer method for family-based association studies. TDT was developed to handle parents-offspring trio data, and is a non-parametric method which does not require a specification disease model or the distribution assumption of the disease. Its validity only depends on the assumption of Mendelian transmissions. The principle of TDT is that, under the null hypothesis of no linkage or no association, the allele

transmitted to the affected offspring will be only determined by Mendel's laws. TDT treats the un-transmitted allele as a control to the transmitted allele and compares the differences between transmitted and un-transmitted alleles from parents to affected offspring. Therefore, only trios with at least one heterozygous parent are informative for the test. The statistical significance is determined by a Chi-square test. A significant result implies that the marker tested and a susceptibility locus for a trait are linked and associated.

At first, TDT was used to test for linkage in the presence of association. In order to reject the null-hypothesis, however, both linkage and association have to be present. Therefore, TDT is also a valid test for association in the presence of linkage (Ewens and Spielman, 2005) and typically used as a test for association now. This dual-alternative hypothesis makes TDT have the advantage to avoid false positive results due to association without linkage, as might happen if there is population stratification or admixture. Thanks to these benign properties, TDT opened the door for family-based association study. However, TDT using only parents-offspring trio data is too restricted in practical situations. Many extensions of TDT have thus been developed in order to address other family structures, such as missing parental information, general pedigrees, multiple affected siblings, and so on.

Missing parental genotypes for the affected siblings are common situations in

complex disease studies particularly the late-onset disorders. A number of methods have been developed to extend the original TDT to handle this problem. The sib TDT (S-TDT) (Spielman and Ewens, 1998) compared the difference of allele frequencies between affected and unaffected siblings using discordant sib-pairs (unrelated nuclear families with only one affected and one unaffected siblings). The SDT test (Horvath and Laird, 1998) generalized S-TDT to multiple affected siblings. Knapp (1999) proposed "reconstruction combined TDT" (RC-TDT), which combined with TDT and S-TDT by reconstructing missing parental genotypes based on siblings' genotypes.

Since various family types may be recruited for a single study, ideally one would like to have a testing method which can handle general pedigrees rather than only parents-offspring trio or only discordant sibpairs. The pedigree disequilibrium test (PDT) (Martin, et al., 2000) combines TDT and S-TDT into a test for general pedigree. PDT treats general pedigrees as independent units and partitions a pedigree into several related subunits of case–parent trios and discordant sib-pairs. The transmission score of each pedigree is a sum of the transmission scores of all related subunits. A score statistic T (asymptotically standard normal distribution) is therefore formed. Although PDT is a valid test of linkage or association for more general pedigree structures, some subtypes of pedigrees are still uninformative, for instance, the affected sib-pairs without parents.

For nuclear families with multiple affected siblings, previously developed tests (e.g.

TDT, S-TDT, and RC-TDT) that treat multiple offspring as independent are not valid if linkage is present. The test for association in the presence of linkage (APL) (Martin, et al., 2003) adjusted this problem by incorporating IBD relationships for affected siblings. The APL can also infer missing parental genotypes in the linkage region by taking the IBD information into consideration.

RRGRESSION AND LIKELIHOOD METHODS

Different from TDT-based methods, methods based on regression or likelihood functions were proposed for family-based association studies. Regression methods possess several attractive properties. For example, covariates like environmental factors or additional marker variables can be easily fitted in the model to test for gene–environment (Schaid, 1999a) or gene–gene interactions. Allelic or genotypic relative risks can also be estimated by regression parameters.

Logistic regression has been used to test for linkage disequilibrium (Schaid, 1996b; Waldman, et al., 1999) in family triads or discordant sibpairs. Log-linear models are also used to construct association tests in family triads (Weinberg, et al., 1998; Weinberg, 1999a) and allow for testing parent-of-origin effects.

The likelihood framework, either likelihood ratio or score tests, was also extended for family-based association studies, especially when there is missing parental data.

When parental data are missing, a likelihood-ratio test is implemented by using the expectation-maximization (EM) algorithm in incompletely genotyped triads (Weinberg, 1999b). Clayton (1999) proposed a likelihood ratio test for incomplete data in the computer software TRANSMIT. The family-based association test (FBAT) (Horvath, et al., 2001; Laird, et al., 2000; Rabinowitz and Laird, 2000) is a score test which treats the offspring genotype as a random variable, conditioning on observed traits and parental genotypes. If parental genotypes are missing, the FBAT is conditioned on sufficient statistics. These strategies keep the FBAT away from making assumptions about the parental allele frequencies, the trait distribution, and the marker allele frequencies. Therefore, it is robust to population stratification and is applicable to many pedigree structures. Several recent researches have been successfully applied by the FABT approach (Smit, et al., 2008).

Quantitative trait association

Although many clinical phenotypes are dichotomized, directly using quantitative measures if available is always encouraged for the sake of statistical power. There are several recent findings which are based on family-based quantitative association studies, such as the association between the nerve growth factor gene and autism spectrum disorder (ASD) (Lu, et al., 2011), the association between ABLIM1 gene and alcohol

dependence (Wang, et al., 2011), and the association between the CHRM2 gene and intelligence (Gosso, et al., 2006). Clearly, the development of family-based association methods for quantitative traits was equally important. For unrelated individuals, if there is population stratification, association tests for quantitative traits have the same potential biases as case–control studies. Family-based tests can be applied to avoid such problems and offered an additional test of linkage. There are several tests proposed for family-based quantitative traits association tests. They can be generally categorized as two types.

The first type is the model-free TDT extension methods. The key of these tests is to examine whether there is a significant difference in a quantitative trait between children who receive and who do not receive the marker allele from a heterozygous parent. For family triads, Allison (1997) and Rabinowitz (1997) first introduced the extension of TDT method for quantitative traits. Xiong et al. (1998) developed a more flexible TDT-like method for quantitative traits if nuclear families with multiple siblings are available. When there is missing parental data, Allison et al. (1999) modified his earlier work to incorporate the absence of parental genotypes. Monks and Kaplan (2000) further extended association tests for quantitative traits to general pedigree structure with or without parental data.

The second type is model-based likelihood/regression framework methods. These

model-based approaches for quantitative trait association tests require the additional assumption of normality for the phenotypic distribution. Fulker et al. (1999) proposed a likelihood-based association test for quantitative traits using sib-pairs data. They partitioned association effects into two orthogonal components, the between-family and the within-family association. Abecasis et al. (2000) further extended their approach for nuclear families with multiple siblings (implemented in the QTDT software: http://www.sph.umich.edu/csg/abecasis/QTDT/). The within-family association parameter is robust to population stratification, and if there is a difference between the between- and within-family association parameters, this can be taken as a sign for stratification. This framework allows not only simultaneous testing linkage and association, but also estimating the additive genetic effects of the marker alleles (Cardon and Abecasis, 2000). The QTDT software is a widely used software that performs association tests for quantitative traits and can implement five association methods, including orthogonal model (OM) (Abecasis, et al., 2000), Monks and Kaplan (MK) Model (Monks and Kaplan, 2000), Fulker Model (Fulker, et al., 1999), Allison's Linear Model (TDTQ5) (Allison, 1997), and Rabinowitz Model (Rabinowitz, 1997).

Gauderman (2003) proposed a regression model which allows for detecting gene-gene and gene-environment interactions. A potential problem with these regression models is that the normality assumption may be violated if the underlying trait

distribution is skewed or non-normal. FBAT (Horvath, et al., 2004; Lange, et al., 2001), or its relatively new software tool PBAT, can also be used for quantitative trait association test. In contrast to QTDT, the FBAT approach is a nonparametric method so that no distributional assumptions for the trait are required.

1.3.2. Population-Based Association Analysis

For qualitative trait, the natural choice of a testing method for SNP genotypes is creating a contingency table and tests for association by the Pearson Chi-squared statistic or Fisher's exact test. The genotypic Pearson and Fisher tests have generally reasonable power over most risk models (recessive, dominant, or overdominant) except that the genotype risks are additive. For human complex disease, it is widely accepted that individual SNPs would contribute roughly additively to disease risk. One way to improve power to detect additive risks is adopting the Cochran–Armitage test (Armitage test) (Armitage, 1955). The Armitage test is more conservative and would sacrifice some power if the genotypic risks are far from additive.

Regression-based analyses such as logistic regression can also be used in the case-control test (Agresti, 2002). Logistic regression offers a more flexible test that can accommodate other covariates. Although their models seem very different in appearance,

there are some connections between the logistic regression models and the Pearson or Armitage test described previously. Both Pearson and Armitage tests are special cases of score tests that correspond to the logistic regression models with no other covariates.

For quantitative (or continuous) traits, linear regression and analysis of variance (ANOVA) are two legitimate choices for testing association. In either test, it assumes the normality of the trait for each genotype. If normality does not hold, either a transformation of the original trait to approximate normality or other regression methods (for example, nonlinear model) could be adopted for testing.

1.3.3. Population Stratification

Without proper processing, population structure can make false positive or spurious association results in population-base analysis. Several statistical analytic strategies are adopted to control such bias if such confounding effect exists.

Genomic Control (GC) (Devlin and Roeder, 1999; Reich and Goldstein, 2001) uses a set of random unlinked SNPs to estimate the inflation in the Armitage test statistic, and then applying this inflation factor to adjust the test statistic. Devlin et al. (2004) proposed a less conservative version GCF procedure. The idea of GC methods is that under population stratification the empirical distribution of Armitage statistics would be inflated

and this inflation factor can be estimated using a set of markers unlinked to disease (null SNPs). GC usually performs well when properly applied, but can be conservative in some extreme settings (Setakis, et al., 2006). Moreover, little knowledge is available on the suitable number of SNPs required for testing, and it can be anti-conservative if insufficient null SNPs are used (Marchini, et al., 2004).

Structured association methods (Hoggart, et al., 2003; Pritchard, et al., 2000; Satten, et al., 2001) are based on the idea of allocating an individual's genome to hypothetical subpopulations, and to test for association conditional on this subpopulation allocation. The STRUCTURE program (Falush, et al., 2003; Pritchard and Rosenberg, 1999; Pritchard, et al., 2000) was developed for subpopulation allocation. These approaches are more complex than genomic control and computationally demanding, because the number of subpopulations is difficult to select. The allocation of the subpopulation is a theoretical construct and the real number of subpopulations cannot be fully answered. Evanno et al. (2005) provided a criterion to determining the number subpopulations.

Patterson et al. (2006) developed an approach using principal components analysis (PCA) to detect population structures and Price et al. (2006) proposed the Eigenstrat method based on PCA. The Eigenstrat method computes principal components for SNPs across the genome. A small number of significant principal components will be calculated and capture the main axes of genetic variation. These significant principal components

would be put in a regression model as covariates to correct for population stratification. PCA provides an effective way to diagnose population structure when large-scale genotype data of a sample of individuals are available. In the past, large numbers of null SNPs have been considered as extra genotyping burden, but it becomes more feasible now in whole genome era.

There are several regression based methods also used to address population structure. Setakis et al. (2006) used logistic regression model for association studies which protects against population structure by treating null SNPs as covariates. Yu et al. (2006) adopted a mixed-model approach that accounts for subpopulation effect as a fixed effect. These approaches can be easily extended to control other confounding such as environmental factors.

1.4.SUMMARY

In this chapter, I have provided as much review as possible on the background of gene mapping studies for human complex diseases and the available, if not all, association tests. Association tests have served as a promising tool to detect genetic variants associated with the qualitative or quantitative traits of interest. In order to gain sufficient statistical power, the requirement of increasing sample size for many genetic association

studies nowadays, such as GWAS, is a costly and time-consuming tendency for recruiting samples. Therefore, efficient utilization of all data available is imperative. This thesis consists with three independent projects, and presents our efforts in developing new association methods that can utilize different combinations of family or population datasets, with the goal of relaxing the recruitment criteria.

Most association methods for quantitative traits are all allelic-based tests and applicable to nuclear families. Since genotype of a marker is a direct observation for an individual, it is desirable to assess association at the genotypic level. Furthermore, some methods are restricted to nuclear families only. For instance, the method proposed by Monks and Kaplan (2000) (MK method) uses only partial information from the general pedigree, which may result in the loss of statistical power. In Chapter Two, we describe our work on the extension of the MK method and the development of a novel program that could perform both allele- and genotype-based association tests in any pedigree structures. The new genotype-based MK method (EMK method) provides test statistic for each genotype and a global test for combined genotypes.

The current family-based association tests are mostly, if not all, based on pedigree structures with parent-offspring triads, parents and multiple affected full siblings, discordant full sibpairs of large size, or extended pedigrees including related nuclear families and sibships. Therefore, full siblings or parents have been the target of

ascertainment in genetic research. In Chapter Three, we describe our work on developing a family-based association test for pedigree including half-sib data (PHAST) to fully utilize all possible information in the family data. This method will benefit understudied populations, especially for late-onset diseases that recruitment is mostly difficult.

Both population-based and family-based designs have their pros and cons. Whether there is an advantage of combining both family and population data in one single test is a question of interest. In current practice, one would probably perform meta-analysis on the p-values obtained from various datasets. However, this runs a risk of under-power in individual dataset if sample sizes are limited to start with. In Chapter Four, we develop a semiparametric additive mixed model (SAMM) for quantitative traits association study which can cope with family and unrelated data simultaneously and control for population stratification through a smooth function. Finally, in Chapter Five, we discuss future directions derived from the current projects.

References

Abecasis, G., Cardon, L. and Cookson, W. (2000) A general test of association for quantitative traits in nuclear families, *The American Journal of Human Genetics*, **66**, 279-292.

Abecasis, G.R., *et al.* (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees, *Nature Genetics*, **30**, 97-101.

Agresti, A. (2002) *Categorical Data Analysis*. Wiley New York.

Allison, D.B. (1997) Transmission-disequilibrium tests for quantitative traits, *American Journal of Human Genetics*, **60**, 676.

Allison, D.B., *et al.* (1999) Sibling-based tests of linkage and association for quantitative traits, *The American Journal of Human Genetics*, **64**, 1754-1764.

Almasy, L. and Blangero, J. (1998) Multipoint quantitative-trait linkage analysis in general pedigrees, *The American Journal of Human Genetics*, **62**, 1198-1211.

Ardlie, K.G., Kruglyak, L. and Seielstad, M. (2002) Patterns of linkage disequilibrium in the human genome, *Nature Reviews Genetics*, **3**, 299-309.

Armitage, P. (1955) Tests for linear trends in proportions and frequencies, *Biometrics*, **11**, 375-386.

Balding, D.J. (2006) A tutorial on statistical methods for population association studies, *Nature Reviews Genetics*, **7**, 781-791.

Becker, K.G. (2004) The common variants/multiple disease hypothesis of common complex genetic disorders, *Medical hypotheses*, **62**, 309-317.

Burton, P.R., *et al.* (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature*, **447**, 661-678.

Cardon, L.R. and Abecasis, G.R. (2000) Regression models for association studies of quantitative trait loci in humans, *GeneScreen*, **1**, 55-57.

Clayton, D. (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission, *The American Journal of Human Genetics*, **65**, 1170-1177.

Clerget-Darpoux, F., Bonaïti-Pellié, C. and Hochez, J. (1986) Effects of misspecifying genetic parameters in lod score analysis, *Biometrics*, 393-399.

Collins, F.S., Guyer, M.S. and Chakravarti, A. (1997) Variations on a theme: cataloging human DNA sequence variation, *Science*, **278**, 1580.

Cordell, H.J. and Clayton, D.G. (2005) Genetic association studies, *The Lancet*, **366**, 1121-1131.

Devlin, B., Bacanu, S.A. and Roeder, K. (2004) Genomic control to the extreme, *Nature Genetics*, **36**, 1129-1130.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies, *Biometrics*, **55**, 997-1004.

Devlin, B., Roeder, K. and Bacanu, S.A. (2001) Unbiased methods for population-based association studies, *Genetic Epidemiology*, **21**, 273-284.

Evanno, G., Regnaut, S. and Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, *Molecular ecology*, **14**, 2611-2620.

Ewens, W.J. and Spielman, R.S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture, *American Journal of Human Genetics*, **57**, 455.

Ewens, W.J. and Spielman, R.S. (2005) What is the significance of a significant TDT?, *Human heredity*, **60**, 206-210.

Falush, D., Stephens, M. and Pritchard, J.K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, *Genetics*, **164**, 1567.

Fulker, D., *et al.* (1999) Combined linkage and association sib-pair analysis for quantitative traits, *The American Journal of Human Genetics*, **64**, 259-267.

Gauderman, W.J. (2003) Candidate gene association analysis for a quantitative trait, using parent-offspring trios, *Genetic Epidemiology*, **25**, 327-338.

Gosso, M., *et al.* (2006) Association between the CHRM2 gene and intelligence in a sample of 304 Dutch families, *Genes, Brain and Behavior*, **5**, 577-584.

Greenland, S. and Rothman, K.J. (1998) *Modern epidemiology*. Lippincott-Raven.

Halpern, J. and Whittemore, A.S. (2000) Multipoint linkage analysis, *Human heredity*, **49**, 194-196.

Hennekens, C.H., Buring, J.E. and Mayrent, S.L. (1987) *Epidemiology in medicine*. Lippincott Williams & Wilkins.

Hill, W. and Robertson, A. (1968) Linkage disequilibrium in finite populations, *TAG Theoretical and Applied Genetics*, **38**, 226-231.
Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits, *Nature Reviews Genetics*, **6**, 95-108.

Hoggart, C.J., *et al.* (2003) Control of confounding of genetic associations in stratified populations, *The American Journal of Human Genetics*, **72**, 1492-1504.

Horvath, S. and Laird, N.M. (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data, *The American Journal of Human Genetics*, **63**, 1886-1897.

Horvath, S., Xu, X. and Laird, N.M. (2001) The family based association test method: strategies for studying general genotype--phenotype associations, *European journal of human genetics: EJHG*, **9**, 301.

Horvath, S., *et al.* (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics, *Genetic Epidemiology*, **26**, 61-69.

Hunter, D.J. (2005) Gene–environment interactions in human diseases, *Nature Reviews Genetics*, **6**, 287-298.

Knapp, M. (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test, *The American Journal of Human Genetics*, **64**, 861-870.

Kong, A. and Cox, N.J. (1997) Allele-sharing models: LOD scores and accurate linkage tests, *The American Journal of Human Genetics*, **61**, 1179-1188.
Kong, A., *et al.* (2009) Parental origin of sequence variants associated with complex diseases, *Nature*, **462**, 868-874.

Kruglyak, L., *et al.* (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach, *American Journal of Human Genetics*, **58**, 1347.

Laird, N.M., Horvath, S. and Xu, X. (2000) Implementing a unified approach to family-based tests of association, *Genetic Epidemiology*, **19**, S36-S42.

Laird, N.M. and Lange, C. (2006) Family-based designs in the age of large-scale gene-association studies, *Nature Reviews Genetics*, **7**, 385-394.

Lander, E.S. (1996) The new genomics: global views of biology, *Science*, **274**, 536.

Lange, K., *et al.* (2001) Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets, *Am J Hum Genet*, **69**, A1886.

Lawrence, R.W., Evans, D.M. and Cardon, L.R. (2005) Prospects and pitfalls in whole genome association studies, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1589-1595.

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

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

Lu, A.T.H., *et al.* (2011) QTL replication and targeted association highlight the nerve growth factor gene for nonverbal communication deficits in autism spectrum disorders, *Molecular Psychiatry*.

Maher, B. (2008) Personal genomes: The case of the missing heritability, *Nature*, **456**, 18.
Marchini, J., *et al.* (2004) The effects of human population structure on large genetic association studies, *Nature Genetics*, **36**, 512-517.

Martin, E.R., *et al.* (2003) Accounting for linkage in family-based tests of association with missing parental genotypes, *The American Journal of Human Genetics*, **73**, 1016-1026.

Martin, E.R., *et al.* (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test, *The American Journal of Human Genetics*, **67**, 146-154.

Monks, S. and Kaplan, N. (2000) Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus, *The American Journal of Human Genetics*, **66**, 576-592.

Morton, N.E. (1955) Sequential tests for the detection of linkage, *American Journal of Human Genetics*, **7**, 277.

O'Connell, J.R. and Weeks, D.E. (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set–recoding and fuzzy inheritance, *Nature Genetics*, **11**, 402-408.

O'Connell, J.R. (2001) Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm, *Human heredity*, **51**, 226-240.

Ott, J. (1989) Statistical properties of the haplotype relative risk, *Genetic Epidemiology*, **6**, 127-130.

Ott, J. (1999) *Analysis of human genetic linkage*. Johns Hopkins Univ Pr.

Ott, J., Kamatani, Y. and Lathrop, M. (2011) Family-based designs for genome-wide association studies, *Nature Reviews Genetics*, **12**, 465-474.

Paterson, A.D., Naimark, D.M.J. and Petronis, A. (1999) The analysis of parental origin of alleles may detect susceptibility loci for complex disorders, *Human heredity*, **49**, 197-204.

Patterson, N., Price, A.L. and Reich, D. (2006) Population structure and eigenanalysis, *PLoS genetics*, **2**, e190.

Price, A.L., *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics*, **38**, 904-909.

Pritchard, J.K. and Rosenberg, N.A. (1999) Use of unlinked genetic markers to detect population stratification in association studies, *The American Journal of Human Genetics*, **65**, 220-228.

Pritchard, J.K., *et al.* (2000) Association mapping in structured populations, *The American Journal of Human Genetics*, **67**, 170-181.

Rabinowitz, D. (1997) A transmission disequilibrium test for quantitative trait loci, *Human heredity*, **47**, 342-350.

Rabinowitz, D. and Laird, N. (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information, *Human heredity*, **50**, 211-223.

Reich, D.E. and Goldstein, D.B. (2001) Detecting association in a case-control study while correcting for population stratification, *Genetic Epidemiology*, **20**, 4-16.

Risch, N. and Teng, J. (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling, *Genome Research*, **8**, 1273-1288.

Risch, N.J. (2000) Searching for genetic determinants in the new millennium, *Nature*, **405**, 847-856.

Roberts, S.B., *et al.* (1999) Replication of linkage studies of complex traits: an examination of variation in location estimates, *The American Journal of Human Genetics*, **65**, 876-884.

Satten, G.A., Flanders, W.D. and Yang, Q. (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model, *The American Journal of Human Genetics*, **68**, 466-477.

Schaid, D.J. (1996b) General score tests for associations of genetic markers with disease using cases and their parents, *Genetic Epidemiology*, **13**, 423-449.

Schaid, D.J. (1999a) Case-parents design for gene-environment interaction, *Genetic Epidemiology*, **16**, 261-273.

Sengul, H., Weeks, D.E. and Feingold, E. (2001) A survey of affected-sibship statistics for nonparametric linkage analysis, *The American Journal of Human Genetics*, **69**, 179-190.

Setakis, E., Stirnadel, H. and Balding, D.J. (2006) Logistic regression protects against population structure in genetic association studies, *Genome Research*, **16**, 290-296.

Smit, L.A.M., *et al.* (2008) CD14 and toll-like receptor gene polymorphisms, country living, and asthma in adults, *American journal of respiratory and critical care medicine*, 200810-201533OCv200811.

Spielman, R.S. and Ewens, W.J. (1996) The TDT and other family-based tests for linkage disequilibrium and association, *American Journal of Human Genetics*, **59**, 983.

Spielman, R.S. and Ewens, W.J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test, *The American Journal of Human Genetics*, **62**, 450-458.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics*, **52**, 506.

Terwilliger, J.D. and Ott, J. (1994) *Handbook of human genetic linkage*. Johns Hopkins Univ Pr.

Van Steen, K., *et al.* (2005) Genomic screening and replication using the same data set in family-based association testing, *Nature Genetics*, **37**, 683-691.

Veal, C.D., *et al.* (2002) Family-Based Analysis Using a Dense Single-Nucleotide Polymorphism–Based Map Defines Genetic Variation at PSORS1, the Major Psoriasis-Susceptibility Locus, *The American Journal of Human Genetics*, **71**, 554-564.

Waldman, I., Robinson, B. and Rowe, D. (1999) A logistic regression based extension of the TDT for continuous and categorical traits, *Annals of human genetics*, **63**, 329-340.

Wang, K.S., *et al.* (2011) Polymorphisms in ABLIM1 are Associated with Personality Traits and Alcohol Dependence, *Journal of Molecular Neuroscience*, 1-7.

Wang, W.Y.S., *et al.* (2005) Genome-wide association studies: theoretical and practical concerns, *Nature Reviews Genetics*, **6**, 109-118.

Weeks, D.E. and Lathrop, G.M. (1995) Polygenic disease: methods for mapping complex disease traits, *Trends in Genetics*, **11**, 513-519.

Weinberg, C. (1999b) Allowing for missing parents in genetic studies of case-parent triads, *The American Journal of Human Genetics*, **64**, 1186-1193.

Weinberg, C., Wilcox, A. and Lie, R. (1998) A log-linear approach to case-parent–triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting, *The American Journal of Human Genetics*, **62**, 969-978.

Weinberg, C.R. (1999a) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads, *The American Journal of Human Genetics*, **65**, 229-235.

Whittaker, J. and Morris, A. (2001) Family-based tests of association and/or linkage, *Annals of human genetics*, **65**, 407-419.

Whittemore, A.S. and Halpern, J. (1994) A class of tests for linkage using affected pedigree members, *Biometrics*, 118-127.
Xiong, M., Krushkal, J. and Boerwinkle, E. (1998) TDT statistics for mapping quantitative trait loci, *Annals of human genetics*, **62**, 431-452.

Yu, J., *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nature Genetics*, **38**, 203-208.

CHAPTER 2

EMK: A novel program for family-based allelic and genotypic association tests on quantitative traits

Yen-Wei Li, Eden R. Martin, and Yi-Ju Li*

Annals of Human Genetics (2008) 72 (Pt 3): 388-396

Abstract

The QTDT program is a widely-used program for analyzing quantitative trait data, but the methods mainly test allelic association. Since the genotype of a marker is a direct observation for an individual, it is of interest to assess association at the genotypic level. In this study, we extended the allele-based association method developed by Monks and Kaplan (MK method) to genotype-based association tests for quantitative traits. We implemented a novel extended MK (EMK) program that can perform both allele- and genotype- based association tests in any pedigree structure. To evaluate the performance of EMK, we utilized simulated pedigree data and real data from Li, et al. 2003, who studied GSTO1 and GSTO2 genes in Alzheimer disease (AD). Both allele- and genotype-based EMK methods (allele-EMK and geno-EMK) showed correct type I error for various pedigree structures and admixture populations. The geno-EMK method showed comparable power to the allele-EMK test. By treating age-at-onset (AAO) as a quantitative trait, the EMK program was able to detect significant associations for rs4925 in GSTO1 (P = 0.006 for allele-EMK and P = 0.009 for geno-EMK), and rs2297235 in GSTO2 (P = 0.005 for allele-EMK and P = 0.009 for geno-EMK), which are consistent with the findings of Li et al. 2003.

Keywords: family-based association, quantitative trait, age at onset, Alzheimer Disease

2.1.INTRODUCTION

Since the development of the Transmission Disequilibrium Test (TDT) (Spielman, et al., 1993), family-based association methods have played an important role in mapping genes for complex human diseases. The original TDT was applicable only to parent-offspring triads for qualitative traits. Since then, many extensions of TDT have been proposed to incorporate various pedigree structures (Boehnke and Langefeld, 1998; Curtis, 1997; Martin, et al., 1997; Martin, et al., 2000; Spielman and Ewens, 1998) as well as quantitative traits (Allison, 1997; Rabinowitz, 1997; Schaid and Rowland, 1999).

To date, QTDT and FBAT are the two primary programs that perform association tests for quantitative traits for family data. In particular, the QTDT implements five association methods (Abecasis, et al., 2000; Allison, 1997; Fulker, et al., 1999; Monks and Kaplan, 2000; Rabinowitz, 1997), in which the likelihood-based orthogonal model (OM) (Abecasis, et al., 2000) and the TDT-based Monks and Kaplan (MK) method (Monks and Kaplan, 2000) are feasible in wider pedigree structures than the other methods. The OM method tests the additive genetic effect of the quantitative trait loci (QTL) in general pedigrees with or without parental data. It requires normally distributed quantitative trait data, which may not always reflect real data. On the other hand, the MK method applies to nuclear families with and without parental data but has fewer

restrictions on the trait distribution. The MK method also has the advantage of providing information about the direction of the allelic effect- for instance, whether the allele increases or decreases the trait's value.

Most association methods for quantitative traits, including OM and MK, provide evidence of association for an allele rather than a genotype at a marker locus. The development of Genotype-PDT for qualitative traits (Martin, et al., 2003) has set an example of the advantage of assessing the genotypic association. Testing for genotype-based association has more power than testing for allele-based association in some genetic models. For instance, Genotype-PDT was found to have higher power than allele-based PDT under recessive and dominant models (Martin, et al., 2003). The results of genotype association tests may help us understand the underlying genetic model of the susceptibility gene or genetic modifier by evaluating genotype-specific effects.

Methods and programs for testing genotypic association between markers and quantitative traits and for family-based association analysis of quantitative traits are lacking. Nuclear family data are still the focus of the application for most methods. When extended pedigrees are available, not all data within the extended pedigree is used, which may result in some loss of statistical power. We built on the advantages of the MK method by extending it to general pedigree data and to test for genotypic association. Simulation studies were conducted to evaluate the validity of our proposed extensions of

the MK method (EMK methods). We also developed a computer program that implements both allele-based and genotype-based EMK methods for real data analysis.

2.2.METHODS

The MK method utilizes the following two types of informative nuclear families: families with known parental genotypes and at least one heterozygous parent; and families without parental genotypes but with at least two siblings with different genotypes. Our version of allele-based EMK (allele-EMK) and genotype-based EMK (geno-EMK) methods are applicable to a general pedigree that contains the same type of informative nuclear families. The test statistics described below are based on the extended general pedigree. They can be simplified for nuclear families if needed.

Allele-based association test (allele-EMK)

Considering a marker locus with two alleles A_1 and A_2, we assume t_i siblings in the *i*th pedigree, where *i = 1, …, F*. We used the notation of Y_{ij} for the quantitative phenotype of the *j*th individual in the *i*th pedigree and $\overline{Y}_i$ for the mean trait value over all non-founders in the *i*th pedigree.

Several genotypic scores were defined in Monks and Kaplan (Monks and Kaplan, 2000) including the following: X^*_{iM} (X^*_{iF}) = 1 if mother (father) is heterozygous at the marker; X^*_{iM} (X^*_{iF}) = 0 if mother (father) is homozygous; X_{ijM} (X_{ijF}) = 1 if allele A_1

was transmitted to the *j*th offspring by the mother (father); and X_{ijM} (X_{ijF}) = 0 if otherwise. For a nuclear family with parental genotype information, a U_i statistic is defined for the *i*th family:

$$U_i = \frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \bar{Y}_i)[X^*_{iM}(X_{ijM} - 0.5) + X^*_{iF}(X_{ijF} - 0.5)]. \qquad (2.1)$$

When parental genotype information is not available, an alternative V_i statistic is used:

$$V_i = \frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \bar{Y}_i)(X_{ijM} + X_{ijF} - \bar{X}_i), \qquad (2.2)$$

where $\bar{X}_i = \frac{1}{t_i}\sum_{j=1}^{t_i}(X_{ijM} + X_{ijF})$ is an average of genotype scores among siblings.

The sum of X_{ijM} and X_{ijF} is the number of A_1 alleles that the *j*th offspring carries. It should be noted that both Equations (2.1) and (2.2) are slightly different from the original definition of U_i and V_i in the MK method since we replaced $\bar{Y}$, the mean trait value over all non-founders in the all pedigrees, with $\bar{Y}_i$, the family-specific mean trait value.

A general association test statistic T_{QPS} was based on the sum of U_i and V_i from families with and without parental genotype data (Monks and Kaplan, 2000). However, T_{QPS} is based on the assumption that the two types of nuclear families are independent. If a general pedigree contains several nuclear families with both types of family structure, the assumption of independence between nuclear families is not valid. Thus, for pedigree

i, we define a family-specific score $T_i = \sum_{j=1}^{F_{iP}} U_{ij} + \sum_{j=1}^{F_{iS}} V_{ij}$, where F_{iP} and F_{iS} are the number of nuclear families with and without parental-genotype information in pedigree *i*. The new allele-EMK test statistic is, therefore, written as

$$T'_{QPS} = \frac{\sum_{i=1}^{F} T_i}{\sqrt{\sum_{i=1}^{F} T_i^2}} \quad ,$$

where T'_{QPS} follows an asymptotic standard normal distribution under the null hypothesis of no linkage disequilibrium.

Genotype-based EMK (geno-EMK) method

In order to construct the genotype-based MK method, we define a random variable X_{ij} to code the observed genotype. Assume that we test genotype A_1A_1. Let X_{ij} = 1 if the observed genotype is A_1A_1; and X_{ij} = 0 if otherwise.

For the case where parental genotypes are available, an estimate of the covariance between the marker genotype and the quantitative trait can be written as random variable U_i for the *i*th nuclear family:

$$U_i = \frac{1}{t}\sum (Y_{ij} - \bar{Y}_i)(X_{ij} - \left(\frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i} \right))$$

where G_i^* is the pool of all possible genotypes (X_{ij}^*) for an offspring based on parental genotype information and S_i is the number of elements in G_i^* .

For the case of no parental genotype information, the covariance for family *i* with t_i siblings is defined as the following:

$$V_i = \frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \overline{Y}_i)(X_{ij} - \overline{X}_i)$$

where the $\overline{X}_i$ is the mean of the t_i values of X_{ij} .

Similarly, for the *i*th general pedigree with F_{iP} and F_{iS} nuclear families with and without parental genotypes, we define $T_i = \sum_{j=1}^{F_{iP}} U_{ij} + \sum_{j=1}^{F_{iS}} V_{ij}$. Under the null hypothesis, we were able to derive $E(U_i) = 0$ for all nuclear families with parental genotypes and $E(V_i) = 0$ for all nuclear families without parental genotypes (see Appendix A and B in supplementary materials). Therefore, for the F independent general pedigrees, we have $E(\sum_{i=1}^{F} T_i) = \sum_{i=1}^{F} E(T_i) = 0$ and $Var(\sum_{i=1}^{F} T_i) = \sum_{i=1}^{F} E(T_i^2)$. Hence, a test statistic can be written as:

$$T_{GPS\_11} = \frac{\sum_{i=1}^{F} T_i}{\sqrt{\sum_{i=1}^{F} T_i^2}}.$$

Under the null hypothesis, T_{GPS_11} follows an asymptotic standard normal distribution. T_{GPS_11} can be easily modified for genotypes A_1A_2 and A_2A_2 .

To obtain an overall assessment of significance at a marker, a global test can be computed as below:

$$T_{global} = \frac{g-1}{g}\left(\sum_{K=1}^{g} T_{GPS\_K}^{2}\right) \quad ,$$

where g is the number of genotypes at a marker. Under the null hypothesis, T_{global} is asymptotic χ^2 distribution with $g-1$ degrees of freedom (Martin, et al., 2003).

Candidate Gene Analysis for Alzheimer disease

Alzheimer disease (AD) is a leading cause of dementia in the elderly and is known to have a complex etiology with strong genetic and environmental components. Many susceptibility genes have been reported to date, but only four AD genes, amyloid precursor protein (APP) (Goate, et al., 1991), presenilin 1and 2 (PS1, PS2) (Levy-Lahad, et al., 1995; Rogaev, et al., 1995; Sherrington, et al., 1995), and apolipoprotein E (APOE) (Corder, et al., 1993), have been confirmed. In addition to these susceptibility genes, we have been interested in mapping genetic modifiers for age-at-onset (AAO) of AD using quantitative trait approaches. We previously reported glutathione S-transferase omega-1 (GSTO1) and GSTO2 as potential AAO genes for AD (Li, et al., 2003; Li, et al., 2006) based on results from the OM and MK methods implemented in the QTDT program. Here, we tested seven single nucleotide polymorphisms (SNPs) from these two studies to compare our proposed EMK program and the QTDT MK methods.

2.3.SIMULATION

A series of computer simulations were used to examine the type I error and power of the allele-EMK and geno-EMK methods under different genetic models and sample sizes. We examined whether the geno-EMK is a valid test for nuclear families and extended general pedigrees. Further, we assessed the validity of the allele-EMK method using simulated two-generation extended pedigree data.

We assume bi-allelic marker A (A_1 and A_2) with population frequencies p_1 and p_2, and a bi-allelic QTL Q (Q_1 and Q_2) with population frequencies q_1 and q_2. The linkage disequilibrium $D = \Pr(A_1Q_1) - p_1q_1$, where $\Pr(A_1Q_1)$ is population haplotype frequency for A_1Q_1. Traits resulting from three QTL genotypes, Q_1Q_1, Q_1Q_2, and Q_2Q_2, are assumed to follow normal distributions. The mean of each genotype-specific trait is defined as $\mu_{11} = a$, $\mu_{12} = d$, and $\mu_{22} = -a$. A common variance σ_G^2 is assumed for all QTL genotypes, where $\sigma_G^2 = 2q_1q_2[a + d(q_2 - q_1)]^2 + (2q_1q_2d)^2$. The residual trait variance is assumed to be σ_E^2. These parameters lead to the broad-sense heritability of $H^2 = \sigma_G^2/(\sigma_G^2 + \sigma_E^2)$ (Falconer, et al., 1996).

Parameters used in the simulations are listed in Table 1. Given allele frequencies of the marker A and QTL Q, and linkage disequilibrium D, four population haplotype frequencies for marker A and QTL Q can be calculated by $P(A_1Q_1) = p_1q_1 + D$, $P(A_1Q_2) = p_1q_2 - D$, $P(A_2Q_1) = p_2q_1 - D$, and $P(A_2Q_2) = p_2q_2 + D$. The haplotypes

of each parent were simulated based on the given haplotype frequencies. We then randomly drew one haplotype from each parent to form two haplotypes for each offspring.

QTLs were simulated with additive, dominant, recessive, and overdominant models. For simplicity, we assumed the dominance effect (d) is the product of additive effect (a) and a scale of dominance (k) ($d = k \times a$). Therefore, the genetic models of recessive, additive, and dominant are reflected by k = -1, 0, and 1, respectively. An overdominant model is the one with $k > 1$ or $k < -1$. We assumed the quantitative traits Y follow normal distributions with corresponding mean and variance, where $Y_{Q_1Q_1} \sim N(a, \sigma_G^2 + \sigma_E^2)$, $Y_{Q_1Q_2} \sim N(k*a, \sigma_G^2 + \sigma_E^2)$, and $Y_{Q_2Q_2} \sim N(-a, \sigma_G^2 + \sigma_E^2)$.

Our simulation studies evaluated various pedigree structures including nuclear families with or without parental information, and two-generation general pedigrees (Figure 1). In each simulation study, there were 200, 500 nuclear families, or 200 general pedigrees for each replicate, and 10,000 replicates. The type I error of the allele-EMK and geno-EMK were estimated under the cases of no association between marker and QTL (linkage disequilibrium coefficient $D = 0$) and statistical power was estimated for $D \neq 0$. We used 0.05 as the significance level for all estimates.

To demonstrate that the EMK tests are not affected by population admixture, we also investigated the type I error rates of the allele-EMK and geno-EMK tests using simulated

admixture population data. We simulated 500 two-sib nuclear families that are a mixture of two equal size subpopulations with different allele frequencies at the marker and QTL. The marker and QTL allele frequencies were 0.3 for the first subpopulation and 0.1 for the second subpopulation.

2.4.RESULTS

Type I Error Rates

Table 2 presents the type I error rates for each genotype, global geno-EMK, allele-EMK, and QTDT MK tests in 200 nuclear families with two and five sibs simulated under different genetic models. Except for the geno-EMK11 genotype tests (P = 0.040 in the recessive model, P = 0.038 in the additive model, and P = 0.039 in the dominant model), the type I error estimates are very close to the nominal significance level of 0.05. The exception is probably the result of the low frequency of the 11 genotype ($P(A_1A_1) = 0.04$) and the small number of observations for these data. Compared to the allele-EMK and geno-EMK tests, the QTDT MK test is consistently conservative, especially for the two-sib nuclear families without parental genotypes case ($P \approx$ 0.041~0.044).

Overall, our error rate estimates in the allele-EMK test are closer to the nominal significance level than those in the geno-EMK global test. Families with two sibs generally show slightly lower type I error rates than those with five sibs and general

pedigree cases. This was expected, because the overall sample size is larger in the five sibs and general pedigree cases than in the two-sib case.

In the case of an admixture population, simulations showed type I error rates close to the nominal significance level in the geno-EMK global test ($P \approx 0.044 \sim 0.052$) and allele-EMK test ($P \approx 0.048 \sim 0.053$). These results demonstrate that the EMK is valid for testing association regardless of whether population substructure exists.

Power Estimates

The statistical power for both geno-EMK and allele-EMK methods was evaluated for all combinations of genetic models, parameters, and pedigree structures. Unlike geno-PDT for testing disease risk (Martin, et al., 2003), we found that geno-EMK has similar power patterns with allele-EMK for quantitative traits simulated under dominant, additive, and recessive models. Here, we present power curves across different degree of linkage disequilibrium (D) (Figures 2 and 3). Interestingly, geno-EMK has higher power than allele-EMK under the overdominant model with k = -2 (Figure 4). Overall, power is very sensitive to the degree of D between marker and QTL, and the availability of parental data. In all cases, maximum power is obtained when the marker is in perfect disequilibrium (e.g. D = 0.16) with the trait locus and when parental genotypes are

available. This is expected because the parental controls provide more accurate estimates of the expected genotype score than sibling controls. Moreover, the tests in general pedigrees show the highest statistical power due to the large sample size per pedigree. For all scenarios, higher power was observed for data simulated under additive and dominant genetic models than for data simulated under the recessive model.

It should be noted that the allele-EMK test has slightly higher power than the original MK under nuclear families without parental genotype (Figure 3 (b)). Therefore, our allele-EMK method may serve as a good alternative for studies with missing parental data, in particular, for the genetic studies of late-onset diseases.

Figure 2 shows the results of geno-EMK power comparisons for all three genotypes and for the global test for data simulated under the additive model. The global test shows a similar power pattern to the main associated genotype, which here is 22. The same pattern was observed in the recessive and dominant models. Therefore, the global test can serve as an initial overall assessment to support the evidence of individual genotype association.

In general, the allele-EMK test has greater power than the geno-EMK global test (Figure 3), except for a few exceptions. First, the genotype specific test may have greater power than the allele-EMK and QTDT MK test in some cases such as the recessive model (k = -1). For example, assume the population mean μ = -1.84 and trait means of

Q_1Q_1, Q_1Q_2, Q_2Q_2 as 2, -2, -2 respectively (Figure 5). Under the assumption of strong LD (D = 0.12), for which maker and QTL genotypes are mostly identical, the geno-EMK 11 test showed more power (92.43%) than 12 (12.66%) and 22 (16.79%) tests under simulations of 500 two-sib nuclear families with parents. This may be because the trait mean of Q_1Q_1 has a much greater difference from μ than the trait means of Q_1Q_2 and Q_2Q_2. Because the allele-based test counts data across two genotypes (11 and 12 for allele 1), the allele-EMK test and the QTDT MK test (power = 64.51% and 74.3% respectively) are less significant.

Second, the geno-EMK test has much higher power than allele-EMK test in data with and without parental genotypes under the overdominant model (Figure 4). For the overdominant model of k = -2 illustrated in Figure 5 (the population trait mean μ = -2.48), the expected trait mean for Q_2Q_2 is closer to μ than Q_1Q_1 and Q_1Q_2. Therefore we found that geno-EMK has higher power in 11 (power = 72.16%) and 12 (power = 92.91%) than in 22 (power = 61.23%) for D = 0.12 in 500 two-sib nuclear families with parents. However, the allele-based tests have much lower power in this case (allele-EMK power = 11.24% and QTDT MK power = 12.77%). This may be because genotypes 11 and 12 have opposite quantitative effects (μ_{11} = 2 and μ_{12} = -4), which diminishes the allelic association signal.

Analysis in Age-At-Onset Data

Table 3 shows the results of the allele-EMK and geno-EMK tests for the seven SNPs in GSTO1 and GSTO2 genes for age-at-onset data in families with Alzheimer disease (AD). There were 711 families in the AD dataset. Li, et al (2003) reported significant findings for rs4925 in GSTO1 (P = 0.023) and rs2297235 in GSTO2 (P = 0.024) based on the QTDT MK test. Our allele-EMK and geno-EMK tests supported the findings of the QTDT MK test with smaller p-values. The allele-EMK test p-values were 0.006 (rs4925) and 0.005 (rs2297235), and the global geno-EMK test p-values were both 0.009 (rs4925 and rs297235). More interestingly, genotype 22 at rs4925 (P = 0.006) in GSTO1 and 11 at rs2297235 (P = 0.007) in GSTO2 are the most significant associated genotypes of the seven SNPs we tested to early age-at-onset of Alzheimer disease. This example shows that our EMK program can handle real data analysis and yield more informative insights into significant results than the existing methods.

2.5.DISUSSION

Family-based association methods have played a central role in candidate gene studies for complex human diseases. Computer programs for performing association tests in real data will become more and more important as we move toward whole genome association studies. In this study, we extended the allele-based MK method to

multi-generation families and developed a genotype association method for quantitative trait based on the framework of the MK method. We evaluated the validity and power of these two new methods by simulating various family datasets and admixture populations. Our simulation studies showed that both geno-EMK and allele-EMK tests have the correct type I error rate for all pedigrees.

The allele-EMK test has slightly higher power than the original MK method under nuclear families without parental genotypes. This may be due to changing the overall trait mean to family specific trait mean in the test statistic (2.1) and (2.2). Since parental genotypes are mostly missing in late onset disease such as Alzheimer disease, the allele-EMK method could serve as a good alternative to the original MK test.

The geno-EMK global test maintains the nominal significance level even in the cases that the type I error of a particular genotype test is too conservative. Moreover, the global test shows similar power to the test of the main associated genotype. Therefore, we recommend using the geno-EMK global test as an initial overall assessment to support the evidence of individual genotype association. For instance, geno-EMK genotype tests have significant findings for rs2164624 in GSTO1 (P = 0.045 for 11 test and P = 0.039 for 12 test), but the global test did not provide a significant result (P = 0.062). Therefore, this SNP may not as important as the other two SNPs (rs4925 and rs2297235). Furthermore, both allele-EMK and QTDT MK tests did not reveal significant results on

rs2164624.

The statistical power is comparable between the global geno-EMK and allele-EMK tests. However, the geno-EMK test for an individual genotype may have higher power than allele-EMK or the original MK method in some cases such as the recessive model. We also found that the geno-EMK test has much higher power than allele-based tests under the overdominant model. Overall, the geno-EMK has the advantage of offering genotype specific association results and can be more power for some genetic models.

Using EMK, significant SNPs rs4925 and rs2297235 and their early AAO genotypes were found, which reproduces Li, et al. (2003) with smaller p-values. It also echoes our previous reports that rs4925 allele 1 (A nucleotide) carriers have a maximum 6.8 year delay of AAO compared to individuals with the 22 genotype (CC) of rs4925 (Li, et al. 2006). Furthermore, the allele-EMK test found rs156697, which is in LD with rs2297235 ($r^2 = 0.78$; Li, et al. (2003)), to be moderately significant ($P = 0.039$), while the MK test did not ($P = 0.113$).

In conclusion, we have shown that our EMK program is a robust tool to analyze quantitative trait in family data. While the QTDT program has the flexibility of choosing different testing methods, we consider that our EMK program will be a better alternative for the MK method. The EMK software for conducting the geno-EMK and allele-EMK is written in C++ and available for UNIX and Windows platforms. It can be publicly

accessed at the Duke Center for Human Genetics Web site (http://wwwchg.duhs.duke.edu/research/software.html).

Acknowledgements

We would like to thank Dr. Andrew Dellinger for his helpful suggestions to improve this paper. This work was supported by a research grant for American Federal for Aging Research (AFAR), a new investigator grant (NIRG-02-3603) and an investigator initiative research grant (IIRG-05-14708) from Alzheimer's association. The Alzheimer data were supported by grants, NS311530, AG021547, AG19757, and AG05128 from NIH.

References

Abecasis, G., Cardon, L. and Cookson, W. (2000) A general test of association for quantitative traits in nuclear families, *The American Journal of Human Genetics*, **66**, 279-292.

Allison, D.B. (1997) Transmission-disequilibrium tests for quantitative traits, *American Journal of Human Genetics*, **60**, 676.

Boehnke, M. and Langefeld, C.D. (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test, *The American Journal of Human Genetics*, **62**, 950-961.

Corder, E., *et al.* (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families, *Science*, **261**, 921.

Curtis, D. (1997) Use of siblings as controls in case-control association studies, *Annals of Human Genetics*, **61**, 319-333.

Falconer, D.S., Mackay, T.F.C. and Frankham, R. (1996) Introduction to Quantitative Genetics (4th edn), *Trends in Genetics*, **12**, 280.

Fulker, D., *et al.* (1999) Combined linkage and association sib-pair analysis for quantitative traits, *The American Journal of Human Genetics*, **64**, 259-267.

Goate, A., *et al.* (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease, *Nature*, **349**, 704-706.

Levy-Lahad, E., *et al.* (1995) Candidate gene for the chromosome 1 familial Alzheimer's disease locus, *Science*, **269**, 973.

Li, Y.J., *et al.* (2003) Glutathione S-transferase omega-1 modifies age-at-onset of Alzheimer disease and Parkinson disease, *Human molecular genetics*, **12**, 3259-3267.

Li, Y.J., *et al.* (2006) Revealing the role of glutathione S-transferase omega in age-at-onset of Alzheimer and Parkinson diseases, *Neurobiology of aging*, **27**, 1087-1093.

Martin, E., *et al.* (2003) Genotype-based association test for general pedigrees: The genotype-PDT, *Genetic epidemiology*, **25**, 203-213.

Martin, E., Kaplan, N. and Weir, B. (1997) Tests for linkage and association in nuclear families, *The American Journal of Human Genetics*, **61**, 439-448.

Martin, E.R., *et al.* (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test, *The American Journal of Human Genetics*, **67**, 146-154.

Monks, S. and Kaplan, N. (2000) Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus, *The American Journal of Human Genetics*, **66**, 576-592.

Rabinowitz, D. (1997) A transmission disequilibrium test for quantitative trait loci, *Human heredity*, **47**, 342-350.

Rogaev, E., *et al.* (1995) Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene, *Nature*, **376**, 775-778.

Schaid, D. and Rowland, C. (1999) Quantitative trait transmission disequilibrium test: allowance for missing parents, *Genetic epidemiology*, **17**, S307.

Sherrington, R., *et al.* (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease, *Nature*, **375**, 754-760.

Spielman, R.S. and Ewens, W.J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test, *The American Journal of Human Genetics*, **62**, 450-458.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics*, **52**, 506.

APPENDIX

Appendix A

Assume t_i children in the *i*th family. Let Y_{ij} denote the trait value for the *j*th child in the *i*th family, and let $\bar{Y}_i$ denote the mean trait value over all offspring in the *i*th family. Define $\mu_i = E(\bar{Y}_i)$ be the trait mean in family *i*. Let $p_{i|r} = \Pr(A_i \mid Q_r)$ be the conditional probability of marker allele A_i given QTL Q_r. Denote the mother's and father's marker and QTL haplotypes as H_{iM1}, H_{iM2} and H_{iF1}, H_{iF2}.

The Expectation of U_i

For the case of one parent with heterozygous genotype ($h_i = 1$), without loss of generality, assume the mother is heterozygous. Let genotype A_1A_1 be the variable of transmission genotype. The expected U_i can be derived as the following:

$$E(U_i) = \frac{1}{t_i}\sum_{j=1}^{t_i} E[(Y_{ij} - \bar{Y}_i)(X_{ij} - \left(\frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i}\right))]$$

$$= E[(Y_{ij} - \bar{Y}_i)(X_{ij} - 1/2)]$$

$$= \sum_{r=1}^{2}\sum_{s=1}^{2}\sum_{l=1}^{2}\sum_{k=1}^{2} E[(Y_{ij} - \bar{Y}_i)(X_{ij} - 1/2) \mid H_{iM1} = A_1Q_r, H_{iM2} = A_2Q_s, H_{iF1} = A_1Q_l, H_{iF2} = A_1Q_k]$$
$$\times \Pr(H_{iM1} = A_1Q_r, H_{iM2} = A_2Q_s, H_{iF1} = A_1Q_l, H_{iF2} = A_1Q_k)$$

$$= \sum_{r=1}^{2}\sum_{s=1}^{2}\sum_{l=1}^{2}\sum_{k=1}^{2} E[(Y_{ij} - \bar{Y}_i)(X_{ij} - 1/2) \mid H_{iM1} = A_1Q_r, H_{iM2} = A_2Q_s, H_{iF1} = A_1Q_l, H_{iF2} = A_1Q_k]$$
$$\times \left(\frac{p_{1|r}\, p_{2|s}\, p_{1|l}\, p_{1|k}\, q_r q_s q_l q_k}{p_1^3 p_2}\right)$$

$$= \left(\frac{1}{p_1^3 p_2} \right) \times \left[\frac{1}{2}(1-\theta) \right]^2 \times \begin{bmatrix} (p_{1|1}^2 p_{1|2} p_{2|2} q_1^2 q_2^2)(2a) + (p_{1|1} p_{1|2}^2 p_{2|1} q_1^2 q_2^2)(-2a) \\ + (p_{1|1}^3 p_{2|2} q_1^3 q_2)(a-d) + (p_{1|1} p_{1|2}^2 p_{2|2} q_1 q_2^3)(a+d) \\ + (p_{1|1}^2 p_{1|2} p_{2|1} q_1^3 q_2)(-a+d) + (p_{1|2}^3 p_{2|1} q_1 q_2^3)(-a-d) \end{bmatrix}$$

$$\approx \frac{(1-2\theta)aD}{4 p_1 p_2} - \frac{(1-2\theta)dD}{4 p_1^3 p_2}[\, p_1^2 (q_1 - q_2) + 2 p_1 D\,]$$

$$= \frac{(1-2\theta)D}{4 p_1 p_2}[a - d(q_1 - q_2)] - \frac{(1-2\theta)dD^2}{2 p_1^2 p_2} \quad .$$

For the case of two parents with heterozygous genotypes ($h_i = 2$), the $E(U_i)$ can be derived as the following:

$$E(U_i)=\frac{1}{t_i}\sum_{j=1}^{t_i}E[(Y_{ij}-\bar{Y}_i)(X_{ij}-\left(\frac{\sum_{X_{ij}^*\in G_i^*}X_{ij}^*}{S_i}\right))]$$

$$=E[(Y_{ij}-\bar{Y}_i)(X_{ij}-\sum_{X_{ij}^*\in G_i^*}X_{ij}^*/4)]$$

$$=\sum_{r=1}^{2}\sum_{s=1}^{2}\sum_{l=1}^{2}\sum_{k=1}^{2}E[(Y_{ij}-\bar{Y}_i)(X_{ij}-\sum_{X_{ij}^*\in G_i^*}X_{ij}^*/4)\,|\,H_{iM1}=A_1Q_r,H_{iM2}=A_2Q_s,H_{iF1}=A_1Q_l,H_{iF2}=A_2Q_k]$$

$$\times\Pr(H_{iM1}=A_1Q_r,H_{iM2}=A_2Q_s,H_{iF1}=A_1Q_l,H_{iF2}=A_2Q_k)$$

$$=\sum_{r=1}^{2}\sum_{s=1}^{2}\sum_{l=1}^{2}\sum_{k=1}^{2}E[(Y_{ij}-\bar{Y}_i)(X_{ij}-1/4)\,|\,H_{iM1}=A_1Q_r,H_{iM2}=A_2Q_s,H_{iF1}=A_1Q_l,H_{iF2}=A_2Q_k]$$

$$\times\left(\frac{p_{1|r}p_{2|s}p_{1|l}p_{2|k}q_rq_sq_lq_k}{(p_1p_2)^2}\right)$$

$$=\left(\frac{1}{p_1^2p_2^2}\right)\times\left[\frac{1}{2}(1-\theta)\right]^2\times\frac{1}{2}\begin{bmatrix}p_{1|1}^2p_{2|1}p_{2|2}q_1^3q_2(a-d)+p_{1|1}p_{1|2}p_{2|1}^2q_1^3q_2(-a+d)\\+p_{1|1}^2p_{2|1}p_{2|2}q_1^3q_2(a-d)+p_{1|1}^2p_{2|2}^2q_1^2q_2^2(2a-d)\\+p_{1|1}p_{1|2}p_{2|1}p_{2|2}q_1^2q_2^2(d)+p_{1|1}p_{1|2}p_{2|2}^2q_1q_2^3(a+d)\\+p_{1|1}p_{1|2}p_{2|1}^2q_1^3q_2(-a+d)+p_{1|1}p_{1|2}p_{2|1}p_{2|2}q_1^2q_2^2(d)\\+p_{1|2}^2p_{2|1}^2q_1^2q_2^2(-2a-d)+p_{1|2}^2p_{2|1}p_{2|2}q_1q_2^3(-a-d)\\+p_{1|1}p_{1|2}p_{2|2}^2q_1q_2^3(a+d)+p_{1|2}^2p_{2|1}p_{2|2}q_1q_2^3(-a-d)\end{bmatrix}$$

$$\approx\frac{(1-2\theta)}{8(p_1p_2)^2}\{2Dap_1p_2+Dd[2p_1p_2(q_2-q_1)+(p_1-3p_2)D]\}$$

$$=\frac{(1-2\theta)D}{4p_1p_2}[a-d(q_1-q_2)]+\frac{(1-2\theta)(p_1-3p_2)dD^2}{8(p_1p_2)^2}\quad .$$

The expectation of U_i for a family with at least one heterozygous parent:

$$E(U_i) = \Pr(h_i = 1 \mid h_i = 1 \text{ or } h_i = 2) \times E(U_i \mid h_i = 1) + \Pr(h_i = 2 \mid h_i = 1 \text{ or } h_i = 2) \times E(U_i \mid h_i = 2)$$

$$= \left[\frac{4p_1p_2(1-2p_1p_2)}{4p_1p_2(1-p_1p_2)} \right] \times \left[\frac{(1-2\theta)D}{4p_1p_2}[a - d(q_1 - q_2)] - \frac{(1-2\theta)dD^2}{2p_1^2p_2} \right]$$

$$+ \left[\frac{4p_1^2p_2^2}{4p_1p_2(1-p_1p_2)} \right] \times \left[\frac{(1-2\theta)D}{4p_1p_2}[a - d(q_1 - q_2)] + \frac{(1-2\theta)(p_1 - 3p_2)dD^2}{8(p_1p_2)^2} \right]$$

$$= \frac{(1-2\theta)D}{4p_1p_2}[a - d(q_1 - q_2)] - \frac{(1-2\theta)dD^2}{4p_1^2p_2(1-p_1p_2)}(4p_1^2 - 5p_1 + 4)$$

$$\approx \frac{(1-2\theta)D}{4p_1p_2}[a - d(q_1 - q_2)] \quad .$$

Because $E(U_i)$ is a function of D, under the null hypothesis of no linkage disequilibrium ($D = 0$), we have $E(U_i) = 0$.

Appendix B

The Expectation of V_i

For the case of one parent with heterozygous genotype

$$E(V_i) = E\left[\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \bar{Y}_i)(X_{ij} - \bar{X}_i)\right]$$

$$= E\left[\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \bar{Y}_i)(X_{ij} - \frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i} + \frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i} - \bar{X}_i)\right]$$

$$= E\left[\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \bar{Y}_i)(X_{ij} - \frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i})\right] + E\left[\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \bar{Y}_i)(\frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i} - \bar{X}_i)\right]$$

$$= E(U_i) + \frac{1}{2}E(Y_{i1} - \bar{Y}_i) - E[(Y_{i1} - \bar{Y}_i)\bar{X}_i]$$

$$= E(U_i) + \frac{1}{2}E(Y_{i1} - \bar{Y}_i) - \frac{1}{t_i}E[(Y_{i1} - \bar{Y}_i)X_{i1}] - \frac{t_i - 1}{t_i}E[(Y_{i1} - \bar{Y}_i)X_{i2}]$$

$$= E(U_i) - \frac{1}{t_i}E[(Y_{i1} - \bar{Y}_i)(X_{i1} - \frac{1}{2})]$$

$$= E(U_i) - \frac{1}{t_i}E(U_i) \quad = \frac{t_i - 1}{t_i}E(U_i) \quad .$$

For the case of two parents with heterozygous genotypes

$$E(V_i) = E\left[\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \overline{Y}_i)(X_{ij} - \overline{X}_i)\right]$$

$$= E\left[\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \overline{Y}_i)(X_{ij} - \frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i} + \frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i} - \overline{X}_i)\right]$$

$$= E\left[\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \overline{Y}_i)(X_{ij} - \frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i})\right] + E\left[\frac{1}{t_i}\sum_{j=1}^{t_i}(Y_{ij} - \overline{Y}_i)(\frac{\sum_{X_{ij}^* \in G_i^*} X_{ij}^*}{S_i} - \overline{X}_i)\right]$$

$$= E(U_i) + \frac{1}{4}E(Y_{i1} - \overline{Y}_i) - E[(Y_{i1} - \overline{Y}_i)\overline{X}_i]$$

$$= E(U_i) + \frac{1}{4}E(Y_{i1} - \overline{Y}_i) - \frac{1}{t_i}E[(Y_{i1} - \overline{Y}_i)X_{i1}] - \frac{t_i - 1}{t_i}E[(Y_{i1} - \overline{Y}_i)X_{i2}]$$

$$= E(U_i) - \frac{1}{t_i}E[(Y_{i1} - \overline{Y}_i)(X_{i1} - \frac{1}{4})]$$

$$= E(U_i) - \frac{1}{t_i}E(U_i) = \frac{t_i - 1}{t_i}E(U_i) \quad .$$

We can know that $E(V_i)$ is a function of $E(U_i)$. Thus we have $E(V_i) = 0$ under the null hypothesis of no linkage disequilibrium.

Table 1. Parameters used in the simulation study

Parameters used in the simulation study		
Marker allele frequency	$P(A_1)$	0.2, 0.5
QTL allele frequency	$P(Q_1)$	0.2, 0.5
Scale of Dominant model ($k = d / a$)*		-2, -1.5, -1, 0, 1, 1.5, 2
Number of sibling size		2, 5
Number of families simulated		200, 500
Heritability (H^2)		0.1
Number of iterations		10,000

* a: additive effect; d : dominant effect.

Table 2. Type I error rates for data simulated from 200 families under different genetic models and family structures

			With Parents		Without Parents	
Model	**Method***		**2 sibs**	**5 sibs**	**2 sibs**	**5 sibs**
Recessive	**Geno-EMK**	**G11**	0.042	0.046	0.040	0.045
		G12	0.050	0.045	0.051	0.051
		G22	0.053	0.045	0.048	0.050
		Global	0.051	0.047	0.048	0.050
	Allele-EMK		0.050	0.048	0.048	0.050
	QTDT	**MK**	0.049	0.044	0.044	0.046
Additive	**Geno-EMK**	**G11**	0.046	0.049	0.038	0.043
		G12	0.051	0.049	0.051	0.049
		G22	0.052	0.050	0.049	0.050
		Global	0.051	0.048	0.047	0.052
	Allele-EMK		0.050	0.049	0.046	0.050
	QTDT	**MK**	0.047	0.046	0.042	0.047
Dominant	**Geno-EMK**	**G11**	0.042	0.045	0.039	0.044
		G12	0.050	0.049	0.049	0.048
		G22	0.052	0.051	0.047	0.052
		Global	0.050	0.050	0.044	0.051
	Allele-EMK		0.051	0.048	0.047	0.051
	QTDT	**MK**	0.046	0.047	0.042	0.048

* G11 = geno-EMK 11 test; G12 = geno-EMK 12 test; G22 = geno-EMK 22 test; Global = geno-EMK global test.

Table 3. QTDT MK (MK), allele-EMK (A-EMK) and geno-EMK (Global, G11, G12, and G22) test results for GSTO1 and GSTO2 in AD families

Gene	dbSNP no. *		MK	A-EMK	Global	G11	G12	G22
GSTO1	**rs11191972**	**(SNP5)**	0.281	0.091	0.239	0.090	0.882	0.251
	rs2164624	**(SNP6)**	0.117	0.107	0.062	0.045	0.039	0.770
	rs4925	**(SNP7)**	0.023	0.006	**0.009**	0.400	**0.015**	**0.006**
	rs1147611	**(SNP8)**	0.132	0.087	0.147	0.066	0.138	0.670
GSTO2	**rs2297235**	**(SNP9)**	0.024	0.005	**0.009**	**0.007**	0.022	0.240
	rs157077	**(SNP10)**	0.380	0.059	0.150	0.129	0.690	0.073
	rs156697	**(SNP11)**	0.113	**0.039**	0.097	0.050	0.119	0.398

* SNPs 5 – 11 listed in Li et al. 2003.

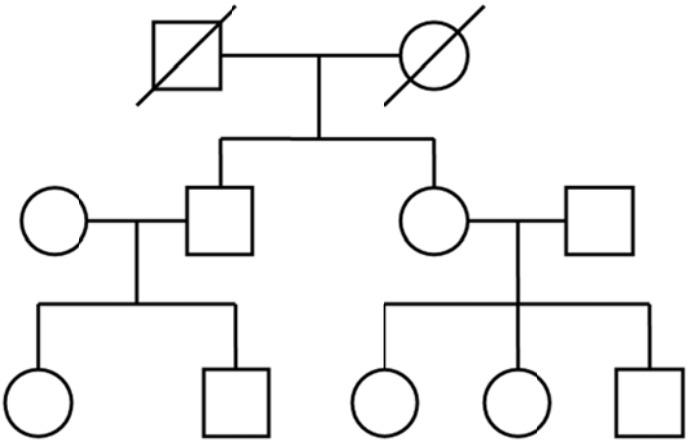

Figure 1. General pedigrees used in simulations.

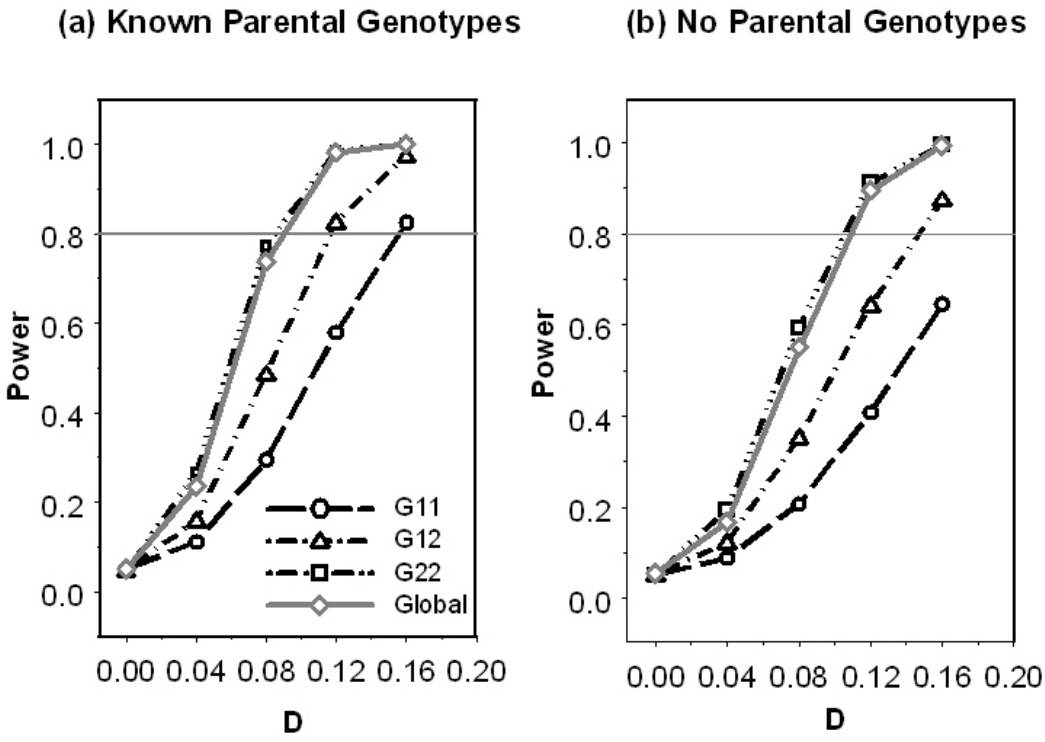

Figure 2. Power comparison among different genotypes under 500 two-sib nuclear families in additive model. Both marker and QTL allele frequencies were 0.2. The heritability was 0.1 and additive genetic effect a = 2.

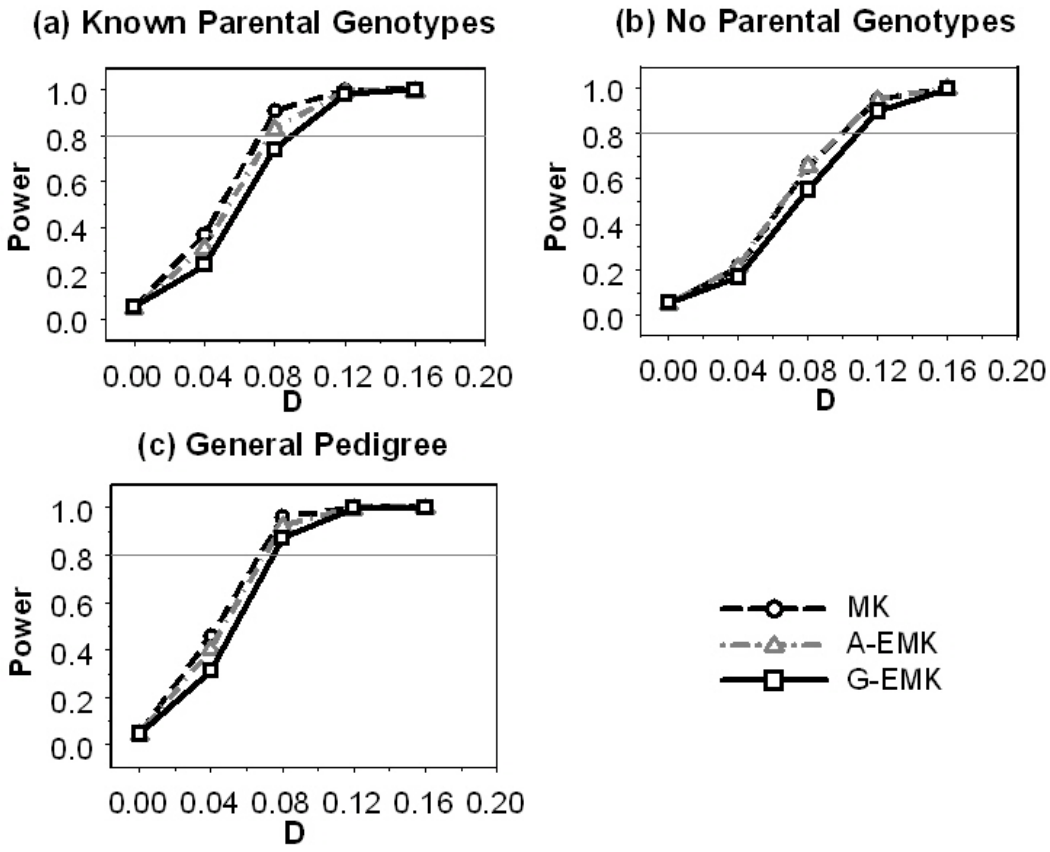

Figure 3. Power comparison among QTDT MK (MK), allele-EMK (A-EMK), and geno-EMK (G-EMK) global test under 500 two-sib nuclear families and 200 general pedigrees in additive model. Both marker and QTL allele frequencies were 0.2. The heritability was 0.1.

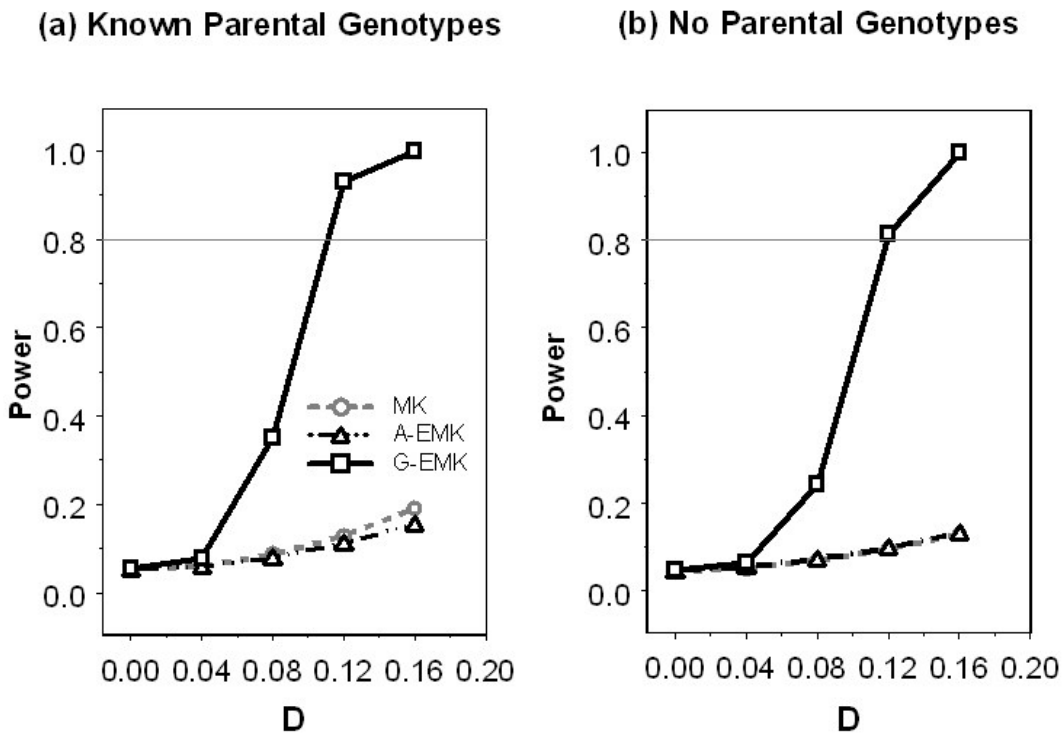

Figure 4. Power comparison among QTDT MK (MK), allele-EMK (A-EMK), and geno-EMK (G-EMK) global test under 500 two-sib nuclear families in overdominant model (k = -2). Both marker and QTL allele frequencies were 0.2. The heritability was 0.1.

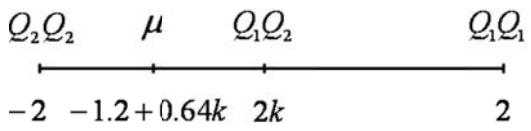

Figure 5. Relationship of genotypic trait means for QTL genotype , , and , and population mean . The additive genetic effect a = 2.

CHAPTER 3

A Family-based Association Method for Pedigree Including Half-Sib Data

Yen-Wei Li and Yi-Ju Li*

In press, Journal of Biometrics & Biostatistics

Abstract

Family datasets could provide good resources for association studies as an initial investigation or a replication study. The current family-based association tests analyze data only from full sibships of a nuclear family or extended pedigrees of related nuclear families. In order to exert more information in the family data, we propose a “Pedigrees with Half-sibs Association Test” (PHAST) to accommodate half-siblings if they are available. PHAST adopts the idea of transmission score from the Pedigree Disequilibrium Test (PDT) to construct the test statistic. The difference is that it utilizes the identity-by-descent (IBD) information of the marker between sibling pairs (full or half sibs) to construct an allelic transmission statistic. If parental genotypes are missing, EM algorithm is used to infer parental genotypes and compute transmission scores for all possible scenarios. The computer simulation results suggested that our new method has valid type I error rates under varied family structures. Our method could have more power than PDT and FBAT when the sample size of half-sibs increases, especially the families without parental genotypes. In conclusion, our method can serve as an alternative method of the existing family-based association tests. Furthermore, it can relax the ascertainment criteria for studying late onset diseases since limited siblings are available.

Key words:

Family-based association; Half-siblings; Identity by descent; Late onset disease

3.1.INTRODUCTION

Association studies have been a predominant approach for analyzing densely spaced genetic markers (e.g. single nucleotide polymorphisms (SNPs)) for detecting genetic variants that may lead to susceptibility genes or genetic modifiers for traits of interest (disease or quantitative traits). Since the development of transmission/disequilibrium test (TDT) (Spielman, et al., 1993), family-based study design was viewed as a robust approach because of less chances of encountering false positive associations due to population stratification existed in the population samples. However, the development of several prominent association methods such as GENOMIC CONTROL (Bacanu, et al., 2002; Devlin and Roeder, 1999), STRUCUTRE (Pritchard, et al., 2000), EIGENSTRAT (Price, et al., 2006) methods has lessened the concern and changed our view of the case-control study design significantly. As the results, many genome wide association studies (GWAS) utilized easily obtained population-based samples rather than family-based samples.

While case-control design seems to dominate the current practice of the association study, particularly GWAS, familial samples still remain some advantages. First, many familial samples were collected during the era of whole genome linkage scans. These family datasets provide good resources for association studies as either an initial

investigation or a replication dataset for other association findings. Second, family data are more robust for detecting genotyping errors (e.g. Mendelian errors) than population-based samples. Third, family data have higher accuracy in inferring haplotypes and confirming the role of rare variants to disease susceptibility. Finally, as the next-generation sequencing is becoming an important approach to pinpoint the causal variant, familial samples have the advantage for the first pass of sequencing effort to discover both common and rare variants. Overall, familial samples will still remain important in genetic studies of human complex diseases.

Many family-based association methods have been developed in the past decade. The TDT was the pioneer of all methods, which tests for association in the presence of linkage using parent-offspring triads. Many extensions of the TDT method were developed for various pedigree structures afterward. The sibling TDT (S-TDT) (Spielman and Ewens, 1998) and several other tests (Boehnke and Langefeld, 1998; Horvath and Laird, 1998; Knapp, 1999) were designed to accommodate discordant sibpairs without parental genotype data. More general methods such as the pedigree disequilibrium test (PDT) (Martin, et al., 2000), and the family-based association test (FBAT) (Rabinowitz and Laird, 2000) can accommodate multiple affected offspring.

The current family-based association tests share a characteristic, which is to analyze data only from full sibships of a nuclear family, such as parent-offspring triads, parents

and multiple affected full siblings, discordant full sibpairs of large size, or extended pedigrees of related nuclear families. Therefore, full siblings or parents have been the target of ascertainment in genetic research. In this study, we explore a new method that can relax the recruitment criteria such as accommodating half-siblings in addition to full siblings if they are available. We also allow our method to handle multiple affected siblings that may exist from pedigrees recruited for linkage studies.

To accommodate these different family structures, we developed a *Pedigrees with Half-sibs Association Test* (*PHAST*) to fully exert all possible information in the family data. The development of PHAST was based on the framework of the pedigree disequilibrium test (PDT) but with several new features: (1) applicable to both full siblings or combinations of full and half siblings family data; (2) inferring parental genotypes when they are missing rather than using siblings information to compute test statistics; and (3) accounting for linkage when parental genotypes are missing. Although this new method may have similar properties as the existing methods, it will largely benefit ascertainment process, particularly for aging related diseases when limited siblings are available to ascertain.

3.2.METHODS

For simplicity, we start with the pedigree structure of affected sibpair (ASP) with parents and concordant half sibpair (HSP) with parents (Figure 6A). Families with more than two affected siblings could be addressed by following the same procedures. We assume that the marker tested is a biallelic marker with a risk allele 1. Our method adopts the concept of allelic transmission presented in PDT but utilizes the identity-by-descent (IBD) information of the marker between a pair of individuals to construct test statistic. For any ASP family, if the two siblings share 2 or 1 IBD at the target marker, we treat the ASP family as a whole unit to compute the allelic transmission score. If the two siblings share 0 IBD at the target marker, we split the ASP family to two independent case-parent trios. The same rules are applied in HSP family except that no 2 IBD sharing between half sibpairs will occur. The mathematical forms of this strategy are described below for different family structures.

Concordant sibpairs with parents:

Define a random variable X for allelic transmission score, which is computed as the number of different copies allele 1 transmitted to the two affected sibs minus the number of allele 1 non-transmitted.

For the ith family (F_i),

$$X_i = \sum_{k=0}^{2} X_{ik} P(IBD = k \mid G_p, G_s) I(F_i \notin Trio) + X_{iT} I(F_i \in Trio), \quad (3.1)$$

where

X_{i0} = (# of allele 1 transmitted to G_{S1}) – (# of allele 1 not transmitted to G_{S1})

+ (# of allele 1 transmitted to G_{S2}) – (# of allele 1 not transmitted to G_{S2}).

X_{i1}, X_{i2} = (# of different allele 1 transmitted in F_i) – (# of allele 1 not transmitted in F_i).

X_{iT} is the number of allele 1 transmitted minus the number of allele 1 non-transmitted for a case-parent trio (*Trio*). G_S = (G_{S1}, G_{S2}) is the marker genotypes for sibling 1 and 2, and G_p is the parental genotypes, that is G_p = (G_{P1}, G_{P2}) for ASP family and G_p = (G_{P1}, G_{P2}, G_{P3}) for HSP family. $P(IBD = k \mid G_p, G_s)$ is the probability that the two affected siblings share *k* IBD, *k* = 0, 1, or 2 (Appendix). Various computer programs can estimate IBD probability. We used MERLIN (Abecasis, et al., 2002) for estimating IBD probabilities at a particular marker.

The value of *X* is observed by counting the number of copes of allele 1 in the affected siblings. Under the null hypothesis of no association between the marker and disease alleles, *X* should have an expected value of 0. To make computation easy for accounting both ASP and HSP units, we looked into the different status of IBD. Under 0 IBD, the two copies of markers of each sibling inherited from totally different source which indicates these two siblings are independent in allelic transmission. Therefore, X_{i0}

is formulated to be the summation of two PDT case-parent trios as the total allelic transmission and sharing score for both ASP and HSP. Under 1 IBD, in order to assure the expectation of statistic X to be 0 under the null hypothesis, the number of "different" copies of alleles transmitted must be equal to the number of alleles non-transmitted. For 1 IBD sharing HSP, in genotypes (G_{S1}, G_{S2}), there are three different copies of allele transmitted (*S1* and *S2* sharing one identical copy from the common parent *P2*) and three alleles remain non-transmitted in (G_{P1}, G_{P2}, G_{P3}). However, this is not the case in ASP family. For 1 IBD sharing ASP, there are three different copies of allele transmitted but the number of allele nontransmitted is always one. In order to get correct transmission statistic, we transform the original two parents ASP into a pseudo HSP family (Figure 6A) by repeating one parent *P1* as *P3*. For instance, assume (G_{P1}, G_{P2}) with genotypes (11, 12) and (G_{S1}, G_{S2}) with genotypes (11, 11), if *S1* and *S2* share 1 IBD, the transformed pseudo HSP family is (G_{P1}, G_{P2}, G_{P1}) with genotypes (11, 12, 11) and the corrected allelic transmission score will be $X_{i1} = 3 - 2 = 1$ (Table 4). Under 2 IBD, which will only occur in ASP unit, it will always be equal between the number of different copies of alleles transmitted and the number of alleles nontransmitted. Therefore, for the 2 IBD sharing ASP family, we can calculate transmission statistic directly without any transformation. In the previous example, if *S1* and *S2* share 2 IBD, G_{S1} and G_{S2} are with the same pair of alleles transmitted from the parents *P1* and *P2*. The different copies allele 1 transmitted

are 2, but not 4, and X_{i2} = 2- (# of allele 1 not transmitted in F_i) = 2 – 1 = 1.

Discordant Sibpairs with parents or parent-offspring triads:

For full and half sibpairs with different disease status (discordance sibpairs, DSP) (Figure 6B, 6C) or simple parent-offspring triads, there is only one affected sibling in each family. The random variable $X = X_{iT}$ (the last part of equation (3.1)) and will follow the PDT framework.

More than two affected siblings:

We also formulate PHAST method for taking into account multiple affected siblings (> two affecteds). The same inference principles described for concordance sibpairs are applied. In the example of nuclear family with three affected siblings (Figure 6D), we identify the IBD status between each pair of siblings first. Split the nuclear family into two independent units if sharing 0 IBD, transform nuclear family to pseudo HSP family if sharing 1 IBD, and then count the allelic transmission score number *X*. The possible *X* scores for nuclear family with three affected siblings are listed in Table 5.

Missing parents --Inferring parental genotypes

When parental genotypes are missing, PDT method does not infer parental genotypes but

utilizes information from discordant sibpairs. However, the allelic sharing scores may vary by the number of available sibling samples genotyped. In our proposed method, we took a step further to infer parental genotypes and compute allelic transmission scores for all possible parental genotypes. At the same time, in order to accommodate linkage between a maker and disease locus, the probability of affected siblings sharing *k* allele IBD, $Z_k = P(IBD = k \mid G_p)$, need to be taken into account. We used the Expectation Maximization (EM) algorithm (Dempster, et al., 1977) to estimate the conditional probability $P(G_p \mid G_s)$ and IBD parameters Z_k. The probability of parental genotypes based on offspring genotypes is formulated as

$$P(G_p \mid G_s) = \frac{P(G_p)\sum_{k=0}^{2} P(G_s \mid G_p, IBD = k)Z_k}{P(G_s)}.$$

Then, the expected random variable in family *i* can be estimated by $X_i = \sum_{j \in C} P(G_{pj} \mid G_{sj}) X_j$. For ASP and HSP without parental genotypes, the value of X_j is calculated by equation (3.1) consistent with G_{sj} and the *j*th set of possible parental genotypes G_{pj}. For missing parental genotypes DSP cases, X_j is the number of allele 1 transmitted minus the number of allele 1 non-transmitted for each inferred parental genotypes, as defined in PDT method under the scenario of known parental genotypes.

The PHAST test statistic

We define a general association statistic D_i based on all ASP, HSP, and DSP subunits in pedigree *i*.

For the *i*th pedigree, where *i = 1, …, N*, a family-specific score

$$D_i = \sum_{j=1}^{F_{iA}} X_{ij} + \sum_{j=1}^{F_{iH}} X_{ij} + \sum_{j=1}^{F_{iD}} X_{ij} ,$$

where F_{iA}, F_{iH}, and F_{iD} are the number of ASP, HSP, and DSP families in pedigree *i*.

Under the null hypothesis, we can derive

$$E(D_i) = \sum_{j=1}^{F_{iA}} E(X_{ij}) + \sum_{j=1}^{F_{iH}} E(X_{ij}) + \sum_{j=1}^{F_{iD}} E(X_{ij}) = 0 .$$

Therefore, the PHAST statistic can be written as

$$T = \frac{\sum_{i=1}^{N} D_i}{\sqrt{\sum_{i=1}^{N} D_i^2}},$$

where *T* follows an asymptotic standard normal distribution under the null hypothesis of no linkage disequilibrium.

3.3.SIMULATION STUDIES

A series of computer simulations was implemented to study the validity of PHAST. For each simulation, 10,000 replicates were generated to estimate type I errors and statistical power. A nominal significance level of 0.05 was used for all estimates. We compared

PHAST with two alternative methods: FBAT (Rabinowitz and Laird, 2000) and PDT (Martin, et al., 2000).

Assume a bi-allelic disease locus *A* with alleles A_1 and A_2 (allele frequencies p_1 and p_2) and a single marker *M* with alleles M_1 and M_2 (allele frequencies q_1 and q_2). Linkage disequilibrium (LD) between the disease locus and the marker was set as $D = P(A_1M_1) - p_1q_1$, where $P(A_1M_1)$ is population haplotype frequency for A_1M_1.

To generate simulation data, four population haplotype frequencies for disease locus *A* and marker *M* were calculated by $P(A_1M_1) = p_1q_1 + D$, $P(A_1M_2) = p_1q_2 - D$, $P(A_2M_1) = p_2q_1 - D$, and $P(A_2M_2) = p_2q_2 + D$. The haplotypes for parental population were generated based on these frequencies. We assume random mating in the population and form two haplotypes for each offspring by randomly drawing one haplotype from each parent. Genetic markers were simulated under the assumption of complete linkage to the disease locus. Three genetic models (recessive, additive, and dominant) were considered through different disease penetrances. The model parameters are given in Table 6. Disease phenotypes were simulated based on disease locus genotypes and their corresponding penetrances.

The type I error was studied under the null hypothesis of no association between the disease and marker alleles ($D = 0$). We generated replicate samples N = 200 and N = 500 of families with different disease models and five types of family structures, ASP, HSP,

DSP, discordant half sibpairs (DHSP) with and without parental genotypes, and nuclear family with three affected siblings, in type I error simulations. To evaluate the power and examine the half sibpairs power contribution, three combinations of different types of family structures were used: (1) 200 families with different ratios of DSP to DHSP, (2) 200 families with different ratios of ASP to HSP, and (3) 200 nuclear families with three affected siblings.

3.4.RESULTS

Type I Error Rates

Tables 7 and 8 present the type I error rates for PHAST, PDT, and FBAT tests in 200 HSP, ASP, and DSP with and without parental genotypes for different simulated genetic models. In the cases that parental genotypes are known, Tables 7 and 8 show that type I error estimates for most tests are very close to the nominal significance level of 0.05. For the scenario of concordant sibpairs without parental genotypes (ASP and HSP), since PDT and FBAT cannot address this type of data, no estimates were obtained for both programs, and a nominal level of type I error estimates was obtained for PHAST.

Table 9 shows type I error rates for data simulated from 200 HSP families under different ratios of 1 IBD to 0 IBD families. Both PHAST and PDT are very robust under different

ratios of 1 IBD to 0 IBD families. However, FBAT tends to have inflated type I error when the ratio of 1 IBD cases are high and conservative type I error when the 1 IBD ratio is low.

Overall, we show that PHAST has correct type I error estimates under different scenario of family structure, and can handle missing parents for concordant sibpairs (full or half sibs). Families with parental genotypes generally show slightly lower type I error rates than those without parents cases. This was expected because the overall sample size is large when parental genotypes are available.

Power Estimates

We also carried out simulations for all combinations of genetic models and different pedigree structures to evaluate the statistical power for PHAST method. Because similar power patterns were found in dominant, additive, and recessive models, we only present the results of additive model here.

Generally, power will increase when the degree of linkage disequilibrium (D) between marker and disease locus increases. For the family structure with parental genotypes cases, PHAST has similar statistical power with PDT and FBAT. For example, when D = 0.21 (the maximum LD scenario for marker and disease allele frequencies at 0.3) for 200 HSP with parents, the power of PHAST is 0.602 and powers of PDT and

FBAT are 0.602 and 0.606. Although the statistical power under the maximum LD was not strong, this is mostly due to the small size of data and the lack of parental genotypes.

Figure 7 shows the results of power comparison among PHAST, PDT, and FBAT under different ratios of DSP without parents to DHSP without parents. A set of 200 families of difference combinations of DSP and DHSP without parents were simulated for each replicate. It is noted that in this simulation study, PDT and FBAT do not use data from DHSP without parents. Therefore, only PHAST method can utilize the full dataset. The results in Figure 7 show that PHAST has increasing power as the proportion of DHSP without parents increases.

Figure 8 shows the results of power comparison among PHAST, PDT, and FBAT under different ratios of concordant ASP without parents to concordant HSP without parents. Similarly, 200 families were simulated for each replicate. The purpose of this simulation is to show that both PDT and FBAT cannot analyze such data while PHAST can handle them to increase some power. For 200 ASPs, PHAST reaches power 0.323 in the maximum disequilibrium (D = 0.21). In the case of 150 ASP and 50 HSP combination, the power of PHAST is 0.188 when D = 0.21. This implies that the power gains from HSP without parents are small due to the limited sibling genotypes available for inferring missing parental genotypes in HSP. The above simulation results for either type of family structure -- siblings without parental genotype information, show that

PHAST will not reach up to 80% power. However, it shows that PHAST can handle more family types and add to the statistical power. We mentioned earlier that we expanded PHAST statistic to handle more than two affected siblings. Our simulation showed that with additional siblings, power to detect risk effects is improved. Figure 9 shows power among PHAST, PDT, and FBAT for 200 families with three affected siblings and known parental genotypes simulated under the additive model. When D = 0.21, the power of PHAST is 0.576, while the power of PDT and FBAT reach 0.735 and 0.641. However, the type I error of PHAST is closest to the significance level of 0.05 (PHAST: 0.049; PDT: 0.043; FBAT: 0.046). The difference in the analysis strategy between PHAST and the other two methods is that PHAST treated the whole family (two parents+3 affecteds) as a whole unit while PDT and FBAT take families with three affected siblings as three independent trios.

3.5.DISCUSSION

In this paper, we proposed the PHAST approach to test for association with the inclusion of half-siblings or more than two affected sibling data. Like other family-based association methods, such as PDT and FBAT, PHAST can also handle the conventional family data structures such as trios, ASP with parents, and DSP. The simulation results

showed that PHAST method has correct type I error under varied family structures. We also studied the properties of the PHAST, PDT, and FBAT test statistics as HSP families contain different ratios of 1 IBD to 0 IBD families. It is important to point out that type I error rates in the FBAT test are inflated as the ratio of half sib-pair sharing 1 IBD increases. Therefore, the power of FBAT method could lead to misinterpretation when the data with half-siblings are used.

We compared our method with two alternative methods, PDT and FBAT. We found the PHAST and PDT version to be, in most cases, highly correlated. Thus in most cases, the PHAST will provide similar power to detect an association over current methods and will, for some specific genetic models, offer a gain in power. Simulations revealed that PHAST could have more power than PDT and FBAT when the sample size of half-siblings increases, especially the families without parental genotypes. Therefore, PHAST can be considered as a useful tool for studying late onset disease.

In summary, the PHAST method can serve as an alternative for the PDT method when either full- or half-siblings, or both, are available. Moreover, PHAST is potentially applicable to genetic studies with special features. For instance, for the aging related diseases, available samples to be recruited may be limited and parents are mostly not available. In this case, if the expansion to half-siblings is possible, PHAST will be able to handle this type of data. Thus, we hope that PHAST can be additional tool for researchers

to facilitate future studies. Accordingly, this method should be able to expand the scope of the current family-based ascertainment strategy, and hopefully add insights into the genetic association study of human complex disease.

Acknowledgements

We are grateful for generous support from a research grant for American Federal for Aging Research (AFAR), a new investigator grant (NIRG-02-3603) and an investigator initiative research grant (IIRG-05-14708) from Alzheimer's association.

References

Abecasis, G.R., *et al.* (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees, *Nature Genetics*, **30**, 97-101.

Bacanu, S.A., Devlin, B. and Roeder, K. (2002) Association studies for quantitative traits in structured populations, *Genetic epidemiology*, **22**, 78-93.

Boehnke, M. and Langefeld, C.D. (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test, *The American Journal of Human Genetics*, **62**, 950-961.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-38.

Devlin, B. and Roeder, K. (1999) Genomic control for association studies, *Biometrics*, **55**, 997-1004.

Horvath, S. and Laird, N.M. (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data, *The American Journal of Human Genetics*, **63**, 1886-1897.

Knapp, M. (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test, *The American Journal of Human Genetics*, **64**, 861-870.

Martin, E.R., *et al.* (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test, *The American Journal of Human Genetics*, **67**, 146-154.

Price, A.L., *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies, *Nature*, **38**, 904-909.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data, *Genetics*, **155**, 945-959.

Rabinowitz, D. and Laird, N. (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information, *Human heredity*, **50**, 211-223.

Spielman, R.S. and Ewens, W.J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test, *The American Journal of Human Genetics*, **62**, 450-458.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM), *American Journal of Human Genetics*, **52**, 506.

APPENDIX

Appendix

$$P(IBD = k \mid G_p, G_s)$$

$$= \frac{P(G_s \mid G_p, IBD = k)P(IBD = k \mid G_p)}{P(G_s \mid G_p)}$$

$$= \frac{P(G_s \mid G_p, IBD = k)Z_k}{P(G_s \mid G_p)}$$

$$= \frac{P(G_s \mid G_p, IBD = k)Z_k}{\sum_{k=0}^{1} P(G_s \mid G_p, IBD = k)Z_k}$$

where

$$Z_k = P(IBD = k \mid G_p)$$

Table 4. X Statistic in ASP family

(G_{P1}, G_{P2})	**(G_{S1}, G_{S2})**	**IBD status**	***X*** **Statistic**
(11, 12)	(11, 11)	0 IBD	X_0 = none
		1 IBD	$X_1 = 3 - 2 = 1$
		2 IBD	$X_2 = 2 - 1 = 1$
(11, 12)	(11, 12)	0 IBD	$X_0 = 1 - 1 = 0$
		1 IBD	$X_1 = 2 - 2 = 0$
		2 IBD	X_2 = none
(11, 12)	(12, 12)	0 IBD	X_0 = none
		1 IBD	$X_1 = 1 - 2 = -1$
		2 IBD	$X_2 = 1 - 2 = -1$
(12, 12)	(11, 11)	0 IBD	X_0 = none
		1 IBD	X_1 = none
		2 IBD	$X_2 = 2 - 0 = 2$
(12, 12)	(11, 12)	0 IBD	X_0 = none
		1 IBD	$X_1 = 2 - 1 = 1$
		2 IBD	X_2 = none
(12, 12)	(11, 22)	0 IBD	$X_0 = 2 - 2 = 0$
		1 IBD	X_1 = none
		2 IBD	X_2 = none

Table 4. (Continued)

(12, 12)	(12, 12)	0 IBD	$X_0 = 0 - 0 = 0$	
		1 IBD	$X_1 = 1 - 2 = -1$	or $2 - 1 = 1$
		2 IBD	$X_2 = 1 - 1 = 0$	
(12, 12)	(12, 22)	0 IBD	X_0 = none	
		1 IBD	$X_1 = 1 - 2 = -1$	
		2 IBD	X_2 = none	
(12, 12)	(22, 22)	0 IBD	X_0 = none	
		1 IBD	X_1 = none	
		2 IBD	$X_2 = 0 - 2 = -2$	
(22, 12)	(22, 22)	0 IBD	X_0 = none	
		1 IBD	$X_1 = 0 - 1 = -1$	
		2 IBD	$X_2 = 0 - 1 = -1$	
(22, 12)	(22, 12)	0 IBD	$X_0 = 1 - 1 = 0$	
		1 IBD	$X_1 = 1 - 1 = 0$	
		2 IBD	X_2 = none	
(22, 12)	(12, 12)	0 IBD	X_0 = none	
		1 IBD	$X_1 = 1 - 0 = 1$	
		2 IBD	$X_2 = 1 - 0 = 1$	

Table 5. X Statistic in the nuclear family with three affected siblings

(G_{P1}, G_{P2})	(G_{S1}, G_{S2}, G_{S3})	X Statistic
(11, 12)	(11, 11 ,11)	$X = 1$
	(11, 11, 12)	$X = 0$
	(11, 12, 12)	$X = 0$
	(12, 12, 12)	$X = -1$
(12, 12)	(11, 11 , 11)	$X = 2$
	(11, 11 , 12)	$X = 1$
	(11, 11 , 22)	$X = 0$
	(11, 12 , 12)	$X = 1$ or 0
	(11, 12 , 22)	$X = 0$
	(11, 22 , 22)	$X = 0$
	(12, 12 , 12)	$X = 0$
	(12, 12 , 22)	$X = -1$ or 0
	(12, 22 , 22)	$X = -1$
	(22, 22 , 22)	$X = -2$
(22, 12)	(22, 22 , 22)	$X = -1$
	(22, 22 , 12)	$X = 0$
	(22, 12 , 12)	$X = 0$
	(12, 12 , 12)	$X = 1$

Table 6. Parameters used in the simulation study

Parameters used in the simulation study	
Disease allele frequency $P(A_1)$	0.3, 0.5
Single Marker allele frequency $P(M_1)$	0.3, 0.5
Two-locus haplotype frequencies $(P_{M1N1}, P_{M1N2}, P_{M2N1}, P_{M2N2})$	(0.3, 0.2, 0.1, 0.4)
GRR2 (=P(affected$\mid A_1A_1$)/ P(affected$\mid A_2A_2$)*	0.15, 0.20
Disease prevalence	0.1
Number of families simulated	200, 500
Number of iterations	10,000

* GRR2: the homozygous genotypic risk ratio

Table 7. Type I error rates for data simulated from 200 concordant half sibpair (HSP) families under different genetic models

200 HSP		**(D= 0, GRR2 = 1.5)**			
		With parental genotypes		**Without parental genotypes**	
Model	**Method**	**P(M)=P(D) =0.3**	**P(M)=P(D) =0.5**	**P(M)=P(D) =0.3**	**P(M)=P(D) =0.5**
Recessive	**PHAST**	0.0507	0.0484	0.0514	0.0429
	PDT	0.0507	0.0484	**NA**	**NA**
	FBAT	0.0503	0.0494	**NA**	**NA**
Additive	**PHAST**	0.0477	0.0499	0.0520	0.0427
	PDT	0.0477	0.0499	**NA**	**NA**
	FBAT	0.0476	0.0517	**NA**	**NA**
Dominant	**PHAST**	0.0493	0.0464	0.0539	0.0455
	PDT	0.0493	0.0464	**NA**	**NA**
	FBAT	0.0489	0.0498	**NA**	**NA**

Table 8. Type I error rates for data simulated from 200 concordant full (ASP) and discordant full sibpair (DSP) families under different genetic models

200 ASP		**(D= 0, GRR2 = 1.5, P(M)=P(D)=0.3)**			
		With parental genotypes		**Without parental genotypes**	
Model	**Method**	**ASP**	**DSP**	**ASP**	**DSP**
Recessive	**PHAST**	0.0468	0.0496	0.0508	0.0539
	PDT	0.0468	0.0478	**NA**	0.0492
	FBAT	0.0479	0.0488	**NA**	0.0492
Additive	**PHAST**	0.0503	0.0521	0.0576	0.0545
	PDT	0.0503	0.0484	**NA**	0.0480
	FBAT	0.0513	0.0514	**NA**	0.0480
Dominant	**PHAST**	0.0520	0.0497	0.0557	0.0526
	PDT	0.0520	0.0505	**NA**	0.0461
	FBAT	0.0518	0.0500	**NA**	0.0461

Table 9. Type I error rates for data simulated from 200 concordant half sibpair (HSP) families under different ratios of 1 IBD to 0 IBD families

200 HSP	**(D= 0, GRR2 = 1.5, P(M)=P(D)=0.3)**			
# 1 IBD families	**# 0 IBD families**	**PHAST**	**PDT**	**FBAT**
200	0	0.0450	0.0450	0.1179
150	50	0.0491	0.0491	0.0835
100	100	0.0497	0.0497	0.0499
50	150	0.0503	0.0503	0.0211
0	200	0.0486	0.0486	0.0023

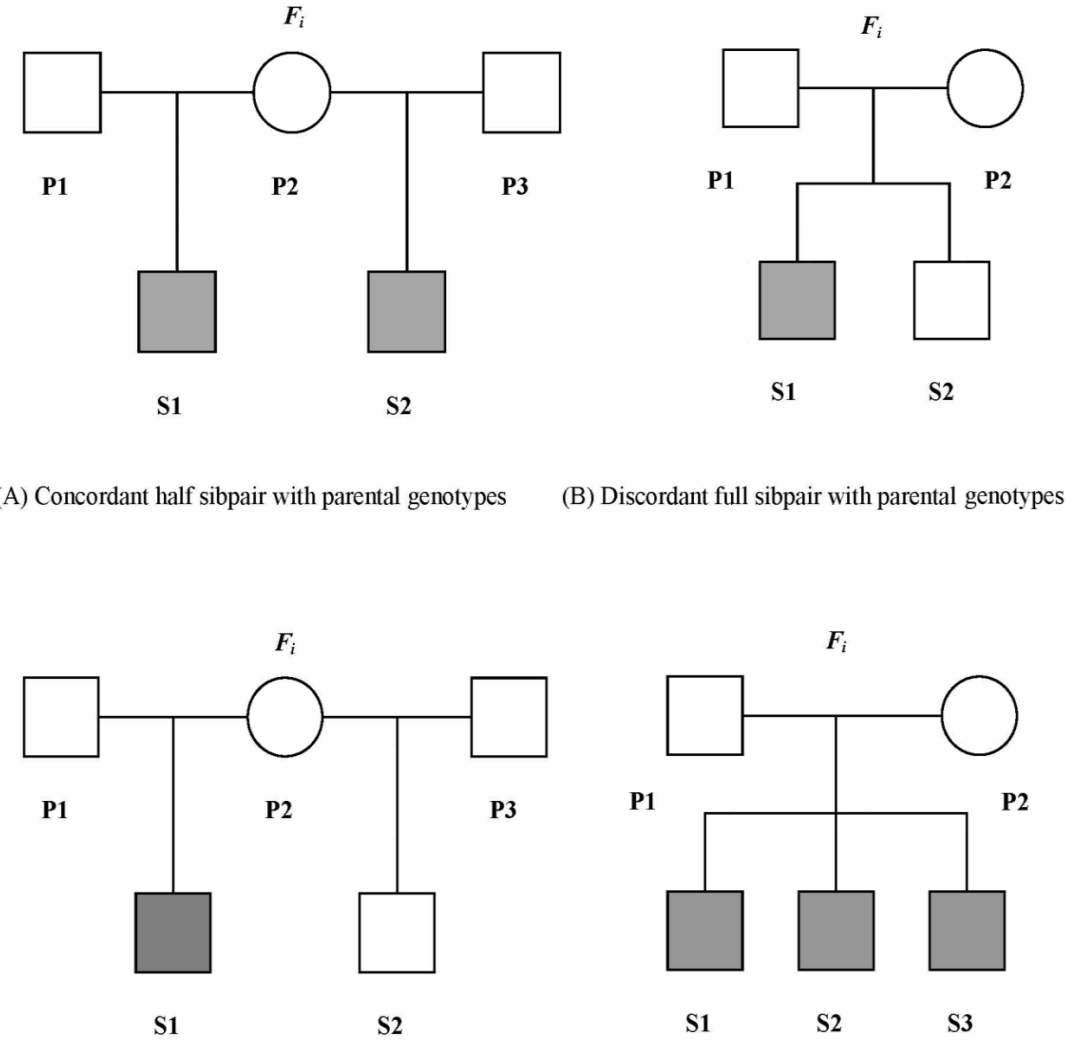

Figure 6. Pedigree Structures

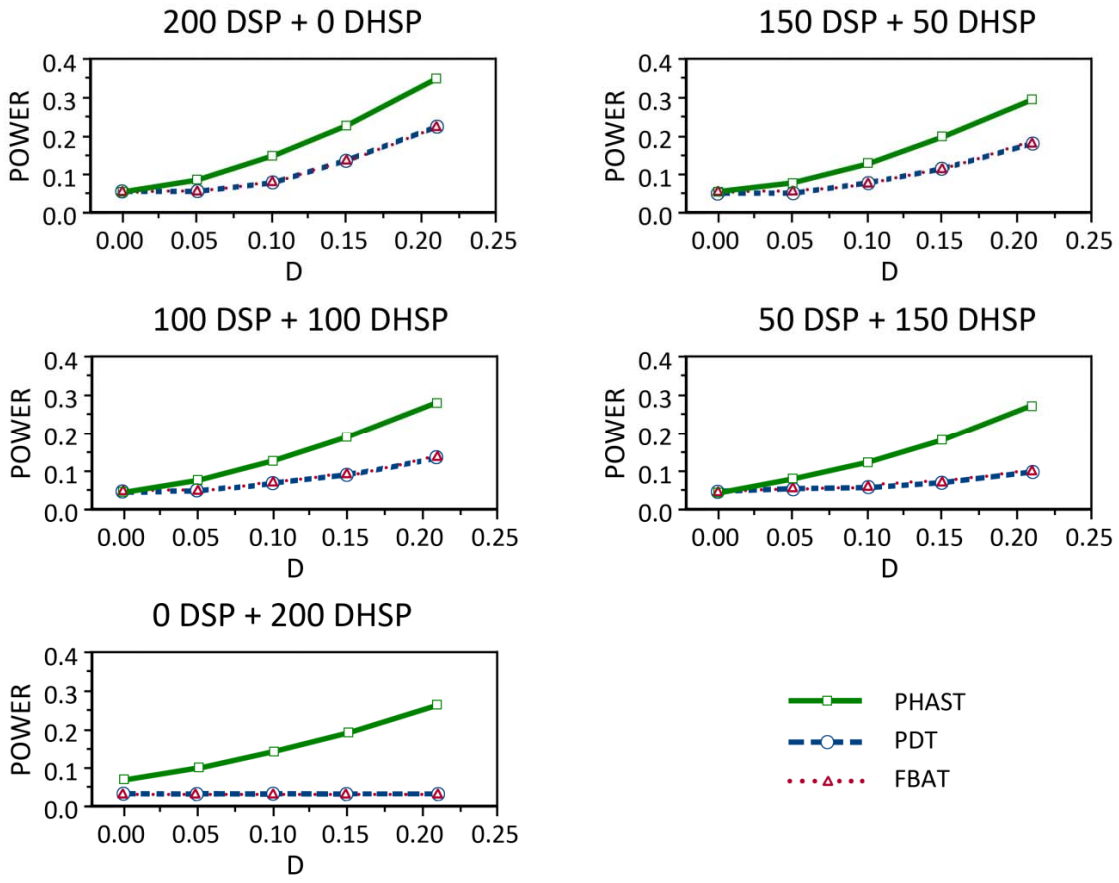

Figure 7. Power comparison among different association methods under different ratios of discordant full sibpairs (DSP) to discordant half sibpairs (DHSP) without parental data in both cases.

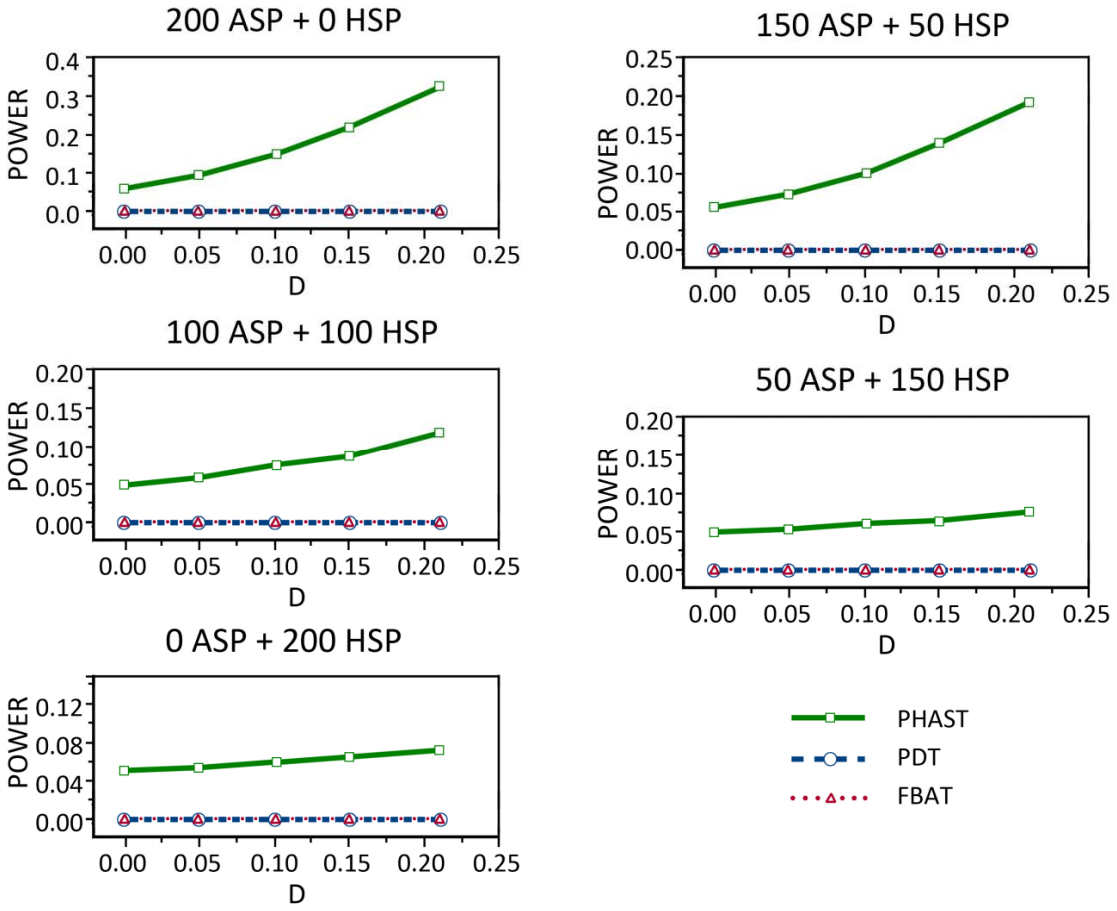

Figure 8. Power comparison among different association methods under different ratios of concordant full sibpairs (ASP) to concordant half sibpairs (HSP) without parental data in both cases.

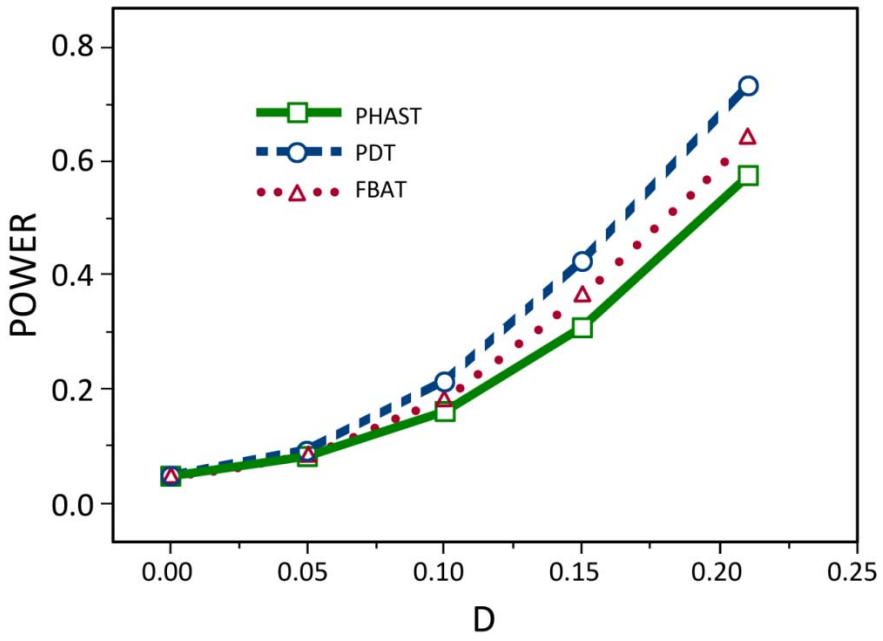

Figure 9. Power comparison among different association methods for 200 known parental genotypes nuclear family with three affected siblings under additive genetic model. Both maker and disease allele frequencies were 0.3. The heritability was 0.1 and GRR2 = 2.0.

CHAPTER 4

Association Test for Family and Unrelated Samples on Quantitative Traits: a semi-parametric mixed model

Abstract

Population and familial samples are two primary data structures frequently used for association studies. When samples in either or both types of data are insufficient, independently analyzing population or family data may result in poor statistical power. To combine population and familial samples as a single dataset, one needs to consider the correlations between relatives in the family dataset and population stratification from unrelated samples. In this Chapter, we propose a semiparametric additive mixed model (SAMM) for association tests using such combined dataset for quantitative traits. Our method doesn't require restrictive assumptions like HWE or rare diseases. It allows general pedigrees with or without parental genotypes, and accounts for population stratification if it exists. The performance of the SAMM was examined through simulations in data sets containing general pedigree and unrelated samples with admixture population. The results suggested that our new method has valid type I error rates and sufficient power under varied data structures.

4.1. INTRODUCTION

The genetic association analysis has served as a promising and useful tool for searching disease variants or modifiers for qualitative or quantitative traits of interest. Population and familial samples are two primary data structures for association studies, and are often used independently. Much of pros and cons of these two study designs have been discussed in Chapter 1.

In practice, one may have insufficient of samples in either or both types of data. This may result in poor statistical power if population and family data are analyzed independently. Although it is still manageable to analyze both related and unrelated data by conventional statistical methods, some underlying issues may not be properly considered. We decide to develop an association method which can properly cope with such mixed samples. A few challenges for this method development is that we need to consider the correlations between relatives in the family dataset (Spielman and Ewens, 1996) and population stratification from unrelated samples simultaneously.

Nagelkerke, et al. (2004) used a likelihood-based approach to pool family trios and unrelated cases and controls. Although easily implemented, their strategy requires strong assumptions of Hardy-Weinberg equilibrium (HWE), random mating, and a multiplicative model of allele effect on disease. Epstein, et al. (2005) proposed an

approach SCOUT that modifies the Nagelkerke's (2004) method to allow for more flexible modeling of allelic effects and less restrictions about the distribution of parental mating-types and genotypes, but it is only applicable to rare diseases. Later on, Thornton (2007), Guo, et al. (2009), and Dudbridge (2008) proposed methods to accommodate both types of data. However, Thornton's (2007) method, SCOUT (Epstein, et al., 2005), CHRR (Guo, et al., 2009), and UNPHASED (Dudbridge, 2008) all make an assumption that no population stratification existed in the data (i.e. homogeneous population sampling). Therefore, one will need to examine the population structure first and exclude samples that are in other subpopulations. This may result in the sample size reduction. A family-case-control (FamCC) method developed by Zhu, et al. (2008) took account of population stratification and could use nuclear families with multiple siblings. One caveat of the FamCC method is that it requires parental genotype data, which is often inaccessible for late-onset diseases.

Here, we propose a semiparametric additive mixed model (SAMM) to examine associations between candidate markers and quantitative traits. Our method doesn't require restrictive assumptions like HWE or rare diseases, allows flexible general pedigrees (combination of different types of sibships or trios) with or without parental genotypes, and accounts for population stratification if it exists. A random variable in the SAMM is used to account for the pedigree correlation. The SAMM controls for

population stratification through a smooth function with a genetic background variable. We examine the performance of the SAMM through simulations in data sets containing general pedigree and unrelated samples with admixture population. The slm function in R package "assist" (Wang, 2010) was used for fitting our semiparametric linear mixed effects model.

4.2. METHOD

Consider a biallelic marker with alleles A_1 and A_2. Suppose that a data consists of N independent units. We denote the notation of Y_{ij} (i = 1, …, N, j = 1, … , n_i) for the quantitative phenotype of the jth individual in the ith unit. Assume that $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$ is the vector of quantitative traits for the ith unit. There are n_i members in the ith independent units. We assume that the data with N independent units was made up of n pedigrees and m unrelated samples ($N = n+m$). Therefore, the total number of individuals in the data is $\sum_{i=1}^{N} n_i = \sum_{i=1}^{n} n_i + m$. For the unrelated sample, because there is only one member in the ith unit, $Y_i = (Y_{i1})$. Let's take one example. Assume a pseudo data containing 2 trios and 2 unrelated samples. The quantitative traits of these 4 (=N) independent units can be written as $Y_1 = (Y_{11}, Y_{12}, Y_{13})$, $Y_2 = (Y_{21}, Y_{22}, Y_{23})$, Y_3, and Y_4. The total number of individuals in this data is 8.

The relationship between the candidate marker and the quantitative trait Y_i can be formulated by a semi-parametric mixed model:

$$Y_i = \mu(t_i) + G_i\alpha + Z_i b_i + e_i, \qquad (4.1)$$

where $\mu(t_i)$ is an unknown smooth function of the genetic background variable t_i. G_i and Z_i are design matrices for α and b_i respectively. α is the fixed effect and $\alpha = \alpha_a a_{ij} + \alpha_d d_{ij}$, where a_{ij} = (# of A_1 alleles -1), α_a is the additive genetic effect, d_{ij} = I (genotype g_i = A_1A_2), and α_d is the dominance genetic effect. e_i is the residual and $e_i \sim N(0, \sigma_e^2 I_{n_i \times n_i})$. b_i is a random effect and $b_i \sim N(0, D_i(\theta))$. D_i is the genetic variance covariance matrix, has elements

$$D_{ijk} = \begin{cases} \sigma_a^2 + \sigma_G^2 & If \;\; j = k \\ \pi_{ijk}\sigma_a^2 + 2\phi_{ijk}\sigma_G^2 & If \;\; j \neq k \end{cases},$$

where σ_a^2 is the additive genetic variance due to the members between families, σ_G^2 is the polygenetic variance, π_{ijk} is the proportion of alleles shared IBD at marker locus between individuals j and k in pedigree i, and ϕ_{ijk} is the kinship coefficient between individuals j and k in pedigree i, and the set of parameters, $\theta = \{\sigma_a^2, \sigma_G^2, \pi_{ijk}, \phi_{ijk}\}$.

For the null hypothesis of no association between the maker and trait, we are interested in testing $H_0 : \alpha_a = 0$ versus $H_a : \alpha_a \neq 0$. We also care about estimating the additive genetic variance σ_a^2. The amount of the variability reflects the degree of variation in the QTL.

Estimate genetic background variable t_i:

The method EIGENSTRAT (Price, et al., 2006) was taken here to estimate genetic background variables t_i. t_i is the first principal component score of the *i*th sample. The EIGENSTRAT software uses principal components analysis to detect sample structures and has been widely used in many studies (See Section 1.3.3).

Estimate parameters (α, θ, and σ_e^2) and test for association :

By using *K*th-order smoothing spline (Zhang and Lin, 2003), the semi-parametric model has a linear mixed model form (APPENDIX). We can get the maximum likelihood estimate (MLE) for the fixed effects α and apply a score test for testing $H_0 : \alpha_a = 0$ versus $H_a : \alpha_a \neq 0$. The variance components (θ , σ_e^2) can be estimated by the restricted maximum likelihood (REML).

4.3. COMPUTER SIMULATION

A series of computer simulations were used to examine the type I error and power of the new method. We assume bi-allelic marker A (A_1 and A_2) with population frequencies p_1 and p_2, and a bi-allelic QTL Q (Q_1 and Q_2) with population frequencies q_1 and q_2. The linkage disequilibrium D is defined as the deviation between observed and expected

haplotype frequency between marker and QTL alleles, for instance, $D = \Pr(A_1Q_1) - p_1q_1$, where $\Pr(A_1Q_1)$ is population haplotype frequency for A_1Q_1. Traits resulting from three QTL genotypes, Q_1Q_1, Q_1Q_2, and Q_2Q_2, are assumed to follow normal distributions. We assumed the quantitative traits Y follow normal distributions with corresponding mean and variance. Parameters used in the simulations are listed in Table 1.

We simulated both marker and QTL allele frequencies of 0.2 for family samples; 0.1 and 0.3 for unrelated samples in population 1 and population 2. Given allele frequencies of the marker A and QTL Q, and linkage disequilibrium D, four population haplotype frequencies for marker A and QTL Q can be derived by $P(A_1Q_1) = p_1q_1 + D$, $P(A_1Q_2) = p_1q_2 - D$, $P(A_2Q_1) = p_2q_1 - D$, and $P(A_2Q_2) = p_2q_2 + D$.

For family samples, the haplotypes of each parent were simulated based on the given haplotype frequencies. We randomly drew one haplotype from each parent to form two haplotypes for each offspring. For unrelated samples in different populations, the haplotypes of each individual were simulated based on the given haplotype frequencies according to the corresponding allele frequencies. QTLs were simulated with additive (k = 0), dominant (k = 1), recessive (k = -1), and overdominant models ($k > 1$ or $k < -1$).

Our simulation studies evaluated different combinations of various data structures including nuclear families with one or two sibs, unrelated samples in population 1, and

unrelated samples in population 2 (Table 10). For each simulation study, 1,000 replicates were generated to estimate type I error and statistical power. The type I error were estimated under the cases of no association between marker and QTL (correlation coefficient $r^2 = 0$) and statistical power was estimated for $r^2 \neq 0$. We used 0.05 as the significance level for all estimates.

4.4. SIMULATION RESULTS

The type I error and statistical power were evaluated for the SAMM method. We compare the type I error to the FBAT method. Table 11 presented the type I error rates for 200 trios and 200 unrelated samples simulated under different genetic models (dominant, additive, and recessive). The type I error estimates for most tests are close to the nominal significance level of 0.05. The exception is for the recessive model in these datasets (P = 0.060). This is probable the result of small observations for these data. In most cases, we found that the FBAT has inflated type I error estimates. This result was expected, because the FBAT isn't designed to address mixed samples with admixture population.

The statistical power for the SAMM method was also evaluated for all combinations of genetic models, parameters, and data structures. We present only the results of additive model here (Figure 10), because similar power pattern was found in dominant, additive,

and recessive models. Generally, power is sensitive to the degree of correlation coefficient (r^2). In all cases, the power increases when the (r^2) between marker and quantitative trait locus increases. Both SAMM and FBAT provide reasonable power when the trait and marker locus are in perfect disequilibrium (r^2= 1). It seems that the FBAT has a little more power than the SAMM in many cases. However, it is necessary to point out that type I error rates in the FBAT test are inflated, the power of the FBAT method could lead to misinterpretation when the data with mixed family and unrelated samples.

4.5. DISCUSSION

In many genetic association studies, data contains both population and familial samples. When properly considered, better power can be achieved by combining both types of data. Current available methods all have strong assumptions that limit the practical application. Moreover, most of developed methods for combining data suit for qualitative traits. In this Chapter, we proposed the SAMM approach to test for associations between candidate markers and quantitative traits. The simulation results showed that the SAMM method does not cause inflated type I error when family and unrelated samples are combined, and has correct type I error rates under admixture data structures. The simulation also showed that the SAMM method provided sufficient power under varied data structures.

So far, we have developed and tested the validity of the SAMM method through a series of computer simulations. We will perform meta-analysis on the p-values obtained from family and population datasets to study the effect of combining both family and population data in one test. We will also compare the SAMM with other well-developed methods such as QTDT, GEE, and FamCC to evaluate the performance of our method. Finally, the SAMM will be applied for real data analysis.

Some future extensions for the SAMM method will also be planed. For the qualitative trait problem, we can generalize model (4.1) as $g(Y_i) = \mu(t_i) + G_i\alpha + Z_i b_i + e_i$, where g is a link function. As Y is the binary or binomial response, like disease status for example, we can denote $g(Y_i)$ = logit (Y_i) and then perform the analysis as the quantitative traits case. When there is a concern of the gene-environment dependence in the study population, we will address this problem by modifying model (4.1) with additional environmental variables as covariates to detect the gene-environment effects.

In summary, the SAMM method can serve as a flexible and robust tool for association studies with data including family and unrelated samples. Nowadays, GWAS dataset and many other researches are based on both family and population designs. We hope that the SAMM can be additional tool for researchers to facilitate and speed up association studies and add insights into the genetic association study of human complex

disease.

Acknowledgements

This work was supported by an investigator initiative research grant (IIRG-05-14708) from Alzheimer's association. We are grateful for generous funding.

References

Dudbridge, F. (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data, *Human heredity*, **66**, 87-98.

Epstein, M.P., *et al.* (2005) Genetic association analysis using data from triads and unrelated subjects, *The American Journal of Human Genetics*, **76**, 592-608.

Guo, C.Y., *et al.* (2009) Combined haplotype relative risk (CHRR): a general and simple genetic association test that combines trios and unrelated case - controls, *Genetic epidemiology*, **33**, 54-62.

Nagelkerke, N.J.D., *et al.* (2004) Combining the transmission disequilibrium test and case–control methodology using generalized logistic regression, *European journal of human genetics*, **12**, 964-970.

Price, A.L., *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies, *Nature genetics*, **38**, 904-909.

Spielman, R.S. and Ewens, W.J. (1996) The TDT and other family-based tests for linkage disequilibrium and association, *American Journal of Human Genetics*, **59**, 983.

Thornton, T. and McPeek, M.S. (2007) Case-control association testing with related individuals: a more powerful quasi-likelihood score test, *The American Journal of Human Genetics*, **81**, 321-337.

Wang, Y. (2010) *Smoothing Splines: Methods and Applications*. Chapman & Hall/CRC.

Zhang, D. and Lin, X. (2003) Hypothesis testing in semiparametric additive mixed models, *Biostatistics*, **4**, 57.

Zhu, X., *et al.* (2008) A unified association analysis approach for family and unrelated samples correcting for stratification, *The American Journal of Human Genetics*, **82**, 352-365.

APPENDIX

Appendix

The *K*th-order smoothing spline estimator $\mu(t_i)$ can be expressed as

$$\mu(t_i) = \sum_{k=1}^{K} \delta_k \gamma_k(t) + \sum_{l=1}^{r} c_l R(t, t_l^0),$$

where $t^0 = (t_1^0, t_2^0, \ldots, t_r^0)$ is a vector of ordered distinct t_i and by μ the vector of $\mu(t)$ evaluated by t^0 (without loss of generality, assume $0 < t_1^0 < \cdots < t_r^0 < 1$).

$$\gamma_k(t) = \frac{t^{k-1}}{(k-1)!},$$

$$R(t,s) = \frac{1}{\{(K-1)!\}^2} \int_0^1 (s-u)_+^{K-1} (t-u)_+^{K-1} du .$$

μ can also be represented as following:

$$\mu = T\delta + \Sigma c, \qquad (4.2)$$

where T is an $r \times K$ matrix with the (*l*,*k*)th element equal to $\gamma_k(t_i^0)$, Σ is a positive matrix with the (*l*,*k*)th element equal to $R(t_l^0, t_k^0)$, $\delta = (\delta_1, \ldots, \delta_K)^T$ is a vector of fixed effects, $c = (c_1, c_2, \ldots, c_r)^T \sim N(0, \tau\Sigma^{-1})$ is a vector of random effects with $\tau \geq 0$ being the inverse of the smoothing parameter for the smoothing spline estimate $\mu(t)$.

According to the mixed effect representation (4.2), semi-parametric additive mixed model (4.1) can reduce to a linear mixed model $Y = X\beta + Bc + Zb + e_i$, where $X = (NT, G)$, $B = N\Sigma$, N is the incidence matrix that maps $\{t_i\}$'s into t^0.

$\beta = (\delta^T, \alpha^T)^T$ are the new fixed effects, c and $b = (b_1^T, \ldots, b_{n+m}^T)^T$ are independent new

random effects. Therefore, we can get the MLE of the fixed effects $\hat{\beta}$ and apply a score test for testing the null hypothesis.

For given variance components ($\theta^T, \tau, \sigma_e^2$), the log-likelihood function of β is

$$l(\beta, c; Y) = -\frac{1}{2}\log|V| - \frac{1}{2}(Y - X\beta)^T V^{-1}(Y - X\beta),$$

where $V = diag(V_1, V_2, \ldots, V_{n+m})$ and $V_i = Z_i D_i Z_i^T + \sigma_e^2 I_i$.

The estimator $\hat{\beta}$ and $\hat{c}$ are got by solving

$$\begin{bmatrix} X^T W X & X^T W B \\ B^T W X & B^T W B + \lambda I \end{bmatrix} \begin{bmatrix} \beta \\ c \end{bmatrix} = \begin{bmatrix} X^T W Y \\ B^T W Y \end{bmatrix},$$

where $W = V^{-1} = diag(W_1, \ldots, W_{n+m})$ and $\lambda = 1/\tau$.

The smoothing parameter τ and variance components θ can be estimated by the restricted maximum likelihood (REML). The REML log-likelihood of (τ, θ) is

$$l_R(\tau, \theta; Y) = -\frac{1}{2}\log|V_*| - \frac{1}{2}\log|X^T V_*^{-1} X| - \frac{1}{2}(Y - X\hat{\beta})^T V_*^{-1}(Y - X\hat{\beta}) \quad ,$$

where $V_* = \tau B B^T + V$.

Table 10. Parameters used in the simulation study

Parameters used in the simulation study	
Marker allele frequency for family samples $P(A_1)$	0.2
Marker allele frequency in population 1	0.1
Marker allele frequency in population 2	0.3
QTL allele frequency for family samples $P(Q_1)$	0.2
QTL allele frequency in population 1	0.1
QTL allele frequency in population 2	0.3
Scale of Dominant model $(k = d_q / a_q)^*$	-2, -1.5, -1, 0, 1, 1.5, 2
Number of samples simulated	
(Trios, Population 1, Population 2)	(200, 100, 100)
	(100, 250, 250)
	(0, 400, 400)
(Quads, Population 1, Population 2)	(100, 200, 200)
	(50, 300, 300)
Heritability (H^2)	0.1
Number of iterations	1,000

* a_q : additive effect; d_q : dominant effect.

Table 11. Type I error rates for data simulated from 200 pedigrees (trios) and 200 unrelated samples under different genetic models

Pedigree :P(M)=P(D)=0.2;

Unrelated: Population 1: P(M)=P(D)=0.1;

Population 2: P(M)=P(D)=0.3

Model	SAMM	FBAT
Recessive	0.061	0.062
Additive	0.045	0.055
Dominant	0.049	0.057

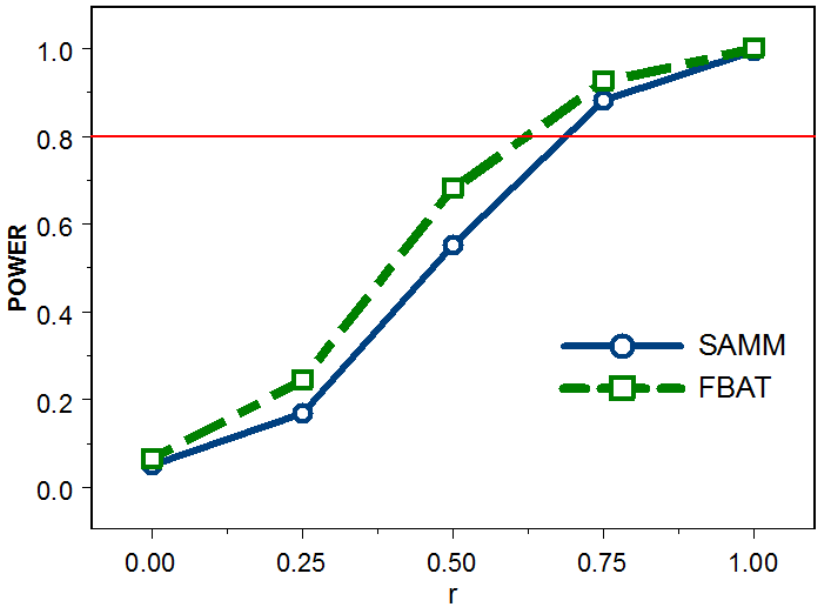

Figure 10. Power comparison between SAMM and FBAT test under 200 trio families, 100 unrelated samples in population 1, and 100 unrelated samples in population 2 in additive model. The heritability was 0.1 and additive effect a = 2.

CHAPTER 5

Summary

Association methods have played a central role in candidate gene studies for human complex diseases. Every year, there are lots of public investments in genetic association studies. Efficient utilization of all data available can not only gain more statistical power but also accelerate research progress by saving time and money. It is especially helpful for GWAS which requires large samples to get significant results.

In summary, we extended the allele-based MK method to multi-generation families and developed a genotype association method for quantitative trait. Our simulation studies showed that both geno-EMK and allele-EMK tests reach correct type I error rate. The statistical power is comparable between the global geno-EMK test and the allele-EMK test. The geno-EMK has the advantage of offering genotype specific association results and can be more powerful for some genetic models such as the recessive and overdominant models. Our simulations and data analysis results indicate the geno-EMK global test maintains nominal significance level even in the cases that the type I error of a particular genotype test is too conservative. Therefore, we recommend using the geno-EMK global test as an initial overall assessment to support the evidence of individual genotype association. We have developed an EMK program that implements both allele-EMK and geno-EMK methods for real data analysis. We have also demonstrated the EMK program by testing the association between the SNP data of GSTO1 and GSTO2 and age-at-onset of Alzheimer disease.

We also proposed the PHAST approach to test for association with the inclusion of half-siblings or more than two affected sibling data. The PHAST can also handle the conventional family data structures such as trios, ASP with parents, and DSP. The simulation results showed that the PHAST method has correct type I error under varied family structures. Simulations revealed that the PHAST could have more power than the PDT and the FBAT when the sample size of half-siblings increases, especially for the families without parental genotypes. Therefore, the PHAST can be considered as a useful tool for studying late onset disease.

We also proposed the SAMM approach for quantitative trait association study. We have developed and tested the validity of the SAMM method through a series of computer simulations. The simulation results showed that the SAMM method does not cause inflated type I error when family and unrelated samples are combined, and has correct type I error rates under admixture data structures. The simulation also showed that the SAMM method provided sufficient power under varied data structures.

We recognize that a more in-depth evaluation of our proposed SAMM needs to be done. In particular, we will compare both type I and power to other approaches in addition to FBAT. For instance, meta-analysis has been used widely to combine dataset. One idea is that we can compare the statistical power between SAMM and the meta-analysis, in which a meta p-value will be obtained from p-values obtained from

family and population datasets, respectively. We will also compare SAMM with other methods such as QTDT, GEE, and FamCC to evaluate the performance of our method. Finally, our goal is to implement the SAMM for real data analysis. Some future extensions for the SAMM method will be planned as well. For the qualitative traits, we can generalize the model as a logistic regression SAMM and perform the analysis as the qualitative traits case. When there is a concern of the gene-environment dependence in the study population, we will address the problem by modifying the SAMM with additional environmental variables as covariates to detect the gene-environment effects.