# VARIANCE COMPONENT SCORE STATISTICS FOR QTL MAPPING

by

**Samsiddhi Bhattacharjee**

BStat, Indian Statistical Institute, Kolkata, India, 2002

MStat, Indian Statistical Institute, Kolkata, India, 2004

Submitted to the Graduate Faculty of
the Department of Human Genetics
Graduate School of Public Health in partial fulfillment
of the requirements for the degree of

**Doctor of Philosophy**

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

**Samsiddhi Bhattacharjee**

It was defended on

**July 3'rd, 2008**

and approved by

Dissertation Advisor: Eleanor Feingold, Ph.D., Associate Professor, Depts. of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Daniel E. Weeks, Ph.D., Professor, Depts. of Human Genetics and Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Michael M. Barmada, Ph.D., Associate Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Bernard J. Devlin, Ph.D., Associate Professor, Department of Psychiatry, School of Medicine, University of Pittsburgh

Eleanor Feingold, Ph.D.

## VARIANCE COMPONENT SCORE STATISTICS FOR QTL MAPPING

Samsiddhi Bhattacharjee, PhD

University of Pittsburgh, 2008

ABSTRACT

Variance Components based models are commonly used for linkage and association mapping of quantitative traits. Score Tests based on these models are generally more robust to various modeling assumptions than the corresponding likelihood ratio tests. They are also computationally much simpler than the likelihood ratio tests, making them the natural choice for whole genome scans, which have become increasingly common with the emergence of high-throughput genotyping technologies. However the popularity of score statistics have been limited, due to several practical issues, such as lack of availability of software and guidelines for choice of score statistic variants. In this dissertation, we develop novel score statistics for both linkage and association mapping, elucidate the theoretical properties of these and of the existing variants, and also compare some of the existing and proposed score variants using simulation. Analytical arguments and simulation results are used to develop guidelines for choice of appropriate score variants under different practical situations.

In this dissertation, we are primarily concerned with identifying robust and powerful score statistics for detecting genetic susceptibility loci for complex diseases by mapping underlying quantitative phenotypes. Unlike Mendelian disorders, complex diseases in humans typically have a large number of modest genetic effects, which cumulatively have a significant impact on the disease. The work in this dissertation is aimed at maximizing the power of genome scans to detect more of these small genetic effects. This is of considerable public health significance, as the identified genetic variants can be followed up to gain important insights into the etiology of the disease, which can further lead to development of screening tests and preventive and therapeutic interventions for complex diseases.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# PREFACE

I am grateful to my advisor Dr. Eleanor Feingold, for her guidance, support and encouragement over the last four years. Her insightful comments and suggestions and her practical approach were extremely helpful in providing direction to my research. Most importantly, she ensured that I met all the deadlines without ever having to push me too hard, which is a considerable feat given my time management skills. Above all, her efficiency and work ethics motivated me to work hard.

I would like to thank Dr.Daniel E. Weeks, Dr. Michael M. Barmada and Dr. Bernie Devlin for serving on my proposal and dissertation committees. Their comments and suggestion helped greatly to improve this dissertation. Particularly, I would like to thank Dr. Daniel E. Weeks, for being an excellent co-mentor. I learnt a lot while working with him on some of the projects. I am grateful to Dr. Weeks, Dr. Candace Kammerer and Dr. Feingold for their help during my job search.

I would like to acknowledge all my friends in Pittsburgh, both in the Department and outside for making the last four years enjoyable. I should particularly mention my office-mates, Ankur, John and Jeesun; former postdocs Indranil, Anbu and co-workers Chia-Ling and Nandita; and some friends outside the department- Pradipta, Suman, Shaswati, Debapriti and Arindam. I should also thank Bodhisattva Sen for some helpful discussions on my research and also for his help and advice during my job search.

I am grateful to Dr. Partha P. Majumder and Dr. Saurabh Ghosh for being my first mentors in the area of Statistical Genetics and also to Dr. Daniel E. Weeks and Dr. Eleanor Feingold for supporting my graduate study and training through their grants. Last but not the least, I am greatly indebted to my family and relatives back in India for their constant love and support. This dissertation is dedicated to my parents for their patience, support and untiring encouragement throughout my academic career.

# 1.0 INTRODUCTION

## 1.1 STRUCTURE OF THIS DISSERTATION.

The broad aim of this dissertation is to identify score statistics for linkage and association mapping of quantitative trait loci (QTLs) that are powerful and at the same time robust to various modeling assumptions. It is broadly divided into four chapters. Below we briefly outline specific issues discussed in each chapter.

In the present chapter 1 - "Introduction," we give some background on QTL mapping and on some of the standard approaches to linkage and association mapping of QTLs. We define score statistics and discuss how they address some of the limitations in the existing approaches for QTL mapping.

In chapter 2 - "Score Tests for QTL Mapping," we describe some of the standard genetic models used for quantitative traits, and discuss the validity of the assumptions required to motivate these models. We also propose a new model that incorporates both linkage and association parameters and as such allows the derivation of both linkage and association tests. We describe the score tests under the standard genetic models as well as those under the proposed model. We also demonstrate some robustness properties of score tests to genetic models and to selected sampling.

In chapter 3 - "Score Tests for Linkage Analysis," we restrict our attention to the standard Variance Components based score statistic for linkage. Several variants of this score have been proposed in the literature with little or no guidelines regarding the best choice in any given scenario. We categorize the existing variants as well as propose some new score variants. We conduct an extensive simulation study to compare the existing and the proposed variants, in terms of robustness of type I error and power under different trait distributions, ascertainment schemes and genetic models. Based on analytical arguments and simulations results we propose general guidelines for choosing appropriate score variants based on study design, normality of phenotype and other considerations.

In chapter 4 - "Score Tests for Association Analysis," we develop novel score-based statistics for family based association mapping of quantitative traits. These statistics try to use maximal information from a family while protecting against stratification. Using simulations, we compare the proposed score statistics against some standard ones. We consider modifications of these statistics to handle missing genotypes and the presence of "known linkage" and derive some recursive and closed-form expressions for conditional genotype moments required in computing these statistics. Finally, we discuss some preliminary ideas to construct score statistics for association that are free of nuisance trait parameters.

In chapter 5 - "Discussion and Future Work," we highlight some of the limitations of the methods described in this dissertation and discuss possible directions of future research on mapping of quantitative traits using score statistics.

## 1.2 QTL MAPPING BACKGROUND

A quantitative trait is any phenotype expressed as a continuum of values, as opposed to binary traits, which take on two values (e.g., affected/unaffected or affected/unknown). Most methods developed for mapping either kind of trait can be used for the other by assigning numerical codes to the binary traits or dichotomizing the quantitative traits, although the specific assumptions of the methods may not be appropriate in all cases. Typically, in humans, most gene mapping methods adhere to the distinction rigorously and, in general neither approach can claim to be superior to the other in all cases. Genes or susceptibility loci mapped using quantitative traits are termed as Quantitative Trait Loci (QTLs), although the QTLs can possibly be identical to loci mapped using the binary disease status, particularly when the disease is defined (or diagnosed) based on the underlying quantitative trait(s). An important advantage of using the underlying quantitative phenotype(s) (instead of the binary end-point trait) is that they are not affected by misdiagnoses due to subjective diagnostic criteria. Also, coding of sub-clinically affected individuals as “unaffected” or “unknown” can adversely affect the power to detect linkage or association, which can be a concern particularly for late-onset diseases. Quantitative traits, on the other hand, allow modeling of the complete observed variation of the phenotype without losing any information. Sometimes, this comes with a price to pay in terms of too many parametric assumptions, for example about the distribution of the phenotype. Another concern may be that modeling the full variation can sometimes lead to genes that control normal variation in the phenotype and that are not involved in the etiology of the disease. This is less of a concern when dealing with families ascertained via affected probands than when dealing with randomly selected families. Sometimes multivariate (quantitative) phenotypes can better approximate the true disease status than a single quantitative trait. In this dissertation, we will confine our discussion to QTL mapping methods with a single continuous phenotype. The phenotype under consideration can be a neutral one such as height, or one underlying a disease such as blood sugar level.

In humans, QTL mapping methods can be classified into two broad classes, namely “linkage” and “association” mapping methods. Linkage mapping tries to locate genes controlling the trait by detecting the cosegregation (limited or no recombination events during meiosis) of a gene and a nearby marker locus within a family. Association mapping, on the other hand, tries to detect coincidence (linkage disequilibrium or LD) of a disease allele with a marker allele in a population, caused by an interplay of historical events such as mutation, founder effects, recombination, admixture etc acting on the population as a whole. While linkage by its definition can only be detected using family data, association mapping can be both “population-based” and “family-based.” In this dissertation, we will primarily be concerned with QTL mapping with family data, using both “linkage” and “association” techniques.

Linkage disequilibrium usually operates over small distances, whereas linkage can theoretically exist between loci located at opposite ends of a chromosome. Hence, association mapping is usually perceived as a more powerful approach for localizing disease genes. But association (or LD) can occur because of factors other than proximity with the disease locus, such as population stratification, admixture or recent mutation or founder effects, which as a result can be potential confounders. Presence of strong linkage, on the other hand, directly correlates with proximity of disease loci. Also, linkage can be detected using fewer markers than association, which typically requires a very dense marker map. This issue is becoming less and less important with emergence of high throughput genotyping technologies. Still, dense genome-wide scans are not yet easily affordable to all investigators. Hence, linkage analysis remains a popular way to narrow down regions on the genome (particularly in absence of biological candidates), often followed by association mapping in the identified candidate regions. At the same time, genome-wide association (GWA) studies have become practical for some investigators and have started to replace the two-step linkage followed by association approach. GWA studies must ensure that the association mapping statistics be designed to protect against confounding factors such as admixture, while being computationally feasible. In this dissertation, we will attempt to identify "linkage" and "association" mapping statistics that are computationally fast, protect against confounders, and are relatively robust to modeling assumptions.

## 1.3 STANDARD METHODS FOR LINKAGE AND ASSOCIATION MAPPING

Linkage analysis methods for quantitative traits can be broadly classified into two categories namely "likelihood-based" and "regression-based." The most popular likelihood-based method performs a likelihood ratio test assuming the Variance Components model (e.g., Lange *et al.* 1976; Falconer 1981; Hopper and Mathews 1982; Amos 1994; Almasy and Blangero 1998), and is known as the Variance Components (VC) approach. This model is discussed in detail in Chapter 2. It is a powerful and flexible approach, as the model can include additive and dominance effects of the major gene as well as other random effects such as polygenic effects and other fixed effects such as covariates. Because of the likelihood setup, arbitrary hypotheses can be tested. In spite of the popularity of this method, it has some disadvantages, namely that it can be computationally intensive, particularly for dense genome scans. More importantly, it is a valid test (i.e, has a correct type I error) only when the assumption of normality is correct and also suffers from considerable loss of power when that assumption is violated (Allison *et al.* 1999). The limitations of the VC method are partly addressed by the regression-based methods. These methods usually exploit the correlation between some function of the trait values and the estimated marker IBDs (e.g., Haseman and Elston 1972; Sham and Purcell 2001; Sham *et al.* 2002) in the presence of linkage between the trait and the marker. These are generally computationally simple and are usually valid tests irrespective of the trait distribution. Unlike the VC approach, these methods are also robust in terms of type I error to selected sampling and to misspecification of trait parameters. However some of these methods such as the Haseman and Elston (HE) regression (Haseman and Elston 1972) are considerably less powerful than the VC, when the normality assumption holds.

Association mapping of QTLs can be population-based or family-based. Population based designs sample unrelated individuals from the population. The most popular population-based design is the "case-control" study design. In this design, the estimated frequency of the marker allele is compared between case and control individuals. In presence of association, i.e. when the marker allele controls the trait or is in LD with the trait predisposing allele, the estimated allele frequencies are expected to be different. This design is more popular for binary disease traits but can be applied to quantitative traits by assigning Case/Control status using thresholding of the continuous trait (e.g., Hegele *et al.* 1999). Population-based approaches by contrast use the full variation of the quantitative trait instead of thresholding (e.g., Allison 1997; Barrett 2002). Population-based approaches are powerful, require small sample sizes and are computationally fast. But since they may detect spurious association due to population substructure, it is not possible to separate spurious associations from real ones in this design.

Family-based designs for association mapping can be constructed to guard against confounding factors. The first family based design, the TDT (Transmission Disequilibrium Test) design, was proposed in the context of binary traits (Terwilliger and Ott 1992; Spielman *et al.* 1993). Ever since, most association mapping methods for QTL mapping have attempted to extend the TDT design for quantitative traits. One such popular method uses a likelihood-based approach similar to the VC (Fulker *et al.* 1999; Abecasis *et al.* 2000; Abecasis *et al.* 2001), implemented in the software QTDT (Abecasis *et al.* 2000). The other popular approach is a non-parametric approach originally proposed by Rabinowitz (1997) and implemented in the software FBAT (Horvath *et al.* 2001). There have been various extensions of the FBAT procedure (*e.g.,* Rabinowitz and Laird 2000; Laird *et al.* 2000; Whittemore and Halpern 2003). These statistics are similar to the regression-based linkage mapping approach and use the correlation between the marker allele transmissions and the phenotype in a way that protects against population stratification effects. Throughout this dissertation, we will use the abbreviations QTDT and FBAT for the methods of Fulker *et al.* (1999) and Rabinowitz (1997), respectively, and their subsequent extensions.

## 1.4 LIMITATIONS ADDRESSED BY SCORE STATISTICS

The score test (Rao and Poti 1946) is a computationally simpler alternative to the likelihood ratio test (LRT), and is asymptotically as powerful as the LRT for local alternatives (*i.e.*, alternatives close to the null hypothesis). Since local alternatives are usually harder to detect, local optimality almost always guarantees an overall powerful statistic. The computational simplicity of the score statistic comes from the fact that it only involves the first derivative of the log-likelihood at the null value of the parameter and does not require the Maximum Likelihood Estimate (MLE) of the parameter. Hence, unlike the LRT, it does not involve maximization of the likelihood and can often be expressed in simple algebraic form. The general form of the score test statistic for a single parameter can be written as follows.

$$T = \frac{S(X)}{\sqrt{n\ I(\theta_0)}} = \frac{\sum_{i=1}^{n} S(x_i)}{\sqrt{n\ I(\theta_0)}},$$

$$\text{where} \quad S(X) = \sum_i \frac{\partial log(L(x_i,\theta))}{\partial\theta}\Big|_{\theta=\theta_0}$$

$$\text{and} \quad I(\theta_0) = -E\left[\sum_i \frac{\partial^2 log(L(x_i,\theta))}{\partial\theta^2}\right]_{\theta=\theta_0}.$$

Here $X = (x_1, \ldots, x_n)$ denotes n independent data points, $\theta$ the unknown parameter, $\theta_0$ its null value and $L(x,\theta)$ is the likelihood function for a single observation. $S(X)$ is known as the "score" and $I(\theta)$ is known as the "information." The score statistic can be completely generalized to multiple parameters including nuisance parameters (Rao 1948). If $\theta$ denotes a vector of parameters of interest and $\nu$ denotes a vector of nuisance parameters, then the score test for $H_0 : \theta = \theta_0$ takes the form

$$T = S_{\theta_0,\hat{\nu}_0}(X)'\ I^{-1}(\theta_0,\hat{\nu}_0)\ S_{\theta_0,\hat{\nu}_0}(X),$$

$$\text{where} \quad S_{\theta_0,\hat{\nu}_0}(X) = \frac{\partial log(L(X,\theta,\nu))}{\partial\theta}\Big|_{\theta=\theta_0,\nu=\hat{\nu}_0}$$

$$\text{and} \quad I(\theta_0,\hat{\nu}_0) = -E\left[\frac{\partial^2 log(L(X,\theta,\nu))}{\partial\theta\ \partial\nu}\right]_{\theta=\theta_0,\nu=\hat{\nu}_0}. \tag{1.4.1}$$

Here $\hat{\nu}_0$ is the MLE of $\nu$ under the constraint $\theta = \theta_0$.

Apart from the attractive properties mentioned above, score statistics also possess another distinct advantage over the LRT, namely that they can be made robust to model violation. The score test statistic $T$ follows an asymptotically $N(0, 1)$ distribution under the null hypothesis. This follows from the specific form of the score statistic shown above and the Central Limit Theorem (CLT). This asymptotic distribution also holds when the assumed model is wrong (by the CLT), provided the denominator contains an empirically estimated variance of the score $S(X)$ instead of the theoretical information. So, the standard normal distribution is generally used to obtain cutoffs or p-values for the score test with an empirical denominator. These cutoffs remain valid, irrespective of the true model as long as the sample size is reasonably large. On the other hand, the significance of the LRT is usually assessed using the fact that asymptotically $-2\log(LRT) \sim \chi^2_1$. But this fact holds only when the assumed model is correct. Thus, the LRT has incorrect type I error whenever the assumed model is wrong unless an empirical null distribution is used, which can further increase its computational complexity. Score statistics however do not guarantee robustness of power; both score and LRT can incur considerable loss of power when the assumed model is grossly violated.

Score statistics become even more useful in the context of disease gene mapping, where it is often more powerful and convenient to sample affected individuals and their relatives. These individuals are usually at the extremes of distribution of the underlying quantitative trait(s). This distorts the distribution of the quantitative trait, making the LRT invalid even if the assumed model correctly represents the overall distribution of the trait. Depending on objectiveness and simplicity of the ascertainment criteria, it may or may not be possible to correct the LRT for ascertainment. On the other hand, the score test can be constructed in a manner (to be discussed in later chapters) such that they have correct type I error even for ascertained samples.

# 2.0 SCORE STATISTICS FOR QTL MAPPING

Before describing the different score statistics used for QTL mapping, we discuss below some of the genetic models under which these score statistics are generally derived.

## 2.1 GENETIC MODEL AT THE TRAIT LOCUS

### 2.1.1 Model Conditional on Trait Genotype

The main focus of this dissertation will be on variants of the score statistics, which were originally derived as efficient scores using a decomposition of the components of the variance of the quantitative phenotype. Below we outline a derivation of the variance-component decomposition using similar ideas as discussed in Tang (2000). An alternative derivation for general multiallelic traits can be found in Lange (2002).

Let us consider a quantitative trait $Y$ controlled by a biallelic major gene having alleles “D” and “d”. Let $p$ and $q = 1 - p$ denote the frequencies of the “D” and “d” alleles. The trait genotype $g_i$ is coded as 0, 1 and 2 for the genotypes “dd,” “dD” and “DD” respectively.
Let us assume the following model for phenotype conditional on the trait genotype.

$$y_i = m + a\ g_i + d\ 1_{\{g_i=1\}} + \epsilon_i, \tag{2.1.1}$$

where $m$ is the baseline effect of the “dd” genotype and “a” is the additive effect of one “D” allele and “d” is the dominance effect. Let $\mu$ and $\sigma^2$ denote the population mean and variance of the trait. The above model can be centered at the trait mean by rewriting it as

$$\begin{aligned}
y_i &= m + a\ g_i + d\ 1_{\{g_i=1\}} + \epsilon_i \\
&= \left[\ m + a\ E(g_i) + d\ E(1_{\{g_i=1\}})\ \right] + a\ [\ g_i - E(g_i)\ ] + d\ \left[\ 1_{\{g_i=1\}} - E(1_{\{g_i=1\}})\ \right] + \epsilon_i \\
&= \mu + a\tilde{g}_i + d\tilde{1}_{\{g_i=1\}} + \epsilon_i
\end{aligned}$$

where $\tilde{g}_i$ and $\tilde{1}_{\{g_i=1\}}$ are centered versions of $g_i$ and $1_{\{g_i=1\}}$. The residuals $\epsilon_i$ are assumed to be uncorrelated with the genotypes and to have mean zero and constant variance $\sigma^2_\epsilon$. The residuals can consist of polygenic effects, other major genes (unlinked and in linkage equilibrium with the QTL modeled above) and environmental effects. We can orthogonalize the above model using a Gram Schmidt orthogonalization procedure, with $< X, Y >= Cov(X, Y) = E(XY)$ as follows.

$$y_i = \mu + \left(a + d\ \frac{< \tilde{g}_i, \tilde{1}_{\{g_i=1\}} >}{< \tilde{g}_i, \tilde{g}_i >}\right) \tilde{g}_i + d \left(\tilde{1}_{\{g_i=1\}} - \frac{< \tilde{g}_i, \tilde{1}_{\{g_i=1\}} >}{< \tilde{g}_i, \tilde{g}_i >}\ \tilde{g}_i\right) + \epsilon_i$$

By noting that under Hardy Weinberg Equilibrium, $g_i \sim Bin(2, p)$, $1_{\{g_i=1\}} \sim Ber(2pq)$ and $Cov(g_i, 1_{\{g_i=1\}}) = 2pq(q-p)$, we obtain

$$\begin{aligned}
y_i &= \mu + [\ a + d\ (q-p)\ ]\ \tilde{g}_i + d\ \left[\ \tilde{1}_{\{g_i=1\}} - (q-p)\ \tilde{g}_i\ \right] + \epsilon_i \\
&= \mu + \alpha\ \tilde{g}_i + \delta\ \tilde{\tilde{1}}_{\{g_i=1\}} + \epsilon_i, \qquad (2.1.2)
\end{aligned}$$

where $\alpha = a + d\ (q-p)$, $\delta = d$ and $\tilde{\tilde{1}}_{\{g_i=1\}}$ is $\tilde{1}_{\{g_i=1\}} - (q-p)\ \tilde{g}_i$, the orthogonal projection of $\tilde{1}_{\{g_i=1\}}$ onto the linear subspace generated by $\tilde{g}_i$. This decomposition does not have any biological interpretation, but provides a mathematically convenient way of extracting the complete variability explained by a linear function of the number of trait alleles (i.e, $g$). The advantage of such a decomposition is that because of the orthogonality, there is significant mathematical simplicity in analyzing a joint model with both the additive and dominance parameters, as discussed later. It also gives an estimate of the actual loss of information when the dominance is ignored.

The total phenotypic variance $Var(Y) = \sigma^2$ can thus be decomposed into an additive genetic variance $\sigma^2_a$, a dominance variance $\sigma^2_d$ and a residual environmental variance $\sigma^2_\epsilon$ as follows.

$$\begin{aligned}
\sigma^2 &= \sigma^2_a + \sigma^2_d + \sigma^2_\epsilon, \\
\text{where}\ \sigma^2_a &= \alpha^2\ Var(\tilde{g}_i) = 2pq\ [a + d(q-p)]^2 \\
\text{and}\ \sigma^2_d &= \delta^2\ Var(\tilde{\tilde{1}}_{\{g_i=1\}}) = 4p^2q^2d^2.
\end{aligned}$$

The last equality can be obtained by using the relation

$$Var(X - E(X|Y]) = Var(X) - Var[\ E(X|Y)\ ] \ = \ <X,X> - \left(1 - \frac{<X,Y>}{<X,X><Y,Y>}\right).$$

Thus $\sigma_a^2$, $\sigma_d^2$ and $\sigma_\epsilon^2$ are known as the variance components of $Y$. The conditional on genotype model (2.1.2) is typically used in the context of association mapping of quantitative traits.

### 2.1.2 Model Conditional on Trait IBD

The variance components derived above can be used to decompose the phenotypic covariance conditional on IBD sharing. To derive this model, let us first consider two relatives with phenotypes $(y_1, y_2)$, genotypes $(g_1, g_2)$, environmental correlation $r$ and IBD sharing proportion $\pi$. Using the ITO matrices (Li and Sacks 1954), and the formula $E[f_1(g_1)f_2(g_2)|\pi] = E_{g_2}[f_2(g_2)E(f_1(g_1) \mid g_2, \pi)]$, it can be shown that

$$\begin{aligned} Cov(g_1, g_2 \mid \pi) &= 2pq\pi \\ Cov(1_{\{g_1=1\}}, g_2 \mid \pi) &= 2pq(q-p)\pi \\ Cov(1_{\{g_1=1\}}, 1_{\{g_2=1\}} \mid \pi) &= 2pq[(q-p)^2\pi + 2pq\ 1_{\{\pi=1\}}. \end{aligned} \qquad (2.1.3)$$

Using the above results and the definitions of $\tilde{g}$, $\tilde{1}_{\{g=1\}}$ and $\tilde{\tilde{1}}_{\{g=1\}}$, it is easy to show that

$$\begin{aligned} Cov(\tilde{g}_1, \tilde{g}_2 \mid \pi) &= 2pq\pi \\ Cov(\tilde{\tilde{1}}_{\{g_1=1\}}, \tilde{g}_2 \mid \pi) &= 0 \\ Cov(\tilde{\tilde{1}}_{\{g_1=1\}}, \tilde{\tilde{1}}_{\{g_2=1\}} \mid \pi) &= (2pq)^2\ 1_{\{\pi=1\}}. \end{aligned}$$

Thus, the unconditional orthogonal decomposition $\tilde{g}$ and $\tilde{\tilde{1}}_{\{g=1\}}$ of the genotypes remains orthogonal after conditioning on IBD. Hence it follows from model (2.1.2) that

$$\begin{aligned} Cov(y_1, y_2 \mid \pi) &= \alpha^2(2pq\pi) + \delta^2(4p^2q^2)1_{\{\pi=1\}} + Cov(\epsilon_1, \epsilon_2) \\ &= \sigma_a^2\ \pi + \sigma_d^2\ 1_{\{\pi=1\}} + r\ \sigma_\epsilon^2. \end{aligned}$$

The above "conditional on IBD" model based on variance component decomposition is used in the context of linkage analysis of quantitative traits.

Let us now introduce a superscript "t" to denote the trait locus. For a pedigree of size $k$, let $y_i$, $g_i^t$ denote the phenotype and trait genotype for individual "$i$." Let $\pi_{ij}^t$ denote trait

IBD (proportion of alleles shared IBD at the trait locus) between individuals "$i$" and "$j$." Let $Y = (y_1, y_2, \dots, y_k)'$ and $g_t = (g_1^t, g_2^t, \dots, g_k^t)$ denote the vectors of genotypes and phenotypes for the pedigree. Similarly let $\Pi_t = ((\pi_{ij}^t))$ denote the $k \times k$ matrix of pairwise IBD sharing proportions for the pedigree. For any pair of individuals "$i$" and "$j$" in the pedigree, we have

$$Cov(y_i, y_j \mid \pi_{ij}^t) = \pi_{ij}^t \sigma_a^2 + 1_{\{\pi_{ij}^t = 1\}} \sigma_d^2 + r_{ij} \sigma_\epsilon^2, \tag{2.1.4}$$

where the residuals $\epsilon_i$ are assumed to be independent of the genotypes and to follow a multivariate normal distribution within the pedigree with mean 0 and dispersion $\Sigma_\epsilon = ((\sigma_\epsilon^2 r_{ij}))$, where $r_{ij}$ is the environmental correlation between individuals "i" and "j." The population mean and dispersion matrix of $Y$ are $\mu_Y = \mu \ \mathbf{1}$ ($\mathbf{1}$ being a vector of $k$ 1's) and $\Sigma_Y = ((\sigma^2 \rho_{ij}))$. Subtracting the expectation from both sides of the above equation, we get

$$Cov(y_i, y_j \mid \pi_{ij}^t) = Cov(y_i, y_j) + \sigma_a^2[\pi_{ij}^t - 2\phi_{ij}] + \sigma_d^2[1_{\{\pi_{ij}^t = 1\}} - \gamma_{ij}^2], \tag{2.1.5}$$

where $\phi_{ij}$ is the kinship coefficient and $\gamma_{ij}^2$ is the Cotterman's coefficient $P(\pi = 1)$ for the pair $(i, j)$. Thus for a pedigree of size $k$ , rewriting in matrix notation, the conditional on IBD model is

$$Cov(Y \mid \Pi_t) = \Sigma_Y + \sigma_a^2[\Pi_t - 2\Phi] + \sigma_d^2[\Pi_t^{(2)} - \Gamma^{(2)}], \tag{2.1.6}$$

where $\Phi$ and $\Gamma^{(2)}$ are $k \times k$ matrices given by $(\Phi)_{ij} = \phi_{ij}$ and $(\Gamma^{(2)})_{ij} = \gamma_{ij}^2$. Note that, standard notation for the Cotterman's coefficient matrices are $\Delta^{(0)}$, $\Delta^{(1)}$ and $\Delta^{(2)}$. We have used $\Gamma^{(2)}$ to avoid confusion with the notation $\Delta$ for LD.

Finally, it should be noted that the conditional on genotype model (2.1.2) and hence the conditional on IBD model (2.1.6) can be extended to incorporate additional variance components such as polygenic additive effects and polygenic dominance effects (see for example Almasy and Blangero 1998). We ignore those components in this dissertation for reasons of clarity. However, for most of the methods discussed here, it is straightforward to obtain extensions which incorporate those components using standard procedures. For the same reason, we will also ignore the dominance component of the major QTL for most of the discussion.

### Orthogonalization for Sibpairs

Since $\Pi_t$ and $\Pi_t^{(2)}$ are not orthogonal, we can orthogonalize model (2.1.6) similarly as we did for genotypes. This orthogonalization would however vary with relationship between the pair of individuals. Hence it provides mathematical and computational simplicity for a two degree of freedom model only when the data consists of pairs with a single relationship type. For example, in case of sibpair or sibship data, denoting $\tilde{\pi}_i^t = \pi_i^t - 1/2$ and $\tilde{1}_{\{\pi_i^t=1\}} = 1_{\{\pi_i^t=1\}} - 1/4$, and using the Gram Schmidt orthogonalization procedure, we have the following orthogonal decomposition of model (2.1.5).

$$\begin{aligned}
Cov(y_i, y_j \mid \pi_{ij}^t) &= Cov(y_i, y_j) + \sigma_a^2 \; \tilde{\pi}_{ij}^t + \sigma_d^2 \; \tilde{1}_{\{\pi_{ij}^t=1\}} \\
&= Cov(y_i, y_j) + (\sigma_a^2 + \sigma_d^2) \; \tilde{\pi}_{ij}^t + \sigma_d^2 \; (\tilde{1}_{\{\pi_{ij}^t=1\}} - \tilde{\pi}_{ij}^t) \\
&\qquad\qquad \left[ \; \because \; < 1_{\{\pi_{ij}^t=1\}}, \pi_{ij}^t > = < \pi_{ij}^t, \pi_{ij}^t > = 1/8 \; \right] \\
&= Cov(y_i, y_j) + (\sigma_a^2 + \sigma_d^2) \; \tilde{\pi}_{ij}^t + \sigma_d^2 \; (1_{\{\pi_{ij}^t=1\}} - \pi_{ij}^t + 1/4) \\
&= Cov(y_i, y_j) + (\sigma_a^2 + \sigma_d^2) \; \tilde{\pi}_{ij}^t + \sigma_d^2 \; [(-1/2) 1_{\{\pi_{ij}^t=1/2\}} + 1/4] \\
&\qquad\qquad \left[ \; \because \; \pi_{ij}^t = 1_{\{\pi_{ij}^t=1\}} + (1/2) 1_{\{\pi_{ij}^t=1/2\}} \; \right] \\
&= Cov(y_i, y_j) + \sigma_a^{2'} \; \tilde{\pi}_{ij}^t + \sigma_d^{2'} \; [(-1/2) \tilde{1}_{\{\pi_{ij}^t=1/2\}}]. \qquad (2.1.7)
\end{aligned}$$

$$\text{Thus,} \qquad Cov(Y \mid \Pi_t) = \Sigma_Y + \sigma_a^{2'} [\Pi_t - 2\Phi] + \sigma_d^{2'}/2 \; [\Delta^{(1)} - \Pi_t^{(1)}], \qquad (2.1.8)$$

where $\sigma_a^{2'} = \sigma_a^2 + \sigma_d^2$ and $\sigma_d^{2'} = \sigma_d^2$ are the additive and dominance variance under this new parametrization. It is easy to verify that $\Pi_t$ and $\Pi_t^{(1)}$ are orthogonal for sibships. Although this orthogonalization provides considerable simplicity in analyzing a 2 d.f. model for sibships, there is one additional complication. The parameters $\sigma_a^{2'}$ and $\sigma_d^{2'}$ are constrained by $\sigma_a^{2'} \geq \sigma_d^{2'}$. This constraint has to be taken into account during model fitting. This is unlike the orthogonalization for genotypes, where the parameters $\alpha$ and $\delta$ in model (2.1.2) (and consequently $\sigma_a^2$ and $\sigma_d^2$) are unconstrained, just like the initial parameters $a$ and $d$ in model (2.1.1). Nevertheless the orthogonality is useful and can be utilized to obtain two degree of freedom statistics for sibship data (e.g., see Tang 2000 and chapter 3 in this dissertation).

## 2.2 GENETIC MODELS AT THE MARKER LOCUS

In this section, we will use the notation of the previous section. In addition, let $g_m$, $\Pi_m$ denote the genotype vector and IBD matrix at a marker (test) locus. We will assume that the marker is biallelic with alleles "A" and "a" with frequencies $p_m$ and $q_m$. Also the marker genotypes $g_i^m$ are coded as number of "A" alleles. Let $v_{g_t} = 2pq$ and $v_{g_m} = 2p_m q_m$ denote $Var(g_i^t)$ and $Var(g_i^m)$

under HWE. Let $\theta$ denote the recombination fraction between the two loci and let $\Delta$ denote the linkage disequilibrium between the alleles "A" and "D" (i.e., $\Delta = P(AD) - P(A)P(D)$).

Most models at a marker locus do not model the parameters $\theta$ and $\Delta$ directly. They use surrogate parameters which implicitly model linkage and association. Below, we outline some of the standard "implicit" models and also propose a new implicit model which is more general than the standard models. This model yields score statistics similar to the standard implicit models under specific assumptions. We will also outline an explicit model parametrized by $\theta$ and $\Delta$ to obtain an heuristic justification for the proposed implicit model.

### 2.2.1 Implicit Models

**2.2.1.1 VC Model for Covariance** The phenotypic covariance conditional on "trait IBD" can be decomposed using variance components as shown in equation (2.1.6). In reality the trait IBD can not be observed. We observe marker genotype data $M$ at the test locus or at multiple markers across the genome. The IBD at the test locus $\Pi_m$ can be calculated (or estimated by $\hat{\Pi}_m = E[\Pi_m \mid M]$) based on the observed genotype data using singlepoint or multipoint methods.

The standard variance components method for linkage analysis of quantitative traits assumes the following model at the marker locus.

$$[Y \mid \hat{\Pi}_m] \sim N(\mu_{Y}, \Sigma_{Y} + v_a[\hat{\Pi}_m - 2\Phi] + v_d[\hat{\Pi}_m^{(2)} - \Delta^{(2)}]). \tag{2.2.1}$$

Note that the covariance is exactly the same as (2.1.6), with $\Pi_t$ replaced by $\hat{\Pi}_m$ and the parameters $(\sigma_a^2, \sigma_d^2)$ replaced by implicit parameters $(v_a, v_d)$. It is assumed that $v_a = v_d = 0 \Leftrightarrow \theta = 0.5$. Hence this model can be used to test for linkage ($H_0 : \theta = 0$ vs. $H_1 : \theta < 0.5$) indirectly by testing $H_0 : v_a = v_d = 0$ vs $H_1 : v_a > 0$ or $v_d > 0$. Generally, however, it is assumed that the dominance effect is negligible ($\delta = 0$ and $v_d = 0$) and the 1 d.f. test $H_0 : v_a = 0$ vs $H_1 : v_a > 0$ is used. It is possible to give an heuristic justification for the implicit parameters $v_a$ and $v_d$ (e.g. Amos 1994, section 2.2.2.2). However the assumption of multivariate normality is not very realistic. In general this distribution is expected to be skewed or multimodal (e.g., Amos 1994) at the trait locus and the marker locus. In fact, the trait locus model (2.1.2) is

$$[Y \mid g_t] \sim N[\mu_{Y} + \alpha\, (g_t - Eg_t) + \delta\, (g_t^{(1)} - Eg_t^{(1)}), \Sigma_\epsilon]. \tag{2.2.2}$$

where $(g_t)_i = g_i^t, {g_t^{(1)}}_i = 1_{\{g_i^t=1\}}\ \forall i = 1, \ldots, k$.

Conditioning on IBD leads to the following mixture normal distributions at the trait and the marker locus.

**Trait Locus:**

$$[Y \mid \Pi_t] \sim \sum_{g_t} [Y \mid g_t] \; P(g_t \mid \Pi_t).$$

**Marker Locus:**

$$[Y \mid \hat{\Pi}_m] \quad \sim \quad \sum_{g_t} \sum_{\Pi_t} [Y \mid g_t] \; P(g_t \mid \Pi_t) \; P_\theta(\Pi_t \mid \hat{\Pi}_m).$$

The distribution $[g_t \mid \Pi_t]$ is expected to be skewed when the trait allele frequency is close to 1 or 0 and multimodal when it is close to 0.5. In spite of the incorrect multivariate normality assumption, the VC model leads to powerful tests for linkage analysis, as discussed in chapter 3. This model also assumes that all of the marker loci, including the test locus, are in linkage equilibrium with the trait, which justifies conditioning on $\hat{\Pi}_m$ as a sufficient statistic for modeling linkage. This assumption is relaxed in the proposed implicit model described in section 2.2.1.4.

#### 2.2.1.2 Model for the Mean

The phenotypic mean conditional on "trait genotype" can be decomposed as shown in equation (2.1.2). In reality the trait genotype cannot be observed. An implicit model conditional on the marker genotypes is given by

$$[Y \mid g_m] \sim N[\mu_{_Y} + \beta \; (g_m - Eg_m) + \gamma \; (g_m^{(1)} - Eg_m^{(1)}), \Sigma_e], \tag{2.2.3}$$

where $(g_m)_i = g_i^m, {g_m^{(1)}}_i = 1_{\{g_i^m = 1\}}$ $\forall i = 1, \ldots, k$. This model is same as the model (2.1.2) with parameters ($\alpha$ and $\delta$) replaced by implicit parameters $(\beta, \gamma)$. It is assumed that $\beta = \gamma = 0 \Leftrightarrow \Delta = 0$. Hence this model can be used to test for association ($H_0 : \Delta = 0$ vs. $H_1 : \Delta > 0$) indirectly by testing $H_0 : \beta = \gamma = 0$ vs $H_1 : \beta > 0$ or $\gamma > 0$. Generally however the dominance effect is assumed to be negligible ($\delta = 0$ and $\delta' = 0$) and the 1 d.f. test $H_0 : \beta = 0$ vs $H_1 : \beta > 0$ is used. This model is used by some standard association mapping methods for quantitative traits such as ANOVA based methods (e.g., Barrett 2002; O'Donnell *et al.* 1998) and FBAT (Rabinowitz 1997; Horvath *et al.* 2001). Although FBAT was originally proposed as a non-parametric approach, some of the statistics implemented in FBAT are equivalent to score statistics derived under the above model (Laird *et al.* 2000; Shih and Whittemore 2002). The implicit parameter $\beta$ (and similarly $\gamma$) can be motivated using an explicit model (see section 2.2.2.2). However model (2.2.3) ignores the observed IBD information at the marker locus. The QTDT model and the proposed implicit model attempt to incorporate IBD information.

Apart from testing for association, a similar model is sometimes used to test for linkage (e.g. Dupuis *et al.* 2007 and section 2.3.2 of this dissertation). These methods use the following mixture normal distribution for $[Y \mid \hat{\Pi}_m]$.

$$[Y \mid g_m] \sim N[\mu_Y + c\, \tilde{g}_m, \Sigma_e], \qquad \text{and} \tag{2.2.4}$$

$$[Y \mid \hat{\Pi}_m] \sim \sum_{g_m} [Y \mid g_m]\ P(g_m \mid \hat{\Pi}_m) \quad \text{(Assuming perfect IBD information)}, \tag{2.2.5}$$

and test for linkage using the parameter $c$. Although the model assumptions are somewhat artificial, this model can be shown to yield similar score statistics as the VC model (2.2.1) (e.g., see Dupuis *et al.* 2007 and section 2.3.2 of this dissertation). Note that there is an implicit assumption of linkage equilibrium ($\Delta = 0$) in this model, as it uses only $\hat{\Pi}_m$ as a sufficient statistic to test for linkage. The genotypes $\tilde{g}_m$ are not used; they act as a latent dummy variable in this model. To see that $c = 0 \Leftrightarrow \theta = 0.5$, we note that the mean and covariance of this model are

$$E(Y \mid \hat{\Pi}_m) = E(\mu_Y + \tilde{g}_m \mid \hat{\Pi}_m) = \mu_Y$$
$$Cov(Y \mid \hat{\Pi}_m) = c^2 v_{g_m} \hat{\Pi}_m + \Sigma_e = \Sigma_Y + c^2 v_{g_m}(\hat{\Pi}_m - 2\Phi).$$

These moments agree with the moments of the VC model (2.2.1) with $v_a$ replaced by $c^2 v_{g_m}$. Thus $c^2$ has a similar interpretation as $v_a$ in the VC model, which implies that a 2-sided test based on model (2.2.5) can be an alternative for the VC model, with the advantage that it models $[Y \mid \hat{\Pi}_m]$ as a mixture-normal instead of a multivariate normal as in the VC model (2.2.1).

#### 2.2.1.3 QTDT Model for Mean and Covariance

The QTDT model (Fulker *et al.* 1999; Abecasis *et al.* 2000) tries to incorporate both marker genotype and IBD information by combining the covariance modeling of the VC approach and mean modeling of the FBAT. Assuming no dominance ($\delta = 0$), this model is

$$[Y \mid g_m, \hat{\Pi}_m] \sim N[\mu_Y + \beta\, (g_m - Eg_m), \Sigma_Y + v_a(\hat{\Pi}_m - 2\Phi)], \tag{2.2.6}$$

where it is assumed that $v_a = 0 \Leftrightarrow \theta = 0.5$ and $\beta = 0 \Leftrightarrow \Delta = 0$. The QTDT model is generally used to test for "association" (using the parameter $\beta$), and sometimes for "linkage" (using the parameter $v_a$). This model should in general be more powerful to detect association than model (2.2.3) when linkage is present. It should also be more powerful to detect linkage than model (2.2.1), whenever an allele at the marker locus is in LD with the trait. However the mean and covariance of the above model are not consistent with each other (see next section 2.2.1.4), which can be a reason for the model not performing as well as expected (see chapter 4). The proposed implicit model described below attempts to eliminate this inconsistency.

**2.2.1.4 Proposed Model for Mean and Covariance** There is an inherent distinction between statistical modeling at a trait locus and modeling at a marker locus, namely that at a trait locus the phenotype only depends on the trait genotype and the environment. The IBD at the trait locus, even if it is known, does not convey any information given the trait genotype. On the other hand, when testing at a marker locus the phenotype is a function of the marker genotype, its IBD sharing and the environment. The marker genotype is always directly observed and the IBD sharing at the marker locus can be estimated based on marker data across the chromosome. Hence ideally, both the genotype and the IBD information should be used for modeling at a marker locus. Most of the standard methods for detecting linkage (or association), however use only the IBD or (only the genotypes respectively), by making suitable assumptions such as no LD (or no linkage), which makes the genotypes (or the IBD) non-informative.

The QTDT model attempts to incorporate both genotype and IBD information using an ad-hoc approach, where the distribution $[Y \mid g_m, \hat{\Pi}_m]$ is assumed to have a mean depending on the genotype $E(Y \mid g_m)$ and a covariance depending on the IBD $Cov(Y \mid \hat{\Pi}_m)$. The modeling of the mean is correct, as the IBD sharing does not affect the marginal distributions and hence the means. However the covariance should be modeled as $Cov(Y \mid g_m, \hat{\Pi}_m)$. To obtain an estimate for this covariance we note that

$$
\begin{aligned}
&Cov(Y \mid \hat{\Pi}_m) = E_{g_m|\hat{\Pi}_m}[Cov(Y \mid g_m, \hat{\Pi}_m)] + Cov_{g_m|\hat{\Pi}_m}[E(Y \mid g_m, \hat{\Pi}_m)] \\
&\Rightarrow \Sigma_Y + v_a(\hat{\Pi}_m - 2\Phi) = E_{g_m|\hat{\Pi}_m}[Cov(Y \mid g_m, \hat{\Pi}_m)] + Cov_{g_m|\hat{\Pi}_m}[\mu_Y + \beta\,(g_m - Eg_m)] \\
&\Rightarrow \Sigma_Y + v_a(\hat{\Pi}_m - 2\Phi) = E_{g_m|\hat{\Pi}_m}[Cov(Y \mid g_m, \hat{\Pi}_m)] + \beta^2 Cov(g_m \mid \hat{\Pi}_m) \\
&\Rightarrow E_{g_m|\hat{\Pi}_m}[Cov(Y \mid g_m, \hat{\Pi}_m)] = \Sigma_Y + v_a(\hat{\Pi}_m - 2\Phi) - \beta^2 Cov(g_m \mid \hat{\Pi}_m). \qquad (2.2.7)
\end{aligned}
$$

Hence if we assume that the covariance $Cov(Y \mid g_m, \hat{\Pi}_m)$ is a constant (i.e., homoscedastic) with respect to $g_m$, then we can estimate that constant covariance by the right hand side of the last equation. This assumption basically means that for each IBD configuration, the covariance of the phenotype is the same for all marker genotypes. The covariance estimate of the QTDT model on the other hand assumes that the covariance is a constant over genotypes (homoscedastic) and also that $\beta = 0$ (i.e. mean is constant over genotypes), which leads to the inconsistency between the mean and covariance estimates when testing for association using $\beta$.

Based on the assumption of homoscedasticity of $Cov(Y \mid g_m, \hat{\Pi}_m)$, we propose the following model for the phenotype $Y$ conditional on marker genotype $g_m$ and estimated marker IBD $\hat{\Pi}_m$.

$$[Y \mid g_m, \hat{\Pi}_m] \sim N[\mu_Y + \beta\ (g_m - Eg_m), \Sigma_Y + v_a(\hat{\Pi}_m - 2\Phi) - \beta^2 v_{g_m}\hat{\Pi}_m], \tag{2.2.8}$$

where under HWE, $Eg_m = 2p_m$ and $v_{g_m} = Var(g_m) = 2p_m q_m$. This model is similar to the QTDT except for the last term, $\beta^2 v_{g_m}\hat{\Pi}_m$, which is subtracted from the covariance. This term follows from equation (2.2.7) and from the fact that $Cov(g_m \mid \Pi_m) = 2p_m q_m \Pi_m = v_{g_m}\Pi_m$, as shown in equation (2.1.3). To see how this model relates to mean and covariance models discussed above, we note that

- When $\beta = 0$ (i.e., no LD) it reduces to the VC model (2.2.1) for covariance. Under this assumption, the QTDT model (2.2.6) also reduces to the VC model.
- When $v_a = 0$ (i.e., marker unlinked) it reduces to the model

  $$[Y \mid g_m, \hat{\Pi}_m] \sim N\ [\ \mu_Y + \beta\ \tilde{g}_m\ ,\ \Sigma_Y - \beta^2 v_{g_m}\hat{\Pi}_m\ ]$$

  which although different from the FBAT model (2.2.3), gives identical score statistics to that model. This is shown in chapter 4. The essential difference between these models is that the FBAT ignores the observed IBD information and assumes $[Y \mid g_m]$ has homoscedastic errors with a constant covariance matrix. Under this assumption the covariance reduces to $\Sigma_e = \Sigma_Y - \beta^2 v_{g_m}(2\Phi)$. On the other hand, model (2.2.8) incorporates the IBD information and assumes homoscedasticity for the errors of $[Y \mid g_m, \hat{\Pi}_m]$ model, which is a slightly weaker assumption for the covariances (i.e. the off-diagonal entries of the variance covariance matrix). Under this assumption, the QTDT model reduces to

  $$[Y \mid g_m, \hat{\Pi}_m] \sim N\ [\ \mu_Y + \beta\ \tilde{g}_m\ ,\ \Sigma_Y\ ],$$

  which once again gives identical scores statistics to the FBAT model (2.2.3).
- When $v_a = \sigma_a^2$ and $\beta = \alpha$ (i.e the test marker is the QTL ), it reduces to the trait locus model (2.2.2), in which the covariance does not depend on IBD. This follows from the definition $\sigma_a^2 = v_{g_m}\alpha^2$ and the fact that $\Sigma_Y - \sigma_a^2(2\Phi) = Cov(Y) - \alpha^2 Cov(g_t) = \Sigma_e$, assuming no dominance.

In this case, the QTDT model reduces to

$$[Y \mid g_m, \hat{\Pi}_m] \sim N\ [\ \mu_{_Y} + \alpha\ \tilde{g}_m\ ,\ \Sigma_{_Y} + \sigma_a^2(\hat{\Pi}_m - 2\Phi)\ ],$$

which is not identical to the trait locus model (2.2.2). In fact the above model depends on the observed IBDs. Thus the parameter $v_a$ of the QTDT model can not be interpreted as $\sigma_a^2$ when the marker locus is the putative QTL.

Note that all of the above models assume that the marker genotype $g_m$ and the estimated marker IBD $\hat{\Pi}_m$ are sufficient for testing linkage and/or association. This assumption is essentially equivalent to assuming that the markers are in linkage equilibrium with each other. When this assumption is violated, one possible approach is to cluster markers (Abecasis and Wigginton 2005) into groups of markers which are in LD with each other but in linkage equilibrium with other groups and then to analyze each group of markers as a whole. Such approaches will not be discussed further in this dissertation.

### 2.2.2 Model Assumptions

In this section, we look at some of the assumptions required to intuitively justify the implicit models described in the previous section. We consider an explicit model parametrized by $\theta$ and $\Delta$ relating the trait and the marker locus, and show that implicit models can be derived from it. We start from the model (2.2.2) at the trait locus, and model $g_t$ as a function of the observed data $g_m$ and $\hat{\Pi}_m$. Note that the model below is based on heuristics and some of the assumptions are at best crude approximations. But our objective in this section is primarily to obtain an intuitive motivation for the implicit models described above.

**2.2.2.1 Explicit Model** To model $[g_t \mid g_m, \hat{\Pi}_m]$, let us first assume the following model for $[\Pi_t \mid \hat{\Pi}_m]$.

$$E(\Pi_t - 2\Phi) = (1 - 2\tilde{\theta})(\hat{\Pi}_m - 2\Phi) \tag{2.2.9}$$

where $\tilde{\theta}$ could be any monotonic function of $\theta$ such that $\tilde{\theta} = 0$ if $\theta = 0$ and $\tilde{\theta} = 0.5$ if $\theta = 0.5$. This is an ad-hoc assumption, which is true for a data set consisting of pairs with only one relationship type. For example for a sibship dataset, this function is $\tilde{\theta} = \theta - \theta^2$, which follows from the joint distribution of $[\pi_t, \pi_m]$ tabulated in Haseman and Elston (1972). For unilineal relative pairs, the linearity $E(\Pi_t \mid \Pi_m)$ obviously holds, as $\Pi_m$ can take only two values 0 and 1/2. The joint distribution of $[\pi_t, \pi_m]$ is given by

| $\Pi_t$ <br> $\Pi_m$ | $0$ | $1/2$ | $[\Pi_t]$ |
|---|---|---|---|
| $0$ | $1-4\phi-f(\theta)$ | $f(\theta)$ | $1-4\phi$ |
| $1/2$ | $f(\theta)$ | $4\phi-f(\theta)$ | $4\phi$ |
| $[\Pi_m]$ | $1-4\phi$ | $4\phi$ | $1$ |

where $\phi$ is the kinship coefficient and $f(\theta)$ is a polynomial of the form $\sum_{i=1}^{n} a_i\theta^i$, such that $f(0) = 0$ and $f(1/2) = 4\phi(1-4\phi)$. The conditional distribution $[\Pi_t \mid \Pi_m]$ and hence the conditional expectation can be obtained from the above table. The conditional expectation has the form of model (2.2.9), with

$$\tilde{\theta} = \frac{f(\theta)}{8\phi(1-4\phi)}.$$

The function $\tilde{\theta}$ for grandparent-grandchild, half sibling and avuncular pairs is $\tilde{\theta} = \theta$, $\tilde{\theta} = 2\theta - 2\theta^2$ and $\tilde{\theta} = 5\theta/2 - 4\theta^2 + 2\theta^3$ respectively, which can be obtained using the joint distribution tables in Amos and Elston (1989). Thus for a general pedigree with multiple relationships, the assumption of a single function $\tilde{\theta}$ is essentially a convenient approximation.

The mean $E(g_t \mid g_m, \hat{\Pi}_m)$ is free of $\hat{\Pi}_m$. Assuming HWE, at each locus, it can be shown that $g_t$ has an exact linear regression on $g_m$ given by

$$E(g_t \mid g_m) = Eg_t \;+\; \tilde{\Delta}\;[g_m - Eg_m],$$

where $\tilde{\Delta} = \Delta/(p_m q_m)$ is a monotonic function of $\Delta$ with $\tilde{\Delta} = 0 \Leftrightarrow \Delta = 0$.

In general, $Var(g_t \mid g_m)$ is not free of $g_m$. In fact, under HWE this variance can be shown to be linear in $g_m$ with a slope of $\tilde{\Delta}^2\;(p_m - q_m) + \tilde{\Delta}\;(q_t - p_t)$, which is non-zero whenever there is LD, except in the trivial cases $p_t = p_m = 1/2$ and $\tilde{\Delta} = \frac{p_t - q_t}{p_m - q_m}$. Thus the errors of the regression of $[g_t \mid g_m]$ are not in general homoscedastic. All the implicit models discussed in the previous section assume homoscedastic errors as a convenient approximation.

To compute the covariance matrix, $Cov(g_t \mid g_m, \hat{\Pi}_m)$, we use the crucial homoscedasticity assumption as we did in section 2.2.1.4 and the conditional mean obtained above. We assume that

$Cov(g_t \mid g_m, \hat{\Pi}_m)$ is constant with respect to $g_m$ for each $\hat{\Pi}_m$. In other words, we assume that for each IBD configuration, $g_m$ affects $g_t$ only through the mean (due to LD) but not the covariance.

$$\begin{aligned}
Cov(g_t \mid g_m, \hat{\Pi}_m) &= E_{g_m|\hat{\Pi}_m}[Cov(g_t \mid g_m, \hat{\Pi}_m)] \\
&= Cov[g_t \mid \hat{\Pi}_m] - Cov_{g_m|\hat{\Pi}_m}[E(g_t \mid g_m)] \\
&= E_{\Pi_t|\hat{\Pi}_m} Cov[g_t \mid \hat{\Pi}_t] - Cov[\tilde{\Delta}\ g_m \mid \hat{\Pi}_m] \\
&= E(v_{g_t}\Pi_t \mid \hat{\Pi}_m) - \tilde{\Delta}^2 v_{g_m}\hat{\Pi}_m \\
&= v_{g_t}[2\Phi + (1-2\tilde{\theta})(\hat{\Pi}_m - 2\Phi)] - \tilde{\Delta}^2 v_{g_m}\hat{\Pi}_m. \qquad (2.2.10)
\end{aligned}$$

The last equality follows by using equation (2.2.9). Let us define $Z = g_t - E(g_t \mid g_m, \hat{\Pi}_m)$. Then, we can write $\tilde{g}_t = \tilde{\Delta}\ \tilde{g}_m + Z$, where Z has mean 0 and covariance $\Sigma_Z$ given by

$$\Sigma_Z = Cov(g_t \mid g_m, \hat{\Pi}_m) = v_{g_t}[2\Phi + (1-2\tilde{\theta})(\hat{\Pi}_m - 2\Phi)] - \Delta^2 v_{g_m}\hat{\Pi}_m.$$

In the following explicit model we make the additional assumption that Z has a multivariate normal distribution. In other words $g_t \mid g_m\hat{\Pi}_m$ has a multivariate normal distribution.

**Explicit Model:**

Here we model $[Y \mid g_m, \hat{\Pi}_m]$ as:

$$\begin{aligned}
[Y \mid g_m, \hat{\Pi}_m] &= \int_{g_t} [\ Y \mid g_t\ ]\ \ [\ g_t \mid g_m, \hat{\Pi}_m\ ] \\
&= \int_{g_t} [\ N(Y;\ \mu_Y + \alpha\ \tilde{g}_t, \Sigma_e)\ ]\ \ [\ N(g_t;\ 2p + \tilde{\Delta}\ \tilde{g}_m, \Sigma_Z)\ ] \qquad (2.2.11)
\end{aligned}$$

#### 2.2.2.2 Derivation of Implicit Models

The mean model (2.2.3) essentially models $[g_t \mid g_m]$ ignoring $\hat{\Pi}_m$. It assumes a linear mean $E(g_t \mid g_m) = E(g_t) + \beta\ \tilde{g}_m$, and a constant covariance for $Cov(g_t \mid g_m)$. Ghosh and De (2007) showed, using the exact distribution of $[g_t \mid g_m]$, that this implicit method has correct type I error (i.e $\beta = 0 \Leftrightarrow \Delta = 0$). However, they also showed that the violation of the homoscedasticity assumption can lead to considerable loss of power.

Next we show that the proposed implicit model (2.2.8) can be derived from the explicit model (2.2.11) described above. Defining $\beta = \alpha\tilde{\Delta}$, we note that

$$E(Y \mid g_m, \hat{\Pi}_m) = E(\mu_Y + \alpha\ \tilde{g}_t) = \mu_Y + \beta\ g_m,$$

which is the mean of the implicit model and by definition $\beta = 0 \Leftrightarrow \Delta = 0$. Similarly defining

$v_a = (1-2\tilde{\theta})\sigma_a^2$, the covariance is

$$\begin{aligned}
Cov(Y \mid g_m, \hat{\Pi}_m) &= \alpha^2 Cov(g_t \mid g_m, \hat{\Pi}_m) + \Sigma_e = \alpha^2 \Sigma_Z + \Sigma_e \\
&= \sigma_a^2 \left[ [2\Phi + (1-2\tilde{\theta})(\hat{\Pi}_m - 2\Phi) \right] - \beta^2 v_{g_m} \hat{\Pi}_m + \Sigma_e \\
&= (\Sigma_e + \sigma_a^2\, 2\Phi) + \sigma_a^2\, (1-2\tilde{\theta})(\hat{\Pi}_m - 2\Phi) - \alpha\tilde{\Delta}^2 v_{g_m} \hat{\Pi}_m \\
&= \Sigma_{_Y} + v_a\, (\hat{\Pi}_m - 2\Phi) - \beta^2 v_{g_m} \hat{\Pi}_m,
\end{aligned}$$

which is same as the covariance of the implicit model and by our definition $v_a = 0 \Leftrightarrow \theta = 0.5$. The multivariate normality of the proposed implicit model follows from our assumptions that $[Y \mid g_t]$ and $[g_t \mid g_m, \hat{\Pi}_m]$ are both normally distributed and the result "if $X \mid Y$ and $Y \mid Z$ are normally distributed, then so is $X \mid Z$."

Thus, we have shown that the explicit model above is one way to justify the implicit parameters in the proposed model (2.2.8). Note however, that the explicit model is only a special case of the proposed implicit model with $v_a = (1-2\tilde{\theta})\sigma_a^2$ and $\beta = \alpha\tilde{\Delta}$. In other words, the validity of the assumptions (made in the explicit model above) provides a "sufficient" condition for the tests using the implicit parameters to be direct (optimally powerful) tests for the parameters $\theta$ and $\Delta$. Even when these assumptions are violated, the tests using $v_a$ and $\beta$ may still capture most of the information in these parameters indirectly.

## 2.3 SCORE TESTS UNDER IMPLICIT MODELS

The commonly used score test for linkage Tang and Siegmund 2001; Putter *et al.* 2002; Wang 2005 is based on the VC model (2.2.1) through the implicit linkage parameter $v_a$. The FBAT statistic (Laird *et al.* 2000; Horvath *et al.* 2001) can also be thought of as a score based on the implicit FBAT model (2.2.3) for the implicit parameter $\beta$. In fact, the statistic as originally proposed in Rabinowitz (1997) for trio data, was motivated as a score statistic. Shih and Whittemore (2002) also show that the FBAT statistic is equivalent to a score test under the assumption of no residual correlation among non-founders.

### 2.3.1 Scores for Proposed Implicit Model (2.2.8)

The score statistics under the standard implicit models (VC and FBAT) can be derived from the proposed implicit model (2.2.8) under appropriate assumptions on the parameters $v_a$ and $\beta$. These and two new score statistics (with fewer assumptions on the parameters) are described below.

#### 2.3.1.1 Tests for Linkage: $H_0 : v_a = 0$ vs $H_1 : v_a > 0$

- **Assume No LD** Under the assumption of no LD ($\beta = 0$), model (2.2.8) reduces to the usual VC model (2.2.1). The score for this model has been derived by various authors (e.g., Tang and Siegmund 2001; Putter *et al.* 2002; Wang 2005) and is given by

$$S_{VC} = vec[\Sigma_Y^{-1}(Y - \mu_Y)(Y - \mu_Y)'\Sigma_Y^{-1} - \Sigma_Y^{-1}]' vec(\hat{\Pi}_m - 2\Phi), \tag{2.3.1}$$

  where *vec* is an operator which vectorizes all the elements of a square matrix in a row-wise order. This score is sometimes also derived under the implicit mixture normal model (2.2.5) as outlined in section 2.3.2 below.

- **Assume Possible LD** In this case the score for model (2.2.8) can be derived in the same way as $S_{VC}$, and using the formula (1.4.1) for the score in the presence of a nuisance parameter,

$$S_{VC,LD} = vec[\tilde{\Sigma}^{-1}(Y - \mu_Y - \hat{\beta}\,\tilde{g}_m)(Y - \mu_Y - \hat{\beta}\,\tilde{g}_m)'\tilde{\Sigma}^{-1} - \tilde{\Sigma}^{-1}]' vec(\Pi_m - 2\Phi), \tag{2.3.2}$$

  where $\tilde{\Sigma} = \Sigma_Y - \hat{\beta}^2\, v_{g_m}\hat{\Pi}_m$ and $\hat{\beta}$ is the MLE of $\beta$ under the null hypothesis of no linkage, i.e under the model

$$[Y \mid g_m, \hat{\Pi}_m] \sim N[\mu_Y + \beta\,\tilde{g}_m, \Sigma_Y - \beta^2 v_{g_m}\hat{\Pi}_m].$$

#### 2.3.1.2 Tests for Association: $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

- **Assume No Linkage** In this case the model (2.2.8) is same as that in the nuisance parameter estimation model above. The score for this model is the same as that for the FBAT model (2.2.3) (derived in chapter 4) given by:

$$S_{FBAT} = (Y - \mu_Y)'\Sigma_Y^{-1}\tilde{g}_m. \tag{2.3.3}$$

  A locally most powerful unbiased (LMPU) test for the two sided hypothesis $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$ can be derived under this model. This test is derived in chapter 4 and is given by

$$S_{FBAT-lmpu} = vec[\Sigma_Y^{-1}(Y - \mu_Y)(Y - \mu_Y)'\Sigma_Y^{-1} - \Sigma_Y^{-1}]' vec(\tilde{g}_m\tilde{g}_m' - v_{g_m}\hat{\Pi}_m). \tag{2.3.4}$$

- **Assume Possible Linkage** Under the assumption of "possible linkage," the score can be derived in the same way as $S_{FBAT}$ and is given by

$$S_{FBAT,linkage} = (Y - \mu_Y)'[\Sigma_Y + \hat{v}_a(\hat{\Pi}_m - 2\Phi)]^{-1}\tilde{g}_m, \tag{2.3.5}$$

where $\hat{v}_a$ is the MLE of $v_a$ under the VC model (2.2.1).

### 2.3.2 Score for the Mixture Normal Model (2.2.5)

The score statistic for linkage is usually derived under the VC model (2.2.1). Tang and Siegmund (2001) outlined a proof that the VC based score test is also the score test under model (2.2.5). But that proof only holds for sibpairs and is based on a Taylor series approximation. Also, they tested the hypotheses $H_0 : \sigma_a^2 = 0$ vs $H_1 : \sigma_a^2 > 0$, although their likelihood was same as (2.2.5) and parametrized by $c$. Dupuis *et al.* (2007) showed this without approximations and gave a general proof of this fact for a class of error distributions (general exponential family). Below we outline an alternative proof for the special case of normally distributed errors. We will show that the locally most powerful unbiased (LMPU) score test for testing $H_0 : c = 0$ versus $H_1 : c \neq 0$ has the same form as the VC based score test.

The likelihood of interest is

$$L_{Y|\hat{\Pi}_m}(c) = P(Y \mid \hat{\Pi}_m, c = c) = \sum_{g_m} L_{Y|g_m}(c)P(g_m \mid \hat{\Pi}_m), \tag{2.3.6}$$

where $L_{Y|g_m}(c)$ is given by equation (2.2.4). We want to test the 2-sided hypotheses $H_0 : c = 0$ against $H_1 : c \neq 0$. The LMPU statistic for linkage (Rao 2002, pp 453-455) is given by

$$S_{LMPU}^{linkage} = \frac{L''_{Y|\hat{\Pi}_m}(0)}{L_{Y|\hat{\Pi}_m}(0)},$$

which after simple algebra (see Appendix A) becomes

$$S_{LMPU}^{linkage} = Var_{g_m|\hat{\Pi}_m}[l'_{Y|g_m}(0)] + E_{g_m|\hat{\Pi}_m}[l''_{Y|g_m}(0)].$$

Using the expressions for $l'_{Y|g_m}(0)$ and $l''_{Y|g_m}(0)$ as given in Appendix A, we get

$$E_{g_m|\hat{\Pi}_m}[l'_{Y|g_m}(0)] = \tilde{Y}' \;\; \Sigma_Y^{-1} \; E[\tilde{g}_m \mid \hat{\Pi}_m] = 0.$$

Therefore we have

$$\begin{aligned}
S_{LMPU}^{linkage} &= Var_{g_m|\hat{\Pi}_m}[l''_{Y|g_m}(0)] \\
&= [\tilde{Y}\ \Sigma_{Y}^{-1}\Sigma_{g_m|\hat{\Pi}_m}\Sigma_{Y}^{-1}\ \tilde{Y}] + [-\tilde{Y}\ \Sigma_{Y}^{-1}\Sigma_g\Sigma_{Y}^{-1}\ \tilde{Y} + trace(\Sigma_{Y}^{-1}\Sigma_g) - trace(\Sigma_{Y}^{-1}E(\tilde{g_m}\tilde{g_m}' \mid \hat{\Pi}_m))] \\
&= \tilde{Y}\ \Sigma_{Y}^{-1}[Cov(g_m \mid \hat{\Pi}_m) - \Sigma_g]\Sigma_{Y}^{-1}\ \tilde{Y} - trace[\Sigma_{Y}^{-1}(Cov(g_m \mid \hat{\Pi}_m) - \Sigma_g)] \\
&= vec[\Sigma_{Y}^{-1}\ \tilde{Y}\tilde{Y}'\ \Sigma_{Y}^{-1} - \Sigma_{Y}^{-1}]' vec[Cov(g_m \mid \hat{\Pi}_m) - \Sigma_g].
\end{aligned}$$

Under HWE, $Cov(g_m \mid \hat{\Pi}_m) = 2p_m q_m \hat{\Pi}_m$ and $\Sigma_g = 4pq\Phi$, where $\Phi$ is the matrix of pairwise kinship coefficients for the pedigree. Hence the above statistic can be further simplified to

$$S_{LMPU}^{linkage} = vec[\Sigma_{Y}^{-1}\ \tilde{Y}\tilde{Y}'\ \Sigma_{Y}^{-1} - \Sigma_{Y}^{-1}]' vec[\hat{\Pi}_m - 2\Phi]. \tag{2.3.7}$$

(Ignoring constants)

From equation (2.3.7) we note that the LMPU statistic for linkage under the mixture normal model (2.2.5) is identical to the commonly used (LMP) score statistic under the variance components(VC) model.

### 2.3.3 Score Tests for the Explicit Model

For the explicit model (2.2.11) or more generally whenever the parameters $v_a$ and $\beta$ can be expressed as $f_1(\sigma_a^2)f_2(\theta)$ and $f_3(\alpha)\ f_4(\Delta)$ respectively (where $f_2$ and $f_4$ are monotonic functions of $\theta$ and $\Delta$), direct score tests for $f_2$ and $f_4$ would be identical to the corresponding score test for implicit models described above. For example, in case of the VC score test, $f_1$ can be absorbed inside $[\hat{\Pi}_m - 2\Phi]$ and the final score would be $f_1\ S_{VC}$, making the standardized score statistic free of the nuisance parameter $f_1$. Similarly it can be shown that the score for $\tilde{\Delta} = 0$ under "no-linkage" would be $f_3\ S_{FBAT}$.

## 2.4 SELECTED SAMPLING

Most family studies for linkage and and association mapping of quantitative traits, particularly traits related to diseases, are conducted using selected sampling designs. Individuals with a disease condition or extreme values of a quantitative trait are ascertained as probands and their families are recruited. This strategy helps to increase the power of the study and also helps to ensure that the sample includes affected individuals with familial disease instead of sporadic disease. Hence, selected sampling designs such as affected sibpairs, extreme discordant and/or extreme concordant sibpairs are common for linkage studies. Similarly, nuclear families with single or multiple affected offspring are frequently used for family-based association studies. Some of the traditional likelihood ratio based methods such as the VC and the QTDT are not robust to selected sampling unless the ascertainment scheme is exactly known and corrected for. This is particularly difficult for quantitative traits if the ascertainment is based on the disease status and hence difficult to translate in terms of the quantitative trait values, even if the scheme is simple and objective. The robustness problem is aggravated when the ascertainment criteria are subjective and complex.

Likelihood ratio based methods can provide biased MLEs and LRT statistics if the ascertainment is ignored. Moreover, the asymptotic $\chi^2$ thresholds for LRT fail to work, resulting in incorrect type I error rates unless empirical or permutation based thresholds are used. Score tests on the other hand do not require the knowledge of the ascertainment scheme for computing the numerator score function, provided nuisance parameter estimates are available. Various authors including Lebrec *et al.* (2004), Wang (2005) and Peng and Siegmund (2006), have shown that the VC model-based score test for linkage, although derived under a "conditional on IBD" model, is identical to the score test under a selective sampling framework. Lebrec *et al.* (2004) proved that the scores of the joint model are same as those for the "conditional on trait" model and the "conditional on IBD" model. However they did not consider an ascertainment scheme in their likelihood. Wang (2005) and Peng and Siegmund (2006) showed that the scores of the joint model with and without ascertainment are identical, however they did not consider "conditional on trait/IBD" models. In other words, they showed that the ascertainment scheme can be ignored for deriving the scores. These results taken together prove that the scores for the "correct" likelihood model $[Y, \hat{\Pi}_m \mid Y \in \mathcal{A}]$ (based on the sampling scheme) or a conditional on trait likelihood $L(\hat{\Pi}_m \mid Y)$ can be derived under the forward VC model $[Y \mid \hat{\Pi}_m]$. Below we derive similar results for the likelihood $[Y, g_m, \hat{\Pi}_m \mid Y \in \mathcal{A}]$ under the proposed implicit and explicit models.

### 2.4.1 Implicit Model

All the score statistics outlined in section 2.3 are based on the "forward" likelihood $L(Y \mid g_m, \hat{\Pi}_m)$. Ideally, based on the study design the likelihood of interest should be $L(Y, g_m, \hat{\Pi}_m \mid Y \in \mathcal{A})$, where $\mathcal{A}$ is the ascertainment scheme. Below, we show that the scores (numerators of the score tests) can be derived by ignoring the ascertainment scheme. Moreover, the score based on the joint model, $[Y, g_m, \hat{\Pi}_m]$, is identical to the score under forward model, $[Y \mid g_m, \hat{\Pi}_m]$ (under which the scores are derived) and the "conditional on trait" model, $L(g_m, \hat{\Pi}_m \mid Y)$ (which is often used used to standardize the scores). We also prove the invariance of the LMPU statistic (2.3.4) under the joint and conditional models (ignoring the ascertainment scheme). From section (1.4), we recall that in the presence of nuisance parameters, scores can be derived by starting from a likelihood with null hypothesis estimates of the nuisance parameters plugged in. Thus, we will assume that the nuisance parameters $\hat{\mu}_Y$ and $\hat{\Sigma}_Y$ are either available or have been estimated under the null hypothesis and plugged in to the likelihood.

We assume that the ascertainment scheme is based only on the phenotype (Peng and Siegmund 2006) and that the distribution of $[g_m, \hat{\Pi}_m]$ does not depend on the implicit parameters $v_a$ and $\beta$ (Lebrec *et al.* 2004). The scores are derived by differentiating the forward likelihood $L(Y \mid g_m, \hat{\Pi}_m)$ with respect to one of the parameters $v_a/\beta$ and holding the other parameter fixed at either 0 or at its MLE under the null. In the latter case, the scores would not be free of the ascertainment scheme, as the obtaining MLE would require knowledge of the ascertainment scheme. So, we will prove the results for $\beta$, fixing $v_a = 0$ as follows. The results for $v_a$ (fixing $\beta = 0$) can be proved similarly.

1. Scores for "forward" likelihood $P(Y \mid g_m, \hat{\Pi}_m)$ and "joint" likelihood $P(Y, g_m, \hat{\Pi}_m)$ are identical. This follows from the fact that

$$P(Y \mid g_m, \hat{\Pi}_m, \beta, v_a = 0) = \frac{P(Y, g_m, \hat{\Pi}_m \mid \beta, v_a = 0)}{P(g_m, \hat{\Pi}_m)}$$

by our assumption that the denominator is free of the parameters. Hence the score (both LMP and LMPU) for the denominator of the right side is zero.

2. Scores for the " conditional on trait" likelihood $P(g_m, \hat{\Pi}_m \mid Y)$ and "joint" likelihood $P(Y, g_m, \hat{\Pi}_m)$

are identical. As before, we note that

$$P(g_m, \hat{\Pi}_m \mid Y, \beta, v_a = 0) = \frac{P(Y, g_m, \hat{\Pi}_m \mid \beta, v_a = 0)}{P(Y \mid \beta, v_a = 0)}.$$

It suffices to show that the score for the denominator likelihood $P(Y \mid \beta, v_a = 0)$ is zero. This likelihood can be rewritten as a mixture likelihood as follows.

$$\begin{aligned} P(Y \mid \beta, v_a = 0) &= \sum P(Y \mid g_m, \hat{\Pi}_m, \beta, v_a = 0)\ P(g_m, \hat{\Pi}_m) \\ L_{Y|v_a=0}(\beta) &= \sum L_{Y|g_m,\hat{\Pi}_m,v_a=0}(\beta)\ P(g_m, \hat{\Pi}_m). \end{aligned}$$

Using the results in Appendix A, the LMP and LMPU scores for the mixture likelihood are given by

$$\begin{aligned} S_{LMP} &= l'_{Y|v_a=0}(0) = \sum l'_{Y|g_m,\hat{\Pi}_m,v_a=0}(0)\ P(g_m, \hat{\Pi}_m) \\ &= E_{g_m,\hat{\Pi}_m}[\tilde{Y}\Sigma_Y^{-1}\tilde{g}_m] \\ &= 0 \\ S_{LMPU} &= l''_{Y|v_a=0}(0) + [l'_{Y|v_a=0}(0)]^2 = Var_{g_m,\hat{\Pi}_m}[l'_{Y|g_m,\hat{\Pi}_m,v_a=0}(0)] + E_{g_m,\hat{\Pi}_m}[l''_{Y|g_m,\hat{\Pi}_m,v_a=0}(0)] \\ &= Var_{g_m,\hat{\Pi}_m}[\tilde{Y}\Sigma^{-1}\tilde{g}_m] + E_{g_m,\hat{\Pi}_m}[-\tilde{Y}'\Sigma_Y^{-1}\Sigma_g\Sigma_Y^{-1}\tilde{Y} + trace(\Sigma_Y^{-1}\Sigma_g) - \tilde{g}_m'\Sigma_Y^{-1}\tilde{g}_m]. \\ &= 0 \end{aligned}$$

Thus both the LMP and LMPU scores for the denominator are zero, proving that the score for the "conditional on trait" model is identical to the score for the joint model. (1) and (2) together imply that the LMP and LMPU scores for the "conditional on trait" model are identical to those of the proposed forward implicit model.

3. Ascertainment can be ignored, i.e., score (LMP) for joint model is same irrespective of conditioning. We note that

$$P(Y, g_m, \hat{\Pi}_m \mid Y \in \mathcal{A}, \beta, v_a = 0) = \frac{P(Y, g_m, \hat{\Pi}_m \mid \beta, v_a = 0)1_{P(Y \in \mathcal{A})}}{P(Y \in \mathcal{A} \mid \beta, v_a = 0)}. \qquad (2.4.1)$$

Peng and Siegmund (2006) showed that the score for the denominator likelihood is zero for the linkage problem starting with a "conditional on IBD" likelihood. Below, we use similar ideas as outlined in their proof to show that the score for the denominator of (2.4.1) is zero.

$$
\begin{aligned}
&\frac{\partial}{\partial\beta}\log P(Y\in\mathcal{A}\mid\beta)=\frac{\partial}{\partial\beta}\log E_Y(1_{Y\in\mathcal{A}}\mid\beta)\\
&=\frac{\partial}{\partial\beta}\log\int_Y 1_{Y\in\mathcal{A}}\;P(Y\mid\beta)\;dY\\
&=\frac{\partial}{\partial\beta}\log\int_Y\sum_{g_m,\hat\Pi_m}\frac{P(Y,g_m,\hat\Pi_m\mid\beta)}{P(Y,g_m,\hat\Pi_m\mid\beta=0)}\;\frac{1_{Y\in\mathcal{A}}\;P(Y,g_m,\hat\Pi_m\mid\beta=0)}{P(Y\in\mathcal{A},g_m,\hat\Pi_m\mid\beta=0)}\;P(Y\in\mathcal{A},g_m,\hat\Pi_m\mid\beta=0)\;dY\\
&=\frac{\partial}{\partial\beta}\log E_{Y,g_m,\hat\Pi_m}\left[\frac{P(Y,g_m,\hat\Pi_m\mid\beta)}{P(Y,g_m,\hat\Pi_m\mid\beta=0)}\;P(Y\in\mathcal{A},g_m,\hat\Pi_m\mid\beta=0)\;\Big|\;Y\in\mathcal{A},\beta=0\right]\\
&=E_{Y,g_m,\hat\Pi_m}\left[l'_{Y,g_m,\hat\Pi_m}(\beta)\;\Big|\;Y\in\mathcal{A},\beta=0\right]\\
&=\frac{E_{Y,g_m,\hat\Pi_m}\left[1_{Y\in\mathcal{A}}\cdot l'_{Y,g_m,\hat\Pi_m}(\beta)\;\Big|\;Y\in\mathcal{A},\beta=0\right]}{P(Y\in\mathcal{A}\mid\beta=0)}\\
&=\frac{E_{Y,g_m,\hat\Pi_m}\left[1_{Y\in\mathcal{A}}\cdot l'_{Y|g_m,\hat\Pi_m}(\beta)\;\Big|\;Y\in\mathcal{A},\beta=0\right]}{P(Y\in\mathcal{A}\mid\beta=0)},
\end{aligned}
\tag{2.4.2}
$$

where the last step follows from (1), i.e., the score for the joint model is same as that for the forward model. In this proof, we have used the fact that the logarithmic derivative can be taken inside the expectation for the normal distribution. The numerator expectation in the last expression can be seen to be zero by conditioning on $Y$. In this case,

$$E(1_{Y\in\mathcal{A}}\;\tilde{Y}'\Sigma_Y^{-1}\tilde{g}_m)=E_Y[1_{Y\in\mathcal{A}}\;\tilde{Y}'\Sigma_Y^{-1}E_{\beta=0}(\tilde{g}_m\mid Y)]=0.$$

Thus the score for the joint model can be obtained by ignoring the ascertainment scheme. Note that result (2) can also be obtained as a special case of (3) with $G=\mathcal{A}$. Also note that by differentiating equation (2.4.2) with respect to $\beta$ again and carrying the derivative inside the integral we get

$$\frac{\partial^2}{\partial\beta^2}\log P(Y\in\mathcal{A}\mid\beta)=E_{Y,g_m,\hat\Pi_m}\left[1_{Y\in\mathcal{A}}\cdot l''_{Y,g_m,\hat\Pi_m}(\beta)\big|Y\in\mathcal{A},\beta=0\right],$$

This implies that the LMPU score given by $l'(\beta=0)^2+l''(\beta=0)$ is not zero, as the expectation of the second derivative (expression given in Appendix A) is non-zero. Thus the ascertainment scheme can not be ignored for the second derivative and hence for the LMPU score.

### 2.4.2 Explicit Model

The scores under the implicit model can also be motivated as scores for the parameters $\tilde{\theta}$ and $\tilde{\Delta}$ under the explicit model. More generally they are also the scores for a subclass of the implicit models for which the parameters $v_a$ and $\beta$ can be interpreted as a product of a linkage/association parameter and a segregation parameter. In these cases the joint, forward and "conditional on trait" likelihoods are all proportional to the likelihood $L(Y, g_m, \hat{\Pi}_m \mid Y \in \mathcal{A})$, as shown below

$$\begin{aligned}
P(Y, g_m, \hat{\Pi}_m \mid Y \in \mathcal{A}, \hat{\mu}_0, \hat{\Sigma}_0, \tilde{\theta}, \tilde{\Delta}) &= \frac{P(Y, g_m, \hat{\Pi}_m \mid \hat{\mu}_0, \hat{\Sigma}_0, \tilde{\theta}, \tilde{\Delta})\ 1_{Y\in\mathcal{A}}}{P(Y \in \mathcal{A} \mid \hat{\mu}_0, \hat{\Sigma}_0)} \\
&\propto P(Y, g_m, \hat{\Pi}_m \mid \hat{\mu}_0, \hat{\Sigma}_0, \tilde{\theta}, \tilde{\Delta}) \\
&= P(Y \mid g_m, \hat{\Pi}_m, \hat{\mu}_0, \hat{\Sigma}_0, \tilde{\theta}, \tilde{\Delta}) P(g_m, \hat{\Pi}_m) \\
&= P(g_m, \hat{\Pi}_m \mid Y, \hat{\mu}_0, \hat{\Sigma}_0, \tilde{\theta}, \tilde{\Delta}) P(Y \mid \hat{\mu}_0, \hat{\Sigma}_0), \\
\text{Thus, } L_{Y, g_m, \hat{\Pi}_m \mid Y\in\mathcal{A}}(\tilde{\theta}, \tilde{\Delta}) &\propto L_{Y, g_m, \hat{\Pi}_m}(\tilde{\theta}, \tilde{\Delta}) \quad \text{[Ascertainment can be ignored]} \\
&\propto L_{Y \mid g_m, \hat{\Pi}_m}(\tilde{\theta}, \tilde{\Delta}) \quad \text{[Forward Model]} \\
&\propto L_{g_m, \hat{\Pi}_m \mid Y}(\tilde{\theta}, \tilde{\Delta}) \quad \text{[Conditional on Trait model].}
\end{aligned}$$

The proportionality of the likelihoods follow from the fact that the marginal distributions $[Y]$ and $[g_m, \hat{\Pi}_m]$ are free from the parameters $\tilde{\theta}$ and $\tilde{\Delta}$. Note that the parameters $\alpha$ and $\sigma_a^2$ appear in all of the above likelihoods but they have been suppressed, as the final standardized scores would be free of them (they act as proportionality constants). Further, the proportionality of the likelihoods implies the identity of the first and second derivatives at any parameter value, including the null hypothesis value. Thus, in this case, the invariance of the scores (LMP and LMPU) to ascertainment scheme and to conditioning on trait or marker data follows simply on the basis of the assumptions that the marginals distribution of the phenotype (and the marker data) are free from the linkage and association parameters and that the ascertainment only depends on one of the variables (in this case the phenotype).

In sections 2.4.1 and 2.4.2 above, we demonstrated the invariance of the numerators of the score statistics under minimal assumptions. Score statistics should be standardized by the variance computed (or estimated) under the appropriate likelihood to obtain the desired score test. Typically, for selected samples, empirical estimates of the "conditional on trait" variance $Var(Score \mid Y)$ are used instead of $Var(Score \mid Y \in \mathcal{A})$. This strategy of using a sufficient statistic $Y$ generally leads to some loss of power unless the ascertainment scheme is completely arbitrary. When the ascertainment scheme is such that a minimal sufficient statistic $T(Y)$ can be obtained (for example, $T(Y) = |Y_1 - Y_2|$ for an EDAC sibpair design), using conditional variance on $T(Y)$ can improve power. The numerator is however invariant to the choice of $T(Y)$. One way to avoid conditioning on sufficient statistics is to use a completely empirical variance, which accounts for arbitrary sampling schemes while being robust. But for small sample sizes, empirical variance estimates may be conservative (see chapter 3).

The null hypothesis estimates $\hat{\mu}_0$ and $\hat{\Sigma}_0$ are essential in evaluating the scores. They should ideally be obtained by maximizing the likelihood

$$\frac{P(Y \mid \mu_Y, \Sigma_Y, v_a = 0, \beta = 0)}{P(Y \in \mathcal{A} \mid \mu_Y, \Sigma_Y, v_a = 0, \beta = 0)} = \frac{N(Y; \mu_Y, \Sigma_Y)}{\int_{\mathcal{A}} N(Y; \mu_Y, \Sigma_Y)}.$$

Sometimes, the exact ascertainment scheme may be complicated but the probands may be known. Peng and Siegmund (2006) suggested just conditioning on the probands to obtain the CPMLE (Conditional on Proband MLE) instead of the CMLE (Conditional MLE) above. However in some cases, probands may also be ill-defined. In such cases estimation of the trait parameters is an open issue. See section 3.5 and chapter 5 for a discussion of possible ways to obtain estimates under arbitrary ascertainment schemes. When the nuisance parameters are estimated wrongly (for example when the ascertainment scheme or probands are specified wrongly), both the LRT and score tests would suffer loss in power but the score test would have correct type I error using asymptotic thresholds, while the LRT would not. Similarly, when the likelihood model is wrong (e.g., normality is violated) the LRT gives incorrect type I error (with asymptotic thresholds), whereas the score test remains robust as the asymptotic normality of the score statistic is based on the Central Limit Theorem.

## 2.5 DISCUSSION

In this chapter, we described some of the genetic models that are used to derive the score tests for linkage and association that are discussed in this dissertation. In section 2.1, we derived an orthogonal decomposition of the trait mean "conditional on genotype" at a biallelic trait locus. Based on this orthogonal decomposition we obtained the variance component decomposition of the trait variance and the trait covariance "conditional on IBD." We discussed orthogonalization of the covariance decomposition for sibships, which is useful in obtaining a two degree of freedom score statistic (e.g., Tang and Siegmund 2001, chapter 3 of this dissertation). Because of the orthogonality these scores can be computed essentially as a simple sum of squares of one degree of freedom scores for the additive and dominance coefficients.

In section 2.2.1, we discussed some of the standard approaches for association and linkage, which model the mean and the covariance of the trait respectively. Most linkage methods tend to assume a model for $[Y \mid \hat{\Pi}_m]$, ignoring LD, whereas association methods typically model $[Y \mid g_m]$ ignoring the IBD information. The assumption of "no LD" in linkage studies may be reasonable and hence the marker genotype data can be ignored. However for association studies the marker IBD information always provides some information through $Cov(g_m \mid \hat{\Pi}_m)$, irrespective of linkage. In the presence of linkage, the IBD information becomes even more relevant. We proposed an implicit model that uses all of the observed data by modeling the distribution $[Y \mid g_m, \hat{\Pi}_m]$, and discussed how it relates to some of the standard approaches under special cases. All of these models use implicit parameters to indirectly model linkage and association. In section 2.2.2 we analyzed some of the assumptions that are required to motivate the implicit parameters starting from an explicit model relating the two loci, parametrized by $\theta$ and $\Delta$.

The explicit model highlights some of the crude approximations required to justify the implicit models. In equation (2.2.11), the normal approximation for $[g_t \mid g_m, \hat{\Pi}_m]$ may not be appropriate as this distribution is discrete. Similarly the homoscedasticity assumption of $Var(g_t \mid g_m, \hat{\Pi}_m)$ and the assumption of a uniform linear regression slope for $E(\Pi_t \mid \hat{\Pi}_m)$ for all relative types are unrealistic. A more accurate (or even exact) modeling of the discrete distribution $[g_t \mid g_m, \hat{\Pi}_m]$ may lead to more powerful statistics, but is generally avoided due to computational complexity and the possibility of confounding due to nuisance parameters. The scores for the parameters $\theta$ and $\Delta$ under the explicit model 2.2.11 would be proportional to those under implicit models (for parameters $v_a$ and $\beta$) and hence free of nuisance parameters. However, this may not be be true if a different explicit model or an exact one is used. Implicit modeling avoids the nuisance parameters $\alpha$ (additive genetic effect) and $p_t$ (trait allele frequency) by using the confounded parameters (in this case $v_a = 2p_tq_t\alpha^2(1-2\tilde{\theta})$ and $\beta = \sqrt{2p_tq_t\alpha^2/2p_mq_m}$). However, in doing so, they lose the ability to estimate the effect size of the locus $\sigma_a^2 = 2p_tq_t\alpha^2$, except under complete linkage and/or complete LD. This is in contrast to parametric linkage analysis methods for binary traits, which model $\theta$ directly and maximize or search over possible values of nuisance parameters like trait penetrances or prevalence (or equivalently relative risks).

In section 2.3, we described a standard score test for linkage, $S_{VC}$, that assumes "no LD" and a standard score test for association $S_{FBAT}$ that assumes "no linkage." We proposed a score test $S_{VC,LD}$ for linkage that uses the marker genotype information allowing for the possibility of LD. This score test would in general be difficult to compute as estimating the LD parameter $\beta$ under null would require a computationally intensive technique such as MLE or Iteratively Reweighted Least Squares. We also proposed a score test $S_{FBAT,linkage}$ for association that uses the IBD information. It requires an estimate of the linkage parameter $\hat{v}_a$ under the null, which can be obtained by maximizing the VC likelihood. In many practical situations, association studies may be conducted on data already used for a previous linkage study using a VC approach. In such cases the nuisance parameter estimates $\hat{v}_a$ are already available. When the association study is conducted on a dense genome scan and linkage parameter estimates are only known for a sparser subset of the markers, obtaining $\hat{v}_a$ from the closest marker for which it is known may provide a reasonable approximation, while maintaining the computational efficiency. However, estimates of $\hat{v}_a$ from the VC model may be biased for selected samples when the ascertainment correction is not possible or inadequate. In such cases, $S_{FBAT}$ should be preferred.

The discussion in the subsequent chapters focuses on the scores $S_{VC}$ and $S_{FBAT}$ for linkage and association mapping respectively. However most of the variants discussed can be easily extended to the scores $S_{VC,LD}$ and $S_{FBAT,linkage}$, provided the nuisance parameters can be estimated under the null. These parameter estimates can be difficult to obtain under complex ascertainment schemes, as they are based on MLE.

Finally, in section 2.4 we demonstrated the invariance of the score statistics to arbitrary ascertainment schemes. The invariance guarantees that the score statistics derived by ignorning ascertainment, remain optimally powerful for arbitrary ascertainment schemes, provided the nuisance parameter estimates are correct. These parameters can either be esimated by taking the ascertainment into account or obtained independently from a random population sample. Further, using appropriate empirical variance estimate in the denominator ensures that even when the nuisance parameter estimates are biased, the score tests preserve correct type I error.

# 3.0 SCORE STATISTICS FOR LINKAGE

This chapter has been published in American Journal of Human Genetics, volume 82, pages 567-582, March 2008 issue (Bhattacharjee *et al.* 2008). A few minor changes and additions have been made to the text that was published in the journal. The journal grants the authors rights to include the article in full or in part in a thesis or a dissertation. I have also obtained the necessary permissions from the the publisher Elsevier to reporoduce the article with modifications.

## 3.1 INTRODUCTION

Recently, a number of new methods have been developed for Quantitative Trait Locus (QTL) mapping in humans using general pedigrees. Most of these are based on score statistics or regression-based statistics, and attempt to achieve the power of the variance component likelihood-based methods (Amos 1994; Almasy and Blangero 1998) while retaining the robustness and computational simplicity of the original Haseman Elston regression (Haseman and Elston 1972). In principle, these methods should be preferred over the traditional Variance Components (VC) approach, which is extremely sensitive to the normality assumption (e.g., Allison *et al.* 1999). These new methods are theoretically expected to be relatively robust to non-normality of the trait distribution and also to selected sampling. QTL mapping in humans is typically employed for studying disease-related traits and hence selected sampling schemes are common, making score statistics the obvious choice. However the literature on these statistics has mostly focused on theoretical development with less attention given to practical issues and implementation. In this paper we address several of the most important practical issues in the computation and use of these statistics.

The score test is a computationally faster, locally most powerful and robust alternative to the likelihood ratio test. In the context of QTL mapping, this test was proposed by a number of authors (e.g., Tang and Siegmund 2001; Wang 2002; Putter *et al.* 2002; Lebrec *et al.* 2004; Wang 2005). The score test statistic is simply the partial derivative of the VC likelihood with respect to the "linkage parameter" evaluated under the null hypothesis (no linkage) and standardized by its null standard deviation or an estimate thereof. In this article, we refer to the unstandardized score as the "score function" or the "numerator," and the standardizing factor as the "denominator". The aforementioned authors used slightly different parameterizations of the VC likelihood to arrive at the same general formula of the score function for an arbitrary pedigree. The score function remains the same under a broad class of ascertainment schemes, namely, ascertainment through phenotype only (Wang 2005; Peng and Siegmund 2006). For sibling pairs the score function reduces to other statistics like the statistic of Sham and Purcell (2001), which were derived independently as direct ways to improve the power of the Haseman-Elston method by incorporating trait squared sums. Similarly, for general pedigrees, an apparently novel statistic (Sham *et al.* 2002) was derived using a reverse regression approach (regression of IBD on trait information). A number of the statistics, including the VC method, score statistics and the reverse regression method (Sham *et al.* 2002) were unified into a common GEE-based framework (Chen *et al.* 2004; Chen *et al.* 2005) . In particular, their calculations imply the exact equivalence of the numerators of the reverse regression statistic (Sham *et al.* 2002) and the score statistic. They also considered the issue of non-Gaussian traits, and proposed a numerator incorporating higher moments, which was shown to be robust to non-normality. They considered some higher moment based statistics in their simulation study, among a number of other statistics including the VC, score statistics and the reverse regression statistic (Sham *et al.* 2002). Although their simulations indicate the superiority of higher moment-based methods for population samples (of Gaussian and non-Gaussian traits), it is not clear whether the higher-moment versions should be preferred over the usual score statistic numerator for selected samples, where accurate trait parameter estimates may not be available.

For the score test to be robust to distributional assumptions, an empirical variance estimate should be used in the denominator to standardize it. This is because using empirical variances ensures that the statistic follows an asymptotically normal distribution (by the central limit theorem) and hence preserves correct type I error even if the assumed model is wrong. A number of different denominator variants have been proposed (e.g., Wang 2005; Sham *et al.* 2002), ranging from partly to fully empirical variance estimates. Some of these are consistent estimators for the null variance of the score statistic, whereas others are consistent for the true variance. Some condition on the trait values whereas others condition on the identity by descent (IBD) information. The choice of an appropriate denominator is an extremely important issue as it directly affects the power of the linkage statistics. There have been some simulation studies, for selected sibling pairs (T.Cuenco *et al.* 2003; Szatkiewicz *et al.* 2003), to investigate denominator variants. For population samples of sibships, some simulations have been conducted (Chen *et al.* 2005), in which, a few denominator variants were considered, among other issues. Here again, a comprehensive evaluation of the denominators is required - particularly for selected samples - to identify the best combinations of numerator and denominator in terms of power and robustness.

Traditionally, most QTL mapping methods neglect the effect of dominance. This is partly because of the computational simplicity under an additive assumption and also because including dominance leads to a loss of power unless the dominance effect is large enough. Two degree of freedom (2 d.f.) score statistics to incorporate dominance have been suggested by a number of authors (e.g., Tang 2000; Wang 2002). The recent simulation study (Chen *et al.* 2005) included a 2 d.f. variance component statistic but not the score statistic. The results of that study indicated that the gain in power of the 2 d.f. VC statistic for a model exhibiting strong dominance may be more than the loss of power when the model is additive. Similar results were reported for a 2 d.f. score statistic in a previous study(Wang 2002). Appropriately constructed 2 d.f. score statistics would allow for dominance, and would retain other attractive properties such as robustness to selected sampling and non-normality. Here we study the performance of 2 d.f. score statistic vis a vis their 1 d.f. counterparts using simulation across a variety of models.

Like most linkage mapping statistics, score statistics require some nuisance parameters, namely the population trait mean, variance and correlation between relative pairs. The higher moment score statistics require two extra nuisance parameters, the skewness and kurtosis of the trait distribution. These parameters, often called the "segregation parameters", are independent of the "linkage parameters," but specifying incorrect values for these parameters may affect the power of the linkage statistic adversely. In a selected sampling situation, or when the sample sizes are small, it is difficult to obtain reliable estimates of these parameters. There have been a few studies (e.g., T.Cuenco *et al.* 2003; Szatkiewicz *et al.* 2003; Peng and Siegmund 2006) on the effect of misspecification of these parameters on the performance of the score statistics. These studies have generally concluded that some statistics are more sensitive than others to parameter misspecification. They also noted that misspecification of parameters (particularly the trait mean) can have a significant effect on the power of the score statistics. Here we conduct simulations to identify statistics robust to parameter misspecification.

An important issue that has not been dealt with in the literature at all is how to combine pedigrees of different types in an overall score statistic for a dataset. Pedigrees of different sizes and structures have different powers to detect linkage, and thus it is natural to think about giving different weights to different pedigrees in an overall statistic. Also in presence of mixed ascertainment schemes (for example the extreme discordant and concordant sib pair design), there may be gain in power by using higher weights on a part of the data set ( say discordant pairs). Theoretically, score statistics for individual pedigrees should simply be added (not weighted) to get a score statistic for the entire data set. This is because the non-standardized scores are on the same linear scale in terms of local power. However in reality, when conducting a genome scan for a QTL, it would be best to get as much power as possible even for non-local alternatives (which the likelihood ratio variance component test achieves at the cost of computational complexity and robustness). A weighted linear combination of pedigree scores may achieve improvement in power for non-local alternatives, while preserving close to optimal power for local alternatives. We address this issue with some analytical calculations as well as limited number of simulations. All of the simulations in this paper focus on nuclear families, but most of the conclusions generalize to extended pedigrees as well (see discussion).

## 3.2 THEORY

### 3.2.1 Notation

Let us consider a dataset consisting of $K$ types of pedigrees with $n_k$ pedigrees of type $k$ for $k = 1, \ldots, K$, each having $s_k$ pedigree members. Let $y_{ki}$, $M_{ki}$ and $\Pi_{ki}$ denote respectively the vector of phenotypes, the marker data and the matrix of estimated pairwise IBD sharing proportions, for the $i$'th family of type $k$. Let $\mu_k$, $\sigma_k^2$ and $\Sigma_{k0}$ denote the population mean vector, variance vector and dispersion matrix of the phenotype for the pedigrees of type $k$. Let $\Phi_k$ denote the matrix of kinship coefficients for a family of type $k$. We also assume that each pedigrees of type $k$ are selected according to selection criterion $\mathcal{A}_k$ defined purely through its phenotypic data (Wang 2005). Throughout this section, we have omitted the subscript $i$ from expressions such as $Var(vec(\Pi_{ki}))$ which do not depend on $i$, but only on the structure of the pedigree.

### 3.2.2 Numerators

A number of authors (e.g., Wang 2005) have shown that the score statistic for the null hypothesis of "no additive effect of the QTL" under the standard variance components model (for selected and unselected samples) is

$$S_{ki} = v'_{ki} vec(\Pi_{ki} - 2\Phi_{ki}), \tag{3.2.1}$$

where

$$v_{ki} = vec[\Sigma_{k0}^{-1}(y_{ki} - \mu_k)(y_{ki} - \mu_k)'\Sigma_{k0}^{-1} - \Sigma_{k0}^{-1}],$$

and *vec* is an operator which vectorizes the super-diagonal elements of a square matrix in a row-wise order. Under the null hypothesis of no additive variance, the scores $S_{ki}$ have mean zero and variance $E[v'_k Var(vec(\Pi_k))v_k \mid y_k \in \mathcal{A}_k]$. This variance can be estimated using the "conditional on trait value" approach (Lebrec *et al.* 2004) by $v'_{ki} Var(vec(\Pi_k))v_{ki}$. Thus the score test for no additive variance is a one-sided test based on the standardized statistic:

$$T = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k} v'_{ki} vec(\Pi_{ki} - 2\Phi_{ki})}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} v'_{ki} Var(vec(\Pi_k))v_{ki}}, \tag{3.2.2}$$

which has a standard normal distribution under the null. The $Var(vec(\Pi_k))$ in the denominator can be estimated either empirically or using simulation, or using partially empirical methods such as the "imputation" method (Sham *et al.* 2002). This test statistic can also be expressed as a GEE-based score test (Chen *et al.* 2005). As in equation (7) of Chen *et al.* (2005),

$$T = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k} D_{ki}^{a'} G_{k0}^{-1} U_{ki}^{0}}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} U_{ki}^{0'} G_{k0}^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & Var(vec(\Pi_k)) \end{pmatrix} G_{k0}^{-1} U_{ki}^{0}}, \tag{3.2.3}$$

where

$$\begin{aligned} U_{ki}^{0'} &= \begin{bmatrix} (y_{ki} - \mu_k)' & \{(y_{ki} - \mu_k)^2 - \sigma_k^2\}' & vec\{(y_{ki} - \mu_k)(y_{ki} - \mu_k)' - \Sigma_{k0}\}' \end{bmatrix}, \\ D_{ki}^{a'} &= \begin{bmatrix} \mathbf{0}' & \mathbf{0}' & vec(\Pi_{ki} - 2\Phi_k)' \end{bmatrix}, \end{aligned}$$

and $G_{k0}$ is the null Gaussian working covariance matrix of $U^0_{ki}$. By comparing equations (3.2.2) and (3.2.3), we note that $v_{ki}$ consists of the last $\binom{s_k}{2}$ elements of $G^{-1}_{k0}U^0_{ki}$. Thus $v_{ki}$ is a transformed version of the original phenotype vector, using the Gaussian working covariance matrix. We call $v_{ki}$ as the lower moment transformed phenotype.

The GEE formulation was used to construct a new GEE-based robust alternative to the score test (Chen *et al.* 2005), which uses a covariance matrix involving higher moments (skewness and kurtosis) of the phenotype. In analogy with $v_{ki}$, we define $h_{ki}$, as the last $\binom{s_k}{2}$ elements of $M^{-1}_{k0}U^0_{ki}$ , where $M_{k0}$ is the higher moment working covariance matrix (Chen *et al.* 2005). Then a higher moment score test statistic, as in equation (11) of Chen *et al.* (2005) may be simply written as,

$$T = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k} h'_{ki} vec(\Pi_{ki} - 2\Phi_{ki})}{\sum_{k=1}^{K}\sum_{i=1}^{n_k} h'_{ki} Var(vec(\Pi_k))h_{ki}}, \tag{3.2.4}$$

We call $h_{ki}$ the higher moment transformed phenotype.

### 3.2.3 Denominators

For both the lower moment transformed phenotype $v_{ki}$ and the higher-moment transformed phenotype $h_{ki}$, we can conceive of different test statistic denominators, depending on how the null variance of the numerator is estimated. The score function for the unconditional likelihood of the data is same as that based on the likelihood conditioned on trait value or that conditioned on the IBD information (e.g., Lebrec *et al.* 2004). This means that the statistic remains a valid score statistic (for the appropriate likelihood) irrespective of whether a conditional or unconditional variance estimator is used. The unconditional variance of the score function can be decomposed in two ways as shown below. Note that we have dropped all the family subscripts in the expressions below for clarity. Conditioning on trait values we get

$$\begin{aligned}
&Var[v' vec(\Pi - 2\Phi) \mid y \in \mathcal{A}] \\
&\quad = Var_{y|y\in\mathcal{A}}[E\{v' vec(\Pi - 2\Phi) \mid v, y \in \mathcal{A}\}] + E_{y|y\in\mathcal{A}}[Var\{v' vec(\Pi - 2\Phi) \mid v, y \in \mathcal{A}\}] \\
&\quad = Var_y[v' E\{vec(\Pi - 2\Phi) \mid v, y \in \mathcal{A}\} \mid y \in \mathcal{A}] + E_y[v' Var\{vec(\Pi - 2\Phi) \mid v, y \in \mathcal{A}\} v \mid y \in \mathcal{A}]
\end{aligned}$$

Under null this reduces to:

$$= \ 0 \ + E_y[v' Var_\Pi\{vec(\Pi)\} v \mid y \in \mathcal{A}] \tag{3.2.5}$$

(Variance Conditional on Trait)

On the other hand, conditioning on the IBD vector gives

$$
\begin{aligned}
&Var[v' vec(\Pi - 2\Phi) \mid y \in \mathcal{A}] \\
&\quad = Var_{\Pi|y\in\mathcal{A}}[E\{v' vec(\Pi - 2\Phi) \mid \Pi, y \in \mathcal{A}\}] + E_{\Pi|y\in\mathcal{A}}[Var\{v' vec(\Pi - 2\Phi) \mid \Pi, y \in \mathcal{A}\}] \\
&\quad = Var_{\Pi}[vec(\Pi - 2\Phi)' E\{v \mid \Pi, y \in \mathcal{A}\} \mid y \in \mathcal{A}] + E_{\Pi}[vec(\Pi - 2\Phi)' Var\{v \mid \Pi, y \in \mathcal{A}\} vec(\Pi - 2\Phi) \mid y \in \mathcal{A}]
\end{aligned}
$$

Under null this reduces to:

$$
= Var_{\Pi}[vec(\Pi - 2\Phi)' E\{v \mid y \in \mathcal{A}\}] + E_{\Pi}[vec(\Pi - 2\Phi)' Var\{v \mid y \in \mathcal{A}\} vec(\Pi - 2\Phi)]
$$

Further, under "no selection" this reduces to:

$$
= \ 0 \ + E_{\Pi}[vec(\Pi - 2\Phi)' Var_{\Pi} v vec(\Pi - 2\Phi)] \tag{3.2.6}
$$

(Variance Conditional on IBD)

Note that equation (3.2.5) always gives the correct null variance, whereas equation (3.2.6) gives an under-estimate of the null variance (and hence inflated type I error) under selected sampling. Depending on which variable is conditioned upon, there can be a number of approaches for constructing the denominator. Also in each case, the means and variances appearing in Equations (3.2.5) and (3.2.6) can be estimated in different ways, leading to different denominator variants as summarized below.

1. *Conditional on Trait Value Approach.*

   In this approach, the variance of the score function is computed conditional on the trait values as in Equation (3.2.5). This makes the statistic robust to selected sampling. The variance of $vec(\Pi_{ki})$ in the denominator can be estimated in a number of different ways, as follows.

   - SCORE.NULL.CT (Variance Conditional on Trait under NULL) This statistic uses a conditional on the trait approach with an empirical variance of $vec(\Pi_{ki})$ centered at its null expectation:

     $$\sigma^2_{NULL.CT} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} v'_{ki} \hat{\Sigma}_k^{NULL.CT} v_{ki}$$

     where

     $$\hat{\Sigma}_k^{NULL.CT} = \frac{1}{n_k} \sum_{i=1}^{n_k} vec(\Pi_{ki} - 2\Phi_k) vec(\Pi_{ki} - 2\Phi_k)'.$$

- SCORE.CT (Variance Conditional on Trait) This statistic also uses a conditional on the trait approach with empirical variance of $vec(\Pi_{ki})$ centered at its sample mean. By its construction, SCORE.CT is expected to have higher power than SCORE.NULL.CT, for samples ascertained using multiple probands, i.e., whenever $E(\Pi_k \mid y \in \mathcal{A}_k) \neq 2\Phi_k$ under the alternative:

$$\sigma^2_{CT} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} v'_{ki}\hat{\Sigma}^{CT}_k v_{ki} \tag{3.2.7}$$

where

$$\hat{\Sigma}^{CT}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k} vec(\Pi_{ki} - \overline{\Pi}_k)vec(\Pi_{ki} - \overline{\Pi}_k)'.$$

We also considered a higher moment version, HM.CT, of this statistic. This statistic uses the higher moment numerator as in Equation (3.2.4) and the following denominator:

$$\sigma^2_{HM.CT} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} h'_{ki}\hat{\Sigma}^{CT}_k h_{ki}$$

Note that the above definitions of SCORE.CT and HM.CT don't work when there is only one pedigree of a particular type in a dataset. In that case, the sample variance of $vec(\Pi_{ki})$ around its sample mean is zero for that pedigree type. To overcome this problem an empirical variance around the null expectation, i.e., $\hat{\Sigma}^{NULL.CT}_k$ is used for such pedigree types. Thus SCORE.CT reduces to SCORE.NULL.CT when there is one pedigree of each type in the dataset.

- SCORE.MERLIN (MERLIN-REGRESS type denominator) This statistic uses the imputed variance estimate of the IBD (Sham *et al.* 2002) as implemented in the software MERLIN-REGRESS (i.e., difference of the prior and posterior variances):

$$\sigma^2_{MERLIN} = \sum_{k=1}^{K}\sum_{i=1}^{n_k} v'_{ki}\hat{\Sigma}^{MERLIN}_{ki} v_{ki}$$

where

$$\hat{\Sigma}^{MERLIN}_{ki} = Var(vec(\tilde{\Pi}_k)) - Var(vec(\tilde{\Pi}_{ki}) \mid M_{ki})$$

where $\tilde{\Pi}_{ki}$ denotes the (unobserved) true IBD matrix. We also included the higher moment version HM.MERLIN of this statistic discussed as "HM-R" in Chen *et al.* (2005). This statistic uses the higher moment numerator as in Equation (3.2.4) and the following denominator:

$$\sigma^2_{HM.MERLIN} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} h'_{ki} \hat{\Sigma}_{ki}^{MERLIN} h_{ki}$$

- SCORE.MERLIN.AV (MERLIN-REGRESS type denominator with an Averaged Variance) We considered a modified version of the SCORE.MERLIN estimator (iii):

$$\sigma^2_{MERLIN.AV} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} v'_{ki} \hat{\Sigma}_{k}^{MERLIN.AV} v_{ki}$$

where

$$\begin{aligned} \hat{\Sigma}_{k}^{MERLIN.AV} &= Var(vec(\tilde{\Pi}_k)) - \frac{1}{n_k} \sum_{i=1}^{n_k} Var(vec(\tilde{\Pi}_{ki}) \mid M_{ki}), \\ &= \frac{1}{n_k} \sum_{i=1}^{n_k} \hat{\Sigma}_{ki}^{MERLIN}. \end{aligned}$$

Both SCORE.MERLIN and SCORE.MERLIN.AV are motivated by the decomposition:

$$\begin{aligned} Var(vec(\tilde{\Pi}_k)) &= Var[E(vec(\tilde{\Pi}_{ki}) \mid M_{ki})] + E[Var(vec(\tilde{\Pi}_{ki}) \mid M_{ki})] \\ &= Var(vec(\Pi_k)) + E[Var(vec(\tilde{\Pi}_{ki}) \mid M_{ki})] \end{aligned}$$

Hence, note that the averaged-variance estimate is expected to give a more accurate estimate of $Var(vec(\Pi_k))$ in general, but reduces to the usual estimate when there is exactly one pedigree of each type in the sample (i.e., $n_k = 1, \ \forall \ k = 1, \ldots, K$). Also, note that the denominator variance estimates of $vec(\Pi_{ki})$ for both SCORE.MERLIN and SCORE.MERLIN.AV can theoretically turn out to be negative for the individual pedigree types, particularly when there are few pedigrees of that type in the sample. However, except in the case of extremely small sample size, the overall denominator would turn out to be positive.

2. *Unconditional Variance approach*

In this approach, the variance of the score function is computed unconditionally, i.e., without conditioning on trait or IBD information.

- SCORE.NULL.EV (Fully Empirical Variance of the score function around its NULL mean, i.e., 0). It was discussed as "score-R" in Chen *et al.* (2005):

$$\hat{\sigma}^2_{NULL.EV} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} S_{ki}^2.$$

- SCORE.EV (Fully Empirical Variance of the score function around its sample mean.) This is expected to have slightly higher power than SCORE.NULL.EV:

$$\hat{\sigma}^2_{EV} = \sum_{k=1}^{K} \frac{n_k}{(n_k - 1)} \sum_{i=1}^{n_k} (S_{ki} - \overline{S}_k)^2.$$

When there is only one pedigree of a particular type, the empirical variance for that pedigree type is computed around the null mean (0) of the score. Thus, SCORE.EV reduces to SCORE.NULL.EV when there is exactly one pedigree of each type.

3. *Variance Conditional on IBD Approach*

- SCORE.NAIVE (Naïve Estimator of Variance) This statistic uses a naïve estimator of variance for the GEE-based score test. It was discussed as "score" in Chen *et al.* (2005). This statistic uses conditioning on IBD as in Equation (3.2.6) with theoretical variance of $v_k$. It is expected to have incorrect type I error for selected samples and also for non-Gaussian traits:

$$\hat{\sigma}^2_{NAIVE} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} D_{ki}^{a'} G_{k0}^{-1} D_{ki}^{a}.$$

We also considered the higher moment version HM.NAÏVE of this statistic discussed as "HM" in Chen *et al.* (2005). It is expected to be slightly more robust in terms of both type I error and power for non-normal traits, but would still have incorrect type I error for selected samples.

This statistic uses a higher moment numerator as in Equation (3.2.4) and the following denominator:

$$\hat{\sigma}^2_{HM.NAIVE} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} D^{a'}_{ki} M^{-1}_{k0} D^{a}_{ki}.$$

- SCORE.CIBD (Empirical Variance Conditional on IBD) This statistic uses the conditional on IBD approach, with variance of the transformed trait $Var(v_{ki})$ estimated empirically centered at the sample mean. This variance is expected to be relatively robust to distributional assumptions (more specifically to misspecification of the working covariance matrix for GEE). However, it can still have incorrect type I error for selected samples:

$$\sigma^2_{CIBD} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} vec(\Pi_{ki} - 2\Phi_k)' \hat{\Sigma}^{CIBD}_k vec(\Pi_{ki} - 2\Phi_k)$$

where

$$\hat{\Sigma}^{CIBD}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (v_{ki} - \overline{v}_k)(v_{ki} - \overline{v}_k)'.$$

Note that as for SCORE.CT, the denominator empirical estimate of $Var(v_{ki})$ for a particular pedigree type becomes zero when there is one pedigree of that type. In such cases, the null expectation of $v_{ki}$ (i.e., 0) is used to center the empirical variance for that pedigree type.

4. *Approach 4: Minimum Variance Approach*

- SCORE.MAX (Maximum of SCORE.CT and SCORE.EV) We note that all the denominators considered above (except $\hat{\sigma}^2_{EV}$) are consistent estimators of the null variance of the numerator (provided each $n_k$ tends to infinity). $\hat{\sigma}^2_{EV}$ being fully empirical, it estimates the true variance of the numerator. In general, the smaller the denominator of the test statistic (under the alternative), the higher is the power of the statistic. It is difficult to decide a priori whether the null or alternative variance is smaller, as this depends on the genetic model. We propose the statistic SCORE.MAX with a standard numerator as in Equation (3.2.3) and the denominator

$$\hat{\sigma}^2_{MAX} = min(\hat{\sigma}^2_{CT}, \hat{\sigma}^2_{EV}).$$

This statistic is effectively a simple maximum of SCORE.CT and SCORE.EV whenever the numerator score is positive. In particular it is equivalent to the simple maximum in terms of both type I error and power any level of significance smaller than 0.5.

Note that this statistic is expected to have correct type I error asymptotically, as the null and true variances are *equal* under the null. At the same time, it should maintain optimal power under all genetic models. However, for small sample sizes, it is expected to have slightly elevated type I error.

### 3.2.4 Dominance

For sibship data, because of the orthogonality of $\pi$ (true IBD between a pair of sibs) and $1_{\pi=0.5}$ (indicator that the pair shares one allele IBD), two orthogonal scores may be obtained and combined easily to form a 2d.f. statistic (Tang 2000; Tang and Siegmund 2001). Following Tang (2000), we define a 2 d.f. score statistic for sibships, as follows. Let $Z_1$ and $Z_2$ be the Z-scores corresponding to the scores for the additive variance ($\alpha$) and dominance variance ($\delta$) respectively. Thus,

$$Z_1 = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k} v'_{ki} vec(\Pi_{ki} - 2\Phi_k)}{\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{n_k} v'_{ki}\hat{\Sigma}_k^{CT} v_{ki}}} \text{ and } Z_2 = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n_k} v'_{ki} vec(\Delta_k - \Pi_{ki}^{(1)})}{\sqrt{\sum_{k=1}^{K}\sum_{i=1}^{n_k} v'_{ki}\hat{\Sigma}_k^{CT(1)} v_{ki}}},$$

where $\Pi_{ki}^{(1)}$ and $\Delta_k$ are the estimated and expected matrix of pairwise probabilities of sharing 1 allele IBD, for the $i^{th}$ pedigree of type $k$.

$\hat{\Sigma}_k^{CT}$ is given by Equation (3.2.7) as before and $\hat{\Sigma}_k^{CT(1)}$ is given by:

$$\hat{\Sigma}_k^{CT(1)} = \frac{1}{n_k - 1}\sum_{i=1}^{n_k} vec(\Pi_{ki}^{(1)} - \overline{\Pi}_k^{(1)}) vec(\Pi_{ki}^{(1)} - \overline{\Pi}_k^{(1)})'.$$

Combining these two Z-scores, subject to the constraint $0 \leq \delta \leq \alpha$, gives the 2 d.f. statistic SCORE.2DF.CT, defined as

$$SCORE.2DF.CT = \begin{cases} Z_1^2 + Z_2^2 & if\ 0 \leq Z_2 \leq (1/\sqrt{2})Z_1 \\ Z_1^2 & if\ Z_2 \leq 0 \leq Z_1 \\ (\sqrt{2/3}Z_1 + \sqrt{1/3}Z_2)^2 & if\ (-1/\sqrt{2})Z_2 \leq Z_1 \leq \sqrt{2}Z_2 \\ 0 & otherwise \end{cases}$$

The higher moment version, HM.2DF.CT, of this statistic can be analogously defined with the higher moment transformed phenotype $h_{ki}$ in the numerator instead of $v_{ki}$. For extended pedigrees, the orthogonal decomposition does not hold, so a two-parameter score statistic would be needed. The information matrix would involve $Cov(\Pi_k, \Pi_k^{(2)})$, which can be estimated empirically.

Note that SCORE.2DF.CT and HM.2DF.CT can run into similar problems as SCORE.CT and HM.CT when the sample consists of only one pedigree of a type, in which case they are modified similarly.

### 3.2.5 Weighting of Pedigrees

Real data often includes pedigrees of different sizes and structures. In such cases, it may be desirable to give appropriate weights to each pedigree type so as to obtain maximum power. The advantage of the likelihood ratio test statistic (Variance Components) is that the weighting is automatic, since the likelihood ratio is evaluated at the maximum likelihood alternative. The score statistic, by contrast, is designed to be locally optimal near the null hypothesis, and under the null hypothesis all pedigrees are weighted equally (or equivalently, standardized scores are weighted in proportion to their null standard deviations). Hence in most of the score statistic literature, equal weighting of pedigree-wise score statistics has been suggested. However, under alternatives away from the null it is quite possible that more power can be obtained by using a score statistic with unequal weighting of different pedigrees. For purists who might object that a weighted score statistic is no longer a score statistic, we point out that the object we call the "score statistic" is only approximately the true score anyway. Strictly speaking, the score function (3.2.1) is derived under a normal model (conditional on IBD). This is not a very realistic model (as the trait should have a mixture distribution when conditioned on IBD), but it is used as a convenient approximation. The same score function can be shown to have some optimality properties under a mixture-normal model

(Tang 2000; Dupuis *et al.* 2007, Section 2.3.2), and is hence generally accepted. Still however in most circumstances the assumption of "normal" or "mixture normal" would fail and hence the statistic (3.2.1) is no longer technically the score function. Similarly the higher moment score function is based on a GEE with an arbitrarily chosen working covariance matrix. When the data violate the higher moment working covariance structure, this statistic is no longer a "GEE-based score statistic." Lastly, when population trait parameters are misspecified (e.g., for an ascertained sample) the above statistics are no longer score statistics and may no longer be additive. Weighting of score statistics may be useful even when the distributional assumption holds. Local optimality ensures that the statistic has optimal power to detect weak effects. The variance component (VC) test is optimal for all alternatives (when the assumed model holds). However it has the disadvantage of being computationally complex and non-robust. By weighting pedigrees, it may be possible to increase the non-local power of the score statistic while retaining most of the local power and robustness properties.

*Notation:* Let $\sigma_a^2$ denote the additive variance and let $\alpha = \sigma_a^2/2$. Let us consider $n_1$ pedigrees of type 1 and $n_2$ pedigrees of type 2. Let$\mu_{0i}$, $\mu_{\alpha,i}$, $\sigma_{0i}^2$ and $\sigma_{\alpha,i}^2$ be the null ($H_0 : \sigma_a^2 = 0$) and alternative ($H_1 : \sigma_a^2 > 0$) means and variances of the score function respectively for pedigrees of type $i = 1,\ 2$. Similarly, we define $m_{\alpha,i}$, $v_{\alpha,i}^2$ to be the means and variances of the standardized score statistic (i.e., centered and scaled to have mean 0 and variance 1). Then, provided $n_1$ and $n_2$ are large, the asymptotic optimal weight for linearly combining the standardized Z-scores from the two types of pedigrees is given by the following expression (Sengul *et al.* 2007):

$$\begin{aligned} w &= \frac{m_{\alpha,2}/v_{\alpha,2}^2}{m_{\alpha,1}/v_{\alpha,1}^2} \\ &= \frac{(\mu_{\alpha,2}-\mu_{0,2})\sigma_{0,2}\sigma_{\alpha,1}^2}{(\mu_{\alpha,1}-\mu_{0,1})\sigma_{0,1}\sigma_{\alpha,2}^2} \end{aligned} \tag{3.2.8}$$

Therefore the optimal weight for the non-standardized score functions is given by:

$$\begin{aligned} w' &= \frac{(\mu_{\alpha,2}-\mu_{0,2})\sigma_{\alpha,1}^2}{(\mu_{\alpha,1}-\mu_{0,1})\sigma_{\alpha,2}^2} = \frac{\mu_{\alpha,2}\sigma_{\alpha,1}^2}{\mu_{\alpha,1}\sigma_{\alpha,2}^2} \\ &= \frac{m_2^2 + 2\alpha m_2^3 + \alpha^2 m_2^4 + (\alpha^2/2)s_2^2}{m_2^2} \times \frac{m_1^2 + 2\alpha m_1^3 + \alpha^2 m_1^4 + (\alpha^2/2)s_1^2}{m_1^2} \end{aligned} \tag{3.2.9}$$

where $m^j = E\{trace[(\Sigma^{-1}A_\pi)^j]\}$ and $s^j = Var\{trace[(\Sigma^{-1}A_\pi)^j]\}$ and subscripts 1 and 2 denote pedigrees of type 1 and 2 respectively. The matrices $\Sigma$ and $A_\pi$ have been defined in Appendix B. The above expressions for moments of the score function under population sampling have been derived in Appendix B. Note that the above formula converges to $w' = 1$ for local alternatives ($\alpha$ close to 0) but not in general. The two weights $w$ and $w'$ defined above are termed as the“standardized optimal weight” and the “non-standardized optimal weight” respectively in the rest of this article.

## 3.3 METHODS

### 3.3.1 Simulation

We conducted a simulation study to compare the performance of score statistic variants for nuclear sibships. Our simulation scheme is similar to that described in T.Cuenco *et al.* (2003). A single biallelic quantitative trait and a single marker with 8 equifrequent alleles were simulated. The recombination distance between the two loci was taken as $\theta = 0.5$ and $\theta = 0$ for simulations under the null and alternative hypothesis respectively.

*Genetic models:* The genetic models used are similar to those in T.Cuenco *et al.* (2003) with a decreased locus specific heritability of 0.15. The details of the models are summarized in Tables 3.1 and 3.2. For the first five models ($1 - 5$), the trait has a mean depending on genotype plus a normally distributed environmental component. The models $1' - 5'$ and $1'' - 5''$ are non-Gaussian models simulated by subjecting the traits simulated under models $1 - 5$ to the transformations $x|x|$ and $x^3$ respectively. Both these sets of models as well as model 3 (rare recessive trait) are expected to depart substantially from the normality assumption. Note that our genetic models do not incorporate polygenic effects explicitly. For our purposes, polygenes can be considered to be a part of the shared environment within the family and hence their effect is modeled by considering environmental correlation between relatives.

*Selection Schemes:* We simulated samples under the following ascertainment schemes - POP (population sampling), SINGLE (single proband sampling with one sib in the top 10% of the trait distribution), ED (extreme discordant sampling with one sib in the top 10% and one in the bottom 10%), EC (extreme concordant sampling with two sibs in the top 10%), EDAC3 (3-corner extreme discordant & concordant sampling with every sibship having a discordant pair at a 12% threshold or a “high concordant” pair at a 4% threshold), MDAC3 (same as EDAC3 with thresholds of 24% and 8% for discordant and concordant pairs respectively). Thus, we defined a “discordant” (or “concordant”) sibship as one having *at least one* discordant (or concordant) sib pair. These

Table 3.1: **Genetic Models: Defining Parameters.**

| Model Parameters: | *Model 1* | *Model 2* | *Model 3* | *Model 4* | *Model 5* |
|---|---|---|---|---|---|
| *Type of inheritance* | Additive | Dominant | Recessive | Additive | Dominant |
| *Locus heritability* | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 |
| *Allele frequency* | 0.1 | 0.1 | 0.1 | 0.5 | 0.5 |
| *Trait means* | -1,0,1 | 0,1,1 | 0,0,1 | -1,0,1 | 0,1,1 |
| *Environmental SD* | 1.010 | 0.934 | 0.237 | 1.683 | 1.031 |
| *Environmental correlation* | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 |

Table 3.2: **Genetic Models: Population Trait Parameters.**

| Models | Parameters | | | | |
|---|---|---|---|---|---|
| | *Mean* | *SD* | *Correlation* | *Skewness* | *Kurtosis* |
| *Normal Models* | | | | | |
| 1 | -0.80 | 1.095 | 0.288 | 0.110 | 0.058 |
| 2 | 0.19 | 1.013 | 0.286 | 0.092 | 0.011 |
| 3 | 0.01 | 0.257 | 0.257 | 0.572 | 2.138 |
| 4 | 0.00 | 1.826 | 0.288 | 0.000 | -0.023 |
| 5 | 0.75 | 1.118 | 0.275 | -0.067 | -0.015 |
| *Non-normal:* $x\|x\|$ | | | | | |
| 1’ | -1.49 | 6.758 | 0.244 | -1.660 | 6.419 |
| 2’ | 0.33 | 3.379 | 0.247 | 1.151 | 9.094 |
| 3’ | 0.01 | 0.023 | 0.241 | 5.821 | 65.848 |
| 4’ | 0.00 | 32.531 | 0.250 | -0.069 | 8.001 |
| 5’ | 1.41 | 6.894 | 0.234 | 1.726 | 6.257 |
| *Non-normal:* $x^3$ | | | | | |
| 1” | -3.22 | 55.940 | 0.182 | -3.783 | 26.989 |
| 2” | 0.69 | 18.719 | 0.191 | 3.649 | 48.395 |
| 3” | 0.01 | 0.022 | 0.222 | 12.387 | 207.990 |
| 4” | 0.06 | 524.930 | 0.180 | 0.051 | 36.345 |
| 5” | 3.11 | 58.087 | 0.180 | 3.759 | 28.926 |

ascertainment schemes have been discussed before in the context of sibpairs (T.Cuenco *et al.* 2003; Szatkiewicz *et al.* 2003). It is possible to define other notions of concordant and discordant sibships, such as by standard deviation of the sibship trait values (Tang 2000), but we consider the above definitions to be more realistic, as sibships are often ascertained through an affected sib or an affected sib-pair.

*Family Sizes:* Most of our simulations were done using sibships of size 4 without parental phenotype information. Parental genotype information was used to estimate IBD sharing between siblings. We did limited simulations with sibships of size 2 and 6, but there were no qualitative differences in the results, except for the expected effects of the increased and decreased sample size respectively. Hence we report only results for sibships of size 4.

*Sample Sizes:* As the objective of our simulation experiments was to compare the statistics to each other, the absolute value of power was not considered to be relevant. We chose the sample sizes arbitrarily to keep the power within a reasonable range (i.e., not too high or low) to facilitate comparison across statistics. The sample sizes for the normally distributed data were 450 families for POP samples, 100 for SINGLE, 150 for MDAC3 and 50 each for ED, EC and EDAC3. The corresponding sample sizes for data transformed using $x|x|$ were 750 (POP), 200 (SINGLE), 300 (MDAC3) and 100 (ED, EC and EDAC3) and those for data transformed using $x^3$ were 1000 (POP), 300 (SINGLE), 500 (MDAC3) and 200 (ED, EC and EDAC3).

We used 1,000 and 10,000 replicates to estimate the power and type I error respectively at a significance level of 0.01. For computing the analytical thresholds, the asymptotic null distributions of the statistics were used. The null distribution of the 1 d.f. statistics is asymptotically , which was used to obtain two-sided p-values. The null distribution of the 2 d.f. statistics is asymptotically a mixture of $\chi^2_2$, $\chi^2_1$ and 0 in the ratio $\psi_0/2\pi : 1/2 : (\pi - \psi_0)/2\pi$, where $\psi_0 = \tan^{-1}(1/\sqrt{2})$ (Tang 2000), which was used to obtain one-sided p-values. For all the type I error and power simulations, the trait parameters were set at their known true values (as given in Table 3.2). The estimated type I errors for the schemes POP and ED have been summarized in Table 3.4(A-B). The type I errors for the other sampling schemes have been summarized in the Supplementary Table F1 in Appendix F. The estimated powers of some of the above statistics have been summarized in Tables 4.3(A-F). The powers of all the statistics have been summarized in the Supplementary Table F2 in Appendix F.

Table 3.3: **Sensitivity Analysis: Misspecified Parameters**

| | *Model 2* | | | *Model 2'* | | | *Model 2"* | | | *Model 4* | | | *Model 4'* | | | *Model 4"* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Parameter: | True | Lower | Upper | True | Lower | Upper | True | Lower | Upper | True | Lower | Upper | True | Lower | Upper | True | Lower | Upper |
| *Mean* | 0.19 | -0.80 | 1.20 | 0.33 | -2.70 | 3.30 | 0.69 | -3.30 | 4.70 | 0.00 | -2.00 | 2.00 | 0.00 | -10.00 | 10.00 | 0.06 | -40.00 | 40.00 |
| *Variance* | 1.03 | 0.03 | 2.03 | 3.38 | 0.40 | 6.40 | 18.72 | 3.70 | 33.70 | 3.33 | 0.33 | 6.33 | 32.53 | 12.53 | 52.53 | 524.93 | 324.93 | 724.93 |
| *Correlation* | 0.29 | 0.10 | 0.50 | 0.25 | 0.10 | 0.40 | 0.19 | 0.05 | 0.35 | 0.29 | 0.10 | 0.50 | 0.25 | 0.10 | 0.40 | 0.18 | 0.05 | 0.35 |
| *Skewness* | 0.09 | -0.90 | 1.10 | 1.15 | -2.80 | 3.20 | 3.65 | -16.40 | 23.60 | 0.00 | -1.00 | 1.00 | -0.07 | -5.00 | 5.00 | 0.05 | -25.00 | 25.00 |
| *Kurtosis* | 0.01 | -2.00 | 2.00 | 9.09 | 3.10 | 15.10 | 48.40 | -11.60 | 108.40 | -0.02 | -2.00 | 2.00 | 8.00 | -2.00 | 18.00 | 36.35 | -13.70 | 86.30 |

### 3.3.2 Sensitivity Analysis

To evaluate the robustness of the statistics to misspecification of population trait parameters, we carried out sensitivity analysis using simulation. For these simulations, we chose four selection schemes (POP, ED, EC and EDAC) and 6 models (2, $2'$,$2''$ and 4, $4'$,$4''$). The five trait parameters (namely mean, variance, correlation, skewness and kurtosis) were in turn set at two arbitrary wrong guesses on either side of the true value, while holding the other four parameters fixed at their true values. The misspecified parameter values have been listed in Table 3.3. Power was then estimated based on the same 1,000 replicates of data, for each combination of parameter values. This process was repeated for all the combinations of models and selection schemes. SCORE.NAÏVE and HM.NAÏVE have theoretically incorrect type I error when parameters are incorrect. SCORE.CIBD has theoretically incorrect type I error for selected samples. So, these three statistics were dropped from this analysis. The results of the sensitivity analysis have been summarized in Figures 3.1 and 3.2.

### 3.3.3 Weighting

As described in the previous section, Equation (3.2.9) can be used to derive optimal weights for sibships of various sizes for different alternative values of the parameter (under population sampling.) We plotted the optimal weights, as a function of heritability ($h^2$) for sibships of sizes 3, 4, 5 and 6 with respect to sibpairs (Figure 3.3). For sibships of size 3 versus sibpairs, we also plotted the behavior of the analytical power curve (Sengul *et al.* 2007) of SCORE.NAÏVE for different values of $h^2$ (Figure 3.4). When we have an ascertained sample (for example, an EDAC sample), Equation (3.2.9) no longer holds. But Equation (3.2.8) can be used to derive the optimal weight for discordant pairs with respect to concordant pairs, where the means and variances are conditional on the ascertainment scheme and can be obtained by numerical integration. Alternatively, power can also be estimated using simulation over a grid of different weights. Figure 3.5 shows the simulation-based power of SCORE.CT for a mixed sample of 20 extreme discordant pairs (one sib in each of higher and lower 10% tails) and 30 extreme concordant pairs (both sibs in the top 10% tail), as a function of the non-standardized weight of a discordant pair with respect to a concordant pair.

## 3.4 RESULTS

### 3.4.1 Simulation Results

The type I errors for the Population and Extreme Discordant sampling schemes have been tabulated in Table 3.4(A-B) and for other sampling schemes in the Supplementary Table F1 in Appendix F. Most of the statistics have close to correct type I error even for the smallish sample sizes that we used. The type I errors for SCORE.NAIVE and HM.NAVE are highly inflated for non-normal as well as selected samples. Similarly, in some cases, the type I error of SCORE.CIBD are inflated for selected samples. Theoretically, all three of these statistics have inflated type I error for selected samples. On the other hand, SCORE.NULL.EV and SCORE.EV have highly conservative type I error. The SCORE.MAX statistic has negligibly inflated type I errors, compared to SCORE.CT. All the statistics except HM.CT and HM.MERLIN have slightly incorrect type I error, in most cases, for the highly skewed models $3'$ and $3''$. The higher moment statistics in general give better type I errors than their lower moment counterparts particularly for the non-normal models. In most cases however, the difference is marginal.

The estimated power for all the models and sampling schemes are summarized in Table 4.3(A-F). SCORE.NAÏVE, HM.NAÏVE and SCORE.CIBD have been dropped from the power tables 4.3(B-F), as they have theoretically incorrect type I error for selected samples. To facilitate comparison, we have also dropped SCORE.NULL.CT, SCORE.NULL.EV and SCORE.MERLIN.AV from the power tables 4.3(B-F). SCORE.CT and SCORE.EV are consistently (and sometimes significantly) more powerful than SCORE.NULL.CT and SCORE.NULL.EV respectively, while the type I errors are negligibly higher. SCORE.MERLIN.AV has also been dropped, as it fails to provide significant

Table 3.4: **Type I Error**

| **(A) Population** | Genetic Model | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **$1'$** | **$1''$** | **2** | **$2'$** | **$2''$** | **3** | **$3'$** | **$3''$** | **4** | **$4'$** | **$4''$** | **5** | **$5'$** | **$5''$** |
| **SCORE.NAÏVE** | 0.011 | **0.026** | **0.063** | 0.01 | **0.029** | **0.087** | **0.033** | **0.209** | **0.295** | 0.011 | **0.024** | **0.065** | 0.012 | **0.026** | **0.072** |
| **SCORE.CIBD** | 0.011 | 0.011 | 0.011 | 0.009 | 0.012 | 0.013 | 0.014 | **0.015** | **0.015** | 0.011 | 0.011 | 0.012 | 0.012 | 0.013 | 0.012 |
| **SCORE.NULL.CT** | 0.011 | 0.011 | 0.011 | 0.009 | 0.012 | 0.013 | 0.014 | **0.015** | **0.015** | 0.011 | 0.011 | 0.012 | 0.012 | 0.013 | 0.012 |
| **SCORE.CT** | 0.011 | 0.011 | 0.011 | 0.009 | 0.012 | 0.013 | 0.014 | **0.015** | **0.015** | 0.011 | 0.011 | 0.012 | 0.012 | 0.013 | 0.012 |
| **SCORE.NULL.EV** | 0.007 | 0.007 | **0.005** | **0.005** | **0.005** | **0.004** | 0.006 | **0.002** | **0.001** | 0.006 | 0.006 | **0.005** | 0.008 | 0.007 | 0.006 |
| **SCORE.EV** | 0.007 | 0.007 | **0.005** | 0.006 | **0.005** | **0.005** | 0.006 | **0.002** | **0.001** | 0.007 | 0.007 | **0.005** | 0.008 | 0.008 | 0.006 |
| **SCORE.MERLIN** | 0.011 | 0.011 | 0.011 | 0.009 | 0.012 | 0.012 | 0.013 | **0.016** | 0.013 | 0.011 | 0.011 | 0.011 | 0.012 | 0.013 | 0.012 |
| **SCORE.MERLIN.AV** | 0.011 | 0.011 | 0.011 | 0.009 | 0.012 | 0.013 | 0.014 | **0.016** | **0.015** | 0.011 | 0.012 | 0.012 | 0.012 | 0.012 | 0.012 |
| **HM.NAÏVE** | 0.011 | **0.025** | **0.061** | 0.01 | **0.031** | **0.073** | **0.033** | **0.22** | **0.299** | 0.011 | **0.021** | **0.055** | 0.012 | **0.024** | **0.066** |
| **HM.MERLIN** | 0.011 | 0.011 | 0.01 | 0.009 | 0.011 | 0.01 | 0.013 | 0.012 | 0.013 | 0.011 | 0.011 | 0.011 | 0.012 | 0.013 | 0.012 |
| **HM.CT** | 0.011 | 0.012 | 0.01 | 0.009 | 0.011 | 0.01 | 0.013 | 0.012 | 0.014 | 0.011 | 0.011 | 0.011 | 0.012 | 0.013 | 0.011 |
| **SCORE.MAX** | 0.011 | 0.011 | 0.012 | 0.009 | 0.013 | **0.015** | **0.015** | **0.016** | **0.015** | 0.011 | 0.013 | 0.014 | 0.012 | 0.014 | 0.014 |
| **SCORE.2DF.CT** | 0.011 | 0.011 | 0.011 | 0.01 | 0.012 | 0.012 | 0.013 | **0.018** | **0.015** | 0.01 | 0.012 | 0.011 | 0.011 | 0.012 | 0.012 |
| **HM.2DF.CT** | 0.011 | 0.012 | 0.011 | 0.01 | 0.011 | 0.014 | 0.014 | **0.017** | **0.019** | 0.01 | 0.012 | 0.012 | 0.011 | 0.012 | 0.013 |
| **(B) Extreme Discordant** | | | | | | | | | | | | | | | |
| **SCORE.NAÏVE** | **0.178** | **0.148** | **0.133** | **0.174** | **0.179** | **0.155** | **0.225** | **0.314** | **0.341** | **0.168** | **0.191** | **0.155** | **0.164** | **0.145** | **0.125** |
| **SCORE.CIBD** | **0.015** | **0.015** | **0.015** | **0.016** | **0.015** | 0.012 | **0.016** | **0.017** | **0.015** | **0.015** | **0.015** | 0.013 | 0.013 | 0.014 | 0.013 |
| **SCORE.NULL.CT** | 0.011 | 0.011 | 0.011 | 0.012 | 0.012 | 0.011 | 0.013 | **0.016** | 0.014 | 0.011 | 0.012 | 0.011 | 0.01 | 0.01 | 0.01 |
| **SCORE.CT** | 0.012 | 0.012 | 0.012 | 0.013 | 0.013 | 0.011 | 0.014 | **0.016** | 0.014 | 0.012 | 0.012 | 0.012 | 0.011 | 0.01 | 0.011 |
| **SCORE.NULL.EV** | **0.005** | 0.006 | **0.005** | **0.005** | 0.007 | **0.005** | **0.002** | **0.002** | **0.001** | **0.005** | 0.006 | **0.005** | **0.004** | **0.005** | **0.005** |
| **SCORE.EV** | 0.007 | 0.007 | 0.006 | 0.008 | 0.009 | **0.005** | **0.005** | **0.002** | **0.002** | 0.008 | 0.008 | 0.006 | 0.007 | 0.007 | **0.005** |
| **SCORE.MERLIN** | 0.012 | 0.012 | 0.012 | 0.013 | 0.013 | 0.01 | **0.015** | **0.016** | **0.016** | 0.013 | 0.011 | 0.012 | 0.011 | 0.01 | 0.01 |
| **SCORE.MERLIN.AV** | 0.012 | 0.012 | 0.011 | 0.012 | 0.012 | 0.011 | 0.014 | **0.016** | **0.015** | 0.013 | 0.012 | 0.012 | 0.011 | 0.01 | 0.011 |
| **HM.NAÏVE** | **0.178** | **0.109** | **0.065** | **0.174** | **0.139** | **0.085** | **0.212** | **0.295** | **0.312** | **0.169** | **0.144** | **0.092** | **0.164** | **0.114** | **0.06** |
| **HM.MERLIN** | 0.012 | 0.01 | 0.011 | 0.013 | 0.012 | 0.011 | 0.014 | **0.016** | 0.012 | 0.013 | 0.012 | 0.01 | 0.011 | 0.011 | 0.011 |
| **HM.CT** | 0.012 | 0.011 | 0.012 | 0.013 | 0.013 | 0.011 | 0.014 | **0.015** | 0.012 | 0.013 | 0.012 | 0.01 | 0.011 | 0.011 | 0.011 |
| **SCORE.MAX** | 0.012 | 0.013 | 0.014 | 0.014 | **0.015** | 0.013 | **0.015** | **0.017** | **0.016** | 0.013 | 0.014 | 0.014 | 0.012 | 0.012 | 0.012 |
| **SCORE.2DF.CT** | 0.01 | 0.011 | 0.011 | 0.013 | 0.012 | 0.01 | 0.014 | **0.016** | **0.015** | 0.012 | 0.011 | 0.009 | 0.01 | 0.011 | 0.011 |
| **HM.2DF.CT** | 0.01 | 0.011 | 0.011 | 0.013 | 0.013 | 0.011 | **0.015** | **0.018** | **0.016** | 0.012 | 0.012 | 0.009 | 0.01 | 0.011 | 0.012 |

Note: Type I error values departing by 0.005 or more, from the nominal value 0.01 are highlighted in bold.

improvement of power over SCORE.MERLIN under most genetic models and selection schemes. In fact, it has slightly reduced power in many cases. The detailed results with all the statistics are given in the Supplementary Table F2 in Appendix F.

For all the models and schemes, the unconditional empirical variance denominator SCORE.EV performs poorly. It has low power and a conservative type I error, which can be attributed to the smallish sample sizes. In their simulations, Chen *et al.* (2005) observed similar behavior for *SCORE.NULL.EV* (denoted as “score-R” in their paper).

For population samples, under normal models (1, 2, 4 and 5) all the statistics perform essentially identically. SCORE.NAÏVE, HM.NAÏVE and SCORE.CIBD have similar power to the other statistics. As noted previously (Chen *et al.* 2005), the higher moment (HM) statistics perform at par with the lower moment (LM) statistics in this case.

For population samples under non-normal models, SCORE.NAÏVE and HM.NAÏVE have inflated type I error. The HM statistics show improvement in power for only some cases, which disagrees with the previous conclusion of Chen *et al.* (2005) that HM statistics are always better

Table 3.5: **Power Results**

| (A) **Population** | Genetic Model | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **$1'$** | **$1''$** | **2** | **$2'$** | **$2''$** | **3** | **$3'$** | **$3''$** | **4** | **$4'$** | **$4''$** | **5** | **$5'$** | **$5''$** |
| **SCORE.NAÏVE** | **0.74** | | | **0.73** | | | **0.65** | | | **0.75** | | | **0.74** | | |
| **SCORE.CIBD** | **0.74** | **0.39** | **0.14** | **0.73** | **0.78** | 0.45 | 0.53 | **0.96** | **0.94** | **0.76** | **0.69** | 0.31 | **0.74** | **0.44** | **0.15** |
| **SCORE.CT** | **0.74** | **0.39** | **0.14** | **0.73** | **0.78** | 0.45 | 0.53 | **0.96** | **0.94** | **0.76** | **0.69** | 0.31 | **0.74** | **0.44** | **0.15** |
| **SCORE.EV** | 0.67 | 0.35 | 0.11 | 0.68 | 0.73 | 0.4 | 0.24 | 0.75 | 0.73 | 0.7 | 0.65 | 0.32 | 0.7 | 0.41 | 0.11 |
| **SCORE.MERLIN** | **0.74** | **0.39** | **0.14** | **0.73** | **0.78** | 0.45 | 0.53 | **0.97** | **0.95** | **0.75** | **0.68** | 0.31 | **0.74** | **0.44** | **0.16** |
| **HM.NAÏVE** | **0.74** | | | **0.73** | | | **0.62** | | | **0.75** | | | **0.74** | | |
| **HM.MERLIN** | **0.74** | 0.37 | **0.14** | **0.73** | 0.74 | **0.48** | 0.51 | 0.92 | 0.91 | **0.75** | 0.67 | 0.34 | **0.75** | 0.42 | **0.16** |
| **HM.CT** | **0.74** | 0.36 | **0.14** | **0.73** | 0.74 | **0.49** | 0.51 | 0.9 | 0.91 | **0.76** | 0.67 | 0.33 | **0.74** | 0.42 | **0.16** |
| **SCORE.MAX** | **0.74** | **0.41** | **0.16** | **0.74** | **0.81** | **0.5** | 0.53 | **0.98** | **0.96** | **0.76** | **0.71** | **0.37** | **0.75** | **0.46** | **0.18** |
| **SCORE.2DF.CT** | 0.71 | 0.37 | **0.14** | **0.72** | 0.75 | 0.43 | **0.62** | **0.98** | **0.97** | 0.72 | 0.66 | 0.29 | **0.76** | **0.45** | 0.15 |
| **HM.2DF.CT** | 0.71 | 0.34 | 0.12 | **0.72** | 0.7 | 0.45 | 0.58 | 0.93 | 0.91 | 0.72 | 0.62 | 0.32 | **0.76** | 0.42 | **0.16** |
| (B) **Single Proband Ascertainment** | | | | | | | | | | | | | | | |
| **SCORE.CT** | **0.69** | **0.78** | 0.54 | **0.7** | **0.78** | 0.53 | 0.79 | **0.99** | **0.99** | **0.38** | **0.43** | 0.24 | **0.2** | **0.19** | **0.12** |
| **SCORE.EV** | 0.59 | 0.71 | 0.49 | 0.59 | 0.74 | 0.52 | 0.4 | 0.93 | 0.92 | 0.29 | 0.37 | 0.18 | 0.13 | 0.16 | 0.09 |
| **SCORE.MERLIN** | **0.69** | 0.78 | 0.53 | **0.69** | **0.78** | 0.55 | 0.8 | **1** | **0.99** | **0.38** | **0.43** | 0.23 | **0.2** | **0.19** | **0.12** |
| **HM.MERLIN** | **0.69** | **0.81** | **0.66** | **0.69** | 0.73 | **0.58** | 0.78 | **0.99** | **0.99** | **0.38** | 0.38 | 0.22 | **0.2** | **0.17** | **0.11** |
| **HM.CT** | **0.69** | **0.8** | **0.65** | **0.69** | 0.73 | 0.57 | 0.76 | **0.98** | **0.98** | **0.38** | 0.37 | 0.22 | **0.2** | **0.17** | **0.11** |
| **SCORE.MAX** | **0.7** | **0.81** | 0.61 | **0.7** | **0.8** | **0.61** | 0.79 | **1** | **0.99** | **0.39** | **0.45** | **0.28** | **0.21** | **0.2** | **0.13** |
| **SCORE.2DF.CT** | 0.66 | 0.76 | 0.52 | 0.65 | 0.75 | 0.51 | **0.85** | **1** | **0.99** | 0.36 | 0.4 | 0.21 | **0.21** | **0.2** | **0.11** |
| **HM.2DF.CT** | 0.66 | 0.78 | 0.63 | 0.65 | 0.7 | 0.53 | **0.83** | **0.99** | **0.99** | 0.36 | 0.35 | 0.21 | **0.22** | **0.18** | **0.11** |
| (C) **Extreme Discordant** | | | | | | | | | | | | | | | |
| **SCORE.CT** | **0.59** | 0.78 | 0.74 | **0.59** | **0.81** | 0.85 | **0.15** | 0.77 | 0.92 | **0.25** | **0.77** | 0.87 | **0.53** | **0.68** | 0.7 |
| **SCORE.EV** | 0.48 | 0.7 | 0.72 | 0.52 | 0.78 | 0.85 | 0.04 | 0.43 | 0.75 | 0.18 | 0.73 | 0.85 | 0.46 | 0.64 | 0.7 |
| **SCORE.MERLIN** | **0.6** | 0.79 | 0.74 | **0.59** | **0.81** | 0.85 | **0.15** | 0.78 | 0.93 | **0.23** | **0.77** | 0.87 | 0.52 | 0.67 | 0.69 |
| **HM.MERLIN** | **0.59** | **0.82** | **0.84** | **0.59** | **0.81** | **0.9** | 0.14 | 0.7 | 0.89 | 0.15 | **0.77** | **0.91** | 0.52 | **0.69** | **0.77** |
| **HM.CT** | **0.59** | **0.81** | **0.85** | **0.59** | 0.8 | **0.9** | **0.15** | 0.68 | 0.87 | 0.14 | **0.77** | **0.91** | **0.52** | **0.7** | **0.76** |
| **SCORE.MAX** | **0.6** | **0.79** | 0.79 | **0.6** | **0.83** | **0.89** | **0.15** | 0.82 | **0.95** | **0.25** | **0.79** | **0.89** | **0.55** | **0.71** | **0.76** |
| **SCORE.2DF.CT** | 0.55 | 0.75 | 0.71 | 0.56 | 0.79 | 0.85 | **0.18** | **0.88** | **0.97** | 0.22 | 0.74 | 0.84 | **0.54** | **0.69** | 0.71 |
| **HM.2DF.CT** | 0.55 | **0.79** | 0.81 | 0.56 | 0.77 | **0.88** | **0.17** | 0.77 | 0.91 | 0.15 | 0.73 | **0.89** | **0.54** | **0.7** | **0.77** |
| (D) **Extreme Concordant** | | | | | | | | | | | | | | | |
| **SCORE.CT** | **0.61** | **0.75** | 0.69 | **0.55** | **0.68** | 0.57 | 0.81 | **0.99** | **1** | **0.23** | **0.26** | 0.22 | **0.12** | **0.1** | 0.09 |
| **SCORE.EV** | 0.48 | 0.63 | 0.61 | 0.4 | 0.62 | 0.55 | 0.46 | 0.88 | **0.98** | 0.13 | 0.18 | 0.18 | 0.07 | 0.07 | 0.07 |
| **SCORE.MERLIN** | **0.6** | **0.74** | 0.69 | **0.53** | **0.68** | 0.58 | 0.81 | **0.99** | **1** | **0.22** | **0.26** | 0.23 | **0.11** | **0.1** | 0.09 |
| **HM.MERLIN** | **0.6** | **0.76** | **0.74** | **0.53** | 0.63 | **0.65** | 0.81 | **0.99** | **1** | **0.22** | 0.25 | **0.25** | **0.12** | **0.1** | **0.11** |
| **HM.CT** | **0.6** | **0.75** | **0.73** | **0.53** | 0.63 | **0.65** | 0.79 | **0.98** | **1** | **0.22** | 0.25 | **0.24** | **0.11** | **0.1** | **0.11** |
| **SCORE.MAX** | **0.62** | **0.77** | **0.75** | **0.55** | **0.7** | **0.64** | 0.81 | **0.99** | **1** | **0.23** | **0.28** | **0.27** | **0.13** | **0.11** | **0.1** |
| **SCORE.2DF.CT** | 0.57 | 0.71 | 0.65 | 0.51 | 0.65 | 0.54 | **0.86** | **1** | **1** | 0.19 | 0.25 | 0.21 | **0.12** | **0.11** | **0.09** |
| **HM.2DF.CT** | 0.57 | 0.72 | 0.69 | 0.51 | 0.61 | 0.61 | **0.85** | **0.99** | **1** | 0.19 | 0.23 | 0.23 | **0.12** | **0.11** | **0.12** |
| (E) **EDAC-3 Corner** | | | | | | | | | | | | | | | |
| **SCORE.CT** | **0.6** | 0.73 | 0.66 | **0.55** | **0.71** | 0.64 | 0.78 | **0.99** | **1** | **0.44** | **0.57** | 0.45 | **0.38** | **0.29** | **0.18** |
| **SCORE.EV** | 0.49 | 0.66 | 0.62 | 0.46 | 0.65 | 0.61 | 0.48 | 0.92 | **0.98** | 0.35 | 0.51 | 0.42 | 0.3 | 0.24 | 0.14 |
| **SCORE.MERLIN** | **0.6** | 0.73 | 0.66 | **0.55** | **0.71** | 0.63 | 0.78 | **1** | **1** | **0.44** | 0.56 | 0.46 | **0.37** | **0.29** | **0.18** |
| **HM.MERLIN** | **0.61** | **0.77** | **0.8** | **0.54** | 0.66 | 0.62 | 0.79 | **0.99** | **1** | **0.44** | 0.5 | 0.4 | **0.37** | 0.22 | 0.13 |
| **HM.CT** | **0.61** | **0.76** | **0.8** | **0.55** | 0.66 | 0.61 | 0.79 | **0.99** | **1** | **0.45** | 0.5 | 0.4 | **0.38** | 0.22 | 0.13 |
| **SCORE.MAX** | **0.61** | **0.74** | 0.71 | **0.56** | **0.74** | **0.71** | 0.78 | **1** | **1** | **0.46** | **0.59** | **0.51** | **0.39** | **0.32** | **0.2** |
| **SCORE.2DF.CT** | 0.56 | 0.7 | 0.61 | 0.52 | 0.68 | 0.59 | **0.85** | **1** | **1** | 0.42 | 0.51 | 0.41 | **0.4** | **0.29** | 0.17 |
| **HM.2DF.CT** | 0.56 | 0.74 | **0.77** | 0.52 | 0.63 | 0.56 | **0.84** | **0.99** | **1** | 0.42 | 0.45 | 0.37 | **0.4** | 0.22 | 0.13 |
| (F) **MDAC-3 Corner** | | | | | | | | | | | | | | | |
| **SCORE.CT** | **0.74** | 0.73 | 0.5 | **0.69** | **0.85** | 0.64 | 0.59 | **0.98** | **0.98** | **0.63** | **0.68** | 0.44 | **0.56** | **0.42** | **0.2** |
| **SCORE.EV** | 0.66 | 0.69 | 0.48 | 0.62 | 0.81 | 0.62 | 0.25 | 0.86 | 0.92 | 0.56 | 0.64 | 0.41 | 0.5 | 0.38 | 0.17 |
| **SCORE.MERLIN** | **0.74** | 0.73 | 0.5 | **0.68** | **0.85** | 0.65 | 0.58 | **0.98** | **0.99** | **0.63** | **0.68** | 0.44 | **0.57** | **0.42** | **0.19** |
| **HM.MERLIN** | **0.73** | **0.8** | **0.67** | **0.68** | 0.79 | 0.64 | 0.59 | **0.98** | **0.98** | **0.63** | 0.64 | 0.44 | **0.57** | 0.38 | 0.17 |
| **HM.CT** | **0.73** | **0.79** | **0.65** | **0.69** | 0.8 | 0.63 | 0.59 | **0.97** | **0.97** | **0.63** | 0.65 | 0.44 | **0.57** | 0.38 | 0.17 |
| **SCORE.MAX** | **0.74** | 0.75 | 0.56 | **0.7** | **0.86** | **0.72** | 0.59 | **0.98** | **0.99** | **0.63** | **0.7** | **0.5** | **0.57** | **0.44** | **0.22** |
| **SCORE.2DF.CT** | 0.71 | 0.69 | 0.48 | **0.67** | 0.83 | 0.61 | **0.66** | **0.99** | **0.99** | 0.6 | 0.65 | 0.41 | **0.58** | **0.43** | **0.19** |
| **HM.2DF.CT** | 0.71 | 0.75 | 0.62 | **0.67** | 0.76 | 0.6 | **0.66** | **0.97** | **0.97** | 0.6 | 0.6 | 0.4 | **0.58** | 0.38 | 0.17 |

Note: For each model, power values within 3% of the maximum are highlighted in bold.

for non-normal models. Generally, for the $x|x|$ models, which can be thought of as being "relatively less non-normal", the higher moments statistics are worse than their lower moment counterparts. For the "relatively more non-normal" $x^3$ models, there is a marked improvement in the performance of the HM statistics in all the cases.

The relative performance of the statistics follows a similar general pattern for population and selected sampling. The conditional on trait variance SCORE.CT performs as well as SCORE.MERLIN, neither of them being consistently better than the other. The two-degree-of-freedom statistics show some improvement in the dominant model 5 and the recessive model 3, and the transformed versions of these models, but are worse for all the other models. The higher moment extensions of SCORE.CT, SCORE.MERLIN and SCORE.2DF.CT usually perform worse for $x|x|$ models (except $1'$) and better for the $x^3$ models (except $3'$). This is true for all the sampling schemes except EDAC3 and MDAC3, in which the HM statistics are worse for both $x|x|$ and $x^3$ models. The SCORE.MAX statistic is close to optimal in most cases, except for a few cases when the higher moment statistics or the two-degree-of-freedom statistics have higher power.

### 3.4.2 Sensitivity Analysis Results

In Figures 3.1 and 3.2, we have plotted the sensitivity analysis results for the models 2, $2'$ and $2''$ and all four selection schemes, POP, ED, EC and EDAC. The results for the models 4, $4'$ and $4''$ were similar. As seen in Figure 3.1, misspecification of the variance does not affect the power significantly. However, misspecification of the mean or the correlation seems to affect the power of all the statistics considerably. Also as seen in Figure 3.2, misspecification of the skewness and the kurtosis can reduce the power of the higher moment statistics drastically in some cases. There was no perceivable difference in sensitivity among the different LM statistics (or among the HM statistics).

For normal models, power always decreases when parameters are misspecified, as the true parameter values give the optimally powered score statistics. But for non-normal models, in some cases (e.g., under-specification of correlation in model $2''$ for population sampling) power may increase by using wrong parameter values, true scores are not necessarily optimal under these models.

For normal models (e.g., model 2), under population sampling, the effects of mean and correlation are symmetric. In other words, over-specification and under-specification have roughly

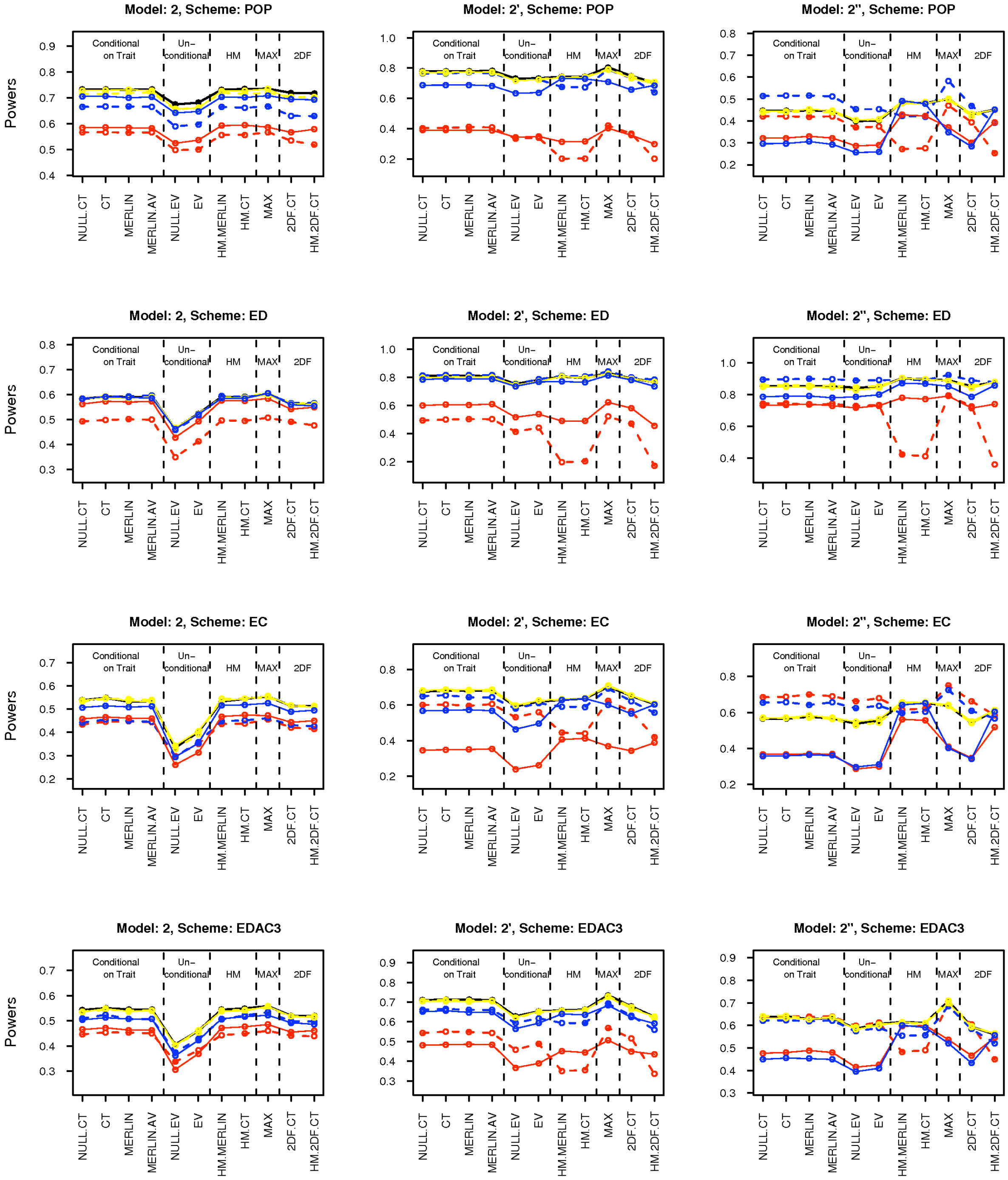

Figure 3.1: **Sensitivity analysis results for mean, variance and correlation.**

Black line gives power for true parameter values. Solid and dashed lines are for over and under specification of parameters respectively. Line colors red, yellow and blue stand for misspecified mean, variance and correlation in that order. Note that the black line roughly coincides with yellow line in almost all cases.

equal effect. However, for non-normal models (e.g., $2'$ and $2''$) or under selected sampling, the effects can be asymmetric. The direction of asymmetry can also change across selection schemes. Also, under-specification of mean and correlation seems to be better than over-specification for LM statistics whereas the order reverses for HM statistics.

For normal models (e.g., model 2), the LM and HM statistics are equally sensitive to mean and correlation. However, the HM statistics have the additional dependence on the skewness and kurtosis parameters, to which they are highly sensitive for these models. For slightly non-normal models (e.g., $2'$), both the LM and HM statistic are highly sensitive to the mean. The HM (respectively LM) statistics are more sensitive to the mean for the ED (respectively EC) scheme. The HM statistics are highly sensitive to skewness and kurtosis, especially to under-specification of these parameters.

For highly non-normal data (e.g., $2''$), the LM statistics are highly sensitive to mean and correlation, especially to over-specification of these parameters. Under-specification can sometimes provide increase in power. In some cases (e.g., EC and EDAC3), the HM statistics are relatively less affected by mean and correlation. For the ED scheme, the HM statistics are strongly affected by misspecification of mean. However, they are quite stable with respect to skewness and kurtosis for all sampling schemes, under these models.

In summary, misspecification of mean or correlation can have significant effect on the power of both LM and HM statistics. Effects can be asymmetric for skewed models or under selected sampling and the direction of asymmetry is generally different for LM and HM statistics. Misspecification of skewness and kurtosis can have drastic effect on the power of HM statistics particularly for normal and slightly non-normal models. However for highly non-normal models, the HM statistics are stable with respect to skewness and kurtosis and also, in some cases, less sensitive than LM statistics to specification of mean and correlation.

### 3.4.3 Weighting Results

The results of the weighting experiments are summarized in the Figures 3.3, 3.4 and 3.5. As shown in Figure 3.3, for population samples, the optimal weights for the larger sibships (with respect to sibpairs) decrease with increase of heritability. The non-standardized optimal weight also decreases with increasing sibship size. However, as expected, the standardized optimal weights

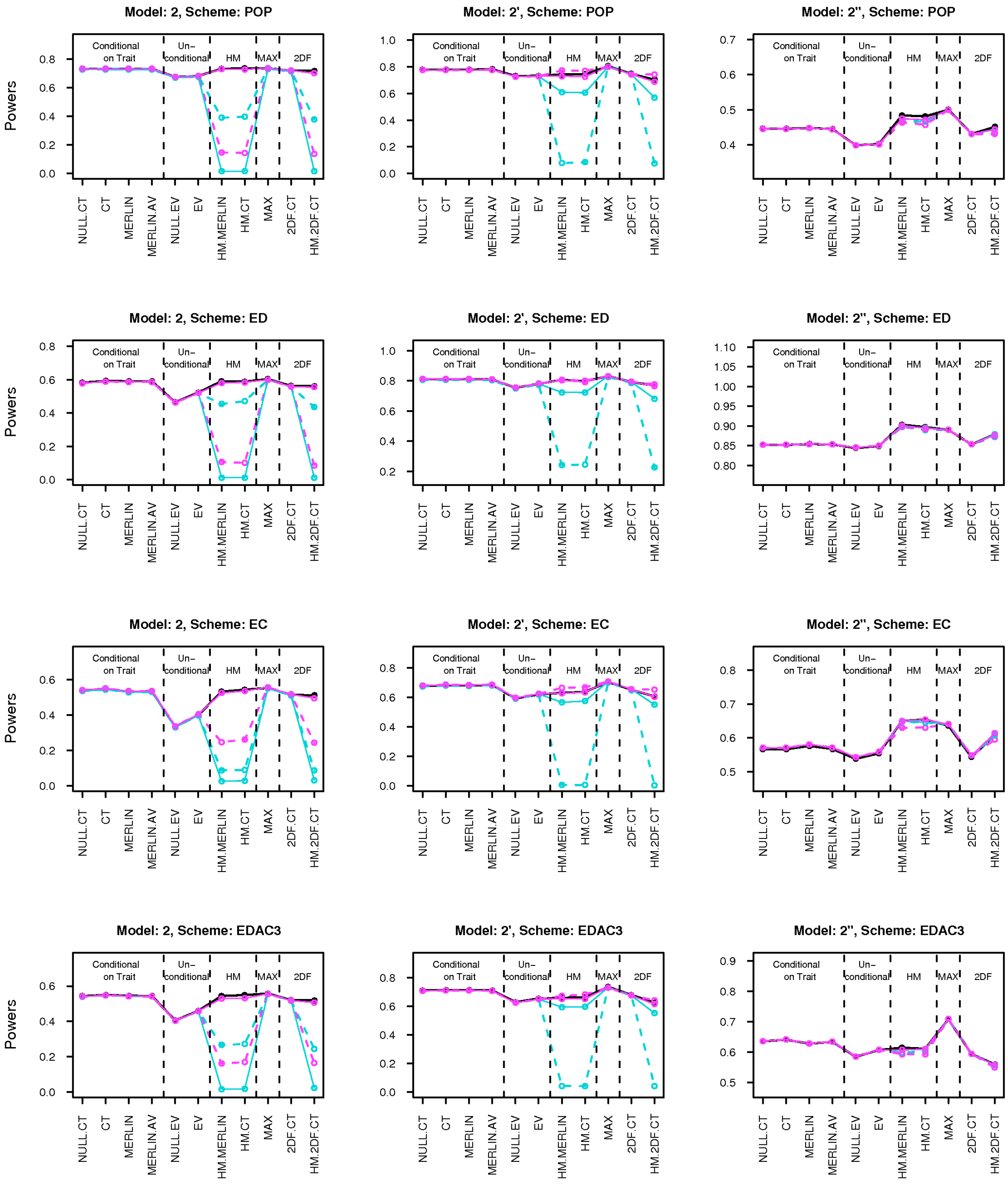

Figure 3.2: **Sensitivity analysis results for skewness and kurtosis.**

Black line gives power for true parameter values. Solid and dashed lines are for over and under specification of parameters respectively. Line colors cyan and magenta stand for misspecified skewness and kurtosis in that order. Note that, the black line coincides with the cyan and magenta lines for lower moment statistics.

are all greater than 1 and increase with sibship size (larger sibships are more informative and hence the corresponding standardized Z-scores receive higher weight.)

Figure 3.4 shows that the power curves are usually flat to the right of the optimal weight. Since 1 lies on the flatter side of the peak, using a non-standardized weight of 1 does not lead to much loss of power even for large effect sizes.

The power curve in Figure 3.5 is similar to those of Figure 3.4, but the peaks cluster closer to 1. Hence even for EDAC samples there is no obvious gain by using unequal weights on the non-standardized scores for discordant and concordant pairs. Our experiments with mixtures of random pairs and concordant/discordant pairs gave similar results (Data not shown).

## 3.5 DISCUSSION

### 3.5.1 Denominator Variants

We have conducted a comprehensive simulation study of some of the existing variants of score statistics as well as some novel ones. Our study attempted to identify the most robust score-based statistics under various genetic models and sampling schemes. The proposed conditional on trait variance (SCORE.CT) outperformed the empirical variance denominator (SCORE.EV), which has been suggested by many articles on score statistics. SCORE.EV appears to have a highly conservative type I error for small sizes and hence low power. This fact, also observed previously (Chen *et al.* 2005) is probably due to the fact that the scores (being a quadratic function of the trait values) are considerably skewed and hence it requires large sample sizes for the Central Limit Theorem to apply. Whereas when we condition on the trait, the IBD vector has a symmetric distribution around its expectation (under the null) and hence the Central Limit Theorem is applicable for smaller sample sizes. SCORE.CT also matches the power of SCORE.MERLIN in most cases and sometimes exceeds it. These two statistics differ only in the computation of the variance of the IBD vector in the denominator. SCORE.MERLIN uses the method of imputation (Sham *et al.* 2002) and requires the joint distribution of pair-wise IBDs for its computation. Limited experiments suggested that computation of SCORE.MERLIN can be slow for large pedigrees with uninformative markers or many ungenotyped individuals (data not shown). On the other hand, SCORE.CT is easier and much faster to compute as it involves a simple empirical variance.

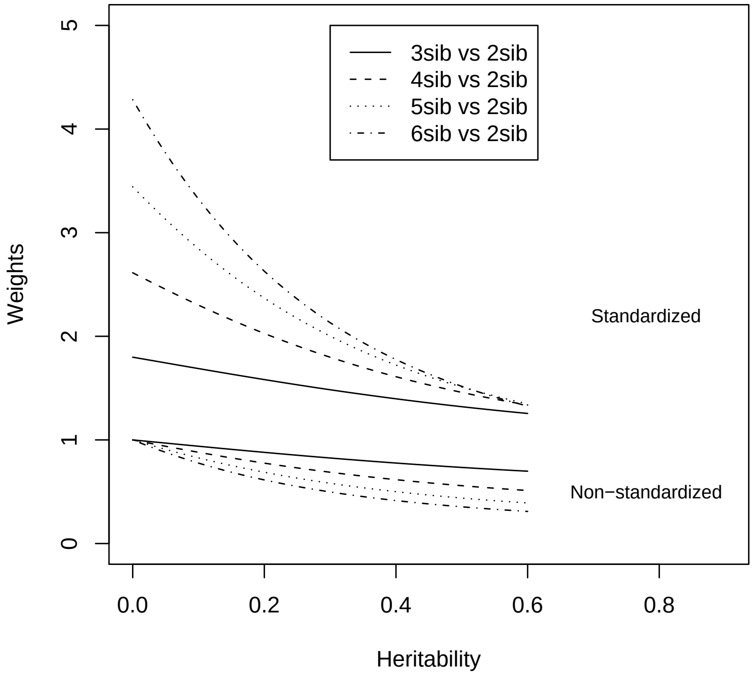

Figure 3.3: **Analytical optimal weights for sibships.**

Plot of asymptotic optimal weights (analytical) for sibships of sizes 3, 4, 5 and 6 (with respect to sibship of size 2) as a function of heritability. The lower cluster of plots shows the optimal weights for non-standardized scores while the upper shows those for standardized scores.

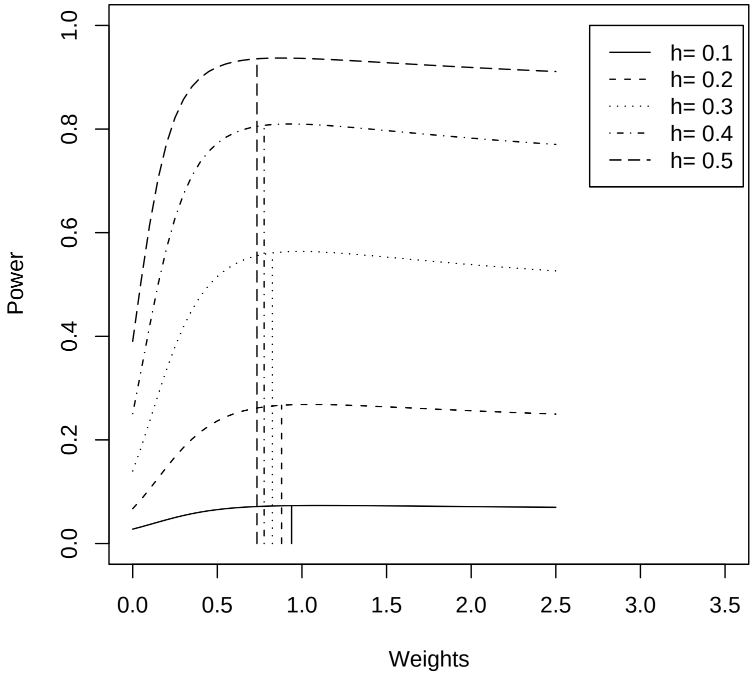

Figure 3.4: **Analytical power curves (of SCORE.NAÏVE) for 3sibs.**

Approximate analytical power curves for a population sample with 100 sibships of size 3 and 100 sibpairs. Power is plotted as a function of non-standardized weight of 3sibs with respect to 2sibs. Curves are shown for five different values of heritability ($h^2$). The vertical lines show asymptotic optimal weights for each value of $h^2$.

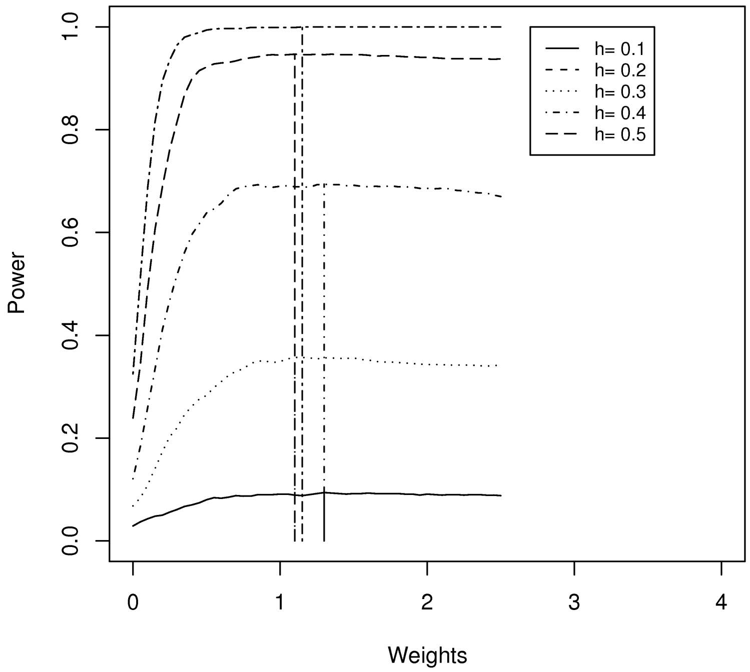

Figure 3.5: **Empirical power curves (of SCORE.CT) for EDAC pairs.**

Plot of simulation-based power for a combined sample of 20 discordant pairs and 30 concordant pairs. Power is plotted as a function of non-standardized weight of discordant with respect to concordant pairs. Curves are shown for five different values of heritability ($h^2$). The vertical lines show the actual optimal weights based on simulation, for each value of $h^2$.

The conditional on IBD statistics, SCORE.NAÏVE and HM.NAÏVE, were shown to have incorrect type I error under most circumstances. In the cases when they have correct type I error (normal traits and population samples) they dont provide any perceivable improvement in power over the conditional on trait statistics. Conditioning on IBD may be used only for population samples and in that case, SCORE.CIBD should be preferred over these two statistics as it maintains correct type I error for non-normal samples and close to optimal power. We do not in general recommend the use of any of these statistics.

Although the SCORE.EV statistic has sub-optimal power, it can be used to construct the SCORE.MAX statistic, which is the best overall statistic in our simulations. It gives significant improvement in power over SCORE.CT in many cases, with negligible inflation in type I error. We did limited simulations with empirical cutoffs (data not shown) to confirm that the power increase is sustained even after correcting for the slightly inflated type I error rate. It was outperformed only in some cases by the 2DF statistics and the higher moment statistics. It would be easy to construct higher moment and 2DF versions of the SCORE.MAX statistic and use them when appropriate.

### 3.5.2 Numerator Variants

Chen *et al.* (2005) proposed the higher moment numerator for score statistics and performed a similar simulation study for population samples. In this study, we were able to validate some of their results for population samples and test them for selected samples as well as a number of different non-normal models. They concluded that higher moment (HM) statistics were always as good as the lower moment (LM) ones and significantly better for all non-normal samples. Our results contradicted this conclusion. For the models we considered, the HM statistics were better than the LM versions only in some cases for the highly non-normal models. Also, their performance is quite unstable because of their dependence on two additional parameters (skewness and kurtosis). In practical situations, the HM statistics should be used only when the data are highly non-Gaussian and reasonably good estimates of skewness and kurtosis parameters are available.

The dominance based 2 d.f. statistics usually have lower power than the 1 d.f. statistics except for completely dominant or recessive models. It has been previously noted that the increase in power (by incorporating dominance) for dominant models is more than the decrease in power for additive models (Wang 2002; Chen *et al.* 2005). There is not enough evidence in our simulations to support this. It holds for the recessive model (3) but not for the dominant models (2 and 5). We recommend that these statistics be used in practice only when there is reason to suspect presence of highly dominant or recessive genetic variants.

### 3.5.3 Parameter Sensitivity

Parameter sensitivity is an extremely important issue for QTL mapping statistics. Although the trait parameters are nuisance parameters (with respect to the hypothesis of linkage), they can have a significant influence on power. They can be estimated fairly accurately for population samples, using a Maximum Likelihood estimation (MLE) approach. For selected samples, if the selection scheme is simple and the proband is known, the MLE can still be used. When the selection scheme is slightly complicated but the proband or probands are known, the Conditional MLE (CMLE) approach (Peng and Siegmund 2006) can be used. However, in reality many studies involve complicated ascertainment criteria with multiple and ill-defined probands. In such cases, we have no way to obtain parameter estimates and we need the statistics to be as robust as possible to wrongly specified parameters.

Our sensitivity analysis results suggest that for normal traits as well as slightly skewed traits, lower moment statistics should be preferred over higher moment ones, because of the latter's strong dependence on the two additional parameters: skewness and kurtosis. On the other hand, for highly non-Gaussian traits, the HM statistics have higher power in most cases and are stable with respect to skewness and kurtosis. Hence, for these models, HM statistics should be preferred. The asymmetric effects in many cases suggest the use of over-estimates or under-estimates of the parameters. But the direction of asymmetry may vary according to sampling scheme and direction of skewness of the model. Hence, proper formulation of these strategies would require a more exhaustive study of different non-Gaussian models and ascertainment schemes.

Note that, for our sensitivity analysis, we used extreme deviations from the true parameters values. This was done to consider a worst-case practical scenario when there is no prior information on the trait and the sample consists only of ascertained pedigrees. However, because of the wide fluctuations of power range under such extreme misspecification, we might have missed subtler differences in sensitivity among the individual LM (and HM) statistics.

### 3.5.4 Weighting of Pedigrees

The results of our weighting experiments show that for population samples, equal weighting of sibships of different sizes gives close to optimal power irrespective of the effect sizes. Similarly, for EDAC samples, equal weighting of non-standardized scores for discordant and concordant pairs is adequate. The results may not be completely generalizable to bigger and more complex pedigrees, or to other sampling schemes and non-normal traits. However, the methods outlined here are quite general, and can be used to study the effects of weighting more exhaustively. For example, this method can be used to study the possibility of weighting for non-normal samples or misspecified parameters. In fact, the formula (3.2.8) for optimal weight always holds for any statistic. The alternative means and variances of the statistic can be derived using the GEE form (as in the numerator of Equation (3.2.3)) for a general misspecified working covariance matrix.

The optimal weights as obtained above would be a function of the true size of the genetic effect, which is completely unknown. Hence, the best one can do is to select a weight that seems to work well for all or most alternatives. Also, this approach has the disadvantage of depending on the model (or working covariance matrix) assumed for calculating the moments. Another option, when sample size for each kind of pedigree is reasonably large is to use a part of the data (for each pedigree type) to estimate the alternative means and variances of the score function (using empirical estimates at each marker). This gives an optimally weighted statistic at each marker, which has increased power for detecting linkage. Similar empirical approaches could also be used to obtain parameter values that maximize power of the statistics. These approaches would work even in complicated ascertainment scenarios or when normality or higher moment assumption is deemed inaccurate. However there would be a simultaneous reduction in sample size, which would tend to reduce power. Which of these effects would dominate would depend among other factors on the sample size.

### 3.5.5 Limitations

There are of course some limitations in this study. Our simulation study considered only nuclear phenotypes without parental phenotype information. Although we expect the broad conclusions for the different groups of statistics (conditional on trait or IBD or unconditional) to hold for extended pedigrees as well, the specific details may vary. For example, in the case of datasets with larger pedigrees, SCORE.CT may reduce to SCORE.NULL.CT, as each pedigree type may be represented by a single pedigree. Also the parameter dependence of all the statistics would increase for larger pedigrees, with pairwise correlations between relatives being required. The relative performance of higher moment statistics with respect to lower moment ones may change in that scenario. Also, most of our results were based on simulations with moderately informative markers (8 equifrequent alleles). However, we did limited experiments (data not shown) for markers with very high and low informativity (20 and 2 equifrequent alleles respectively), and observed similar results.

Some score-based statistics in the literature have been omitted from our study. For example, we did not consider the sibship score variance (Wang 2002), discussed in Chen *et al.* (2005) as "score-S." This variance assumes the independence of sibpair IBDs, which holds only for perfectly informative markers. Because of computational limitations we were not able to consider some variance component (VC) based statistics such as Conditional VC statistic (Sham *et al.* 2000) and the semiparametric VC approach (Diao and Lin 2005). Note however that the former is not applicable for non-normal models while the latter would fail for selected samples.

The non-normal models we used were based on the hypothesis that the original trait has a mixture normal distribution and we observe the trait on a different scale. Hence, the final trait value was transformed. We considered this model to be realistic although some authors prefer to use models with non-normal errors. For example in the Chen *et al.* (2005) only the unshared environmental component was squared. We conducted limited simulations with chi-square residual models (data not shown) and got similar results to those of Chen *et al.* (2005). Also, one approach to dealing with non-normal traits is to apply a normalizing transformation (e.g., Wang 2002) to the traits and then apply variance components or standard score based approaches. We have not included this approach in our comparison as it does not fit into the score statistic framework. However as indicated by the results of Chen *et al.* (2005), this is a promising approach and deserves further investigation, particularly for population samples. For selected samples, such an approach can be used if a normalizing transformation for the trait is known a-priori from a previous population-based study.

### 3.5.6 Software

Currently there is a dearth of publicly available software implementing the score based statistics, which, because of their inherent robustness should be the method of choice for linkage mapping of quantitative traits. We have implemented most of the statistics discussed here and also other sibpair-specific statistics (some of which are discussed in T.Cuenco *et al.* 2003) in the user-friendly software QTL-ALL (QTL Analysis and Linkage Library). QTL-ALL recommends appropriate statistics based on the study design. Figure 3.6 shows a decision tree for choosing appropriate score statistics for sibships under different scenarios. The software implements some methods to increase speed by avoiding inversion of large matrices. These are outlined in Appendix C. QTL-ALL (Mukhopadhyay *et al.* , unpublished data) is available freely from our website (`http://watson.hgen.pitt.edu/register/`).

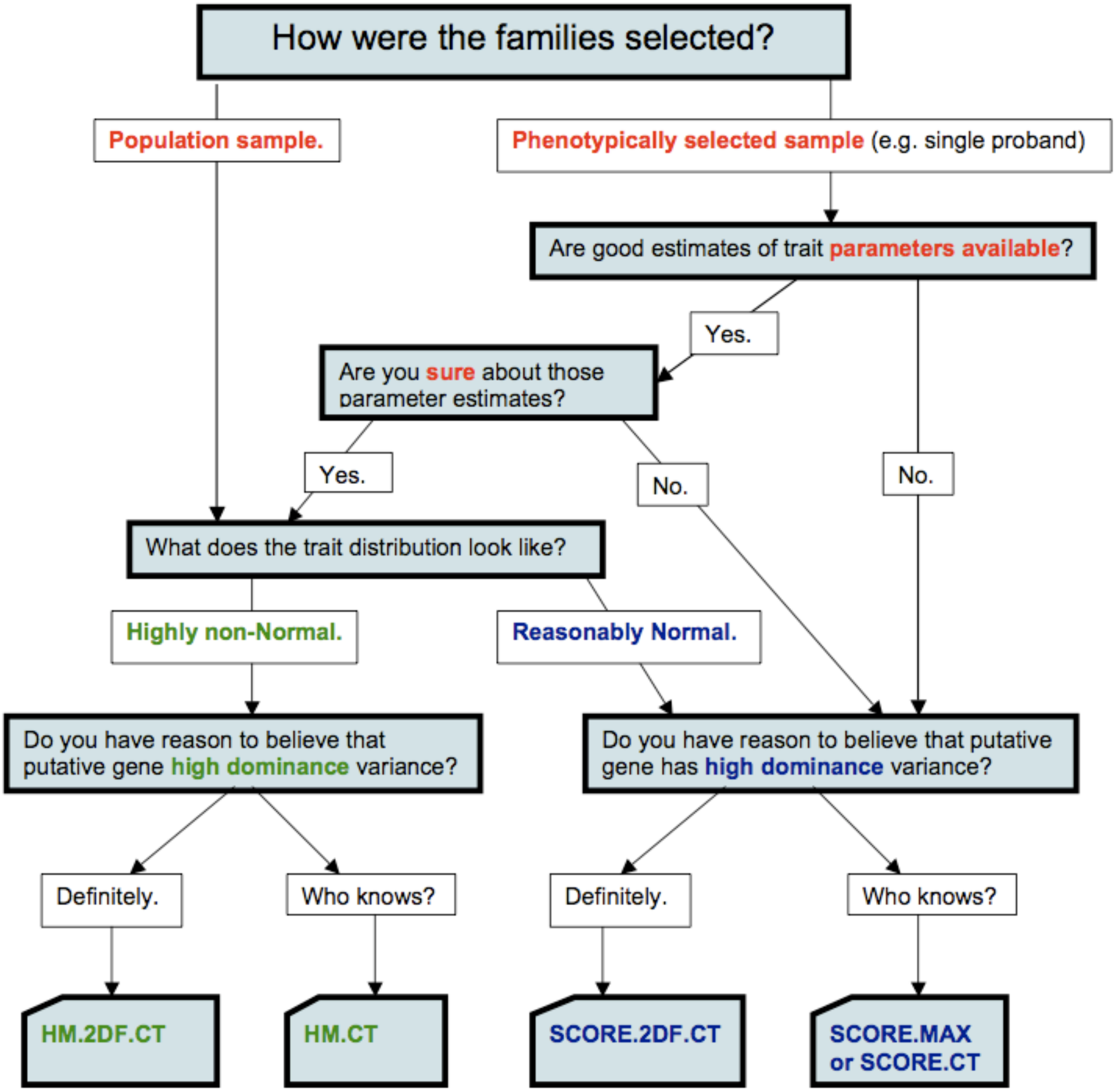
How were the families selected?
Population sample.
Phenotypically selected sample (e.g. single proband)
Are good estimates of trait parameters available?
Yes.
Are you sure about those parameter estimates?
Yes.
No.
No.
What does the trait distribution look like?
Highly non-Normal.
Reasonably Normal.
Do you have reason to believe that putative gene high dominance variance?
Do you have reason to believe that putative gene has high dominance variance?
Definitely.
Who knows?
Definitely.
Who knows?
HM.2DF.CT
HM.CT
SCORE.2DF.CT
SCORE.MAX or SCORE.CT

Figure 3.6: **Choice of Score Statistics for QTL Linkage Analysis with Sibships.**

# 4.0 SCORE STATISTICS FOR ASSOCIATION

Most family-based tests of association for quantitative traits are extensions of the Transmission Disequilibrium Test (Spielman *et al.* 1993; Terwilliger and Ott 1992). They condition upon parental genotypes to protect against population stratification and generally ignore parental phenotypes. Both of these factors contribute to loss of power of these tests relative to population-based or unconditional family-based tests. To improve power, all the available data including parental phenotypes should be used when confounding factors such as age or cohort specific differences are not suspected. We derive novel likelihood-based score statistics which have improved power to detect association in families, while protecting against population sub-structure and phenotype-based ascertainment. We discuss possible modifications of these statistics for incorporating IBD information and handling non-normally distributed traits and compare the performance of the proposed statistics to some of the standard family-based tests of association. We also address some computational issues arising in constructing the proposed statistics.

## 4.1 INTRODUCTION

In this section we give some background on two commonly used family-based association mapping methods for quantitative traits, FBAT and QTDT, and discuss some of the outstanding practical issues in the applicability of these statistics.

### 4.1.1 FBAT

The FBAT is a class of family-based tests of association that is robust to population stratification. It is quite general and can handle different kinds of phenotypes including binary, quantitative, censored and multiple traits. The FBAT statistic is motivated as an extension to the Transmission Disequilibrium Test or TDT (Spielman *et al.* 1993). Like the TDT, FBAT conditions on founder genotypes to protect against population stratification. The statistic was originally proposed for trio data by Rabinowitz (1997) and subsequently extended to handle nuclear families and extended pedigrees (e.g., Laird *et al.* 2000; Rabinowitz and Laird 2000; Lange *et al.* 2004) in the software packages FBAT and PBAT. The FBAT statistic has the following general form (Laird and Lange 2006)

$$\frac{\sum_{family:i}\sum_{non-founder:j} T'_{ij}[X_{ij} - E(X_{ij}|S_i)]}{\sum_i \sum_j \sum_{j'} T_{ij}T_{ij'}Cov(X_{ij}, X_{ij'} \mid S_i, T_{ij}, T_{ij'})},$$

where $T_{ij}$ and $X_{ij}$ are coded versions of the phenotype and genotypes of the $j^{th}$ non-founder in the $i^{th}$ family and $S_i$ is a sufficient statistic for the genetic information in the founders. For quantitative traits, the phenotypes are usually coded as $T_{ij} = E(Y_{ij} - \mu_{ij})$, where $\mu_{ij}$ are offsets usually chosen as the phenotype mean. The marker genotypes are usually coded according to a hypothesized genetic model (for example a coding of $\{aa, Aa, AA\} \rightarrow \{0, 1, 2\}$ would correspond to an additive model). Conditioning on the sufficient statistic for the founder genotypes makes the statistic robust (in terms of type I error) to population stratification as well as misspecification of the genetic model. Further, conditioning on the phenotype makes it robust to ascertainment. However, the conditioning on phenotypes and founder genotypes only guarantee robustness of type I error. Typically FBAT is considerably less powerful compared to population-based association studies, such as matched case-control studies which also protect against stratification to a certain extent. In this chapter, we investigate the possibility of improving the power of the FBAT statistic by changing the form of the numerator and/or by relaxing the conditioning on sufficient statistics.

Originally the FBAT numerator was motivated as a score function (Rabinowitz 1997), under certain models for trio data. For nuclear families or extended pedigrees, the above form of the numerator is usually motivated as a natural measure of association between the trait and the genotype. Note that the numerator simply measures the sample covariance between the trait and the genotype, which implicitly assumes “no residual environmental correlation.” But except for trio data (for which the FBAT was originally proposed), this assumption is generally unrealistic. For example, if we use a normal model for $[Y \mid g_m]$, the scores have the same form as above with $T_i = \Sigma_Y^{-1}(Y_i - \mu_Y)$ (e.g., Laird *et al.* 2000; Whittemore and Halpern 2003 and section 4.2.1 of this dissertation) instead of $T_i = Y_i - \mu_Y$ as usually recommended by FBAT. Thus the usual coding is optimal only under the assumption of uncorrelated environments. The FBAT software

(Horvath *et al.* 2001) suggests using the Gaussian scores $\Sigma_Y^{-1}(Y_i - \mu_Y)$, but this coding is not implemented in the software. For ascertained samples it may be difficult to obtain reliable estimates of the population mean and dispersion matrices of the phenotype. However, in this chapter, we have restricted our attention to this form of the numerator, assuming those parameter estimates are available (possibly from a previous population sample).

The choice of the offsets $\mu_{ij}$ is also an extremely important issue. FBAT allows different choices of the offsets, including the (weighted) sample mean of the phenotype. As mentioned above, the score function based on a Gaussian likelihood uses $\mu = \mu_Y$, the true population mean. In general, when the true population mean is known and the assumed model is correct, the score test gives a more powerful test than the FBAT (with $\mu_{ij} = \overline{Y}$). Note however that, for population samples, these two tests would be equivalent (as $\overline{Y} \approx \mu_Y$). Also, for selected samples, the trait mean may often be difficult to estimate. In section 4.2.8, we discuss how the choice of the offset affects the FBAT statistic and possible ways to construct statistics free of the trait parameters.

The numerator of the FBAT statistic is $Y_i'(g_i - E(g_i \mid S_i)$. It is designed to detect the alternative $H_{LA}$ of "linkage AND association." The null hypothesis for the test can be $H_{00}$ (no linkage and no association), $H_{L0}$ (linkage but no association) or $H_{0A}$ (association but no linkage). The choice of the null hypothesis affects the choice of the sufficient statistic $S_i$. When testing against $H_{00}$ or $H_{0A}$, the sufficient statistic can be the founder genotypes. Under both of these null hypotheses, the phenotype does not affect the mean of $g_i$, conditional on the founder genotypes. However, when testing against $H_{L0}$, the null hypothesis expectation $E(g_i \mid g_F, Y)$ (where $g_F$ denotes the founder genotypes) would in general depend on the recombination fraction $\theta$ which is unknown. As a way to get rid of the dependence on $\theta$, FBAT uses the expectation $E(g_i \mid g_F, \hat{\Pi}_m)$, where $\hat{\Pi}_m$ is the estimated IBD at the marker locus. Thus, when testing against $H_{L0}$, $S_i$ should consist of founder genotypes as well as estimated marker IBDs. The choice of the null hypothesis should generally depend on the design of the study. For a *de novo* genome scan $H_{00}$ may be appropriate. But under certain situations, it is known *a priori* that the marker is linked (e.g., fine mapping under a linkage peak) or associated (e.g., validation of an association signal obtained using a population-based study). Ideally one should condition on $g_F$ and $\hat{\Pi}_m$ as this guarantees correct type I error under all three null hypotheses, but more conditioning usually means less power. Hence it is customary to use the type-1 null hypothesis $H_{00} \cup H_{0A}$ (no linkage) for most purposes, except for fine mapping under a linkage peak, in which case the type-2 null hypothesis $H_{00} \cup H_{L0} \cup H_{0A}$ (no linkage or no association) is used. In the following sections, we propose statistics for testing both the type-1 and type-2 null hypothesis. However, our simulation studies are restricted to statistics that test the type-1 null hypothesis.

The FBAT statistic, like the TDT, ignores founder phenotype information. One of the reasons for this is that in many cases founders belong to a different age group or cohort. If the distribution of the phenotype varies with generation or age, analyzing all the phenotypes jointly may lead to spurious associations or to attenuation of an existing association. Nevertheless, for some phenotypes, the investigator may be able to rule out confounding due to generation effects or remove these effects for example by regressing out age. Also, in a multigenerational pedigree, founders who marry-in generally belong to the same cohort as their spouses and should probably be used. When founder phenotype information is available and generational biases are absent or removed, it may be possible to improve the power of the FBAT statistic by incorporating that information. This is because founders convey information through their environmental correlation with the non-founders. The FBAT statistic ignores the correlation structure of the family, and as a result founder phenotypes are non-informative.

Similarly, the FBAT always conditions on all the founder genotypes to protect against population stratification. In some situations, we may have some information regarding the nature of stratification in the population. We consider one such situation, in which there are possibly multiple strata in the population but there is strong assortative mating within each stratum. We show that, in this case, it is possible to improve the power of the FBAT statistic substantially by incorporating the founder genotype information partially instead of conditioning on all the information. Although this comes at the price of detecting certain types of markers that are associated but not linked to the trait, this is not a significant shortcoming considering that such markers if any are expected to be rare (see section 4.4 for a discussion of this issue).

Sometimes it may be reasonable to assume that there is no population stratification, but we may still want to use a family-based association test. In this case, using an FBAT type test that conditions on founder genotypes leads to considerable loss of power. Hence an unconditional score test should be used in this case. We propose extensions of the FBAT for the above mentioned scenarios in section 4.2 and compare some of those using simulations in section 4.2.9.

### 4.1.2 QTDT

Unlike the FBAT, the QTDT (Fulker *et al.* 1999; Abecasis *et al.* 2000; Abecasis *et al.* 2000) is a likelihood ratio test for association using family data that protects against stratification. The FBAT model just uses the marker genotype information, ignoring the IBD information at the locus. QTDT on the other hand incorporates both the genotypes and IBD, and as such should be more powerful. It uses a likelihood ratio test (LRT) based on the likelihood

$$[Y \mid g_m, \hat{\Pi}_m] \sim N[\mu_Y + a_b\ g_b + a_w\ g_w, \Sigma_Y + v_a\ (\hat{\Pi}_m - 2\Phi)],$$
$$\text{where}\ \ g_b = E(g_m \mid g_F)\ \ \text{and,}\ \ g_w = g_m - E(g_m \mid g_F),$$

and where the mean is "conditional on genotype" and the variance is "conditional on IBD." Here $g_b$ and $g_w$ constitute an orthogonal decomposition of the marker genotypes into between family and within family components. The $a_w$ parameter can be used to test for association, whereas a test of $a_b = a_w$ can be used to test for presence of stratification. Being an LRT, it can easily accommodate other parameters such as dominance effects, polygenic effects and environmental covariates. Also, the model is quite flexible in that it can test for any of the four null hypotheses discussed in the previous section as it separates the linkage and association parameters. Usually, the type-2 null hypothesis is tested using the parameter $a_w$, while estimating the linkage parameter $v_a$ under both the null and alternative.

In spite of its flexibility, QTDT has a number of disadvantages compared to FBAT. It protects only against "between family stratification," unlike FBAT, which is robust to arbitrary ascertainment schemes. This is because it uses the founder genotype information ($g_b$) as a surrogate for the stratum, with the implicit assumption that founders in the same family come from the same stratum. Also, it is quite computationally intensive, particularly for selected samples as the asymptotic chi-square thresholds fail. In such cases, permutations conditional on the observed inheritance vectors are used to obtain the null distribution of the statistic. Also, unlike the FBAT and the score tests it is not robust to non-normality of the phenotype. Permutation-based thresholds would be required for non-normal traits. Also the construction of the model is slightly ad-hoc and has some inconsistency (see section 2.2.1.4). Nevertheless QTDT is a popular approach for quantitative traits because of its flexibility. We used the QTDT statistic as implemented in the "qtdt" software (Abecasis *et al.* 2000; Abecasis *et al.* 2001) for some of the simulation comparisons in section 4.2.9.

Purcell *et al.* (2005) proposed likelihood ratio tests similar to QTDT that attempt to incorporate parental phenotypes. Unlike the QTDT, their procedure was only proposed for nuclear families and it does not attempt to incorporate IBD information. The method was further extended to binary traits using a liability threshold model, which is currently implemented in the "–parenTDT" option of the PLINK software package (Purcell *et al.* 2007; Purcell 2008). The original model as proposed for quantitative traits uses similar ideas as those in the next section to incorporate parental phenotypes and parental "genotype-phenotype" correlation (see section 4.2.4 for a comparison with the statistics proposed here). However the method has not been implemented for quantitative traits and hence could not be included in our simulation study.

## 4.2 METHODS

### 4.2.1 Score Tests for Type-I Null Hypothesis

In this section we derive four different score statistics for "type-I" null hypotheses, i.e., under the assumption of "no linkage". Because of this assumption, these statistics do not require IBD information. Score statistics incorporating IBD information will be discussed in section 4.2.2.

**4.2.1.1 Model and Notation** In this section we will use the notation of chapter 2. We consider the implicit mean model (2.2.3) for vector of quantitative traits $Y$ and marker genotypes $g_m$ observed on a pedigree of size $k$, with the additional assumption that the dominance effect is negligible (i.e., $d = 0$). We use the type-1 null hypothesis $H_0:$ No linkage.

$$Y = \mu_{_Y} + \beta\,(g_m - E_{g_m}) + e \text{ , where } \quad e \;\sim\; N(0, \Sigma_e). \tag{4.2.1}$$

$\Sigma_e$ is the unknown environmental covariance matrix, which can be written as $\Sigma_e = \Sigma_Y - \beta^2 \Sigma_{g_m}$. Under HWE, we have $E_{g_m} = 2p_m\mathbf{1}$ and $\Sigma_{g_m} = 4pq\Phi$. Further, let us assume that the pedigree has $L$ "founders" and $Q$ "non-founders." We partition the phenotype and genotype vectors into "founder" and "non-founder" parts as $Y = (Y'_{_F}, Y'_{_N})'$ and $g_m = (g'_{_F}, g'_{_N})'$. Let $(Y_{_F}, Y_{_N}) \in \mathcal{A}$ be the ascertainment scheme. Let us also partition the covariance matrices into founder and non-founder parts as

$$\Sigma_Y = \begin{pmatrix} \Sigma_{FF} & \Sigma_{FN} \\ \Sigma_{NF} & \Sigma_{NN} \end{pmatrix}, \quad \Sigma_{g_m} = \begin{pmatrix} \Sigma^g_{FF} & \Sigma^g_{FN} \\ \Sigma^g_{NF} & \Sigma^g_{NN} \end{pmatrix}, \quad \Sigma_e = \begin{pmatrix} \Sigma^e_{FF} & \Sigma^e_{FN} \\ \Sigma^e_{NF} & \Sigma^e_{NN} \end{pmatrix}.$$

Let $\Sigma_{N/F}$ denote the Schur complement of $\Sigma_Y$ of the non-founder covariance matrix with respect to the founder covariance matrix, i.e., $\Sigma_{N/F} = \Sigma_{NN} - \Sigma_{NF}\Sigma^{-1}_{NN}\Sigma_{FN}$. Finally we define "corrected" non-founder phenotype and genotype vectors as follows

$$Y_{N/F} = Y_N - \Sigma_{NF}\Sigma^{-1}_{NN}\ Y_F \quad \text{and,}\ \ g_{N/F} = g_N - E(g_N \mid g_F).$$

Below we derive four different score statistics that protect against different types of population stratification.

**4.2.1.2 No Stratification: SCORE.NS** This statistic does not protect against any stratification. In this case, the likelihood model of interest is $L(Y, g_m \mid \mathcal{A})$. As shown in section 2.4, the score for this likelihood can be obtained from the likelihood $L(g_m \mid Y)$, which does not require knowledge of the ascertainment scheme $\mathcal{A}$. Thus, by conditioning on a sufficient statistic $Y$ for the ascertainment, we protect the score statistic against arbitrary ascertainment. As shown before in section 2.4, the score for the reverse likelihood $L(g_m \mid Y)$ is same as the score obtained from the forward likelihood,

$$L_{Y|g_m}(\beta) \propto \frac{\exp\{-\frac{1}{2}(\tilde{Y} - \beta\ \tilde{g}_m)'[\Sigma_Y - \beta^2\ \Sigma_{g_m}]^{-1}\ (\tilde{Y} - \beta\ \tilde{g}_m)\}}{|\Sigma_Y - \beta^2\ \Sigma_{g_m}|^{\frac{1}{2}}}.$$

As shown in Appendix A, the score for this likelihood is given by

$$l'_{Y|g_m}(0) = \tilde{Y}'\ \ \Sigma^{-1}_Y\ \ \tilde{g}_m. \tag{4.2.2}$$

The standardized score statistic $SCORE.NS$ is constructed using the above score function standardized by a "conditional on trait" variance. Thus, if we have observed data $(Y_i, g_{m_i}),\ for\ i = 1, \ldots, n$ for $n$ pedigrees having the same structure, then we define

$$SCORE.NS = \frac{\sum_i \tilde{Y}'_i\ \ \Sigma^{-1}_Y\ \ \tilde{g}_{m_i}}{\sqrt{\sum_i \tilde{Y}'_i\ \ \Sigma^{-1}_Y\ \ \Sigma_{g_m}\ \ \Sigma^{-1}_Y\ \ \tilde{Y}_i}}, \tag{4.2.3}$$

where we have used the fact that under the null, assuming HWE, $Cov(g \mid \mathcal{A}) = \Sigma_{g_m} = 4p_m q_m \Phi$. When pedigrees of different sizes and structures are present, the definition of the statistic can be modified as done in chapter 3 for score statistics for linkage analysis. Note that $SCORE.NS$ easily extends to handling general data sets with either families or unrelated individuals or both. All unrelated individuals in a sample can be thought of as a big pedigree where all individuals are independent i.e $\Sigma_{Y} = \Sigma_{g_m} = 0$. If a data set contains unrelated individuals only, $SCORE.NS$ essentially reduces to a linear regression of the phenotype on the marker genotypes, centered at the population means $\mu_{Y}$ and $2p_m$.

**4.2.1.3 Arbitrary Stratification: SCORE.AS** In the presence of unknown stratification in the sample, the minimal sufficient statistic for the allele frequencies in the population is the vector of founder genotypes $g_{F}$. Using a likelihood conditional on $g_{F}$ thus leads to score statistics that are protected against arbitrary stratification. The likelihood of interest is $L(Y, g_{N} \mid g_{F}, \mathcal{A})$. Ideally, to obtain the score under this likelihood we should condition on a minimal sufficient statistic for the ascertainment scheme $\mathcal{A}$. When $\mathcal{A}$ is completely unknown, the minimal sufficient statistic is $Y$. Conditioning on $Y$ leads to some loss of information in this case because of the simultaneous conditioning on $g_{F}$ (unlike the no stratification case, where the score conditional on $\mathcal{A}$ is identical to that conditional on $Y$). The score for $L(g_{N} \mid g_{F}, Y)$ leads to $SCORE.FP$, which is discussed in the next section. Here we discuss $SCORE.AS$ which, like FBAT, uses the idea of conditioning on $g_{F}$ but ignores the founder phenotype data. This approach should be used whenever founder phenotypes are unknown or there are possible generation specific differences in the phenotype. In this case, the likelihood of interest is $L(g_{N} \mid g_{F}, Y_N)$, as $Y_N$ are minimal sufficient for unknown ascertainment scheme $\mathcal{A}$. The score for this likelihood can be derived as follows

$$\begin{aligned}
L_{g_{N}|g_{F},Y_N}(\beta) &= P(g_{N} \mid g_{F}, Y_N) \\
&= \frac{P(Y_N, g_{F}, g_{N})}{P(Y_N, g_{F})} \\
&= \frac{P(Y_N \mid g_{N})P(g_{N} \mid g_{F})}{\sum_{g_{N}} P(Y_N \mid g_{N})P(g_{N} \mid g_{F})} \\
&\propto \frac{L_{Y_N|g_{N}}(\beta)}{\sum_{g_{N}} L_{Y_N|g_{N}}(\beta)P(g_{N} \mid g_{F})}
\end{aligned}$$

$$\begin{aligned}
l_{g_{N}|g_{F},Y_N}(\beta) &= l_{Y_N|g_{N}}(\beta) - \log\{\sum_{g_{N}} L_{Y_N|g_{N}}(\beta)P(g_{N} \mid g_{F})\} \\
Score(g_{N} \mid g_{F}, Y_N) &= Score(Y_N \mid g_{N}) - \{\sum_{g_{N}} Score(Y_N \mid g_{N})P(g_{N} \mid g_{F})\},
\end{aligned}$$

where we have used $P(x)$ to denote p.m.f or p.d.f for discrete and continuous variables respectively. The last step uses the fact that score for a mixture likelihood is a mixture of the scores, which can be shown easily by differentiating the log-likelihood for the mixture distribution. The score for the likelihood $[Y_N \mid g_{\scriptscriptstyle N}]$ is $\tilde{Y}'_{\scriptscriptstyle N}\ \Sigma^{-1}_{\scriptscriptstyle NN}\ \tilde{g}_{\scriptscriptstyle N}$ (analogous to 4.2.2). Thus the score function simplifies to

$$\begin{aligned} Score(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F}, Y_N) &= \tilde{Y}'_{\scriptscriptstyle N}\ \Sigma^{-1}_{\scriptscriptstyle N N}\ [\ \tilde{g}_{\scriptscriptstyle N} - \sum_{g_{\scriptscriptstyle N}} \tilde{g}_{\scriptscriptstyle N}\ P(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F})\ ] \\ &= \tilde{Y}'_{\scriptscriptstyle N}\ \Sigma^{-1}_{\scriptscriptstyle N N}\ [\ g_{\scriptscriptstyle N} - E(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F})\ ]. \end{aligned}$$

The standardized score $SCORE.AS$, is obtained by standardizing the above score function using a "conditional on trait" variance. Thus, for a dataset $(Y_i, g_{m_i}),\ for\ i = 1, \ldots, n$ on $n$ pedigrees of the same type, we define:

$$SCORE.AS = \frac{\sum_i \tilde{Y}'_{N_i}\ \ \Sigma^{-1}_{\scriptscriptstyle NN}\ \ [g_{\scriptscriptstyle N_i} - E(g_{\scriptscriptstyle N_i} \mid g_{\scriptscriptstyle F_i})]}{\sqrt{\sum_i \tilde{Y}'_{N_i}\ \ \Sigma^{-1}_{\scriptscriptstyle NN}\ \ Cov(g_{\scriptscriptstyle N_i} \mid g_{\scriptscriptstyle F_i})\ \ \Sigma^{-1}_{\scriptscriptstyle NN}\ \ \tilde{Y}'_{N_i}}}, \qquad (4.2.4)$$

where we have used the fact that under the null, assuming HWE, $Cov(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F}, Y_N) = Cov(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F})$. The computation of $Cov(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F})$ will be discussed in section 4.2.5.

#### 4.2.1.4 Arbitrary Stratification: SCORE.FP

When founder phenotypes are available, they can be used to improve the power of $SCORE.AS$. In this case we consider the likelihood $L(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F}, Y)$ as discussed in the previous section. $g_{\scriptscriptstyle F}$ and $Y = (Y_{\scriptscriptstyle N}, Y_{\scriptscriptstyle F})$ serve as minimal sufficient statistics for marker allele frequencies (assuming unknown stratification) and selection (assuming unknown ascertainment scheme) respectively. The score for this likelihood is obtained as follows

$$\begin{aligned} L_{g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F}, Y}(\beta) &= P(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F}, Y) \\ &= \frac{P(Y, g_{\scriptscriptstyle F}, g_{\scriptscriptstyle N})}{P(Y, g_{\scriptscriptstyle F})} \\ &= \frac{P(Y \mid g_m) P(g_m \mid g_{\scriptscriptstyle F})}{\sum_{g_m} P(Y \mid g_m) P(g_m \mid g_{\scriptscriptstyle F})} \\ &\propto \frac{L_{Y \mid g_m}(\beta)}{\sum_{g_m} L_{Y \mid g_m}(\beta) P(g_m \mid g_{\scriptscriptstyle F})} \end{aligned}$$

$$\begin{aligned} l_{g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F}, Y}(\beta) &= l_{Y \mid g_m}(\beta) - \log\{\sum_{g_m} L_{Y \mid g_m}(\beta) P(g_m \mid g_{\scriptscriptstyle F})\} \\ Score(g_{\scriptscriptstyle N} \mid g_{\scriptscriptstyle F}, Y) &= Score(Y_m \mid g_m) - \{\sum_{g_m} Score(Y \mid g_m) P(g_m \mid g_{\scriptscriptstyle F})\}. \end{aligned}$$

The score for the likelihood $[Y \mid g_m]$ is $\tilde{Y}' \ \Sigma_{Y}^{-1} \ \tilde{g}_m$ (from 4.2.2). Thus the score function simplifies to

$$\begin{aligned} Score(g_N \mid g_F, Y) &= \tilde{Y}' \ \Sigma_{Y}^{-1} \ [\ \tilde{g}_m - \sum_{g_m} \tilde{g}_m \ P(g_m \mid g_F) \ ] \\ &= \tilde{Y}' \ \Sigma_{Y}^{-1} \ [\ g_m - E(g_m \mid g_F) \ ]. \end{aligned}$$

Note that the founder component of $[g_m - E(g_m \mid g_F)]$ is $[g_F - E(g_F \mid g_F)] = 0$. By partitioning $Y$ and $\Sigma_Y$ into founder and non-founder components and using the standard inversion formula for inverse of partitioned matrix $\Sigma_Y$, the score can be simplified to

$$Score(g_N \mid g_F, Y) = \tilde{Y}'_{N/F} \ \Sigma_{N/F}^{-1} \ [\ g_N - E(g_N \mid g_F) \ ],$$

which is similar to $SCORE.AS$, but with the non-founder phenotypes regressed on the founder phenotypes. The residuals $Y_{N/F}$ and the residual covariance matrix $\Sigma_{N/F}$ are used instead of the uncorrected non-founder phenotype $Y_N$ and covariance matrix $\Sigma_{NN}$. Thus the founder phenotypes act as covariates in reducing the variability of the non-founder phenotypes due to environmental factors. Thus $SCORE.FP$ is expected to have higher power than $SCORE.AS$. However, the increase in power tends to be modest (see simulation results in section 4.3). The power improvement is expected to increase with increase in number of founders in the family or increase in environmental correlation between founder and non-founders.

The standardized score $SCORE.FP$ is obtained by standardizing the above score function using a "conditional on trait" variance. Thus, for a dataset $(Y_i, g_{m_i}),\ for\ i = 1, \ldots, n$ on $n$ pedigrees of the same type, we define

$$SCORE.FP = \frac{\sum_i \tilde{Y}'_i \ \ \Sigma_{Y}^{-1} \ \ [g_{m_i} - E(g_{m_i} \mid g_{F_i})]}{\sqrt{\sum_i \tilde{Y}'_i \ \ \Sigma_{Y}^{-1} \ \ Cov(g_{m_i} \mid g_{F_i}) \ \ \Sigma_{Y}^{-1} \ \ \tilde{Y}_i}}, \tag{4.2.5}$$

where we have used the fact that under the null, assuming HWE, $Cov(g_m \mid g_F, Y) = Cov(g_m \mid g_F)$. Also note that $Cov(g_m \mid g_F)$ has zeros except for the last $Q \times Q$ block of non-founder covariances $Cov(g_N \mid g_F)$.

#### 4.2.1.5 Between Family Stratification: SCORE.FPG

Here we assume that founder phenotypes are known and also that stratification is "between family" only. In such a case, and if both founder phenotypes and genotypes can be used construct a score test that derives information from founder genotype-phenotype correlation and is, as a result, significantly more powerful than

$SCORE.AS$ and $SCORE.FP$. Such an assumption may be reasonable when there are possibly multiple strata in the population but there is strong assortative mating, so that founders in each family come from the same stratum. In this case the founder marker genotypes in a family are independently distributed as $Bin(2, p_{m,s})$, where $p_{m,s}$ is the marker allele frequency in the stratum $s$ to which the family belongs (assuming HWE in each stratum). Thus the minimal sufficient statistic for the stratum allele frequency is the founder genotype mean $\overline{g}_{_F}$. The ascertainment scheme is again assumed to be unknown, so that all phenotypes $Y$ constitute the minimal sufficient statistic. Thus the likelihood of interest in $L(g_m \mid Y, \overline{g}_{_F})$. The score for this likelihood is given by

$$\begin{aligned}
L_{g_m|\overline{g}_{_F},Y}(\beta) &= P(g_m \mid \overline{g}_{_F}, Y) \\
&= \frac{P(Y, g_m, \overline{g}_{_F})}{P(Y, \overline{g}_{_F})} \\
&= \frac{P(Y \mid g_m)P(g_m \mid \overline{g}_{_F})}{\sum_{g_m} P(Y \mid g_m)P(g_m \mid \overline{g}_{_F})} \\
&\propto \frac{L_{Y|g_m}(\beta)}{\sum_{g_m} L_{Y|g_m}(\beta)P(g_m \mid \overline{g}_{_F})}
\end{aligned}$$

$$\begin{aligned}
l_{g_m|\overline{g}_{_F},Y}(\beta) &= l_{Y|g_m}(\beta) - \log\{\sum_{g_m} L_{Y|g_m}(\beta)P(g_m \mid \overline{g}_{_F})\} \\
Score(g_m \mid \overline{g}_{_F}, Y) &= Score(Y_m \mid g_m) - \{\sum_{g_m} Score(Y \mid g_m)P(g_m \mid \overline{g}_{_F})\}.
\end{aligned}$$

The score for the likelihood $[Y \mid g_m]$ is $\tilde{Y}'\ \Sigma_{_Y}^{-1}\ \tilde{g}_m$. (from 4.2.2). Thus the score function simplifies to

$$\begin{aligned}
Score(g_m \mid \overline{g}_{_F}, Y) &= \tilde{Y}'\ \Sigma_{_Y}^{-1}\ [\ \tilde{g}_m - \sum_{g_m} \tilde{g}_m\ P(g_m \mid \overline{g}_{_F})\ ] \\
&= \tilde{Y}'\ \Sigma_{_Y}^{-1}\ [\ g_m - E(g_m \mid \overline{g}_{_F})\ ] \\
&= \tilde{Y}'\ \Sigma_{_Y}^{-1}\ [\ g_m - \overline{g}_{_F}\ \mathbf{1}\ ],
\end{aligned}$$

where we have used the fact that $E(g_i^m \mid \overline{g}_{_F}) = \overline{g}_{_F}$ for all individuals "i" in the pedigree (proved in section 4.2.5). The standardized score $SCORE.FPG$ is obtained by standardizing the above score function using a "conditional on trait" variance. For a dataset $(Y_i, g_{m_i}),\ for\ i = 1, \ldots, n$ on $n$ pedigrees of the same type, we define:

$$SCORE.FPG = \frac{\sum_i \tilde{Y}_i'\ \ \Sigma_{_Y}^{-1}\ \ [\ g_{m_i} - E(g_{m_i} \mid \overline{g}_{F_i})\ ]}{\sqrt{\sum_i \tilde{Y}_i'\ \ \Sigma_{_Y}^{-1}\ \ Cov(g_{m_i} \mid \overline{g}_{F_i})\ \ \Sigma_{_Y}^{-1}\ \ \tilde{Y}_i}}, \qquad (4.2.6)$$

where we have used the fact that under the null, assuming HWE, $Cov(g_m \mid \overline{g}_{_F}, Y) = Cov(g_m \mid \overline{g}_{_F})$. The computation of $Cov(g_m \mid \overline{g}_{_F})$ will be discussed in section 4.2.5.

It can be shown that the additional information captured by $SCORE.FPG$ is essentially a measure of the "genotype-phenotype correlation" among the founders in each family.

### 4.2.2 Incorporating IBD information

Most approaches for association mapping ignore the IBD information at the marker locus that may be available from other markers across the chromosome. As discussed in section 2.2.1.4, ideally all the available information, i.e., $Y$, $g_m$ and $\hat{\Pi}_m$ should be modeled, irrespective of the type of test (linkage or association). IBD information helps in reducing the variability of $g_m$ by modeling $Cov(g_m \mid \hat{\Pi}_m)$ even in absence of linkage with the trait locus. Below, we discuss two ways of incorporating IBD information in the score tests for association described above.

**4.2.2.1 Testing Type-2 Null Hypothesis** The four score statistics discussed in the previous section can be modified to include the null hypothesis $H_{L0}$ : Linkage but No association by additionally conditioning on the observed IBD $\hat{\Pi}_m$ to obtain the null mean and variance of each statistic. This is because $(Y, \hat{\Pi}_m)$ is sufficient for the linkage parameter, and so the mean and variances become free of the coefficient of recombination $\theta$. This approach is recommended by FBAT, but only for the purposes of fine mapping under a linkage peak, as the additional conditioning leads to loss of information and power. The modified versions of the score statistics described above to test for the type-2 null hypothesis are

$$\begin{aligned}
SCORE.IBD.NS &= \frac{\sum_i \tilde{Y}_i' \ \ \Sigma_Y^{-1} \ \ \tilde{g}_{m_i}}{\sqrt{\sum_i \tilde{Y}_i' \ \ \Sigma_Y^{-1} \ \ Cov(g_m \mid \hat{\Pi}_m) \ \ \Sigma_Y^{-1} \ \ \tilde{Y}_i}} \\
SCORE.IBD.AS &= \frac{\sum_i \tilde{Y}_{N_i}' \ \ \Sigma_{NN}^{-1} \ \ [g_{N_i} - E(g_{N_i} \mid g_{F_i}, \hat{\Pi}_{m_i})]}{\sqrt{\sum_i \tilde{Y}_{N_i}' \ \ \Sigma_{NN}^{-1} \ \ Cov(g_{N_i} \mid g_{F_i}, \hat{\Pi}_{m_i}) \ \ \Sigma_{NN}^{-1} \ \ \tilde{Y}_{N_i}}} \\
SCORE.IBD.FP &= \frac{\sum_i \tilde{Y}_i' \ \ \Sigma_Y^{-1} \ \ [g_{m_i} - E(g_{m_i} \mid g_{F_i}, \hat{\Pi}_{m_i})]}{\sqrt{\sum_i \tilde{Y}_i' \ \ \Sigma_Y^{-1} \ \ Cov(g_{m_i} \mid g_{F_i}, \hat{\Pi}_{m_i}) \ \ \Sigma_Y^{-1} \ \ \tilde{Y}_i}} \\
SCORE.IBD.FPG &= \frac{\sum_i \tilde{Y}_i' \ \ \Sigma_Y^{-1} \ \ [\ g_{m_i} - E(g_{m_i} \mid \overline{g}_{F_i}, \hat{\Pi}_{m_i})\ ]}{\sqrt{\sum_i \tilde{Y}_i' \ \ \Sigma_Y^{-1} \ \ Cov(g_{m_i} \mid \overline{g}_{F_i}, \hat{\Pi}_{m_i}) \ \ \Sigma_Y^{-1} \ \ \tilde{Y}_i}},
\end{aligned}$$

where we have used the fact that $[g_N \mid g_F, Y, \hat{\Pi}_m] = [g_N \mid g_F, \hat{\Pi}_m]$ under the null hypothesis of no association (irrespective of linkage). The computation of the conditional means and variances involved in the above formulas will be discussed in section 4.2.5.

**4.2.2.2 Modeling IBD information** The four score statistics derived in section 4.2.1 assume the mean model (4.2.1), which ignores IBD information and models the distribution $[Y \mid g_m]$. The reverse likelihoods used to derive the scores are $L(g_m \mid Y)$, $L(g_N \mid g_F, Y_N)$, $L(g_N \mid g_F, Y)$, $L(g_m \mid \overline{g}_F, Y)$. Instead of ignoring the observed IBD information, scores can be derived by considering the likelihoods $L(g_m, \hat{\Pi}_m \mid Y)$, $L(g_N, \hat{\Pi}_m \mid g_F, Y_N)$, $L(g_N, \hat{\Pi}_m \mid g_F, Y)$, $L(g_m, \hat{\Pi}_m \mid \overline{g}_F, Y)$. Let us first consider the likelihood $L(g_N, \hat{\Pi}_m \mid g_F, Y)$. We note that

$$L(g_N, \hat{\Pi}_m \mid g_F, Y) \propto L(g_N \mid g_F, Y, \hat{\Pi}_m)\; L(\hat{\Pi}_m \mid Y, g_F).$$

Under "no linkage," the second component, $L(\hat{\Pi}_m \mid Y, g_F)$ becomes $L(\hat{\Pi}_m)$. So, the likelihood is essentially proportional to the first component, $L(g_N \mid g_F, Y, \hat{\Pi}_m)$. The score for this likelihood is given by

$$\begin{aligned}
L_{g_N|g_F,Y,\hat{\Pi}_m}(\beta) &= P(g_N \mid g_F, Y, \hat{\Pi}_m) \\
&= \frac{P(Y, \hat{\Pi}_m, g_N, g_F)}{P(Y, \hat{\Pi}_m, g_N)} \\
&= \frac{P(Y \mid \hat{\Pi}_m, g_N, g_F)\; P(g_N \mid g_F, \hat{\Pi}_m)}{\sum_{g_N} P(Y \mid \hat{\Pi}_m, g_F, g_N)\; P(g_N \mid g_F, \hat{\Pi}_m)} \\
&\propto \frac{L_{(Y|\hat{\Pi}_m,g_N,g_F)}(\beta)}{\sum_{g_N} L_{(Y|\hat{\Pi}_m,g_N,g_F)}(\beta) P(g_N \mid g_F, \hat{\Pi}_m)}.
\end{aligned}$$

$$Score(g_N \mid g_F, Y, \hat{\Pi}_m) = Score(Y \mid \hat{\Pi}_m, g_N, g_F) - E_{g_N|g_F,\hat{\Pi}_m}(Y \mid \hat{\Pi}_m, g_N, g_F). \qquad (4.2.7)$$

The model $[Y \mid \hat{\Pi}_m, g_N, g_F]$ is essentially the proposed implicit model (2.2.8) under the assumption of no linkage. The score for this model is

$$\begin{aligned}
L_{Y|g_m,\hat{\Pi}_m}(\beta) &\propto \frac{\exp\{-\frac{1}{2}(\tilde{Y} - \beta\ \tilde{g}_m)'[\Sigma_Y - \beta^2\ v_{g_m}\hat{\Pi}_m]^{-1}\ (\tilde{Y} - \beta\ \tilde{g}_m)\}}{|\Sigma_Y - \beta^2\ v_{g_m}\hat{\Pi}_m|^{\frac{1}{2}}} \\
l_{Y|g_m,\hat{\Pi}_m}(\beta) &\propto -\frac{1}{2}(\tilde{Y} - \beta\ \tilde{g}_m)'\ [\Sigma_Y - \beta^2\ v_{g_m}\hat{\Pi}_m]^{-1}\ (\tilde{Y} - \beta\ \tilde{g}_m) - \frac{1}{2}\log|\Sigma_Y - \beta^2\ v_{g_m}\hat{\Pi}_m| \\
l'_{Y|g_m,\hat{\Pi}_m}(\beta) &= \beta\ trace\left([\Sigma_Y - \beta^2\ v_{g_m}\hat{\Pi}_m]^{-1}\ v_{g_m}\hat{\Pi}_m\right) + (\tilde{Y} - \beta\ \tilde{g}_m)'[\Sigma_Y - \beta^2\ v_{g_m}\hat{\Pi}_m]^{-1}\ \tilde{g}_m + \\
&\quad \frac{1}{2}(\tilde{Y} - \beta\ \tilde{g}_m)'\ [\Sigma_Y - \beta^2\ v_{g_m}\hat{\Pi}_m]^{-1}\ (\beta\ v_{g_m}\hat{\Pi}_m)\ [\Sigma_Y - \beta^2\ v_{g_m}\hat{\Pi}_m]^{-1}\ (\tilde{Y} - \beta\ \tilde{g}_m) \\
l'_{Y|g_m,\hat{\Pi}_m}(0) &= \tilde{Y}'\ \Sigma_Y^{-1}\ \tilde{g}_m,
\end{aligned}$$

which is the same as the usual score from the mean model (4.2.1), which ignores IBD information. The score in equation (4.2.7) is then given by

$$\begin{aligned}
Score(g_N \mid g_F, Y, \hat{\Pi}_m) &= \tilde{Y}'\ \Sigma_Y^{-1}\ \tilde{g}_m\ -\ E_{g_N|g_F,\hat{\Pi}_m}(\tilde{Y}'\ \Sigma_Y^{-1}\ \tilde{g}_m) \\
&= \tilde{Y}'\ \Sigma_Y^{-1}\ [g_m - E(g_m \mid g_F, \hat{\Pi}_m)],
\end{aligned}$$

which is same as the numerator of $SCORE.IBD.FP$. Similarly it can be shown that the numerators of $L(g_N, \hat{\Pi}_m \mid g_F, Y_N)$ and $L(g_m, \hat{\Pi}_m \mid \overline{g}_F, Y)$ are the same as those of the "type-2 statistics" $SCORE.IBD.AS$ and $SCORE.IBD.FPG$ respectively. For the likelihood $L(g_m, \hat{\Pi}_m \mid Y)$ the numerator is same as that for $SCORE.NS$ and $SCORE.IBD.NS$, as this likelihood is proportional to $L(Y \mid g_m, \hat{\Pi}_m)$. However, under the type-1 null hypothesis, the denominators should be evaluated as variance of the scores conditional on $g_F$ , similar to the type-1 statistics and unlike the denominators in the type 2 statistics, which use conditioning on $g_F$ and $\hat{\Pi}_m$. The standardized score test for testing the type-1 null, under "no stratification" is thus the same as $SCORE.NS$. The standardized score test (that models IBD) corresponding to $SCORE.FP$, has the form

$$\frac{\sum_i \tilde{Y}_i' \ \ \Sigma_Y^{-1} \ \ [g_{m_i} - E g_{m_i} \mid g_{F_i}, \hat{\Pi}_{m_i}]}{\sqrt{\sum_i \tilde{Y}_i' \ \ \Sigma_Y^{-1} \ \ \{Cov(g_{m_i} \mid g_{F_i}) - Cov_{\hat{\Pi}_m}[E(g_{m_i} \mid g_{F_i}, \hat{\Pi}_{m_i})]\} \ \ \Sigma_Y^{-1} \ \ \tilde{Y}_i}},$$

which has a different denominator from both $SCORE.FP$ and $SCORE.IBD.FP$. Similar formulas can be obtained for scores corresponding to type-1 statistics $SCORE.AS$ and $SCORE.FPG$. However, the denominator variances involve $Cov(\hat{\Pi}_m)$ (shown in section 4.2.5), which has to be computed empirically.

Score statistics incorporating linkage such as those derived from the implicit model assuming "possible linkage," would generally require the MLE of $v_a$ (as in equation 2.3.5). Obtaining the MLE is computationally intensive and also requires knowledge of the exact ascertainment scheme. Although the above statistics are derived ignoring the IBD information (or assuming "no linkage") they are reasonably powerful to detect association in presence of linkage. In fact, the type-1 statistics have similar power to QTDT (see simulation results in 4.3), a model which incorporates IBD explicitly and estimates the linkage parameters. However, the power of these statistics would depend on the type of null distribution used to obtain the means and variances required to standardize the scores. The choice of the type of null hypothesis is an important issue as it has an effect on the power to detect the alternative of interest "linkage AND association." The FBAT software (Horvath *et al.* 2001) recommends using the type-1 statistics for most purposes except for fine mapping under a linkage peak, as conditioning on IBD tends to reduce power (Rabinowitz and Laird 2000). Also, it is not obvious how the power of the scores that model IBD (derived in this section) would compare with respect to that of type 1 and type 2 statistics. We will not address these issues in this dissertation, and restrict our simulation comparisons to the type-1 statistics described in section 4.2.1.

### 4.2.3 Statistic Summary

Below we comparate the properties of the type-1 and type-2 statistics described in this section, as well as standard approaches FBAT and QTDT. The properties are also summarized in Table 4.2.3.

- $SCORE.NS$ and $SCORE.IBD.NS$ give measures of total association in a sample, and are comparable in power to a population-based study with the same number of individuals, provided the phenotype mean and covariance matrix (for a family) are specified correctly. $SCORE.NS$ uses the null hypothesis $H_{00}$ while $SCORE.IBD.NS$ uses $H_{00} \cup H_{L0}$. Both of them have power to detect association irrespective of linkage (i.e $H_{0A}$ or $H_{LA}$).

- $FBAT$ (and its type 2 equivalent) measure association "within a family" due to linkage. In fact, these statistics do not have any power to detect association in absence of linkage ($H_{0A}$). The type-1 statistic uses the null hypothesis of "no linkage" (with or without association). The type-2 statistic uses the type-2 null hypothesis $H_{LA}^c$ (i.e either no association or no linkage).
- $SCORE.AS$ and $SCORE.FP$ (and their type-2 equivalents) are similar in spirit to the FBAT type-1 and type-2 statistics, with two differences. Firstly, $FBAT$ takes an offset parameter $\mu$ and assumes residual environmental correlation among non-founders is zero. $SCORE.AS$ and $SCORE.FP$, on the other hand, require the population trait mean $\mu_Y$ and covariance $\Sigma_Y$ parameters. Secondly, while the FBAT statistics ignore founder phenotypes, $SCORE.FP$ provides a way of incorporating them when they are available and generation-specific biases are not suspected.
- $QTDT$ measures association "within a family" due to linkage. It uses the type-2 null hypothesis $H_{LA}^c$ i.e., it has correct type I error irrespective of whether linkage is present. However, unlike FBAT, it is only protected against between family stratification. It estimates nuisance parameters $\mu_Y$ and $\Sigma_Y$. For selected samples $QTDT$ gives correct type I error only if permutations are used to obtain the empirical distribution.
- $SCORE.FPG$ (and $SCORE.IBD.FPG$) measure "total association within a family," unlike $SCORE.AS$ and $SCORE.FP$ which measure only the part of the "within family association" that is due to linkage. In particular, it derives information from founder genotype-phenotype correlation from each family. Like $SCORE.NS$, $SCORE.FPG$ uses the null hypothesis $H_{00}$ while $SCORE.IBD.FPG$ uses $H_{00} \cup H_{L0}$. Both of them have power to detect association in the presence or absence of linkage.

Table 4.1: **Comparison of Statistics**

| | $SCORE.NS$ | $SCORE.AS$ | $FBAT$ | $SCORE.FP$ | $SCORE.FPG$ | $QTDT$ |
|---|---|---|---|---|---|---|
| Formula | $\tilde{Y}'\Sigma_Y^{-1}\tilde{g}$ | $\tilde{Y}'_N \Sigma_{NN}^{-1} g_{N/F}$ | $\tilde{Y}'_N [g_N - E(g_N \mid g_F)]$ | $\tilde{Y}'\Sigma_Y^{-1}[g_m - E(g_m \mid g_F)]$ | $\tilde{Y}'\Sigma_Y^{-1}[g_m - \overline{g}_F]$ | $\frac{N[\hat{\mu}_Y + \hat{a}_b\ g_b + \hat{a}_w\ g_w, \hat{\Sigma}_Y + \hat{v}_a\ (\hat{\Pi}_m - 2\Phi)]}{N[\hat{\mu}_Y + \hat{a}_b\ g_b, \hat{\Sigma}_Y + \hat{v}_a\ (\hat{\Pi}_m - 2\Phi)]}$ |
| Stratification Protection | None | Arbitrary | Arbitrary | Arbitrary | Between Family | Between Family |
| Null Hypothesis | T1: $H_{00}$<br>T2:$H_{00} \cup H_{L0}$ | T1:$H_{00} \cup H_{0A}$<br>T2: $H_{LA}^c$ | T1: $H_{00} \cup H_{0A}$<br>T2: $H_{LA}^c$ | T1: $H_{00} \cup H_{0A}$<br>T2: $H_{LA}^c$ | T1: $H_{00}$<br>T2: $H_{00} \cup H_{L0}$ | $H_{00} \cup H_{L0}$ |
| Condition on | $Y$ | $Y_N,\ g_F$ | $Y_N,\ g_F$ | $Y,\ g_F$ | $Y,\ \overline{g}_F$ | $g_F$ |
| Uses Founder Phenotypes | Yes | No | No | Yes | Yes | No |
| Founder Genotypes | Measures $g_F - Y_F$ correlation | Conditions on $g_F$ | Conditions on $g_F$ | Conditions on $g_F$ | Measures $g_F - Y_F$ correlation | Conditions on $g_F$ |
| Measures | Total Association in sample | Association due to linkage within family | Association due to linkage within family | Association due to linkage within family | Total Association within family | Association due to linkage within family |
| Detects | Association & Linkage | Association | Association | Association | Association & Linkage | Association |

### 4.2.4 Comparison with parenTDT

Purcell *et al.* (2005) proposed a model similar to QTDT that can incorporate parental phenotypes and parental genotype-phenotype correlation. They considered the likelihood $[Y \mid g_m] \sim N[\tilde{\mu}, \Sigma_Y]$, where

$$\tilde{\mu} = \begin{pmatrix} u + c\ \overline{g}_F + d\ (g_F - \overline{g}_F) \\ m + b\ \overline{g}_F + w\ (g_N - \overline{g}_F) \end{pmatrix}. \tag{4.2.8}$$

The parameters $u$ and $m$ are overall trait means allowing for a generational difference. The parameters $b$ and $c$ capture stratum effects assuming between family stratification and also allow for a generational difference. These parameters can be used to test for presence of stratification. $d$ and $w$ together capture total within family association (similar to $SCORE.FPG$) allowing for a generational difference. If parental phenotypes are ignored, only parameters $m$, $b$ and $w$ are modeled thus making the mean model identical to QTDT. They also considered regressing out founder phenotypes and using the QTDT type mean model (similar to $QTDT - FP$). They proposed several tests (Tests A-G, Purcell *et al.* 2005, pp 251) based on constraints on the parameters $b$, $c$,

$w$ and $d$. The score statistics discussed above can be thought of as corresponding to the following likelihood ratio tests under the parenTDT model (4.2.8).

- $SCORE.NS$ - $H_0 : u = m,\ b = c = w = d = 0 \ \ vs \ \ H_1 : u = m,\ b = c = w = d > 0$
- $SCORE.AS$ - Use only offspring phenotypes, $H_0 : m,\ b,\ w = 0 \ \ vs \ \ H_1 : m,\ b,\ w > 0$
- $SCORE.FP$ - Same as $SCORE.AS$, regress out parental phenotypes.
- $SCORE.FPG$ - $H_0 : u = m,\ b = c,\ w = d = 0 \ \ vs \ \ H_1 : u = m,\ b = c,\ w = d > 0$

This method, as originally proposed, does not incorporate IBD information but extension to a QTDT type model or the proposed implicit model (2.2.8) is straightforward. Also, this method was proposed for nuclear families but can be easily extended to handle general pedigrees. In spite of the generality and flexibility of this method, it has not been implemented into software for quantitative traits to our knowledge. An extension to binary traits is available using the "–parenTDT" options in the PLINK software package (Purcell *et al.* 2007; Purcell 2008). PLINK currently uses an *ad hoc* procedure "QFAM" (Purcell 2008) for family based association mapping of quantitative traits that corrects for family relationships using permutations.

### 4.2.5 Computing Conditional Moments

For nuclear families, FBAT uses an exhaustive enumeration of all transmissions consistent with Mendelian segregation to obtain the null distribution under no linkage, i.e., $[g_{N} \mid g_{F}]$. To generate the empirical distribution of $[g_{N} \mid g_{F}, \hat{\Pi}_m]$, FBAT uses a permutation-based algorithm. The algorithm randomly chooses neither, one or both of the parents and switches, in all offsprings, the transmitted alleles from those parents. The exact distributions thus obtained are then used to obtain the mean and variance of the denominator of the statistic. Extended pedigrees are broken up into nuclear families by FBAT. PBAT (Lange *et al.* 2004) uses the permutation-based R-L (Rabinowitz-Laird) algorithm (see Rabinowitz and Laird 2000, section 4.2.6 of this dissertation) to obtain these distributions exactly. This algorithm is quite general, in that it can handle missing founder genotypes, but at the same time it is computationally intensive. In sections 4.2.5.1 through 4.2.5.3, we derive some closed form expressions for the means and covariances under these distributions under the assumption of no missing founders. These formulas do not require the exact conditional distributions and therefore offer considerable efficiency in the computation of the null means and variances of the FBAT and the score statistics described in the previous section for situations where all the founders are genotyped. In section 4.2.6, we consider the case of missing founder genotypes and discuss possible modifications of the R-L algorithm to obtain the conditional distributions. We will assume a non-inbred pedigree with kinship coefficient matrix $\Phi$. Let $\Pi_m$ and $\hat{\Pi}_m = E(\Pi_m \mid M)$ denote the true unobserved IBD and the estimated IBD at the marker locus. We partition $\Phi$, $\Pi_m$ and $\hat{\Pi}_m$ into founder and non-founder blocks as

$$\Phi = \begin{pmatrix} \Phi_{FF} & \Phi_{FN} \\ \Phi_{NF} & \Phi_{NN} \end{pmatrix}, \quad \textit{and} \quad \Pi_m = \begin{pmatrix} \Pi_{FF} & \Pi_{FN} \\ \Pi_{NF} & \Pi_{NN} \end{pmatrix}, \quad \hat{\Pi}_m = \begin{pmatrix} \hat{\Pi}_{FF} & \hat{\Pi}_{FN} \\ \hat{\Pi}_{NF} & \hat{\Pi}_{NN} \end{pmatrix}.$$

Also, for the $n^{th}$ non-founder and $f^{th}$ founder, we will use the lower case symbols $\phi$, $\pi$ and $\hat{\pi}$ along with the subscripts $fN$, $nF$ and $nf$ to denote respectively the $f^{th}$ row (written as column vector), the $n^{th}$ column and the $(n, f)^{th}$ entry of these matrices.

#### 4.2.5.1 Conditional Means:

Below we compute the conditional means required for computing the numerators of the score statistics proposed in section 4.2.1 and 4.2.2.1.

$\mathbf{E(g_N \mid g_F)}$

It is known (e.g., Abecasis *et al.* 2000) that

$$\begin{aligned} E(g_n \mid g_F) &= 2\phi'_{nF} g_F \quad \text{and} \\ E(g_N \mid g_F) &= 2\Phi_{NF} g_F. \end{aligned}$$

This relation is easy to prove recursively for each member of a pedigree assuming it holds for that person's parents.

$\mathbf{E(g_N \mid \overline{g}_F)}$

To obtain the mean conditional on $\overline{g}_F$ we note that $g_f$s are iid binomial random variables. As a result, $[g_F \mid \overline{g}_F]$ has a multivariate hypergeometric distribution $HG(g_F; m = L\overline{g}_F, N = 2L, n = 2)$. We know that mean of the multivariate hypergeometric distribution is given by $E(g_{f_i}) = (nm/N)$. Therefore,

$$\begin{aligned} E(g_f \mid \overline{g}_F) &= \overline{g}_F \quad \text{and,} \\ E(g_n \mid \overline{g}_F) &= E_{g_F \mid \overline{g}_F}[E(g_n \mid g_F)] \\ &= 2\phi'_{nF}(\overline{g}_F\ \mathbf{1}) \\ &= \overline{g}_F \qquad (\because\ \phi'_{nF}\mathbf{1} = 1/2). \end{aligned}$$

$\mathbf{E(g_N \mid g_F, \Pi_m)}$

Conditioning on "true IBD" gives

$$\begin{aligned} E(g_n \mid g_F, \Pi_m) &= \pi'_{nF} g_F \quad \text{and,} \\ E(g_N \mid g_F, \Pi_m) &= \Pi_{NF} g_F. \end{aligned} \tag{4.2.9}$$

When the true IBD $\Pi_m$ is unknown, we use the fact that $[g_N \mid g_F, M] = [g_N \mid g_F, \hat{\Pi}_m]$, which follows from our assumption that all the markers are in linkage equilibrium with each other.

$$\begin{aligned} E(g_n \mid g_F, \hat{\Pi}_m) &= E(g_n \mid g_F, M) \\ &= E_{\Pi_m|M}[E(g_n \mid g_F, \Pi_m)] \\ &= \hat{\pi}'_{nF} g_F \quad \text{and} \\ E(g_N \mid g_F, \hat{\Pi}_m) &= \hat{\Pi}_{NF} g_F. \end{aligned}$$

The formulas 4.2.9 can be justified intuitively by the fact that the two founders who transmit their alleles to a non-founder act as that person's parents. To prove it more rigorously, let $f_1$ and $f_2$ denote the two founders who transmitted their alleles to "n". Also suppose $f_1$ and $f_2$ transmit the $i_1^{th}$ and $i_2^{th}$ alleles, where $(i_1, i_2) \in \{1, 2\}$. The probability of each such transmission is $(1/8)1_{\pi_{nf_1}=\pi_{nf_2}=1/2}$. Thus summing over all possible transmissions we have

$$\begin{aligned} E(g_n \mid g_F, \Pi_m) &= \sum_{f_1} \sum_{f_2 \neq f_1} \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} (g_{f_1,i_1} + g_{f_2,i_2})((1/8)1_{\pi_{nf_1}=\pi_{nf_2}=1/2}) \\ &= (\frac{1}{2}) \sum_{f_1} \sum_{f_2 \neq f_1} \sum_{i_1=1}^{2} g_{f_1,i_1} 1_{\pi_{nf_1}=\pi_{nf_2}=1/2} \quad \text{(By symmetry)} \\ &= (\frac{1}{2}) \sum_{f_1} g_{f_1} \; 1_{\pi_{nf_1}} \sum_{f_2 \neq f_1} 1_{\pi_{nf_2}=1/2} \\ &= (\frac{1}{2}) \sum_{f_1} g_{f_1} \; 1_{\pi_{nf_1}=1/2} \quad \text{(Only two founders share an allele IBD)} \\ &= \sum_{f_1} g_{f_1} \; \pi_{nf_1} \\ &= \pi'_{nF} g_F. \end{aligned}$$

$\mathbf{E(g_N \mid \overline{g}_F, \Pi_m)}$

$E(g_n \mid \overline{g}_F, \Pi_m)$ and $E(g_n \mid \overline{g}_F, \hat{\Pi}_m)$ can be obtained similarly as $E(g_n \mid \overline{g}_F)$ using the hypergeometric distribution. We condition on $g_{_F}$ and use the facts that $E(g_{_F} \mid \overline{g}_{_F}) = \overline{g}_{_F}\mathbf{1}$ and $\pi'_{_{nF}}\mathbf{1} = 1$, to get

$$\begin{aligned} E(g_n \mid \overline{g}_{_F}, \Pi_m) &= \overline{g}_{_F}\mathbf{1} \\ E(g_n \mid \overline{g}_{_F}, \hat{\Pi}_m) &= \overline{g}_{_F}\mathbf{1}. \end{aligned}$$

**4.2.5.2 Conditional Variances:** Here we derive expressions for the conditional variances of the scores some of which can be derived as special cases of the expressions for covariance derived in the next section.

$\mathbf{Var(g_N \mid g_F)}$

Let us first derive the conditional variance $Var(g_n \mid g_{_F})$. For a trio pedigree, if "p" and "m" denote the two parents of "n," then we know that $[g_n \mid g_p, g_m] \sim Ber(g_p/2) + Ber(g_m/2)$, a sum of two independent Bernoulli random variables (namely the indicators of the "A" allele being transmitted from each parent). Hence we get

$$\begin{aligned} Var(g_n \mid g_p, g_m) &= g_p(2-g_p)/4 + g_p(2-g_p)/4 \\ E(g_n^2 \mid g_p, g_m) &= Var(g_n \mid g_p, g_m) + (g_p+g_m)^2/4 \\ &= \frac{1}{2}(g_p + g_m + g_p g_m) \\ E(g_n^2 \mid g_{_F}) &= \frac{1}{2}E[(g_p + g_m + g_p g_m) \mid g_{_F}] \\ &= 2\phi'_{nF_p}g_{F_p} + 2\phi'_{nF_m}g_{F_m} + 8\ \phi'_{nF_p}g_{F_p}\ \phi'_{nF_m}g_{F_m}, \end{aligned}$$

where, $F_p$ and $F_m$ denote the founders of "n" in the paternal and maternal sides respectively. In the last step, we used the fact that $\phi_{pf} = \phi_{mf} = 2\phi_{nf}$. Using the above relation and the formula for the conditional mean in the previous section, we get

$$\begin{aligned} Var(g_n \mid g_{_F}) &= E(g_n^2 \mid g_{_F}) - E^2(g_n \mid g_{_F}) \\ &= \{2\phi'_{nF_p}g_{F_p} + 2\phi'_{nF_m}g_{F_m} + 8\ \phi'_{nF_p}g_{F_p}.\phi'_{nF_m}g_{F_m}\} - \{4(\phi_{nF_p}g_{F_p} + \phi_{nF_m}g_{F_m})^2\} \\ &= \{4\phi'_{nF_p}g_{F_p}.\phi'_{nF_p}\mathbf{2} + 4\phi'_{nF_m}g_{F_m}\phi'_{nF_m}\mathbf{2}\} - \{4\phi'_{nF_p}g_{F_p}.\phi_{nF_p}g'_{F_p} + 4\phi'_{nF_m}g_{F_m}.\phi'_{nF_m}g_{F_m}\} \\ &= 4g'_{F_p}\phi_{nF_p}\phi'_{nF_p}(\mathbf{2} - g_{F_p}) + 4g'_{F_m}\phi_{nF_m}\phi'_{nF_m}(\mathbf{2} - g_{F_m})\}, \end{aligned} \qquad (4.2.10)$$

where we have used the fact that $\phi'_{_{nF}}\mathbf{1} = 1/2$ (total kinship coefficient with founders), which follows from the definition of kinship coefficient.

$\mathbf{Var(g_N \mid g_F, \Pi_m)}$

Next let us derive the conditional variance given the true IBD, i.e., $Var(g_n \mid g_F, \Pi_m)$. As before, summing over all possible transmissions from founders, we get

$$\begin{aligned}
E(g_n^2 \mid g_F, \Pi_m) &= \sum_{f_1 \in F} \sum_{f_2 \neq f_1} \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} (g_{f_1,i_1} + g_{f_2,i_2})^2 ((1/8) 1_{\pi_{nf_1} = \pi_{nf_2} = 1/2}) \\
&= E(g_n \mid g_F, \Pi_m) + (\frac{1}{4}) \sum_{f_1 \in F} \sum_{f_2 \neq f_1} \sum_{i_1=1}^{2} \sum_{i_2=1}^{2} g_{f_1,i_1} g_{f_2,i_2} 1_{\pi_{nf_1} = \pi_{nf_2} = 1/2} \\
&\qquad (\because \quad g_{f_i}^2 = g_{f_i}) \\
&= g'_F \pi_{nF} + \{(1/2) \sum_{f_1 \in F} g_{f_1} 1_{\pi_{nf_1} = 1/2} \; (1/2) \sum_{f_2 \in F} g_{f_2} 1_{\pi_{nf_2} = 1/2} - (1/4) \sum_{f_1 \in F} g_{f_1}^2 1_{\pi_{nf_1} = 1/2} \\
&= g'_F \Pi_{nF} + E^2(g_n \mid g_F, \Pi_m) - (1/2) \; g'_F \; Diag(\Pi_{nF}) \; g_F
\end{aligned}$$

$$\begin{aligned}
Var(g_n \mid g_F, \Pi_m) &= g'_F \Pi_{nF} - (1/2) \; g'_F \; Diag(\Pi_{nF}) \; g_F \\
&= g'_F Diag(\Pi^2_{nF})(\mathbf{2}) - g'_F \; Diag(\Pi^2_{nF}) \; g_F \\
&= g'_F \; Diag(\Pi^2_{nF}) \; (\mathbf{2} - g_F),
\end{aligned}$$

where we have used the fact that $2\pi^2_{nf} = \pi_{nf}, \; \forall f \in F$. When the true IBD $\Pi_m$ is unknown, we use the fact $[g_n \mid g_F, M] = [g_n \mid g_F, \hat{\Pi}_m]$ and the identity $Var(Y) = Var[E(Y \mid X)] + E[Var(Y \mid X)]$, to get

$$\begin{aligned}
&Var(g_n \mid g_F, \hat{\Pi}_m) = Var(g_n \mid g_F, M) \\
&\quad = E_{\Pi_m \mid M}[Var(g_n \mid g_F, \Pi_m)] + Var_{\Pi_m \mid M}[E(g_n \mid g_F, \Pi_m)] \\
&\quad = g'_F \hat{\Pi}_m - (1/2) \; g'_F \; Diag(\hat{\Pi}_{nF}) \; g_F + g'_F \; Cov(\pi_{nF} \mid M) \; g_F.
\end{aligned}$$

**4.2.5.3 Conditional Covariances:** Here we derive expressions for the conditional covariance of genotypes of two relatives that are required for computing the denominators of the score statistics proposed in sections 4.2.1 and 4.2.2.1.

$\mathbf{Cov(g_N \mid g_F)}$

Let us first derive the conditional covariance $Cov(g_{n_1}, g_{n_2} \mid g_F)$ for two non-founders $n_1$ and $n_2$. Let $F_C$, $F_1$ and $F_2$ denote the common ancestors of ($n_1$ and $n_2$), the unique ancestors of $n_1$ (i.e not common with $n_2$) and the unique ancestors of $n_2$ respectively. For each common ancestor in $F_C$, there is a unique "Most Recent Common Ancestor" (MRCA) in the pedigree (possibly same as that ancestor). Two common ancestors can share the same MRCA, but a common ancestor can not have two MRCAs. Let $M_C$ denote the set of MRCAs for $n_1$ and $n_2$. Then $g_{n_1}$ and $g_{n_2}$ are independent of $g_{F_C}$ and of each other conditional on $g_{M_C}$. Hence,

$$\begin{aligned} Cov(g_{n_1}, g_{n_2} \mid g_{_F}) &= Cov_{g_{M_C} \mid g_{F_C}}[E(g_{n_1} \mid g_{M_C}, g_{F_1}), E(g_{n_2} \mid g_{M_C}, g_{F_2})] \\ &= Cov_{g_{M_C} \mid g_{F_C}}[(2\phi'_{n_1 F_1} g_{F_1} + 2\phi'_{n_1 M_C} g_{M_C})\ (2\phi'_{n_2 F_2} g_{F_2} + 2\phi'_{n_2 M_C} g_{M_C})] \\ &= Cov_{g_{M_C} \mid g_{F_C}}[2\phi'_{n_1 M_C} g_{M_C}, 2\phi'_{n_2 M_C} g_{M_C}] \\ &= 4\phi'_{n_1 M_C} Cov(g_{M_C} \mid g_{F_C}) \phi_{n_2 M_C} \\ &= 4\phi'_{n_1 M_C} Diag[Var(g_{M_C} \mid g_{F_C})] \phi_{n_2 M_C}. \end{aligned} \tag{4.2.11}$$

The last step follows from the fact that the MRCAs are independent of each other, as each MRCA corresponds to a distinct founder or set of founders. Equation (4.2.11) gives a recursive formula for the conditional covariance $Cov(g_{n_1}, g_{n_2} \mid g_{_F})$. To get a closed-form non-recursive expression, we substitute the expression for $Var(g_{M_C} \mid g_{F_C})$ as derived before in equation (4.2.10). For each MRCA $c$, let $F_{p(c)}$ and $F_{m(c)}$ denote the founders of $c$, in the paternal and maternal sides respectively. Also note that $Var(g_c \mid g_{F_C}) = 0$, if $c$ itself is a founder. Equation (4.2.10) can be written as

$$\begin{aligned} & Cov(g_{n_1}, g_{n_2} \mid g_{_F}) = 4 \sum_{c \in M_c} \phi'_{n_1 c}[Var(g_c \mid g_{F_{p(c)}}, g_{F_{m(c)}})] \phi_{n_2 c} \\ &\quad = 4 \sum_{c \in M_c} [1_{c \notin F}] \phi_{n_1 c}[4\phi'_{cF_{p(c)}} g_{F_{p(c)}} (\mathbf{2} - g_{F_{p(c)}})' \phi_{cF_{p(c)}} + 4\phi'_{cF_{m(c)}} g_{F_{m(c)}} (\mathbf{2} - g_{F_{m(c)}})' \phi_{cF_{m(c)}}] \phi_{n_2 c} \\ &\quad = 4 \sum_{c \in M_c} [1_{c \notin F}] 4\phi'_{n_1 F_{p(c)}} g_{F_{p(c)}} (\mathbf{2} - g_{F_{p(c)}})' \phi_{n_2 F_{p(c)}} + 4\phi'_{n_1 F_{m(c)}} g_{F_{m(c)}} (\mathbf{2} - g_{F_{m(c)}})' \phi_{n_2 F_{m(c)}} \\ &\qquad\qquad\qquad\qquad\qquad\qquad (\ \because \quad \phi_{n_1 F_{p(c)}} = 2\phi_{n_1 c} \phi_{n_1 F_{p(c)}}, \text{ etc}) \end{aligned}$$

$$\begin{aligned} &\quad = 4 \ \sum_{c \in M_c} [1_{c \notin F}] \sum_{f_i \in F_{p(c)}}^{L} \ \sum_{f_j \in F_{p(c)}}^{L} [\phi_{n_1 f_i} g_{f_i} (2 - g_{f_j}) \phi_{n_2 f_j}] + \sum_{f_i \in F_{p(c)}}^{L} \ \sum_{f_j \in F_{p(c)}}^{L} [\phi_{n_1 f_i} g_{f_i} (2 - g_{f_j}) \phi_{n_2 f_j}] \\ &\quad = 4 \ \sum_{i=1}^{L} \sum_{j=1}^{L} M_{(i,j)}(n_1, n_2) . \phi_{n_1 f_i} g_{f_i} (2 - g_{f_j}) \phi_{n_2 f_j} \end{aligned}$$

$$\quad = 4 \ \phi'_{n_1 F} Diag(g_{_F}) \ M(n_1, n_2) \ Diag(\mathbf{2} - g_{_F}) \phi_{n_2 F} \tag{4.2.12}$$

$$\quad = 4 \ g'_{_F} Diag(\phi_{n_1 F}) \ M(n_1, n_2) \ Diag(\phi_{n_2 F}) \ (\mathbf{2} - g_{_F}) \tag{4.2.13}$$

where $M(n_1, n_2)$ is the $L \times L$ matrix with $(i,j)^{th}$ entry given by

$$M_{i,j}(n_1, n_2) = \begin{cases} 1 & \text{if paths } f_i \rightarrow n_1 \text{ and } f_j \rightarrow n_2 \text{ share at least one meiosis,} \\ 0 & \text{otherwise.} \end{cases}$$

In the above derivation, we have used the fact that two ancestors are on the paternal (or maternal side) of an MRCA (who is not a founder) if and only if the paths from the founders share at least one meiosis. In fact, one such meiosis is from the MRCA's father (or mother) to the MRCA. Thus, equation (4.2.13) gives a closed form formula for obtaining the genotype covariances conditional on the founders. But is should be noted that the matrix $M(n_1, n_2)$ changes with $n_1$ and $n_2$, which can be precomputed for a particular pedigree structure. For example, if $(n_1, n_2)$ are siblings in a nuclear family the matrix $M$ has all zero entries, whereas if a pair of grandparents are available then that pair contributes to $M$, making the covariance non-zero. This can also be seen from the recursive expression (4.2.11). In the former case all MRCA's (i.e parents) are founders, so the sibs are uncorrelated, whereas in the latter case one parent is an MRCA but not a founder, and has positive variance, resulting in a positive covariance between the sibs. Figure 4.1 illustrates this idea.

$\mathbf{Cov(g_N \mid \overline{g}_F)}$

To obtain the covariances conditional on $\overline{g}_{_F}$, we use that fact the multivariate hypergeometric distribution with parameters $HG(g_{_F}; n = L\overline{g}_{_F}, N = 2L, m = 2)$ has variance and covariance given by

$$\begin{aligned} Var(g_{f_i} \mid \overline{g}_{_F}) &= \frac{nm(N-m)(N-n)}{N^2(N-1)} \\ &= \frac{\overline{g}_{_F}(2-\overline{g}_{_F})(L-1)}{2L-1} \\ Cov(g_{f_i}, g_{f_j} \mid \overline{g}_{_F}) &= -\frac{nm^2(N-n)}{N^2(N-1)} \\ &= -\frac{\overline{g}_{_F}(2-\overline{g}_{_F})}{2L-1}. \end{aligned}$$

In matrix notation, we can rewrite the above as

$$\begin{aligned} Cov(g_{_F} \mid \overline{g}_{_F}) &= \frac{\overline{g}_{_F}(2-\overline{g}_{_F})}{(2L-1)}[LI - J] \\ &= h\ (LI - J) \quad (\text{where,} \quad h = \tfrac{\overline{g}_{_F}(2-\overline{g}_{_F})}{(2L-1)}) \\ \text{Also,}\ \ E[g_{_F}(\mathbf{2} - g_{_F}) \mid \overline{g}_{_F}] &= E[g_{_F} \mid \overline{g}_{_F}]E[(\mathbf{2} - g_{_F})' \mid \overline{g}_{_F}] - Cov(g_{_F} \mid \overline{g}_{_F}) \\ &= h\ (2L-1)J - h(LI - J) = h\ (2LJ - LI), \end{aligned}$$

where $J = \mathbf{1}\mathbf{1}'$. To obtain $Cov(g_{n_1}, g_{n_2} \mid \overline{g}_F)$ we use the relation $Cov(X,Y) = E_Z[Cov(X,Y \mid Z) + Cov_Z[E(X \mid Z), E(Y \mid Z)]$ by conditioning on $g_F$.

$$
\begin{aligned}
Cov(g_{n_1}, g_{n_2} \mid \overline{g}_F) &= Cov_{g_F \mid \overline{g}_F}(2\phi'_{n_1F} g_F, 2\phi'_{n_2F} g_F) + E_{g_F \mid \overline{g}_F}(g'_F Diag(\phi_{n_1F})\ M\ Diag(\phi_{n_1F})(\mathbf{2} - g_F)) \\
&= 4\phi'_{n_1F}[h\ (LI - J)]\phi_{n_2F} + 4\ trace\{Diag(\phi_{n_1F})\ M\ \phi_{n_1F}\ h\ (2LJ - LI)\} \\
&= 2Lh\ \left\{ \sum_{f \in F} 2\phi_{n_1f}\phi_{n_2f} + \sum_{f_1 \in F} \sum_{f_2 \in F} M_{f_1,f_2}(n_1,n_2) 4\phi_{n_1f}\phi_{n_2f} - \sum_{f \in F} M_{f,f}(n_1,n_2) 2\phi_{n_1f}\phi_{n_2f} \right\} - hJ \\
&= 2Lh\ \left\{ \sum_{f \in F} [1 - M_{f,f}(n_1,n_2)]\ 2\phi_{n_1f}\phi_{n_2f} + \sum_{f_1 \in F} \sum_{f_2 \in F} M_{f_1,f_2}(n_1,n_2) 4\phi_{n_1f}\phi_{n_2f} \right\} - hJ,
\end{aligned}
$$

where we have used the facts $\phi'_{nF}\mathbf{1} = 1/2$ and $trace[ADiag(B)] = trace[Diag(A)Diag(B)]$. The last expression can be simplified to $\Phi_{NN}$ (see appendix D), so that $Cov(g_N \mid \overline{g}_F)$ is given by

$$
\begin{aligned}
Cov(g_N \mid \overline{g}_F) &= 2Lh\ \Phi_{NN} - hJ \\
&= \frac{\overline{g}_F(2 - \overline{g}_F)}{(2L-1)}[2L\ \Phi_{NN} - J].
\end{aligned}
$$

### $\mathbf{Cov(g_N \mid g_F, \Pi_m)}$

In general it is difficult to obtain an expression for $Cov(g_N \mid g_F, \Pi_m)$ in terms of the pairwise IBDs $\Pi_m$. However, if the true IBD configuration (inheritance vector) $C_m$ is known, it can be shown that

$$
\begin{aligned}
Cov(g_{n_1}, g_{n_2} \mid g_F, C_m) &= \pi'_{n_1F} Diag(g_F)\ \tilde{M}(n_1,n_2)\ Diag(\mathbf{2} - g_F)\pi_{n_2F} \\
&= g'_F\ Diag(\pi_{n_1F})\ \tilde{M}(n_1,n_2)\ Diag(\pi_{n_2F})\ (\mathbf{2} - g_F),
\end{aligned}
$$

where $M(n_1,n_2)$ is the $L \times L$ diagonal matrix with $(i,i)^{th}$ entry given by

$$
\tilde{M}_{i,i}(n_1,n_2) = \begin{cases} +1 & \text{if founder } f_i \text{ transmits the same allele to } n_1 \text{ and } n_2, \\ -1 & \text{otherwise.} \end{cases}
$$

A proof of this result has been outlined in Appendix E. Note that in many cases pairwise IBDs would be sufficient to infer the matrix $\tilde{M}$, however this may be difficult for founders who are also MRCA. For example, in the case of a sibling pair family, if the sibs share 1 allele IBD, it would not be possible to infer which of the two parents (MRCAs) transmitted the same allele to both of the sibs. When $C_m$ is unknown, we can use the estimated distribution of the inheritance vectors conditional on the marker data $M$ to obtain

$$Cov(g_{n_1}, g_{n_2} \mid g_{_F}, M) = E_{C_m|M}[Cov(g_{n_1}, g_{n_2} \mid g_{_F}, C_m)] + Cov_{\Pi_m|M}[E(g_{n_1} \mid g_{_F}, \Pi_m), E(g_{n_2} \mid g_{_F}, \Pi_m)]$$
$$= g'_{_F}\ E\left[Diag(\pi_{_{n_1F}})\ \tilde{M}(n_1, n_2)\ Diag(\pi_{_{n_2F}}) \mid M\right]\ (\mathbf{2} - g_{_F}) + g'_{_F}\ Cov(\pi_{_{n_1F}}, \pi_{_{n_2F}} \mid M)\ g_{_F}.$$

**$\mathbf{Cov(g_N \mid \overline{g}_F, \Pi_m)}$**

To obtain $Cov(g_{n_1}, g_{n_2} \mid \overline{g}_{_F}, C_m)$, we use similar ideas as for $Cov(g_{n_1}, g_{n_2} \mid \overline{g}_{_F})$. As before, using the moments of the hypergeometric distribution we have

$$Cov(g_{_{n_1}}, g_{_{n_2}} \mid \overline{g}_{_F}, C_m) = Cov_{g_{_F}|\overline{g}_{_F}}(\pi'_{n_1F} g_{_F}, 2\pi'_{n_2F} g_{_F}) + E_{g_{_F}|\overline{g}_{_F}}[g'_{_F} Diag(\pi_{n_1F})\ \tilde{M}\ Diag(\pi_{n_2F})(\mathbf{2} - g_{_F})]$$
$$= \pi'_{n_1F}[h\ (LI - J)]\pi_{n_2F} + trace\{Diag(\pi_{n_1F})\ \tilde{M}\ Diag(\pi_{n_2F})\ h\ (2LJ - LI)\}$$

$$= Lh\ \left\{ \sum_{f \in F} \pi_{n_1f}\pi_{n_2f} + \sum_{f \in F} \tilde{M}_{f,f}(n_1, n_2)\pi_{n_1f}\pi_{n_2f} \right\} - hJ.$$

Defining $K(n_1, n_2)$ as the matrix $\tilde{M}$ with $+1$ and $-1$ on the diagonal replaced by 1 and 0 respectively, and noting that $\tilde{M} + 1 = 2K$, we get

$$\begin{aligned} Cov(g_{_{n_1}}, g_{_{n_2}} \mid \overline{g}_{_F}, C_m) &= Lh \sum_{f \in F} 2\ K_{f,f}(n_1, n_2)\pi_{n_1f}\pi_{n_2f} - hJ \\ &= Lh\ \pi_{_{n_1,n_2}} - LJ \\ Cov(g_{_N} \mid \overline{g}_{_F}, \Pi_m) &= Lh\ \Pi_{NN} - hJ \\ &= \frac{\overline{g}_{_F}(2 - \overline{g}_{_F})}{(2L - 1)}[L\ \Pi_{NN} - J], \end{aligned}$$

where we have used the fact that

$$\pi_{_{n_1,n_2}} = \sum_{f \in F} 2\ K_{f,f}(n_1, n_2)\pi_{n_1f}\pi_{n_2f},$$

which follows from the definition of $\pi_{_{n_1,n_2}}$. When $\Pi_{_{NN}}$ can not be inferred exactly the estimated matrix $\hat{\Pi}_{_{NN}}$ can be used.

### 4.2.6 Missing Data

The score statistics discussed in this chapter can handle missing (founder or non-founder) phenotypes and missing non-founder genotypes just by restricting $Y$ and $g_m$ (when constructing the statistics) to individuals for which both genotypes and phenotypes are available. However, *SCORE.FP* can use phenotypes of founders that are not genotyped, as founder phenotypes can be thought of as covariates. In either case, the conditional mean and variance of $g_m$ should be computed based on genotypes of all the founders including those that are not phenotyped. When some founder genotypes are missing, the conditional moments are computed conditional on all the available genotypes in the pedigree or a function thereof (as discussed below), including those of founders and non-founders with missing phenotypes. In this section, we focus on partly or completely missing founder genotypes and discuss possible modifications of the score statistics in such cases.

The formulas described in the sections 4.2.5.1 through 4.2.5.3 are useful when all the founders in a pedigree are genotyped. When some or all of the founder genotypes are missing, one possible approach is to impute the missing founder genotypes based on the observed genotypes, the flanking markers and the IBD information. If imputed founder genotypes are used, the formulas described above can be used to obtain the required conditional moments. QTDT uses a rough estimate (average of available siblings) to impute $E(g_n \mid g_{_F})$. FBAT, on the other hand, uses an exact approach of conditioning on the minimal sufficient statistic based on the available data. Below, we give a brief motivation and outline of this approach and discuss possible modifications.

**4.2.6.1 Arbitrary Stratification** Let $g_A$ denote the available genotype data in a pedigree, which includes some non-founders and possibly some founders. Let $p_m$ denote the allele frequencies in the founders, which can be possibly different for each founder in the case of extreme stratification. Ideally we should use conditioning on the minimal sufficient statistic for $p_m$ based on the observed incomplete data $g_A$. We recall that $g_{_F}$ is minimal sufficient for the complete data. Hence, when $g_{_F}$ is partly or fully missing, one natural and intuitive way of obtaining a minimal sufficient statistic for $[g_A \mid p_m]$ is to obtain a minimal sufficient statistic for $g_{_F}$ (which is unknown) based on the likelihood $[g_A \mid g_{_F}]$. Rabinowitz and Laird (2000) showed that such a strategy would (by transitivity) lead to the desired minimal sufficient statistic (for $[g_A \mid p_m]$), provided that the full data minimal sufficient statistic (in this case $g_{_F}$) is also "complete." The completeness would hold only in the case of completely arbitrary stratification. If we restrict to "between family stratification" or "admixture," $g_{_F}$ is still sufficient but not complete. For example, in these cases the function $g_P - g_M$ (where $g_P$ and $g_m$ are the paternal and maternal genotypes in a nuclear family) would have expectation zero for all $p_m$, thus violating the definition of completeness. Thus, if we possess some knowledge ("restriction") of the stratification scheme, the R-L algorithm may not generate the desired conditional distribution of $g_A \mid p_m$. Nevertheless, even for restricted stratification schemes like admixture, whenever $g_F$ is minimal sufficient for the complete data, this algorithm is expected to provide a good approximation to the actual minimal sufficient statistic. However, in case of "between family stratification," the full data minimal sufficient statistic is $\overline{g}_F$ (in stead of $g_F$) and hence the R-L algorithm would have to be modified to use the minimal sufficient statistic for $[g_A \mid \overline{g}_F]$.

- **Type-1 Null Hypothesis** To motivate the R-L algorithm, let $T(\cdot)$ and $T(g_A) = t_0$ denote the minimal sufficient statistic and its observed value respectively. Let $ST_A$ denote the set $T^{-1}(t_0)$ of possible outcomes which give the same value of the minimal sufficient statistic as the observed data. To obtain the mean $E(g_A \mid t_0)$ and covariance matrix $Cov(g_A \mid t_0)$, we need to compute the probability $P(g_a \mid T(g_a) = t_0)$ for all outcomes $g_a$ in the set $ST_A$. Let $SF_A$ denote the set $\{g_F : P(g_A \mid g_F) > 0\}$ of patterns of founder genotypes that are compatible with $g_A$. Let $g_{F_0}$ denote one such pattern in $SF_A$. We recall that by definition of the minimal sufficient statistic, the following two conditions are satisfied.

$$\{g_F : P(g_a \mid g_F) > 0\} = SF_A \qquad \forall\ g_a \in ST_A \qquad \text{(By sufficiency)}$$

$$\frac{P(g_a \mid g_F)}{P(g_A \mid g_F)} = \frac{P(g_a \mid g_{F_0})}{P(g_A \mid g_{F_0})} \qquad \forall\ g_F \in SF_A \text{ and } \forall\ g_a \in ST_A. \tag{4.2.14}$$

The above two conditions can be used to identify $ST_A$ the support of the required distribution $[g_a \mid T(g_a) = t_0]$. For any $g_a$ in this support, the probability is computed as follows

$$\begin{aligned} P(g_a \mid T(g_a) = t_0) &= \sum_{g_F \in SF_A} P(g_a \mid g_F, t_0) P(g_F \mid t_0) \\ &= P(g_a \mid g_{F_0}, t_0) = \frac{P(g_a \mid g_{F_0})}{P(g_a \in ST_A \mid g_{F_0})} \\ &\propto P(g_a \mid g_{F_0}), \end{aligned}$$

where the second step follows due to the fact that $P(g_a \mid g_F, t_0)$ is a constant for all $g_F$ in $SF_A$, under the assumptions of equation (4.2.14). Thus the required conditional distribution can be obtained by assigning an arbitrary compatible pattern $g_{F_0}$ to the founders and using Mendelian transmission rules to compute $P(g_a \mid g_{F_0})$. However, obtaining the support of the distribution is computationally more difficult. It is not simply "all outcomes compatible with $g_{F0}$." The R-L algorithm exhaustively computes the $SF_a = \{g_F : P(g_a \mid g_F) > 0\}$ for each $g_a$ compatible with

$SF_A$ and selects the subset that satisfies $SF_a = SF_A$ and equation (4.2.14). The search space can be significantly restricted by considering each $g_a$ compatible with $g_{F_0}$. Also wherever the pedigree contains a genotyped-couple the pedigree underneath that couple can be pruned out. This would lead to a number of independent sub-pedigrees $P_i$, with disjoint sets of founders $F_i$. Minimal sufficient statistics $T_i$ can be obtained for each $P_i$ based on the observed data $g_{A_i}$ in that pedigree. Then $P(g_{a_i} \mid T(g_{a_i} = T_{0_i})$ can be computed for each sub-part and multiplied to obtain the joint probabilities. Note that the couples used for pruning are not counted in two likelihoods, as the likelihood of the pedigree beneath them conditions on their genotypes.

- **Type-2 Null Hypothesis** Under the type-2 null hypothesis, we need the distribution $P(g_a \mid t_0, C_m)$, where $C_m$ is the (true) inheritance vector. When the marker data are not fully informative the mean $E(g_A \mid t_0, M)$ and covariance $Cov(g_A \mid t_0, M)$ can be obtained by using the formulas

$$E(g_A \mid t_0, M) = E_{C_m|M}[E(g_A \mid t_0, C_m)]$$
$$Cov(g_A \mid t_0, M) = E_{C_m|M}[Cov(g_A \mid t_0, C_m)] + Cov_{C_m|M}[E(g_A \mid t_0, C_m)].$$

To obtain $P(g_a \mid t_0, C_m)$, let $T(\cdot, C_m)$ and $T(g_A, C_m) = t_0$ be the minimal sufficient statistic and its observed value. Let $ST_{A,C_m}$ denote the set $T^{-1}(t_0)$ of outcomes that give the same value of the minimal sufficient statistic as the observed value. We define $SF_{A,C_m}$ as the set $\{g_F : P(g_A \mid g_F, C_m) > 0\}$ of patterns of founder genotypes that are compatible with $g_A$ and $C_m$. Let $g_{F_0,C_m}$ be one such pattern. The conditional genotypes can be obtained as before, using the formula

$$P(g_a \mid t_0, C_m) \propto P(g_a \mid g_{F_0,C_m}, C_m).$$

The support of this distribution can be obtained by searching over $g_a$ compatible with $g_{F_0,C_m}$ and $C_m$, that satisfies the conditions

$$\begin{aligned} \{g_F : P(g_a \mid g_F, C_m) > 0\} = SF_{A,C_m} \qquad & \forall\ g_a \in ST_{A,C_m} \\ \frac{P(g_a \mid g_F, C_m)}{P(g_A \mid g_F, C_m)} = \frac{P(g_a \mid g_{F_0}, C_m)}{P(g_A \mid g_{F_0}, C_m)} \qquad & \forall\ g_F \in SF_{A,C_m} \text{ and } \forall\ g_a \in ST_{A,C_m}. \end{aligned}$$

Note that this formulation is different from that used by the RL-algorithm. To compute conditional distributions in the presence of linkage when some founders are missing, the RL-procedure outlined in Rabinowitz and Laird (2000) ignores the IBD information (if available) and uses the minimal sufficient statistic for $g_F$ and $C_m$ (treating them as missing) based on the observed genotypes $g_A$. The above procedure on the other hand assumes that the IBD pattern $C_m$ is known. When $C_m$ is not available, the distribution $[C_m \mid M]$ (computed by multipoint methods) can be used to weight the means and covariances accordingly. Thus the above procedure uses more of the available marker information for computing the means and variances at each locus, except in the case of single-point IBD estimation. Even for single point analysis, the two procedures differ. The RL-algorithm conditions on the minimal sufficient statistic for IBD (by treating it as missing), so that the distribution under linkage becomes free of IBD. This extra conditioning possibly explains the low power of Type-2 statistics with variances estimated using the RL-algorithm. LAKE *et al.* (2000) observed that nuclear families with ambiguous IBD are not used in the RL-algorithm, resulting in substantial loss of information. Currently, FBAT and PBAT use the mean based on "no linkage" (i.e., under type-1 null) and an empirical variance to avoid this loss of information. The procedure proposed above may provide significant improvement in power for type-2 statistics, although because of the computational burden under ambiguous IBD information, computing the mean under type-1 null and using an empirical covariance "conditional on trait" may still be preferable.

#### 4.2.6.2 Between Family Stratification

In this case, the full data minimal sufficient statistic is given by $\overline{g}_{F}$. When some or all of the founders are untyped, we condition on the minimal sufficient statistic for $\overline{g}_{F}$. Let $T(\cdot)$ and $T(g_A) = t_0$ denote the minimal sufficient statistic for $\overline{g}_{F}$, and its observed value. Let $ST_A$ denote the set $T^{-1}(t_0)$. We define $SF_A$ as the set $\{\overline{g}_{F} : P(g_A \mid \overline{g}_{F}, C_m) > 0\}$ of founder genotype means that are compatible with $g_A$. Let $\overline{g}_{F_0}$ be one such mean. The conditional genotypes can be obtained as before, using the formula

$$P(g_a \mid t_0) \propto P(g_a \mid \overline{g}_{F_0}).$$

The support of this distribution can be obtained by searching over $g_a$ compatible with $\overline{g}_{F_0}$ that satisfies the conditions:

$$\begin{aligned} \{\overline{g}_{F} : P(g_a \mid \overline{g}_{F}) > 0\} = SF_A & \qquad \forall\ g_a \in ST_A \\ \frac{P(g_a \mid \overline{g}_{F})}{P(g_A \mid \overline{g}_{F})} = \frac{P(g_a \mid \overline{g}_{F_0})}{P(g_A \mid \overline{g}_{F_0})} & \qquad \forall\ g_F \in SF_A \text{ and } \forall\ g_a \in ST_A. \end{aligned}$$

Note that the probabilities $P(g_a \mid \overline{g}_{F})$ can be computed using the fact that

$$P(g_a \mid \overline{g}_{F}) = \sum_{g_F} P(g_a \mid g_F)\ P(g_F \mid \overline{g}_{F}),$$

where the second term inside the summation is computed based on the hypergeometric distribution.

Under the type-2 null hypothesis, the algorithm above can be modified similar to the proposal above for arbitrary stratification. To avoid the computational complexity of the above approach, we can use an imputation approach, where the missing $\overline{g}_{_F}$ is estimated by its best linear unbiased estimator (e.g., Bourgain *et al.* 2003):

$$\widehat{\overline{g}_{_F}} = (\mathbf{1}'\Phi_{_{AA}}^{-1}\mathbf{1})^{-1}(\mathbf{1}'\Phi_{_{AA}}^{-1}g_{_A}),$$

where $\Phi_{_{AA}}$ is the kinship coefficient matrix for the genotyped individuals. The imputed value of $\overline{g}_{_F}$ can then be plugged into the mean and covariance formulas, derived in section 4.2.5 for both type-1 and type-2 null hypothesis.

### 4.2.7 Handling Non-normal Traits

In this section we propose a novel score test for association mapping using second derivatives of the likelihood function that is similar to the standard score test for linkage. This statistic can be easily extended to incorporate higher moments (Chen *et al.* 2005) as analogous to linkage scores. However, this statistic was found in limited simulations (data not shown) to be considerably less powerful than the first derivative based score tests.

In association analysis, we are interested in testing the hypotheses $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. We know that the score test gives a locally most powerful (LMP) test for testing a one-sided alternative hypothesis. In general an LMP test for testing a 2-sided hypotheses can not be guaranteed. However, it can be shown that a locally most powerful unbiased (LMPU) test exists, which imposes the additional "unbiasedness" restriction on the test, i.e., all tests for which the power function has a local minimum at the null value $\beta = 0$. The assumption of "unbiasedness" is not very restrictive, as any reasonable test in this case would have a smooth symmetric power function, with a unique global minimum at $\beta = 0$.

It follows from section 2.4 that the LMPU score for the reverse likelihood $[g_m \mid Y]$ is same as that for the forward likelihood

$$L_{Y|g_m}(\beta) \propto \frac{\exp\{-\frac{1}{2}(\tilde{Y} - \beta\ \tilde{g}_m)'[\Sigma_{_Y} - \beta^2\ \Sigma_g]^{-1}\ (\tilde{Y} - \beta\ \tilde{g}_m)\}}{|\Sigma_{_Y} - \beta^2\ \Sigma_g|^{\frac{1}{2}}}. \tag{4.2.15}$$

As shown in Rao (2002) (pp453-455), the LMPU statistic for association is given by

$$S^{assoc}_{_{LMPU}} = \frac{L''_{Y|g_m}(0)}{L_{Y|g_m}(0)},$$

which, after some simple algebra (see appendix A)becomes,

$$S^{assoc}_{LMPU} = l''_{Y|g_m}(0) + [l'_{Y|g_m}(0)]^2. \tag{4.2.16}$$

Substituting expressions for $l'_{Y|g_m}(0)$ and $l''_{Y|g_m}(0)$, given in Appendix A, we get

$$\begin{aligned} S^{assoc}_{LMPU} &= -\ \tilde{Y}'\ \Sigma_Y^{-1}\Sigma_g\Sigma_Y^{-1}\ \tilde{Y} + trace(\Sigma_Y^{-1}\Sigma_g) - trace(\Sigma_Y^{-1}\widetilde{g_m}\widetilde{g_m}') + \tilde{Y}'\ \Sigma_Y^{-1}\widetilde{g_m}\widetilde{g_m}'\Sigma_Y^{-1}\ \tilde{Y} \\ &= \ \tilde{Y}'\ \Sigma_Y^{-1}[\widetilde{g_m}\widetilde{g_m}' - \Sigma_g]\Sigma_Y^{-1}\ \tilde{Y} - trace[\Sigma_Y^{-1}(\widetilde{g_m}\widetilde{g_m}' - \Sigma_g)] \\ &= \ vec[\Sigma_Y^{-1}\ \tilde{Y}\tilde{Y}'\ \Sigma_Y^{-1} - \Sigma_Y^{-1}]'vec[\widetilde{g_m}\widetilde{g_m}' - \Sigma_g]. \end{aligned}$$

Under the biallelic model (4.2.1), assuming random union of gametes (random mating) it can be shown that $Eg_m = \mathbf{2}p_m$ and $\Sigma_g = 4p_m q_m \Phi$. Hence the LMPU statistic can be written as

$$S^{assoc}_{LMPU} = \ vec[\Sigma_Y^{-1}\ \tilde{Y}\tilde{Y}'\Sigma_Y^{-1} - \Sigma_Y^{-1}]'\ vec[(g_m - \mathbf{2}p_m)(g_m - \mathbf{2}p_m)' - 4p_m q_m \Phi]. \tag{4.2.17}$$

Note that this statistic closely resembles the standard VC model based score statistic for linkage (equation 2.3.1), with the marker IBD matrix $\hat{\Pi}_m$ being replaced by a direct measure of genotype similarity $(g_m - \mathbf{2}p_m)(g_m - \mathbf{2}p_m)'$. Chen *et al.* (2005) described a class of GEE based score tests for linkage - the higher moment tests, which allow violation of normality to some extent in that they allow the trait distribution to have non-zero skewness and kurtosis. These methods have been discussed in detail in chapter 3. The form of the LMPU association test being similar to that of the linkage score tests, it can be extended to handle non-normal traits using higher-moment transformation of phenotypes (Chen *et al.* 2005), in an analogous way to the higher moment based linkage scores (3.2.2).

Although the LMPU statistic is derived under the "conditional on genotype" likelihood, it is still optimal (LMPU) under the joint likelihood model or under selected sampling (this is proved in section 2.4), as long as it is standardized by null variance computed under the appropriate model-joint model or "conditional on trait" or "conditional on IBD."

To combine data from independent pedigrees, the individual pedigree scores cannot be added. The first part of the expression (4.2.16) being the second derivative is additive over pedigrees, but the second part is not. Thus, in general $l''_{Y|g_m}(0)$ and $[l'_{Y|g_m}(0)]^2$ need to be computed separately and added to get the LMPU statistics. However, in most cases, the departure from additivity will be small, as the cross product terms in between pedigrees for $[l'_{Y|g_m}(0)]^2$ are uncorrelated.

In spite of these advantages, this statistic may not be very useful in practice. Preliminary simulations indicated that this statistic is considerably less powerful than an FBAT type statistic based on $[l'_{Y|g_m}(0)]^2$, and was not pursued further. Limited simulations also indicated that for the normal distribution, LMPU scores for most problems are highly suboptimal for most non-local alternatives. For example, the LMPU test for the mean ($H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$) is based on $\sum X_i^2 - \sigma^2$, which is known to be a poor test compared to $|\overline{X}|$. The power approaches optimality for alternatives extremely close to the null but drops sharply as the alternative becomes non-local. This may not be true for other families of distributions; hence, properties of the LMPU test for other distributions would need further investigation.

### 4.2.8 Parameter Dependence

In this section, we discuss possible approaches to extend the FBAT statistic to make it free of the population trait parameters, for the simple case of a trio data set. The form of the FBAT statistic in this case is

$$\frac{\sum_{trio:i}(Y_i - \mu_i)[g_i - E(g_i|S_i)]}{\sqrt{\sum_i (Y_i - \mu_i)^2\ Var(g_i \mid S_i, Y_i)}}\ .$$

Ewens *et al.* (in press), have shown that choosing the offset $\mu_i$ as the sample mean essentially reduces the FBAT statistic to a test of "slope" of the regression of $W_i = g_i - E(g_i \mid S_i)$ on $Y_i$. They noted that such a statistic is completely different from the TDT statistic which measures the "intercept" of the same regression. In particular, they showed that for a trio data set, the TDT statistic is exactly equal to $\frac{\overline{W}}{SD(\overline{W})}$. The dependence of $W$ on $Y$ can be measured either using the intercept or the slope (as under the null, both are expected to be zero). The original TDT for binary traits can only measure the intercept as $Y$ is constant (only affected offspring are sampled). On the other hand, for a random population sample, the intercept test would not have any information and a "slope test" (e.g., FBAT with an offset of sample mean) would be optimal. The score test provides an optimally weighted linear combination of these two tests, with data dependent (self-adjusting) weights, as shown below for trios.

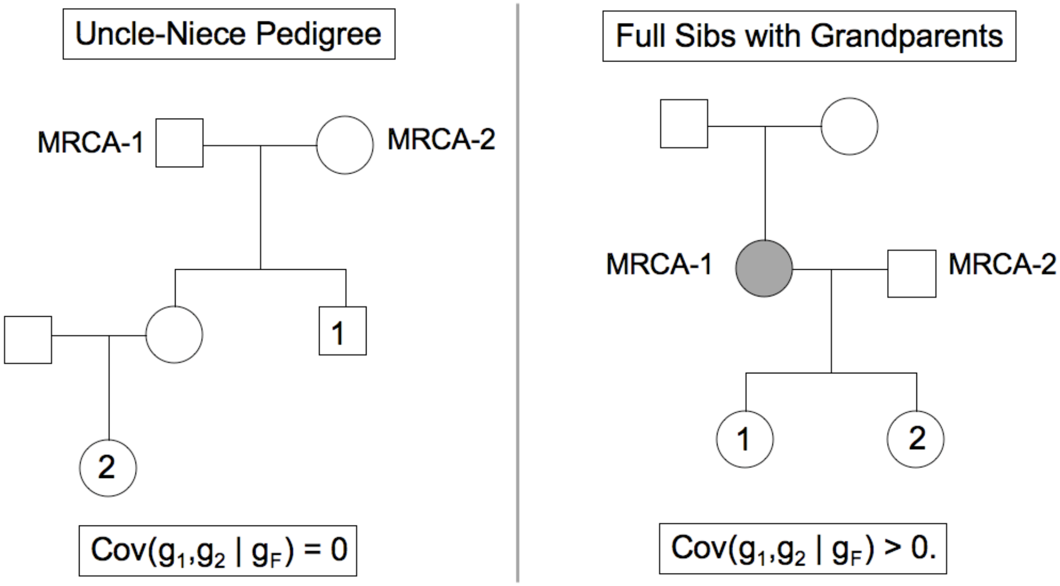

Figure 4.1: **MRCA Founders**

Covariance conditional on founder genotypes. In the first pedigree both the MRCAs of individuals 1 and 2 are founders, whereas in the second pedigree one MRCA (shaded) is a non-founder resulting in a positive covariance.

$$\begin{aligned}
\text{Score Test} &= \frac{\sum_i \{(Y_i - \mu_Y)\ W_i\}}{\sqrt{\sum_i (Y_i - \mu_Y)^2\ \frac{1}{n}\sum_i W_i^2}} \\
&= w_1 .\ \frac{\sum_i \{(Y_i - \overline{Y})\ W_i\}}{\sum_i (Y_i - \overline{Y})^2\ \frac{1}{n}\sum_i W_i^2} + w_2 .\ \frac{\overline{W}}{\sqrt{\frac{1}{n}\sum_i W_i^2}} \\
&= w_1 \text{ Slope Test} + w_2 \text{ Mean Test },
\end{aligned}$$

where the weights $w_1$ and $w_2$ are given by

$$w_1 = \sqrt{\frac{\sum_i (Y_i - \overline{Y})^2}{\sum_i (Y_i - \mu_Y)^2}} \quad \text{and, } w_2 = \frac{\sum_i n(\overline{Y} - \mu_Y)}{\sqrt{\sum_i (Y_i - \mu_Y)^2}},$$

where $w_1^2 + w_2^2 = 1$. Note that under population sampling $\overline{Y} \approx \mu_Y$, so the score reduces to the slope test ($w_2 = 0$). On the other hand, when the offspring have extreme phenotypes (e.g., affected proband sampling) $Y_i \approx \overline{Y}$, so the score test reduces to the intercept test or TDT ($w_2 = 1$). This trade off between slope and intercept tests is analogous to the relationship between "correlation-based" and "IBD-sharing based" tests for linkage (Forrest and Feingold 2000; Szatkiewicz *et al.* 2003; T.Cuenco *et al.* 2003). As in this case, score statistics for linkage provide an optimally weighted linear combination (Szatkiewicz and Feingold 2004), when the population trait parameters are known. When $\mu_Y$ is unknown or cannot be estimated, the choice of the offset becomes important. For discordant sampling schemes, the slope tests ($\mu_i = \overline{Y}$) would tend be powerful, whereas for extremely concordant sampling schemes, conducting simple TDT like tests (ignoring the quantitative trait) may be better.

It is also possible to construct intercept tests that take the observed phenotypes into account. The test based on $\overline{W}$, is strictly speaking, a test of the mean and is equivalent to a test of the intercept only under the assumption that the slope of the regression is zero. An intercept test that takes the slope into account could be obtained by plugging in a slope estimate under the null (i.e. zero intercept) into the regression model. The form of such an intercept test for trios would be

$$\frac{n\left[\overline{W} - \hat{\beta}\ \overline{Y}\right]}{\sqrt{\widehat{Var}(\overline{W}) - \overline{Y}^2\widehat{Var}(\hat{\beta})}} = \frac{n\left[\overline{W} - \hat{\beta}\ \overline{Y}\right]}{\sqrt{\frac{\hat{\sigma}_e^2}{n}\left[\frac{\sum_i (Y_i - \overline{Y})^2}{\sum_i Y_i^2}\right]}} \quad \text{where, } \hat{\beta} = \frac{\sum_i W_i\ Y_i}{\sum Y_i^2} \quad \text{and } \hat{\sigma}_e^2 = \frac{\sum_i W_i - \hat{\beta}\ Y_i}{n-1}.$$

Note that, like the the mean test, the intercept test is free of the trait mean $\mu_Y$. This gives a possible natural extension of the TDT (an intercept test) for quantitative traits.

It is not obvious whether a slope test or an intercept test would have more power to detect association. The score test provides a compromise between the two, but it depends on the trait mean $\mu_Y$, which is often difficult to estimate reliably in selected sample settings. Another possible compromise would be to conduct a (regression-based) two degree of freedom test for both the slope and the intercept. The higher degrees of freedom would tend to reduce the power, particularly when either the intercept or the slope is non-informative (such as population sampling or affected proband sampling). However, the 2df test would have the advantage of being free of the trait mean. For example, in the case of trios, a 2df score test (based on a normal pseudo-likelihood of $[W|Y]$) for the two parameter regression of $W_i$ on $Y_i$ is given by

$$w \left[ \frac{\{\sum_i W_i(Y_i - \overline{Y})\}^2}{(Y_i - \overline{Y})^2 \, \frac{1}{n} \sum_i W_i^2} \right] + \left[ \frac{n^2 \, \overline{W}^2}{\sum_i W_i^2} \right] = w \text{ (Slope Test)}^2 + \text{(Mean Test)}^2, \qquad (4.2.18)$$

where $w = \frac{\sum_i (W_i - \overline{W})^2}{\sum_i W_i^2}$. Note that this statistic has a $\chi^2_2$ distribution asymptotically under the null. $w$ is a decreasing function of the mean of $W$ and an increasing function of its variance. Thus, the 2df statistic puts higher weight on a mean test, when the mean is high, but downweights it when the variability increases. This is in contrast to the score test which weights the tests based on the distortion of the phenotype $(\overline{Y} - \mu_Y)$ due to ascertainment.

Thus intercept-based tests and 2df tests provide ways to construct family based tests of association that are free of the population trait parameters. The mean test can be easily extended for bigger pedigrees as $n \, \overline{W}' Cov(W)^{-1} \overline{W}$, where $Cov(W)$ can be estimated using the formulas derived in section 4.2.5. Similarly, the 2df test proposed here can also be extended to bigger pedigrees using equation (4.2.18). Each of the three parts of the 2df test, namely the slope test statistic, the mean test statistic and the weight $w$ would involve $Cov(W)$. Note that these extensions are free of the trait mean as well as the dispersion matrix, as they model the distribution $[W \mid Y]$, for which the correlation structure is known. The Gaussian pseudo-likelihood assumption used for the 2df test may not be a good approximation for this discrete distribution. On the other hand, the intercept test is completely non-parametric and does not require the Gaussian assumption.

### 4.2.9 Simulation

We conducted a simulation study with nuclear families to compare the score statistics proposed in section 4.2.1 with some standard approaches, in terms of power to detect departures form the type-1 null hypothesis "No linkage and no association." The statistics compared were $SCORE.NS$,

$SCORE.AS$, $SCORE.FP$, $SCORE.FPG$, $QTDT$ (Abecasis *et al.* 2000) and $QTDT - FP$ ($QTDT$ with founder phenotypes regressed out as covariates). $QTDT$ and $QTDT - FP$ were both computed using the software "qtdt" with the command line options "-1 –p-values -wea" and "-1 –p-values -wea –cp" respectively. We implemented the other four statistics in the "R" programming language (R Development Core Team 2008). We did not use the statistics implemented in the "FBAT" software, as the statistic "SCORE.AS" is essentially equivalent to the standard FBAT statistic for quantitative traits. In fact, it is expected to be strictly superior to the FBAT, as it uses the true family correlation structure, while the FBAT assumes "no environmental correlation." Similarly, the $parenTDT$ approach (Purcell *et al.* 2005) could not be included in our simulations as it has not been implemented in publicly available software for quantitative traits. However, our simulation results for the score statistics were roughly consistent with those of Purcell *et al.* (2005) for the corresponding likelihood ratio tests. Below we give an outline of the simulation scheme.

#### 4.2.9.1 Model Parameters

We simulated a biallelic trait locus "t" with alleles "D" and "d"(with allele frequencies $p_t$, $q_t$) and a biallelic marker locus "m" with alleles "A" and "a" (with allele frequencies $p_m, q_m$). The recombination distance between the two loci ($\theta$) and linkage disequilibrium between the "D" and "A" alleles ($\Delta$) were fixed at different values corresponding to the four different hypothesis, as shown below

| | No Association | Association |
|---|---|---|
| No Linkage | $H_{00} : \theta = 0.5,\ \Delta = 0$ | $H_{0A} : \theta = 0.5,\ \Delta = 0.05$ |
| Linkage | $H_{L0} : \theta = 0.01,\ \Delta = 0$ | $H_{LA} : \theta = 0.01,\ \Delta = 0.05$ |

The phenotypes were generated using the trait locus model $Y = m + a\ g_t + e$, with $m = 4$ and $a = 3$. Thus, the phenotypic means for the three genotype classes were $E(Y \mid DD) = m + 2a = 10$, $E(Y \mid Dd) = m + a = 7$ and $E(Y \mid dd) = m = 4$ (i.e. dominance=0). The environmental variance ($\sigma_e^2$), the parent-sibling, the sib-sib and the parental environmental correlations ($r_{ps}$, $r_{ss}$ and $r_{fm}$) were fixed at 9, 0.25 and 0.25 and 0.1 respectively. The trait and marker allele frequencies were varied according to "stratification scheme" as follows.

1. **No Stratification:** Two strata with same trait allele frequency $p_t = 0.2$ but different marker allele frequencies $p_m = (0.4, 0.6)$. All the founders in each family come from the same stratum.

Equal number of families from each stratum. Note that, although there are two marker allele frequency "strata," this scheme is equivalent to a scheme with no strata, as the phenotype distribution is identical in the two strata.

2. **Between Family Stratification:** 2 strata with disease allele frequencies $p_t = (0.1, 0.3)$ and marker allele frequencies $p_m = (0.4, 0.6)$. All founders in each family come from the same stratum. Equal number of families from each stratum.
3. **Admixture:** Two ancestral populations with disease allele frequencies $p_t = (0.1, 0.3)$ and marker allele frequencies $q_t = (0.4, 0.6)$. Each founder in the sample is randomly assigned to one of the ancestral populations with equal probability.

For each combination of the above parameters and stratification scheme, families were simulated using the following design.

**4.2.9.2 Simulation Design** For simulating genotypes, each parent was assigned a stratum membership according to one of the three ascertainment schemes above. For each parent, haplotypes were generated independently using haplotype frequencies for that stratum determined by $\Delta$ ($P(AD) = p_t p_m + \Delta$, etc) and then haplotypes were dropped in the families according to Mendelian transmission rules using a recombination frequency of $\theta$ for each meiosis. Thus a vector of trait genotypes $g_{t_i}$ and marker genotypes $g_{m_i}$ were simulated for the $i^{th}$ family. For each family, the environment was simulated as multivariate normal $e \sim N(0, \Sigma_e)$ with constant correlations (within and across strata) as defined before. The simulated genotypes and environments were then combined as $Y = 4 + 3\ g_t + e$ to obtain to obtain the phenotypes.

For all the three stratification schemes, 200 nuclear families were simulated, each having three offspring. All the simulated families were ascertained (i.e population sampling). It is important to note that, in our simulation model, $\Delta$ refers to the LD within each stratum, and not to the overall LD in the population. Thus, $\Delta$ does not capture spurious LD due to population stratification. It is positive only in the presence of true genetic association.

**4.2.9.3 Nuisance Parameters** The nuisance parameters $\mu_Y$, $\Sigma_Y$ are required for computing all the score statistics, and the marker allele frequency $p_m$ is required for computing $SCORE.NS$. The phenotype mean $\mu_Y$ and covariance matrix $\Sigma_Y$ were empirically estimated from all observed families. The marker allele frequency $p_m$ was estimated from from all the founders in the sample.

This would not be possible for selected samples, as the estimate of $p_m$ in that case would be biased causing $SCORE.NS$ to have incorrect type I error even if there is no stratification in the sample. Similarly, biased estimates of $\mu_Y$ and $\Sigma_Y$ would reduce the power of all the score tests, but type I error would be unaffected as all the statistics condition on the phenotypes. Parameter misspecification was not addressed in our limited simulations.

**4.2.9.4 Power Estimation** 1000 replicates were simulated to estimate the power (or type I error) under all four hypotheses. For the four score statistics, we considered absolute values of the standardized statistics (defined in section 4.2.1) and asymptotic two sided p-values were obtained using the normal distribution. For $QTDT$ and $QTDT - FP$, we used the asymptotic thresholds (i.e., without the "-m" option) based on $\chi^2$ distribution. As we used population samples for our simulations, the asymptotic thresholds are expected to be correct. Power was computed for each statistic as the proportion of replicates with p-values less than the nominal threshold of 0.05.

## 4.3 RESULTS

The type-I error and power results under each of the four hypotheses are summarized in Table 4.3 for all the stratification schemes. Figures 4.2, 4.3 and 4.4 show comparison of power across statistics for the stratification schemes (1), (2) and (3) respectively. Below, we outline the results under the three ascertainment schemes. Note that for each statistic we have used the words "power" and "type I" error according to the type of hypothesis it is designed to detect. For example, the rejection rate under $H_{0A}$ is referred to as power for $SCORE.FPG$ or $SCORE.NS$ while the same rejection rate gives type I error for the other four statistics.

**4.3.0.5 No Stratification** The power results under no-stratification and population sampling are compared in Figure 4.2. All the score statistics including $SCORE.NS$ have correct type-I error under $H_{00}$. As they are based on the type I null hypothesis they have slightly elevated type-I errors under $H_{L0}$. The QTDT statistics are slightly conservative for these hypotheses. Under $H_{0A}$, all the statistics except $SCORE.NS$ and $SCORE.FPG$ show close to nominal type-I error, as these

Table 4.2: **Type I Error and Power Results**

| **No Stratification** | | | | | | |
|---|---|---|---|---|---|---|
| | **SCORE.NS** | **SCORE.AS** | **SCORE.FP** | **SCORE.FPG** | **QTDT** | **QTDT-FP** |
| H00 | 0.042 | 0.055 | 0.051 | 0.05 | 0.015 | 0.035 |
| HL0 | 0.06 | 0.062 | 0.071 | 0.067 | 0.033 | 0.045 |
| H0A | 0.458 | 0.048 | 0.041 | 0.184 | 0.018 | 0.027 |
| HLA | 0.944 | 0.615 | 0.657 | 0.829 | 0.557 | 0.622 |
| **Between Family Stratification** | | | | | | |
| H00 | 0.132 | 0.044 | 0.039 | 0.06 | 0.021 | 0.035 |
| HL0 | 0.121 | 0.059 | 0.061 | 0.058 | 0.032 | 0.049 |
| H0A | 0.744 | 0.053 | 0.057 | 0.18 | 0.022 | 0.039 |
| HLA | 0.987 | 0.61 | 0.657 | 0.846 | 0.53 | 0.609 |
| **Admixture** | | | | | | |
| H00 | 0.133 | 0.034 | 0.038 | 0.052 | 0.028 | 0.055 |
| HL0 | 0.131 | 0.066 | 0.072 | 0.058 | 0.028 | 0.055 |
| H0A | 0.772 | 0.05 | 0.047 | 0.195 | 0.136 | 0.153 |
| HLA | 0.991 | 0.626 | 0.662 | 0.837 | 0.551 | 0.617 |

Type I Error and Power Results for all the statistics under the 3 ascertainment schemes.

statistics can detect association only in presence of linkage. $SCORE.NS$ has the highest power to detect $H_{0A}$, while $SCORE.FPG$ is also reasonably powerful. Under $H_{LA}$, $SCORE.NS$ followed by $SCORE.FPG$ are most powerful statistics. $SCORE.AS$, $SCORE.FP$ and the QTDT statistics have similar power. The QTDT statistics, in spite of being likelihood ratio tests and incorporating IBD information are slightly less powerful than the two score statistics. This is probably due to the inconsistency in QTDT model as discussed in section 2.2.1.4. The statistics incorporating founder phenotypes provide noticeable improvement in power. $SCORE.NS$ is the best overall statistic in this case.

**4.3.0.6 Between Family Stratification** Power results for between family stratification are shown in Figure 4.3. All the statistics except $SCORE.NS$ have correct type-I error under $H_{00}$. $SCORE.NS$ has inflated type-I error as it detects spurious associations due to allele frequency differences across strata. $SCORE.NS$ also has inflated type-I error under $H_{L0}$, while the other score statistics have slightly elevated type I errors as before. $QTDT$ and $QTDT-FP$ have slightly conservative type I errors. $SCORE.FPG$ detects the alternative $H_{0A}$, as it derives power from

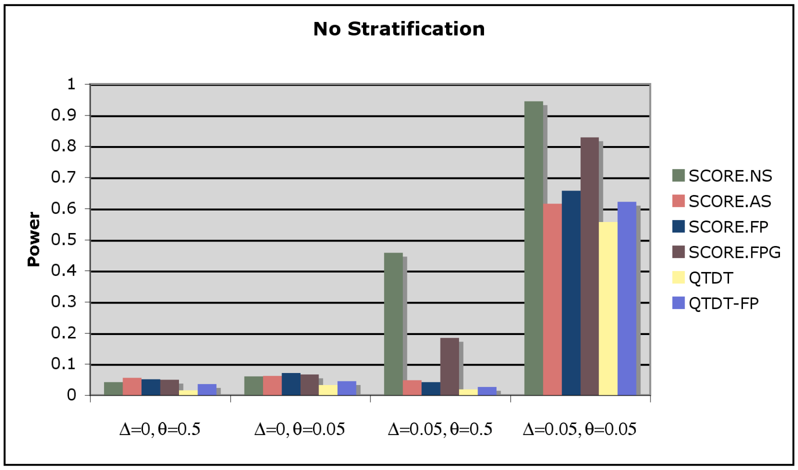

Figure 4.2: **Power comparison under no stratification.**

Power comparison for 200 sibships (of size 3) from a single stratum under the four different hypotheses.

founder genotype-phenotype correlation, which can be detected irrespective of linkage. Under $H_{LA}$, as before, $SCORE.AS$ and $SCORE.FP$ have similar power to the QTDT statistics, with founder phenotypes giving slight improvement in power. $SCORE.NS$ gives the highest power followed by $SCORE.FPG$ which provides considerable power improvement while maintaining correct type I error rates. Thus, $SCORE.FPG$ is the best overall statistic under between family stratification.

**4.3.0.7 Admixture** Power results under admixture are summarized in Figure 4.4. In this case, all the statistics except $SCORE.NS$ give close to nominal type I error rate under $H_{00}$ and $H_{L0}$. $SCORE.FPG$ was expected to have incorrect type I error in this case. But our sample size may have been too small to detect the elevation of type I error. Under our design, the effective sample size for $SCORE.FPG$ to detect spurious association is half the total number of founders, as half of the founder pairs come from the same ancestral population. $SCORE.NS$ has the highest power to detect $H_{0A}$, followed by $SCORE.FPG$. The QTDT statistics show incorrect type I error under this hypothesis, as they are designed to protect against between family stratification but not against admixture. Under $H_{LA}$, as before $SCORE.NS$ and $FPG$ have highest power while the other four statistics have similar power. $SCORE.AS$ and $SCORE.FP$ are slightly more powerful than the

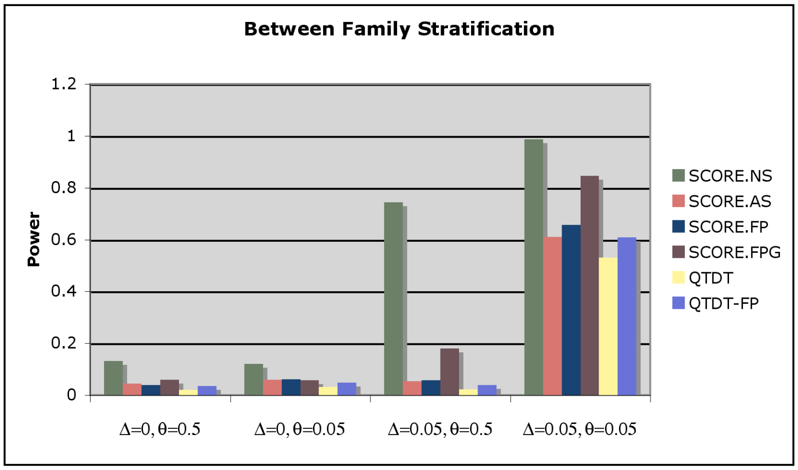

Figure 4.3: **Power comparison under between family stratification.**

Power comparison for 100 sibships (of size 3) each from two strata under the four different hypotheses. Parents in each family come from the same stratum.

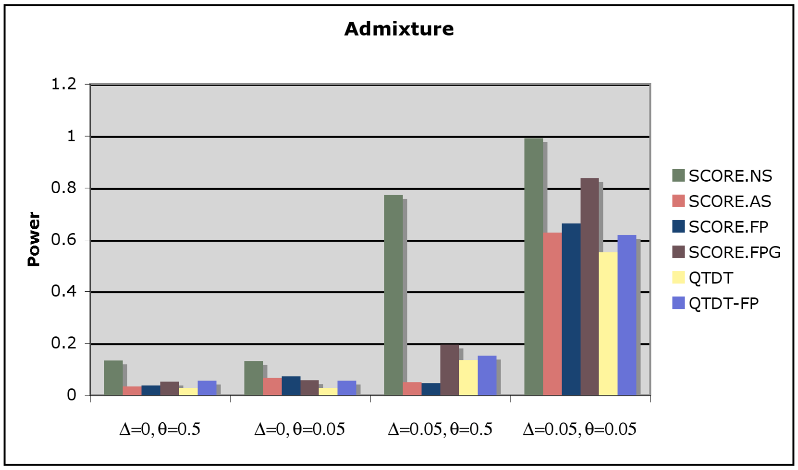

Figure 4.4: **Power comparison under admixture.**

Power comparison for 200 sibships (of size 3) from an admixed population under the four different hypotheses. Parents in each family belong to either of the two strata with probability 1/2.

QTDT statistics. Founder phenotypes provide modest power improvement. It is not clear whether $SCORE.FPG$ should be preferred in this case. Further simulations may be required to understand the effect of admixture on $SCORE.FPG$. For the hypothesis of $H_{LA}$ , $SCORE.FP$ has the highest power with theoretically correct type I error rate.

## 4.4 DISCUSSION

In this chapter, we proposed novel score statistics for association mapping of quantitative traits and compared some of them with two of the standard approaches for family based association mapping of quantitative traits, QTDT and FBAT. Family-based tests of association are designed to protect against population substructure, unlike population-based association tests such as ANOVA-based methods. However, in gaining that robustness, they suffer considerable loss of power to detect association compared to the population-based tests. This power loss can be mostly attributed to the fact that these tests condition on founder genotype information and ignore founder phenotype information. Both of these factors imply that family-based tests can only detect cotransmission of alleles due to linkage as they do not make use of genotype-phenotype correlation among founders which contains most of the LD information.

Ignoring founder phenotypes is often justified, as there may be age or generation specific differences in the phenotypic distribution. However, in many situations this may not be a concern, particularly for founders who marry into a multi-generational pedigree. When founder phenotypes are available and it is reasonable to use them, we proposed two extensions of the FBAT-type test ($SCORE.AS$) that attempt to retrieve the LD information in founders partially. The first of these, $SCORE.FP$, just uses the founder phenotypes, while not making any assumptions about the nature of stratification. As such, it is protected against arbitrary stratification, while providing modest improvement in power due to environmental correlation between founders and non-founders. The second, $SCORE.FPG$, makes the additional assumption that there is only between family stratification (an assumption that is also made by QTDT), and derives information from founder genotype-phenotype correlation (within families) to significantly improve the power to detect association (irrespective of linkage). This statistic is protected against allele frequency differences across strata, as long as all founders in each family come from the same stratum. However, using the genotype-phenotype correlation also means that this statistic has power to detect markers on a different chromosome or those placed far from the QTL on the same chromosome, that are still in LD with the trait. The proposed unconditional score test $SCORE.NS$ as well as population-based methods such as case-control studies also detect markers that are in LD but are not linked to the trait (even when there is no stratification). Such markers are generally undesirable to most investigators, but they are extremely rare. Such associations can be a result of a recent mutation/founder effect, joint selection or due to some unknown epigenetic factors. Arguably, the primary objective of family based association tests is to protect against spurious associations due to systematic allele frequency differences across strata. By measuring only "association in the presence of linkage," they rule out both kinds of spurious association those due to stratification and also those due to unknown genetic factors. However, in doing so, these statistics lose a part of association information that is independent of the linkage information. On the other hand, by relaxing the requirement to protect against these rare markers, considerable power is gained (as seen in our simulations), for detecting markers of interest, i.e., those that are linked and in LD with the trait.

Assuming, that the "no within family stratification" assumption remains valid, $SCORE.FPG$ thus provides a way to significantly improve power of family-based association, while protecting against spurious associations caused by the differing allele frequencies (and disease prevalence) across strata. When it is important to strictly guard against markers that are in LD but are not linked, one possible strategy may be to apply $SCORE.FPG$ for an initial genome scan, and then screen the resulting marker list by $SCORE.AS$ or $SCORE.FP$ to identify and drop the unlinked markers. This strategy preserves higher power at the first stage, while ensuring that the markers retained after the second stage are both linked and associated.

Thus, our simulation results indicated that when founder phenotypes can be used, $SCORE.NS$ and $SCORE.FPG$ should be the preferred statistics provided the assumptions of "no stratification"or "between family stratification only" respectively, are reasonable. $SCORE.FPG$ provides higher power than $QTDT$, which also protects only against between family stratification. When this assumption can not be made, modest improvement can be gained by using $SCORE.FP$; $QTDT - FP$ can not be used in that circumstance as it has inflated type I error. Under "admixture", $SCORE.FPG$ did not show detectable type I error inflation in our simulations. Limited experiments with strongly disassortative mating (parents always come from separate strata), showed significant type I error inflation. This suggests that larger sample sizes (which would increase the number of discordant couples) may lead to incorrect type I errors. An increase in the number of founders in each family may also inflate type I error. Further simulations would be required to judge the extent of the impact of admixture on $SCORE.FPG$.

There were many limitations in our simulation study. We considered population sampling and only a few stratification schemes. Selected sampling can have a significant impact on the performance of the statistics as the nuisance parameter estimates or input values may be biased. Our stratification schemes were based on varying the trait allele frequency parameter. Stratification may be caused by change in trait prevalence with any of the other parameters $m$ (environmental mean), $a$ (relative risk) or $\sigma_e^2$ (environmental variance). It is possible that the statistics would compare differently when strata are generated by changing those parameters. Similarly, a better understanding of the performance of the statistics under admixture would require a realistic simulation from an admixture model using multiple loci. Also, we could not compare the performance of our statistics with the "parenTDT" approach of Purcell *et al.* (2005) owing to lack of software implementation. We expect that the score statistics would be roughly equivalent in terms of power to some of the LRT statistics proposed in Purcell *et al.* (2005) and more robust to ascertained sampling, but a detailed simulation study would be required to confirm this.

In many practical scenarios, if the assumption of strong assortative mating holds, it is likely that stratum membership would be known for each family. In such cases, $SCORE.FPG$ can be improved even further by conditioning on the "stratum founder genotype mean" instead of restricting to each family. Such a statistic would measure total association "within strata" and would be almost as powerful as an unconditional population-based test. However, $SCORE.FPG$ would not work if there is admixture, or even in the simple case where there are two strata but the condition of strong assortative mating does not hold. However, it should be noted that different scores can be used for different families in a data set. For example, if a subset of the families are from an admixed population, then $SCORE.FP$ can be used for those families and $SCORE.FPG$ for the other families; variances of all the families are added together to obtain the denominator of the standardized score statistic.

In this chapter we also derived formulas for conditional moments required to compute the proposed score statistics. When founder genotypes are available, these formulas can provide quick estimates of mean and variances of the score statistics for bigger pedigrees without resorting to gene dropping. When the founder genotypes are completely missing, the R-L algorithm or one of its modifications proposed in section 4.2.6 can be used to derive the distribution (and the moments) of the numerators conditional on the minimal sufficient statistic for the missing genotypes. This algorithm is quite intensive computationally, particularly for the type-2 null hypothesis. Although imputation is generally not recommended, it is becoming increasingly popular with dense genomic scans and LD maps being available. Imputing missing founder genotypes based on the available genotypes in the pedigree at one or multiple loci can help in fast calculation of the score statistics using the formulas derived here. Also, the proposed modification of the R-L algorithm for $SCORE.FPG$ is highly computationally intensive. When some of the founders are available, $SCORE.FPG$ can be constructed by conditioning on the mean genotype of the available founders, thus retrieving the genotype-phenotype correlation in that subset of founders.

Although the type-1 score statistics assume "no linkage" for computing the null distribution, our simulations indicated that the type I error inflation under $H_{L0}$ is not appreciable. Thus it may be adequate to use the type-1 null hypothesis for the proposed score tests as well as FBAT for most purposes. This is particularly relevant in presence of missing founders, for which the R-L algorithm increases the computational burden substantially. Another shortcoming of the R-L algorithm for type-2 statistics as used by FBAT (Rabinowitz and Laird 2000) is that it leads to considerable reduction of power due to loss of nuclear families not informative for IBD. This problem stems from the fact that when IBD information is not perfect, the R-L algorithm conditions on all possible IBD configurations without weighting them based on posterior probabilities of the configuration. Due to this power loss, FBAT and PBAT currently use an empirical variance estimate. The power reduction problem may be remedied by using the modified version of the R-L algorithm for type-2 null hypothesis proposed in section 4.2.6. Simulation studies with different kinds of missing data and IBD configurations would be required to verify this claim and also compare its performance to the empirical variance approach.

Finally, we studied the relationship between score statistics and slope and intercept tests as defined by Ewens *et al.* (in press) and used this distinction to suggest an intercept test for trios that parallels the TDT while incorporating the quantitative traits. We also discussed the possibility of deriving joint 2 d.f slope and intercept tests that would be free of trait nuisance parameters.

Decision trees for choice of family-based association mapping statistics are shown in Figures 4.5 and 4.6 for pedigrees with and without founder phenotype information respectively. Note that we have included only the score statistics and the QTDT-type statistics, as these can handle general pedigrees. For specific types of data sets such as trios, sibpairs and nuclear families there

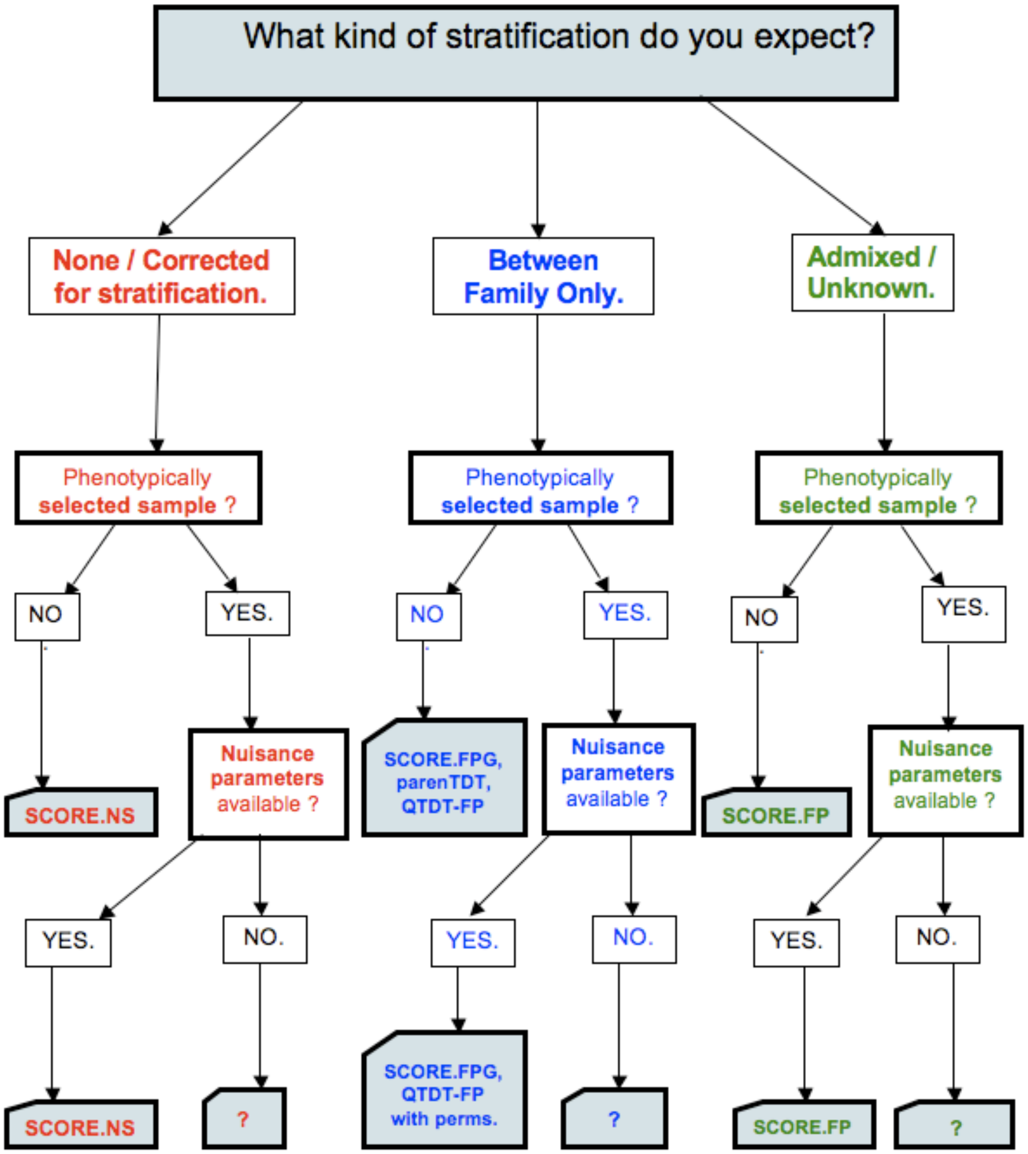
What kind of stratification do you expect?
None / Corrected for stratification.
Between Family Only.
Admixed / Unknown.
Phenotypically selected sample ?
Phenotypically selected sample ?
Phenotypically selected sample ?
NO
YES.
NO
YES.
NO
YES.
SCORE.NS
Nuisance parameters available ?
SCORE.FPG, parenTDT, QTDT-FP
Nuisance parameters available ?
SCORE.FP
Nuisance parameters available ?
YES.
NO.
YES.
NO.
YES.
NO.
SCORE.NS
?
SCORE.FPG, QTDT-FP with perms.
?
SCORE.FP
?

Figure 4.5: **Choice of Statistics for Family Based Association Mapping when Parental Phenotypes are Known.**

are other specialized methods that we did not consider in our flowcharts. These flowcharts give tentative guidelines based on expected theoretical behavior of these statistics. Further simulations would be required to arrive at more precise recommendations. For selected samples with nuisance parameters unavailable, none of the methods considered here are ideal. Hence the corresponding boxes in the flowchart have been marked with a "?". All the score statistics discussed here as well as $QTDT$ (with permutations) would have robust type I error with arbitrary parameter estimates. However, power would be sub-optimal and may be highly sensitive to choice of some of these parameters. FBAT ignores the within family residual correlations and requires only the offset parameters, making it an attractive statistic for these situations. However, it can be argued that $FBAT$ essentially misspecifies the family correlation matrix as the identity matrix, which makes it a sub-optimal statistic. Also, it should be noted that in most cases, choice of "sample mean" as the offset parameter is likely to be a bad choice in selected sampling scenarios (see section 4.2.8). A detailed parameter sensitivity study would be required to assess whether FBAT with a reasonable offset choice is an efficient statistic for selected sampling scenarios when parameter estimates are unknown or unreliable.

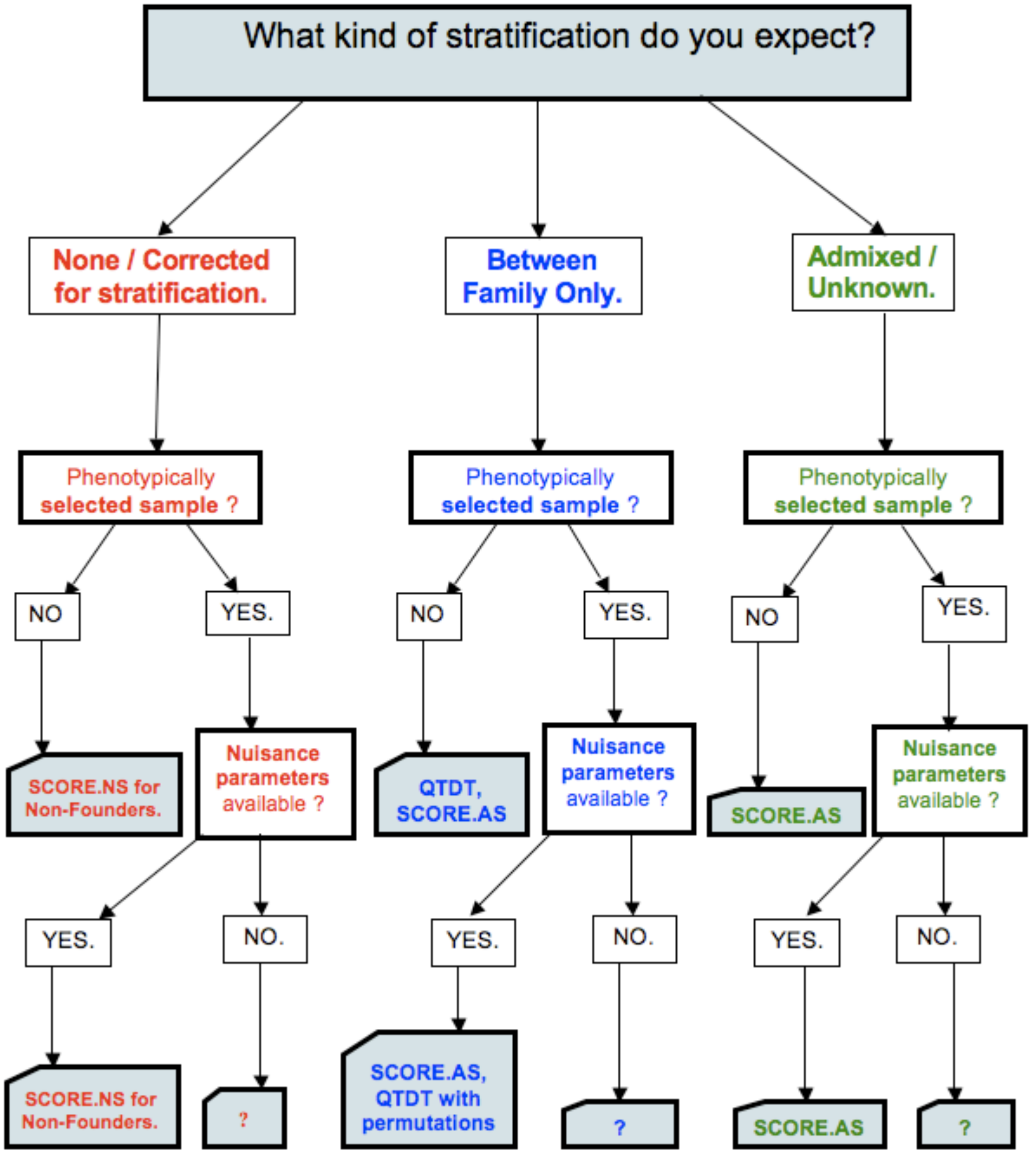
What kind of stratification do you expect?
None / Corrected for stratification.
Between Family Only.
Admixed / Unknown.
Phenotypically selected sample ?
Phenotypically selected sample ?
Phenotypically selected sample ?
NO
YES.
NO
YES.
NO
YES.
SCORE.NS for Non-Founders.
Nuisance parameters available ?
QTDT, SCORE.AS
Nuisance parameters available ?
SCORE.AS
Nuisance parameters available ?
YES.
NO.
YES.
NO.
YES.
NO.
SCORE.NS for Non-Founders.
?
SCORE.AS, QTDT with permutations
?
SCORE.AS
?

Figure 4.6: **Choice of Statistics for Family Based Association Mapping when Parental Phenotype are Unknown.**

# 5.0 DISCUSSION AND FUTURE WORK

In this dissertation, we discussed a number of different existing score statistics and proposed some novel score statistics for linkage and association mapping mapping of quantitative traits. Most of these score statistics are based on one of the implicit models discussed in Chapter 2. The implicit models intuitively capture the parameters of interest, namely linkage and association, using the assumption of a normal distribution for the phenotype which allows mathematical and computational simplicity. In reality, quantitative traits may be skewed or multimodal. There has not been significant progress in the literature for mapping of non-normal traits. When dealing with highly non-Gaussian traits, the likelihood ratio as well as score tests based on a normal assumptions should be used with care. Score statistics should be preferred when the data cannot be transformed to approximate normality, as they guard against false positives. But even when the normality assumption for the phenotype holds, the implicit models may not adequately capture the linkage or association information and suffer loss of power. Methods for linkage mapping of binary traits often model $\theta$ explicitly up to a few nuisance parameters; this is usually difficult for quantitative traits because of the continuity of the trait distribution.

Even under a normal model, the nuisance parameters include relative pair correlations which grow in number with pedigree size. As seen in chapter 2, the variance components model or the proposed implicit model possibly require some unrealistic assumptions such as uniform linearity of $E(\Pi_t \mid \Pi_m)$ and homoscedasticity of $g_t \mid g_m$. For a given normal distribution (of the phenotype), the validity of the implicit models can be checked by obtaining the exact (explicit) distribution of $[g_t \mid g_m, \Pi_m]$ computationally using Mendelian transmission rules. Explicit models that maximize over nuisance parameters may provide considerable improvement in power to detect linkage or association and also provide direct estimates of the parameters. Often, because of the confounding of multiple nuisance parameters, it is difficult to interpret estimates under implicit models. For example, the estimate for "additive genetic variance" under a variance component model for linkage is a valid estimate of that parameter only when there is perfect linkage.

Our proposed implicit model 2.2.8, combines the implicit linkage and association parameters into a single model similar to the VC model used by QTDT. It was used to derive a score statistic for linkage that allow for presence of LD, which has become increasingly relevant with linkage studies being conducted with high density markers. However this score statistic may be computationally intensive, even for a single marker. Similarly it also allows for an association test that incorporates the extent of linkage of a marker. Such tests (e.g., QTDT) should be more powerful than tests that are derived under models assuming no linkage such as the FBAT and its modifications discussed here. However in our simulations in chapter 4, QTDT was consistently less powerful than FBAT. The power may be improved by using model 2.2.8 which is possibly a more consistent model with fewer assumptions than the QTDT model 2.2.6.

Family based association tests are considerably less powerful than population-based tests such as case-control tests, as they do not attempt to incorporate genotype-phenotype correlation. In chapter 4, we proposed some score tests that improve on the power of family-based tests, when founder phenotypes are available and can be used. An important outstanding issue in family based association tests is handling missing founder genotypes in a computationally tractable way. The algorithms currently used, as well the modifications we proposed, are computationally inefficient. Breaking of pedigrees into independent parts as suggested in section 4.2.6 may provide substantial computational speed up in most cases, but requires further study.

Score statistics for both linkage and association that condition on traits lose some power due to conditioning on a sufficient statistic $Y$ that is not minimal sufficient. The numerator of the scores are invariant up to the nuisance parameters. The denominator is $Var(Score|\mathcal{A})$, where $\mathcal{A}$ is the ascertainment scheme. Ideally the variance should be computed conditional on a minimal sufficient statistic: $Var(Score|T(Y))$, where $T(Y)$ is minimal sufficient for $\mathcal{A}$. For most practical ascertainment schemes it may be difficult to obtain a minimal sufficient statistic. Peng and Siegmund (2006) suggested using probands phenotypes as a sufficient statistic for most "regular" ascertainment schemes. It may be possible to obtain similar sufficient statistics specific for each scheme that may not be minimal sufficient but contain less information than $Y$. However even if such statistics are as simple as "range of Y" or "maximum of Y," it may be difficult to derive the theoretical conditional variances of the scores, or empirical estimates for them. If the sample size is sufficiently large, an empirical variance of the numerator can give a reasonably accurate estimate of this variance. In chapter 3, we found empirical variances to be conservative under our simulations. Further experiments with sample size are required decide whether empirical variance estimates are reasonably correct for practical sample sizes.

We did not consider combined "linkage and association" studies (i.e. linkage AND/OR association) in this dissertation. These may be very relevant for investigators who have family data originally collected for linkage studies, but with high density genotypes for all family members. Often, the disease or trait of interest has not been mapped in detail previously, and there is little or no prior information about possible location of genetic susceptibility variants. In such circumstances, the investigator has no reason to prefer either linkage or association methods. A combined linkage and association study may be preferred to get an idea of the regions of interest if any. It is possible to construct 2 d.f. scores for combined linkage and association (i.e $H_0$ : $\beta = 0,\ v_a = 0$) simply by squaring and adding the standardized scores for linkage and association derived under the same model (e.g., the proposed implicit model), fixing the other parameter at the null value. This is because the linkage and association scores are orthogonal, which can be seen by observing that

$$\begin{aligned} Cov(S_L, S_A) &= E[Cov(f_1(Y)\tilde{g}_m, f_2(Y)\ v\tilde{e}c(\Pi_m) \mid Y)] = f(Y)\ Cov(\tilde{g}_m, v\tilde{e}c(\Pi_m)) \\ &= f(Y)\ Cov(E(\tilde{g}_m \mid \Pi_m), vec(\tilde{\Pi}_m)) \\ &= 0 \end{aligned}$$

where $f_1$, $f_2$ and $f$ are some functions of the phenotype, $S_L$ and $S_A$ are the linkage and association scores. It is not clear, how such a statistic would compare in terms of power to 1 d.f. linkage or association scores. However it avoids making assumptions such as no association or no LD, typically used to a certain extent by linkage and association statistics (as the scores for each parameter are derived fixing the other at the null value). However in the presence of population stratification, the association score would involve $[g - f(g_F)]$ and would not be orthogonal to the linkage score. Hence it may be necessary to use empirically estimated information matrices to combine the scores. An alternative strategy to deal with stratification may be to do an "association"( such as $SCORE.NS$) or "combined linkage and association" scan followed by a linkage scan such as VC or a "linkage AND association" scan (such as FBAT) to prune out the spurious associations.

One of the most important outstanding issues for quantitative traits mapping is dealing with non-normal traits. The higher moment score tests (Chen *et al.* 2005) provide a way to improve power of linkage score statistics by incorporating skewness and kurtosis information. However this introduces two additional nuisance parameters to which the power is highly sensitive. These parameters are very difficult to estimate accurately even from population samples with small sample size. Nevertheless when these parameter estimates are available a-priori, they provide significant power improvement. We attempted to construct an association score that could be extended to use higher moment extension, but it turned out to be grossly under-powered. The semiparametric approach of Diao and Lin (2005) can handle arbitrary distributions of traits under a VC model for linkage. Diao and Lin (2006) proposed semiparametric extensions of the family based association tests QTDT, FBAT and PDT (Monks and Kaplan 2000; Martin *et al.* 2000) that can handle arbitrary trait distributions. However the semiparametric approach cannot handle selected samples as it is based on a likelihood ratio test. Further work is needed in this area to develop score tests that retain high power under a large class of trait distributions.

All the score tests discussed in this dissertation depend on the trait distribution parameters $\mu_Y$ and $\Sigma_Y$. The estimation of these nuisance parameters for selected samples is an important issue for all linkage and association score statistics as well as likelihood ratio tests. Score tests are robust to the specification of these parameters, but suffer considerable loss of power when these are wrongly specified as seen in our sensitivity analysis (chapter 3). Estimation of these parameters using conditional likelihoods are feasible only for very simple ascertainment schemes. In reality ascertainment schemes are generally complex and almost always impossible to specify in terms of quantitative traits, as probands are usually ascertained based on disease status. Considering the importance of nuisance parameter estimation, ideally population-based pilot studies of the phenotype should be conducted prior to linkage or association studies using ascertained samples. When this is not feasible, a portion of the data may be used to estimate nuisance parameters as suggested in section 3.5. This approach provides a way of obtaining nuisance parameters estimates from selected samples (with arbitrary unknown ascertainment scheme) that are designed to optimize the power of the relevant statistic. Simulation studies would be required to assess whether such a strategy provides enough power improvement to balance the power loss due to reduction of sample size.

# APPENDIX A

# DERIVATIVES OF THE LIKELIHOOD FOR THE MEAN MODEL

Here we obtained the derivatives of the likelihood for the mean model 2.2.3, which are required for obtaining the $SCORE.NS$ in section 4.2.1 and the linkage score for the mixture normal model (2.2.5). Here we denote $g_m$ and $\hat{\Pi}_m$ by $g$ and $\Pi$ for clarity. For any likelihood function $L_g(a)$, we have the identities

$$\begin{aligned} l'_g(0) &= \frac{L'_g(0)}{L_g(0)} \\ l''_g(0) &= \frac{L''_g(0)}{L_g(0)} - \frac{[L'_g(0)]^2}{[L_g(0)]^2} \\ \text{and hence } \frac{L''_g(0)}{L_g(0)} &= l''_g(0) + [l'_g(0)]^2. \end{aligned}$$

Next let us consider the "conditional on IBD" likelihood (2.2.5). Note that the identities also hold for the conditional likelihood $L_\Pi(a)$. Directly taking logarithms and differentiating we get

$$\begin{aligned} l'_\Pi(0) &= \sum_g l'_g(0)P(g \mid \Pi) \\ &= E_{g|\Pi}[l'_g(0)] \end{aligned}$$

$$\begin{aligned} l''_\Pi(0) &= \frac{L(0)\sum_g L''_g(0)P(g \mid \Pi) - [\sum_g L'_g(0)P(g \mid \Pi)]^2}{L^2(0)} \\ &= \sum_g [l''_g(0) + l'_g(0)^2]P(g \mid \Pi) - [\sum_g l'_g(0)P(g \mid \Pi)]^2 \\ &= E_{g|\Pi}[l''_g(0)] + Var_{g|\Pi}[l'_g(0)]. \end{aligned}$$

Next, we derive $l'_g(a)$ and $l''_g(a)$ explicitly. The likelihood and log-likelihood functions are

$$\begin{aligned} L_g(\beta) &\propto \frac{\exp\{-\frac{1}{2}(\tilde{Y}-\beta\ \tilde{g})'[\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ (\tilde{Y}-\beta\ \tilde{g})\}}{|\Sigma_{_Y}-\beta^2\ \Sigma_g|^{\frac{1}{2}}} \\ l_g(\beta) &\propto -\frac{1}{2}(\tilde{Y}-\beta\ \tilde{g})'\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ (\tilde{Y}-\beta\ \tilde{g})-\frac{1}{2}\log|\Sigma_{_Y}-\beta^2\ \Sigma_g|. \end{aligned}$$

We will need the following matrix identities (Tang 2000), for invertible matrix $G$ and scalar $x$ :

$$\frac{\partial G^{-1}}{\partial x}=-G^{-1}\frac{\partial G}{\partial x}G^{-1} \text{ and } \frac{\partial \log \mid G \mid}{\partial x}=trace(G^{-1}\frac{\partial G}{\partial x}). \tag{A.0.1}$$

Using these identities and the usual chain rule of derivatives, the first and second derivatives of the log likelihood are obtained to be

$$\begin{aligned} l'_g(\beta) &= \beta\ trace\left([\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \Sigma_g\right)+(\tilde{Y}-\beta\ \tilde{g})'[\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \tilde{g} \\ &\quad -(\tilde{Y}-\beta\ \tilde{g})'\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ (\beta\ \Sigma_g)\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ (\tilde{Y}-\beta\ \tilde{g}) \end{aligned}$$

$$\begin{aligned} l''_g(\beta) &= trace\left([\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \Sigma_g\right)+2\beta^2\ trace\left([\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \Sigma_g\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \Sigma_g\right)-\tilde{g}'\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \tilde{g} \\ &+ 2\beta\ (\tilde{Y}-\beta\ \tilde{g})'\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \Sigma_g\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \tilde{g}+2(\tilde{Y}-\beta\ \tilde{g})'\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ (\beta\ \Sigma_g)\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \tilde{g} \\ &- (\tilde{Y}-\beta\ \tilde{g})'\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \Sigma_g\ [\Sigma_{_Y}-2\beta^2\ \Sigma_g]^{-1}\ (\tilde{Y}-\beta\ \tilde{g})-4\beta^2\ (\tilde{Y}-\beta\ \tilde{g})'\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \Sigma_g \\ &\qquad [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ \Sigma_g\ [\Sigma_{_Y}-\beta^2\ \Sigma_g]^{-1}\ (\tilde{Y}-\beta\ \tilde{g}) \end{aligned}$$

Substituting $\beta = 0$, we get

$$\begin{aligned} l'_g(0) &= \tilde{Y}'\ \ \Sigma_{_Y}^{-1}\ \ \tilde{g} \\ l''_g(0) &= -\tilde{Y}'\Sigma_{_Y}^{-1}\Sigma_g\Sigma_{_Y}^{-1}\tilde{Y}+trace(\Sigma_{_Y}^{-1}\Sigma_g)-\tilde{g}'\Sigma_{_Y}^{-1}\tilde{g}. \end{aligned}$$

# APPENDIX B

## MOMENTS OF THE SCORE STATISTIC

Here we derive the null and alternative means and variances of the score statistic for an extended pedigree. It provides an alternative to the more complicated derivation outlined previously in Tang and Siegmund (2001).

Let $Y$ be the phenotype vector for a pedigree with mean 0 (for simplicity) and variance covariance matrix $\Sigma$. Let $A_\pi$ be the matrix given by

$$(A_\pi)_{ij} = 2(\Pi_{ij} - 2\Phi_{ij}),$$

where $\Pi_{ij}$ and $\Phi_{ij}$ are the estimated IBD and kinship coefficient between the $i^{th}$ and $i^{th}$ individuals of the pedigree. The assumed model is $Y \sim N(0, \Sigma)$ where $\Sigma_\pi = \Sigma + \alpha A_\pi$, with $\alpha = \sigma_a^2/2$, and dominance is assumed to be zero. The score statistic can be written as (Tang and Siegmund 2001),

$$S = -\frac{1}{2}[trace(\Sigma^{-1}A_\pi) - trace(\Sigma^{-1}A_\pi\Sigma^{-1}YY')].$$

It is easy to see that null and alternative means are given by $\mu_0 = 0$ and

$$\mu_\alpha = E[E(S \mid \pi)] = (\alpha/2)E\{trace[(\Sigma^{-1}A_\pi)^2]\}.$$

The variance can be computed as follows.

$$\begin{aligned}
Var_{\alpha}(S \mid \pi) &= (1/4) Var[trace(\Sigma^{-1} A_{\pi} \Sigma^{-1} Y Y')] \\
&= (1/4) Var(Y'\Sigma^{-1} A_{\pi} \Sigma^{-1} Y) && [\because \text{ Trace is commutative}] \\
&= (1/4) Var(Y'CC'A_{\pi}CC'Y) && [\because \Sigma \text{ is positive definite, } \Sigma = B'B \text{ and } \Sigma^{-1} = CC',\ C = B^{-1}]. \\
&= (1/4) Var(Y'CP'D_{\lambda}PC'Y) && [C'A_{\pi}C = P'D_{\lambda}P \text{ using spectral decomposition of } C'A_{\pi}C\ ] \\
&= (1/4) Var(Z'D_{\lambda}Z) && [\text{defining } Z = PC'Y\ ]. \\
&= (1/4) \textstyle\sum_{i=1}^{s} Var(\lambda_i Z_i^2) \\
&= (1/4) \textstyle\sum_{i=1}^{s} \lambda_i^2 .2(1+\alpha\lambda_i)^2 && [\because Z \sim N(0, I+\alpha D_{\lambda}) \text{ i.e., } Z_i\text{'s are independent N}(0,\ 1+\alpha\lambda_i)\ ]. \\
&= (1/4) \textstyle\sum_{i=1}^{s} 2(\lambda_i^2 + 2\alpha\lambda_i^3 + \alpha^2\lambda_i^4) \\
&= (1/2)\{trace[(\Sigma^{-1}A_{\pi})^2] + 2\alpha\ trace[(\Sigma^{-1}A_{\pi})^3] + \alpha^2\ trace[(\Sigma^{-1}A_{\pi})^4]\} && .
\end{aligned}$$

Therefore,

$$\sigma_{\alpha}^2 = Var(S) = Var[E(S \mid \pi)] + E[Var(S \mid \pi)]$$

$$= (\alpha^2/4) Var\{trace[(\Sigma^{-1}A_{\pi})^2]\} + (1/2)\{trace[(\Sigma^{-1}A_{\pi})^2] + 2\alpha\ trace[(\Sigma^{-1}A_{\pi})^3] + \alpha^2\ trace[(\Sigma^{-1}A_{\pi})^4]\}$$

Substituting $\alpha = 0$, gives

$$\sigma_0^2 = (1/2) E\{trace[(\Sigma^{-1}A_{\pi})^2]\}.$$

For sibships $\Sigma$ has a simple form (all diagonal elements equal and all off-diagonal elements equal). Thus, a simple expression for $\Sigma^{-1}$ and hence the moments of the score statistic can be obtained (e.g., see Tang 2000).

# APPENDIX C

## LARGE MATRIX INVERSION

Let us consider a pedigree of size $s$. The computation of the MERLIN-REGRESS (SCORE.MERLIN), as originally defined, involves an inversion of a $s(s+1)/2 \times s(s+1)/2$ matrix, of trait squared sums and differences. However, as suggested by the calculations in the appendix of (Chen *et al.* 2005), it suffices to invert the a $s \times s$ dispersion matrix. If $\Omega$ denotes the $s \times s$ trait dispersion matrix, then following the notation of Chen *et al.* (2005), the inverse of the Gaussian working covariance matrix is given by:

$$(G^0)^{-1} \;=\; \begin{pmatrix} \Omega^{-1} & 0 & 0 \\ 0 & \frac{1}{2}\,(\Omega^{-1}_{ij})^2 & \Omega^{-1}_{il}\ \Omega^{-1}_{im} \\ 0 & \Omega^{-1}_{uj}\ \Omega^{-1}_{vj} & \Omega^{-1}_{ul}\ \Omega^{-1}_{vm}\ +\ 2\ \Omega^{-1}_{um}\ \Omega_{vl} \end{pmatrix} \text{ where,}$$

$$G^0 \;=\; \begin{pmatrix} \Omega & 0 & 0 \\ 0 & 2\ (\Omega_{ij})^2 & 2\ \Omega_{il}\ \Omega_{im} \\ 0 & 2\ \Omega_{uj}\ \Omega_{vj} & \Omega_{ul}\ \Omega_{vm}\ +\ 2\ \Omega_{um}\ \Omega_{vl} \end{pmatrix}$$

where $\Omega_{ij}$ and $\Omega^{-1}_{ij}$ are the elements in the $i^{th}$ row and $j^{th}$ column of $\Omega$ and $\Omega^{-1}$ respectively. Direct symbolic multiplication can be used to verify this inverse. This offers significant improvement in computational speed for large pedigrees. Additionally for sibships, the $\Omega$ matrix can be inverted analytically as it has a simple form (all diagonal elements equal the trait variance and all off-diagonal elements equal the sibling trait covariance).

The higher moment working covariance matrix $M^0$ is also a $s(s+1)/2 \times s(s+1)/2$ matrix. This matrix can be inverted by inverting at most one $2s \times 2s$ matrix. This matrix is defined as follows.

$$M^0 = \begin{pmatrix} \Omega & m_3\ I & 0 \\ m_3\ I & 2\ (\Omega_{ij})^2 + m_4\ I & 2\ \Omega_{il}\ \Omega_{im} \\ 0 & 2\ \Omega_{uj}\ \Omega_{vj} & \Omega_{ul}\ \Omega_{vm}\ +\ 2\ \Omega_{um}\ \Omega_{vl} \end{pmatrix}, \text{ where } \quad m_3 = \hat{\gamma}_3 \hat{\sigma}^3 \text{ and } m_4 = \hat{\gamma}_4 \hat{\sigma}^4.$$

Here $\hat{\gamma}_3$, $\hat{\gamma}_4$ and $\hat{\sigma}^2$ denote the estimated trait skewness, kurtosis and variance. It can be shown that $M^0 = G^0 + BDB'$, where

$$B = \begin{pmatrix} m_3\ I & 0 \\ 0 & I \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & I \\ I & m_4\ I \end{pmatrix} \quad \text{and} \quad D^{-1} = \begin{pmatrix} -m_4\ I & I \\ I & 0 \end{pmatrix}.$$

Using the following identity (e.g., Rao 2002),

$$(A + BDB')^{-1} = A^{-1} - A^{-1}B[B'A^{-1}B + D^{-1}]^{-1}B'A^{-1},$$

it suffices to invert the $2s \times 2s$ matrix

$$B'\ (G^0)^{-1}\ B + D^{-1}\ \ =\ \ \begin{pmatrix} m_3^2\ \Omega^{-1}\ -\ m_4\ I & I \\ I & \frac{1}{2}\ (\Omega^{-1}_{ij})^2 \end{pmatrix}.$$

For sibships, the matrices in the diagonal blocks again have the same simple form as $\Omega$ (diagonal elements equal and off diagonal elements equal). Hence, using the theory of partitioned matrices, this matrix (and hence $M^0$) can be inverted analytically.

# APPENDIX D

# CONDITIONAL COVARIANCE FOR SCORE.FPG

Here we prove the following identity used in the derivation of $Cov(g_N \mid \overline{g}_F)$ in the section 4.2.5 (required to obtain denominator of SCORE.FPG):

$$\phi_{n_1 n_2} = \sum_{f \in F} [1 - M_{f,f}(n_1, n_2)]\ 2\phi_{n_1 f}\phi_{n_2 f} + \sum_{f_1 \in F} \sum_{f_2 \in F} M_{f_1,f_2}(n_1, n_2) 4\phi_{n_1 f}\phi_{n_2 f}.$$

To prove this, we consider $1 - \phi_{n_1 n_2}$ the probability that two alleles $a_1$ and $a_2$ drawn randomly from $n_1$ and $n_2$ are not IBD. This event can occur if either (a) $a_1$ and $a_2$ come from two different founders $f_1$ and $f_2$ respectively or (b) $a_1$ and $a_2$ come from the same founder $f$, but are not IBD. In case (a), all founder pairs can contribute except those with $M_{f_1,f_2}(n_1, n_2) = 1$ (i.e., those for which the paths to $n_1$ and $n_2$ have a common meiosis). For any such founder pair, the probability of transmitting the two alleles is $4\phi_{n_1 f}\phi_{n_2 f}$ (product of the two path length as the paths don't coincide in any meiosis). In case (b), note that $f$ has to be an MRCA-founder (if not, the paths from $f$ to $n_1$ and $n_2$ would coincide in all the meioses starting from that founder to the MRCA). For any such founder, the probability of transmitting two non-IBD alleles is $2\phi_{n_1 f}\phi_{n_2 f}$ i.e., the product of path lengths times 1/2, as there is an equal probability that the same allele (IBD) would be transmitted to both $n_1$ and $n_2$. Therefore, we can write $1 - \phi_{n_1 n_2}$ as

$$\begin{aligned} 1 - \phi_{n_1 n_2} &= \sum_{f_1 \in F} \sum_{f_2 \in F, f_2 \neq f_1} [1 - M_{f_1,f_2}(n_1, n_2)] 4\phi_{n_1 f}\phi_{n_2 f} + \sum_{f \in F} [1 - M_{f,f}(n_1, n_2)]\ 2\phi_{n_1 f}\phi_{n_2 f} \\ &= \sum_{f_1 \in F} \sum_{f_2 \in F} [1 - M_{f_1,f_2}(n_1, n_2)] 4\phi_{n_1 f}\phi_{n_2 f} - \sum_{f \in F} [1 - M_{f,f}(n_1, n_2)]\ 2\phi_{n_1 f}\phi_{n_2 f}. \end{aligned}$$

Subtracting both sides from 1, and using the fact that $\sum_{f_1 \in F} \sum_{f_2 \in F} 4\phi_{n_1 f}\phi_{n_2 f} = 1$, we get the required identity.

# APPENDIX E

## COVARIANCE CONDITIONAL ON IBD

Here we prove the result 4.2.14. To derive $Cov(g_{n_1}, g_{n_2} \mid g_F, \Pi_m)$, we first note that the $g_{n1}$ can be written as $g_{f_1,i_1} + g_{f_2,i_2}$, where $f_1$ and $f_2$ are two founders who transmit there $i_1^th$ and $i_2^th$ alleles respectively to $n_1$. Similarly $g_{n2} = g_{f_3,i_3} + g_{f_4,i_4}$. Also, we note that $[g_{f,1} \mid g_f]$ and $[g_{f_2} \mid g_f]$ have hypergeometric distributions $HG(2, g_f, 1)$ and hence have mean $g_f/2$, variance $(g_f/2)(1 - g_f/2)$ and covariance $-(g_f/2)(1 - g_f/2)$. The transmitted and non-transmitted allele have Bernoulli distributions with success probability $g_f/2$ and hence have the same moments. To see that result 4.2.14, we verify it explicitly for all possible IBD configurations for the pair $n_1$ and $n_2$.

**Case-1:** $\pi_{n_1,n_2} = 1$

In this case, we must have $(f_1, i_1) = (f_3, i_3)$ and $(f_2, i_2) = (f_4, i_4)$

$$\begin{aligned} Cov(g_{n_1}, g_{n_2} \mid g_F, \Pi_m) &= Var(g_{f_1,i_1} + g_{f_2,i_2} \mid g_F) \\ &= Var(g_{f_1,i_1} \mid g_{f_1}) + Var(g_{f_2,i_2} \mid g_{f_2}) \\ &= [g_{f_1}(2 - g_{f_1})/4] + [g_{f_2}(2 - g_{f_2})/4] \ . \end{aligned}$$

Note that $g_{f_1,i_1}$ and $g_{f_2,i_2}$ are independent, as the paths $f_1 \rightarrow n_1$ and $f_2 \rightarrow n_1$ (and similarly the paths $f_1 \rightarrow n_1$ and $f_2 \rightarrow n_1$) cannot intersect, due to our assumption that $f_1$ and $f_2$ transmit 2 distinct alleles to $n_1$ (similarly $n_2$).

**Case-2:** $\pi_{n_1,n_2} = 1/2$ (One founder common).

In this case, we have $(f_1, i_1) = (f_3, i_3)$.

$$\begin{aligned} Cov(g_{n_1}, g_{n_2} \mid g_F, \Pi_m) &= Cov(g_{f_1,i_1} + g_{f_2,i_2}, g_{f_1,i_1} + g_{f_4,i_4} \mid g_F) \\ &= Var(g_{f_1,i_1} \mid g_{f_1}) \\ &= g_{f_1}(2 - g_{f_1})/4 \ . \end{aligned}$$

Note that the transmissions from $f_2$ and $f_4$ are independent of the transmissions from $f_1$, as ($f_1$ and $f_2$) are coancestors of $n_1$, and ($f_1$ and $f_4$) are coancestors of $n_2$. $f_2$ and $f_4$ can however intersect at an MRCA, who then transmits different alleles independently to $n_1$ and $n_2$ (by our assumption that $\pi_{n_1,n_2} = 1/2$).

**Case-3:** $\pi_{n_1,n_2} = 1/2$ (Both founders common).
In this case, we have $(f_1, i_1) = (f_3, i_3)$ and $(f_2, i_2) = (f_4, 2 - i_4)$.

$$\begin{aligned} Cov(g_{n_1}, g_{n_2} \mid g_F, \Pi_m) &= Cov(g_{f_1,i_1} + g_{f_2,i_2}, g_{f_1,i_1} + g_{f_2,2-i_2} \mid g_F) \\ &= Var(g_{f_1,i_1} \mid g_{f_1}) + Cov(g_{f_2,i_2}, g_{f_2,2-i_2} \mid g_{f_2}) \\ &= [g_{f_1}(2 - g_{f_1})/4] - [g_{f_2}(2 - g_{f_2})/4] \ . \end{aligned}$$

Here again the transmissions from $f_1$ and $f_2$ are independent, as the paths to $n_1$ and $n_2$ can not intersect ($f_2$ in this case must be an MRCA).

All the covariances derived above satisfy equation 4.2.14. The remaining three cases corresponding to $\pi_{n_1,n_2} = 0$ can be verified similarly.

# APPENDIX F

# LINKAGE: SUPPLEMENTARY TABLES

Table F1: **Detailed Type I Error Results**

| **Single Proband** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Genetic Model | | | | | | | | | | | | | | |
| | $\mathbf{1}$ | $\mathbf{1'}$ | $\mathbf{1''}$ | $\mathbf{2}$ | $\mathbf{2'}$ | $\mathbf{2''}$ | $\mathbf{3}$ | $\mathbf{3'}$ | $\mathbf{3''}$ | $\mathbf{4}$ | $\mathbf{4'}$ | $\mathbf{4''}$ | $\mathbf{5}$ | $\mathbf{5'}$ | $\mathbf{5''}$ |
| **SCORE.NAÏVE** | **0.058** | 0.013 | 0.008 | **0.058** | **0.171** | **0.266** | **0.16** | **0.353** | **0.408** | **0.049** | **0.107** | **0.17** | **0.048** | **0.16** | **0.247** |
| **SCORE.CIBD** | 0.013 | 0.012 | 0.012 | 0.011 | 0.013 | 0.013 | **0.015** | **0.015** | **0.018** | 0.012 | 0.01 | 0.013 | 0.011 | 0.012 | 0.012 |
| **SCORE.NULL.CT** | 0.013 | 0.012 | 0.011 | 0.01 | 0.013 | 0.012 | **0.015** | **0.015** | **0.018** | 0.011 | 0.01 | 0.013 | 0.011 | 0.011 | 0.011 |
| **SCORE.CT** | 0.013 | 0.012 | 0.011 | 0.01 | 0.013 | 0.012 | **0.015** | **0.015** | **0.018** | 0.012 | 0.01 | 0.013 | 0.011 | 0.011 | 0.012 |
| **SCORE.NULL.EV** | 0.006 | **0.005** | **0.003** | **0.004** | 0.007 | **0.005** | **0.004** | **0.002** | **0.002** | **0.005** | **0.005** | **0.004** | **0.005** | 0.006 | **0.005** |
| **SCORE.EV** | 0.007 | **0.005** | **0.004** | **0.005** | 0.007 | 0.006 | **0.004** | **0.003** | **0.002** | 0.006 | **0.005** | **0.004** | 0.006 | 0.007 | **0.005** |
| **SCORE.MERLIN** | 0.013 | 0.012 | 0.011 | 0.01 | 0.013 | 0.012 | **0.015** | **0.016** | **0.018** | 0.012 | 0.01 | 0.012 | 0.011 | 0.011 | 0.012 |
| **SCORE.MERLIN.AV** | 0.012 | 0.012 | 0.011 | 0.01 | 0.013 | 0.012 | 0.014 | **0.015** | **0.018** | 0.012 | 0.01 | 0.013 | 0.011 | 0.011 | 0.012 |
| **HM.NAÏVE** | **0.059** | 0.009 | **0.005** | **0.059** | **0.192** | **0.275** | **0.165** | **0.37** | **0.425** | **0.05** | **0.116** | **0.172** | **0.047** | **0.177** | **0.254** |
| **HM.MERLIN** | 0.013 | 0.012 | 0.012 | 0.011 | 0.011 | 0.01 | 0.014 | 0.012 | 0.014 | 0.012 | 0.01 | 0.011 | 0.011 | 0.01 | 0.011 |
| **HM.CT** | 0.012 | 0.012 | 0.012 | 0.011 | 0.012 | 0.011 | 0.014 | 0.012 | 0.014 | 0.012 | 0.009 | 0.011 | 0.011 | 0.011 | 0.011 |
| **SCORE.MAX** | 0.014 | 0.013 | 0.012 | 0.011 | 0.014 | 0.014 | **0.016** | **0.016** | **0.019** | 0.013 | 0.011 | **0.015** | 0.011 | 0.012 | 0.012 |
| **SCORE.2DF.CT** | 0.014 | 0.01 | 0.012 | 0.011 | 0.014 | **0.015** | **0.015** | **0.017** | **0.019** | 0.012 | 0.011 | 0.012 | 0.012 | 0.012 | 0.011 |
| **HM.2DF.CT** | 0.014 | 0.011 | **0.015** | 0.011 | 0.012 | 0.014 | **0.016** | **0.017** | **0.02** | 0.012 | 0.011 | **0.015** | 0.011 | 0.012 | 0.013 |
| **Extreme Concordant** | | | | | | | | | | | | | | | |
| **SCORE.NAÏVE** | **0.085** | **0.017** | 0.008 | **0.077** | **0.224** | **0.308** | **0.208** | **0.389** | **0.435** | **0.065** | **0.16** | **0.213** | **0.066** | **0.222** | **0.293** |
| **SCORE.CIBD** | **0.016** | **0.015** | **0.016** | 0.014 | 0.014 | 0.013 | **0.016** | 0.014 | 0.014 | 0.014 | **0.016** | 0.013 | 0.013 | 0.013 | 0.01 |
| **SCORE.NULL.CT** | **0.015** | 0.012 | 0.012 | 0.013 | 0.013 | 0.013 | **0.015** | 0.014 | 0.013 | 0.013 | **0.015** | 0.012 | 0.012 | 0.012 | 0.01 |
| **SCORE.CT** | **0.015** | 0.013 | 0.013 | 0.013 | 0.014 | 0.013 | **0.016** | **0.015** | 0.014 | 0.014 | **0.016** | 0.013 | 0.013 | 0.012 | 0.01 |
| **SCORE.NULL.EV** | **0.004** | **0.004** | **0.005** | **0.004** | 0.006 | 0.006 | **0.002** | **0.002** | **0.002** | **0.004** | 0.008 | **0.005** | **0.003** | 0.006 | **0.005** |
| **SCORE.EV** | 0.007 | **0.005** | 0.006 | 0.006 | 0.007 | 0.006 | **0.004** | **0.002** | **0.002** | 0.006 | 0.009 | **0.005** | 0.006 | 0.007 | **0.005** |
| **SCORE.MERLIN** | 0.014 | 0.013 | 0.012 | 0.013 | 0.013 | 0.013 | **0.016** | **0.016** | 0.013 | 0.013 | **0.016** | 0.013 | 0.013 | 0.013 | 0.01 |
| **SCORE.MERLIN.AV** | 0.014 | 0.012 | 0.013 | 0.013 | 0.013 | 0.013 | **0.016** | **0.016** | 0.014 | 0.013 | **0.016** | 0.013 | 0.013 | 0.012 | 0.01 |
| **HM.NAÏVE** | **0.09** | 0.012 | 0.006 | **0.08** | **0.251** | **0.323** | **0.224** | **0.405** | **0.451** | **0.066** | **0.186** | **0.233** | **0.064** | **0.253** | **0.317** |
| **HM.MERLIN** | 0.014 | 0.012 | 0.01 | 0.012 | 0.012 | 0.011 | 0.014 | 0.012 | 0.012 | 0.013 | **0.015** | 0.01 | 0.013 | 0.012 | 0.01 |
| **HM.CT** | 0.014 | 0.012 | 0.011 | 0.013 | 0.012 | 0.011 | **0.015** | 0.012 | 0.012 | 0.013 | 0.014 | 0.01 | 0.013 | 0.012 | 0.01 |
| **SCORE.MAX** | **0.016** | 0.014 | **0.015** | 0.014 | **0.015** | **0.015** | **0.017** | **0.016** | 0.014 | 0.014 | **0.018** | **0.015** | 0.014 | 0.013 | 0.011 |
| **SCORE.2DF.CT** | **0.016** | 0.012 | 0.011 | 0.014 | **0.016** | 0.014 | **0.016** | **0.017** | 0.014 | 0.013 | **0.017** | 0.012 | 0.014 | 0.012 | 0.011 |
| **HM.2DF.CT** | **0.015** | 0.013 | 0.013 | 0.014 | 0.013 | **0.015** | **0.016** | **0.017** | **0.015** | 0.013 | **0.017** | 0.013 | 0.014 | 0.013 | 0.013 |
| **EDAC-3 Corner** | | | | | | | | | | | | | | | |
| **SCORE.NAÏVE** | **0.138** | **0.103** | **0.08** | **0.145** | **0.222** | **0.309** | **0.235** | **0.382** | **0.426** | **0.134** | **0.202** | **0.245** | **0.133** | **0.216** | **0.296** |
| **SCORE.CIBD** | 0.011 | 0.012 | 0.012 | 0.014 | 0.012 | 0.013 | **0.015** | 0.013 | **0.016** | 0.013 | 0.012 | 0.012 | 0.013 | 0.01 | 0.012 |
| **SCORE.NULL.CT** | 0.01 | 0.01 | 0.011 | 0.012 | 0.012 | 0.012 | 0.013 | 0.013 | **0.015** | 0.012 | 0.011 | 0.012 | 0.012 | 0.009 | 0.012 |
| **SCORE.CT** | 0.011 | 0.011 | 0.011 | 0.013 | 0.012 | 0.012 | 0.014 | 0.013 | **0.015** | 0.012 | 0.011 | 0.012 | 0.013 | 0.01 | 0.012 |
| **SCORE.NULL.EV** | **0.004** | **0.005** | **0.004** | **0.005** | **0.005** | **0.004** | **0.004** | **0.002** | **0.002** | **0.004** | 0.006 | **0.005** | **0.005** | 0.006 | **0.005** |
| **SCORE.EV** | 0.006 | **0.005** | **0.005** | 0.007 | 0.006 | **0.005** | **0.005** | **0.003** | **0.002** | 0.007 | 0.007 | **0.005** | 0.008 | 0.007 | 0.006 |
| **SCORE.MERLIN** | 0.011 | 0.011 | 0.011 | 0.013 | 0.011 | 0.013 | **0.015** | 0.013 | **0.015** | 0.013 | 0.012 | 0.011 | 0.012 | 0.01 | 0.012 |
| **SCORE.MERLIN.AV** | 0.011 | 0.011 | 0.011 | 0.013 | 0.012 | 0.012 | **0.015** | 0.013 | **0.015** | 0.012 | 0.012 | 0.012 | 0.012 | 0.01 | 0.012 |
| **HM.NAÏVE** | **0.139** | **0.078** | **0.039** | **0.146** | **0.24** | **0.323** | **0.238** | **0.395** | **0.442** | **0.135** | **0.205** | **0.256** | **0.133** | **0.236** | **0.32** |
| **HM.MERLIN** | 0.011 | 0.01 | 0.011 | 0.013 | 0.012 | 0.011 | 0.014 | 0.012 | 0.012 | 0.013 | 0.011 | 0.011 | 0.012 | 0.012 | 0.011 |
| **HM.CT** | 0.011 | 0.009 | 0.011 | 0.013 | 0.012 | 0.012 | 0.014 | 0.013 | 0.013 | 0.012 | 0.011 | 0.011 | 0.012 | 0.011 | 0.011 |
| **SCORE.MAX** | 0.011 | 0.011 | 0.013 | 0.014 | 0.013 | 0.014 | **0.015** | 0.014 | **0.016** | 0.013 | 0.012 | 0.013 | 0.014 | 0.012 | 0.014 |
| **SCORE.2DF.CT** | 0.01 | 0.01 | 0.01 | 0.013 | 0.012 | 0.013 | **0.015** | **0.015** | **0.018** | 0.012 | 0.011 | 0.014 | 0.013 | 0.012 | 0.013 |
| **HM.2DF.CT** | 0.011 | 0.01 | 0.012 | 0.013 | 0.013 | 0.013 | **0.015** | **0.015** | **0.016** | 0.012 | 0.011 | 0.014 | 0.013 | 0.013 | 0.014 |
| **MDAC-3 Corner** | | | | | | | | | | | | | | | |
| **SCORE.NAÏVE** | **0.067** | **0.042** | **0.037** | **0.068** | **0.107** | **0.19** | **0.122** | **0.308** | **0.38** | **0.063** | **0.081** | **0.119** | **0.061** | **0.112** | **0.178** |

**Continued on next page**

**Table F1 – continued from previous page**

| | **1** | **$1'$** | **$1''$** | **2** | **$2'$** | **$2''$** | **3** | **$3'$** | **$3''$** | **4** | **$4'$** | **$4''$** | **5** | **$5'$** | **$5''$** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SCORE.CIBD** | 0.014 | 0.012 | 0.013 | 0.012 | 0.013 | 0.011 | 0.011 | **0.017** | **0.015** | 0.012 | 0.01 | 0.013 | 0.011 | 0.013 | 0.011 |
| **SCORE.NULL.CT** | 0.013 | 0.012 | 0.013 | 0.012 | 0.013 | 0.011 | 0.01 | **0.017** | **0.015** | 0.011 | 0.01 | 0.013 | 0.011 | 0.012 | 0.011 |
| **SCORE.CT** | 0.014 | 0.012 | 0.013 | 0.012 | 0.013 | 0.011 | 0.01 | **0.017** | **0.015** | 0.011 | 0.01 | 0.013 | 0.011 | 0.012 | 0.011 |
| **SCORE.NULL.EV** | 0.008 | 0.006 | **0.005** | 0.006 | 0.007 | **0.004** | **0.004** | **0.003** | **0.002** | 0.007 | 0.006 | 0.007 | 0.006 | 0.007 | 0.006 |
| **SCORE.EV** | 0.009 | 0.006 | **0.005** | 0.007 | 0.007 | **0.005** | **0.005** | **0.003** | **0.002** | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.006 |
| **SCORE.MERLIN** | 0.013 | 0.012 | 0.014 | 0.012 | 0.013 | 0.01 | 0.01 | **0.016** | 0.014 | 0.012 | 0.01 | 0.013 | 0.011 | 0.013 | 0.011 |
| **SCORE.MERLIN.AV** | 0.014 | 0.012 | 0.013 | 0.011 | 0.013 | 0.011 | 0.011 | **0.018** | **0.015** | 0.012 | 0.01 | 0.013 | 0.011 | 0.013 | 0.011 |
| **HM.NAÏVE** | **0.067** | **0.024** | 0.014 | **0.069** | **0.114** | **0.197** | **0.122** | **0.328** | **0.388** | **0.063** | **0.078** | **0.117** | **0.061** | **0.116** | **0.188** |
| **HM.MERLIN** | 0.014 | 0.012 | 0.011 | 0.011 | 0.01 | 0.01 | 0.011 | 0.013 | 0.012 | 0.012 | 0.01 | 0.012 | 0.01 | 0.011 | 0.011 |
| **HM.CT** | 0.014 | 0.011 | 0.011 | 0.012 | 0.01 | 0.009 | 0.011 | 0.013 | 0.012 | 0.012 | 0.01 | 0.012 | 0.011 | 0.011 | 0.01 |
| **SCORE.MAX** | 0.014 | 0.012 | 0.014 | 0.012 | 0.014 | 0.013 | 0.011 | **0.018** | **0.016** | 0.012 | 0.011 | **0.016** | 0.011 | 0.013 | 0.013 |
| **SCORE.2DF.CT** | 0.012 | 0.01 | 0.013 | 0.011 | 0.013 | 0.011 | 0.012 | **0.017** | **0.016** | 0.011 | 0.009 | 0.014 | 0.012 | 0.013 | 0.012 |
| **HM.2DF.CT** | 0.012 | 0.011 | 0.013 | 0.011 | 0.011 | 0.013 | 0.013 | **0.017** | **0.018** | 0.011 | 0.01 | 0.014 | 0.012 | 0.012 | 0.011 |

Note: Type I error values departing by 0.005 or more, from the nominal value 0.01 are highlighted in bold.

Table F2: **Detailed Power Results**

| **Population** | Genetic Model | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **$1'$** | **$1''$** | **2** | **$2'$** | **$2''$** | **3** | **$3'$** | **$3''$** | **4** | **$4'$** | **$4''$** | **5** | **$5'$** | **$5''$** |
| **SCORE.NAÏVE** | **0.74** | 0.52 | 0.36 | **0.73** | 0.87 | 0.76 | **0.65** | 1 | 1 | **0.75** | 0.78 | 0.57 | **0.74** | 0.56 | 0.38 |
| **SCORE.CIBD** | **0.74** | **0.39** | **0.14** | **0.73** | **0.78** | 0.45 | 0.53 | **0.96** | **0.94** | **0.76** | **0.69** | 0.31 | **0.74** | **0.44** | **0.15** |
| **SCORE.NULL.CT** | **0.74** | **0.39** | **0.14** | **0.73** | **0.78** | 0.45 | 0.53 | **0.96** | **0.94** | **0.76** | **0.69** | 0.31 | **0.74** | **0.44** | **0.15** |
| **SCORE.CT** | **0.74** | **0.39** | **0.14** | **0.73** | **0.78** | 0.45 | 0.53 | **0.96** | **0.94** | **0.76** | **0.69** | 0.31 | **0.74** | **0.44** | **0.15** |
| **SCORE.NULL.EV** | 0.67 | 0.34 | 0.11 | 0.67 | 0.73 | 0.4 | 0.23 | 0.75 | 0.72 | 0.69 | 0.65 | 0.32 | 0.69 | 0.41 | 0.11 |
| **SCORE.EV** | 0.67 | 0.35 | 0.11 | 0.68 | 0.73 | 0.4 | 0.24 | 0.75 | 0.73 | 0.7 | 0.65 | 0.32 | 0.7 | 0.41 | 0.11 |
| **SCORE.MERLIN** | **0.74** | **0.39** | **0.14** | **0.73** | **0.78** | 0.45 | 0.53 | **0.97** | **0.95** | **0.75** | **0.68** | 0.31 | **0.74** | **0.44** | **0.16** |
| **SCORE.MERLIN.AV** | **0.74** | **0.38** | **0.14** | **0.73** | **0.78** | 0.45 | 0.53 | **0.96** | **0.94** | **0.76** | **0.69** | 0.31 | **0.74** | **0.44** | **0.15** |
| **HM.NAÏVE** | **0.74** | 0.46 | 0.31 | **0.73** | 0.83 | 0.75 | **0.62** | 0.99 | 0.99 | **0.75** | 0.74 | 0.57 | **0.74** | 0.53 | 0.35 |
| **HM.MERLIN** | **0.74** | 0.37 | **0.14** | **0.73** | 0.74 | **0.48** | 0.51 | 0.92 | 0.91 | **0.75** | 0.67 | 0.34 | **0.75** | 0.42 | **0.16** |
| **HM.CT** | **0.74** | 0.36 | **0.14** | **0.73** | 0.74 | **0.49** | 0.51 | 0.9 | 0.91 | **0.76** | 0.67 | 0.33 | **0.74** | 0.42 | **0.16** |
| **SCORE.MAX** | **0.74** | **0.41** | **0.16** | **0.74** | **0.81** | **0.5** | 0.53 | **0.98** | **0.96** | **0.76** | **0.71** | **0.37** | **0.75** | **0.46** | **0.18** |
| **SCORE.2DF.CT** | 0.71 | 0.37 | **0.14** | **0.72** | 0.75 | 0.43 | **0.62** | **0.98** | **0.97** | 0.72 | 0.66 | 0.29 | **0.76** | **0.45** | 0.15 |
| **HM.2DF.CT** | 0.71 | 0.34 | 0.12 | **0.72** | 0.7 | 0.45 | 0.58 | 0.93 | 0.91 | 0.72 | 0.62 | 0.32 | **0.76** | 0.42 | **0.16** |
| **Single Proband** | | | | | | | | | | | | | | | |
| **SCORE.NULL.CT** | **0.69** | 0.78 | 0.53 | **0.69** | **0.78** | 0.53 | 0.79 | **0.99** | **0.99** | **0.38** | **0.43** | 0.24 | **0.2** | **0.19** | **0.12** |
| **SCORE.CT** | **0.69** | **0.78** | 0.54 | **0.7** | **0.78** | 0.53 | 0.79 | **0.99** | **0.99** | **0.38** | **0.43** | 0.24 | **0.2** | **0.19** | **0.12** |
| **SCORE.NULL.EV** | 0.55 | 0.71 | 0.48 | 0.56 | 0.72 | 0.51 | 0.37 | 0.91 | 0.91 | 0.27 | 0.36 | 0.17 | 0.11 | 0.15 | 0.08 |
| **SCORE.EV** | 0.59 | 0.71 | 0.49 | 0.59 | 0.74 | 0.52 | 0.4 | 0.93 | 0.92 | 0.29 | 0.37 | 0.18 | 0.13 | 0.16 | 0.09 |
| **SCORE.MERLIN** | **0.69** | 0.78 | 0.53 | **0.69** | **0.78** | 0.55 | 0.8 | **1** | **0.99** | **0.38** | **0.43** | 0.23 | **0.2** | **0.19** | **0.12** |
| **SCORE.MERLIN.AV** | **0.68** | 0.78 | 0.53 | **0.69** | **0.78** | 0.54 | 0.79 | **0.99** | **0.99** | **0.38** | **0.43** | 0.23 | **0.2** | **0.19** | **0.12** |
| **HM.MERLIN** | **0.69** | **0.81** | **0.66** | **0.69** | 0.73 | **0.58** | 0.78 | **0.99** | **0.99** | **0.38** | 0.38 | 0.22 | **0.2** | **0.17** | **0.11** |
| **HM.CT** | **0.69** | **0.8** | **0.65** | **0.69** | 0.73 | 0.57 | 0.76 | **0.98** | **0.98** | **0.38** | 0.37 | 0.22 | **0.2** | **0.17** | **0.11** |
| **SCORE.MAX** | **0.7** | **0.81** | 0.61 | **0.7** | **0.8** | **0.61** | 0.79 | **1** | **0.99** | **0.39** | **0.45** | **0.28** | **0.21** | **0.2** | **0.13** |
| **SCORE.2DF.CT** | 0.66 | 0.76 | 0.52 | 0.65 | 0.75 | 0.51 | **0.85** | **1** | **0.99** | 0.36 | 0.4 | 0.21 | **0.21** | **0.2** | **0.11** |
| **HM.2DF.CT** | 0.66 | 0.78 | 0.63 | 0.65 | 0.7 | 0.53 | **0.83** | **0.99** | **0.99** | 0.36 | 0.35 | 0.21 | **0.22** | **0.18** | **0.11** |
| **Extreme Discordant** | | | | | | | | | | | | | | | |
| **SCORE.NULL.CT** | **0.58** | 0.78 | 0.74 | **0.58** | **0.81** | 0.85 | **0.15** | 0.77 | 0.92 | **0.24** | **0.76** | 0.87 | 0.52 | 0.67 | 0.7 |
| **SCORE.CT** | **0.59** | 0.78 | 0.74 | **0.59** | **0.81** | 0.85 | **0.15** | 0.77 | 0.92 | **0.25** | **0.77** | 0.87 | **0.53** | **0.68** | 0.7 |
| **SCORE.NULL.EV** | 0.42 | 0.67 | 0.7 | 0.46 | 0.75 | 0.84 | 0.03 | 0.38 | 0.74 | 0.15 | 0.7 | 0.85 | 0.39 | 0.61 | 0.69 |
| **SCORE.EV** | 0.48 | 0.7 | 0.72 | 0.52 | 0.78 | 0.85 | 0.04 | 0.43 | 0.75 | 0.18 | 0.73 | 0.85 | 0.46 | 0.64 | 0.7 |
| **SCORE.MERLIN** | **0.6** | 0.79 | 0.74 | **0.59** | **0.81** | 0.85 | **0.15** | 0.78 | 0.93 | **0.23** | **0.77** | 0.87 | 0.52 | 0.67 | 0.69 |
| **SCORE.MERLIN.AV** | **0.59** | 0.78 | 0.73 | **0.59** | **0.81** | 0.85 | **0.16** | 0.77 | 0.92 | **0.23** | **0.77** | 0.86 | **0.52** | **0.68** | 0.69 |
| **HM.MERLIN** | **0.59** | **0.82** | **0.84** | **0.59** | **0.81** | **0.9** | 0.14 | 0.7 | 0.89 | 0.15 | **0.77** | **0.91** | 0.52 | **0.69** | **0.77** |
| **HM.CT** | **0.59** | **0.81** | **0.85** | **0.59** | 0.8 | **0.9** | **0.15** | 0.68 | 0.87 | 0.14 | **0.77** | **0.91** | **0.52** | **0.7** | **0.76** |
| **SCORE.MAX** | **0.6** | **0.79** | 0.79 | **0.6** | **0.83** | **0.89** | **0.15** | 0.82 | **0.95** | **0.25** | **0.79** | **0.89** | **0.55** | **0.71** | **0.76** |
| **SCORE.2DF.CT** | 0.55 | 0.75 | 0.71 | 0.56 | 0.79 | 0.85 | **0.18** | **0.88** | **0.97** | 0.22 | 0.74 | 0.84 | **0.54** | **0.69** | 0.71 |

**Continued on next page**

**Table F2 – continued from previous page**

| | **1** | **1′** | **1″** | **2** | **2′** | **2″** | **3** | **3′** | **3″** | **4** | **4′** | **4″** | **5** | **5′** | **5″** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **HM.2DF.CT** | 0.55 | **0.79** | 0.81 | 0.56 | 0.77 | **0.88** | **0.17** | 0.77 | 0.91 | 0.15 | 0.73 | **0.89** | **0.54** | **0.7** | **0.77** |
| **Extreme Concordant** | | | | | | | | | | | | | | | |
| **SCORE.NULL.CT** | **0.6** | **0.75** | 0.69 | **0.54** | **0.68** | 0.57 | 0.8 | **0.99** | **1** | **0.21** | **0.26** | 0.22 | **0.12** | **0.1** | 0.09 |
| **SCORE.CT** | **0.61** | **0.75** | 0.69 | **0.55** | **0.68** | 0.57 | 0.81 | **0.99** | **1** | **0.23** | **0.26** | 0.22 | **0.12** | **0.1** | 0.09 |
| **SCORE.NULL.EV** | 0.4 | 0.6 | 0.6 | 0.33 | 0.59 | 0.54 | 0.37 | 0.85 | **0.98** | 0.11 | 0.17 | 0.17 | 0.06 | 0.06 | 0.06 |
| **SCORE.EV** | 0.48 | 0.63 | 0.61 | 0.4 | 0.62 | 0.55 | 0.46 | 0.88 | **0.98** | 0.13 | 0.18 | 0.18 | 0.07 | 0.07 | 0.07 |
| **SCORE.MERLIN** | **0.6** | **0.74** | 0.69 | **0.53** | **0.68** | 0.58 | 0.81 | **0.99** | **1** | **0.22** | **0.26** | 0.23 | **0.11** | **0.1** | 0.09 |
| **SCORE.MERLIN.AV** | **0.6** | **0.75** | 0.68 | **0.53** | **0.68** | 0.57 | 0.8 | **0.99** | **1** | **0.22** | **0.26** | 0.22 | **0.11** | **0.1** | 0.08 |
| **HM.MERLIN** | **0.6** | **0.76** | **0.74** | **0.53** | 0.63 | **0.65** | 0.81 | **0.99** | **1** | **0.22** | 0.25 | **0.25** | **0.12** | **0.1** | **0.11** |
| **HM.CT** | **0.6** | **0.75** | **0.73** | **0.53** | 0.63 | **0.65** | 0.79 | **0.98** | **1** | **0.22** | 0.25 | **0.24** | **0.11** | **0.1** | **0.11** |
| **SCORE.MAX** | **0.62** | **0.77** | **0.75** | **0.55** | **0.7** | **0.64** | 0.81 | **0.99** | **1** | **0.23** | **0.28** | **0.27** | **0.13** | **0.11** | **0.1** |
| **SCORE.2DF.CT** | 0.57 | 0.71 | 0.65 | 0.51 | 0.65 | 0.54 | **0.86** | **1** | **1** | 0.19 | 0.25 | 0.21 | **0.12** | **0.11** | **0.09** |
| **HM.2DF.CT** | 0.57 | 0.72 | 0.69 | 0.51 | 0.61 | 0.61 | **0.85** | **0.99** | **1** | 0.19 | 0.23 | 0.23 | **0.12** | **0.11** | **0.12** |
| **EDAC: 3-Corner** | | | | | | | | | | | | | | | |
| **SCORE.NULL.CT** | **0.59** | 0.72 | 0.66 | **0.54** | **0.71** | 0.64 | 0.78 | **0.99** | **1** | **0.43** | 0.56 | 0.45 | **0.37** | 0.28 | **0.17** |
| **SCORE.CT** | **0.6** | 0.73 | 0.66 | **0.55** | **0.71** | 0.64 | 0.78 | **0.99** | **1** | **0.44** | **0.57** | 0.45 | **0.38** | **0.29** | **0.18** |
| **SCORE.NULL.EV** | 0.43 | 0.64 | 0.61 | 0.41 | 0.63 | 0.59 | 0.39 | 0.91 | **0.98** | 0.29 | 0.47 | 0.41 | 0.25 | 0.21 | 0.14 |
| **SCORE.EV** | 0.49 | 0.66 | 0.62 | 0.46 | 0.65 | 0.61 | 0.48 | 0.92 | **0.98** | 0.35 | 0.51 | 0.42 | 0.3 | 0.24 | 0.14 |
| **SCORE.MERLIN** | **0.6** | 0.73 | 0.66 | **0.55** | **0.71** | 0.63 | 0.78 | **1** | **1** | **0.44** | 0.56 | 0.46 | **0.37** | **0.29** | **0.18** |
| **SCORE.MERLIN.AV** | **0.6** | 0.73 | 0.66 | **0.54** | **0.71** | 0.63 | 0.78 | **0.99** | **1** | **0.45** | 0.56 | 0.45 | **0.38** | 0.29 | 0.17 |
| **HM.MERLIN** | **0.61** | **0.77** | **0.8** | **0.54** | 0.66 | 0.62 | 0.79 | **0.99** | **1** | **0.44** | 0.5 | 0.4 | **0.37** | 0.22 | 0.13 |
| **HM.CT** | **0.61** | **0.76** | **0.8** | **0.55** | 0.66 | 0.61 | 0.79 | **0.99** | **1** | **0.45** | 0.5 | 0.4 | **0.38** | 0.22 | 0.13 |
| **SCORE.MAX** | **0.61** | **0.74** | 0.71 | **0.56** | **0.74** | **0.71** | 0.78 | **1** | **1** | **0.46** | **0.59** | **0.51** | **0.39** | **0.32** | **0.2** |
| **SCORE.2DF.CT** | 0.56 | 0.7 | 0.61 | 0.52 | 0.68 | 0.59 | **0.85** | **1** | **1** | 0.42 | 0.51 | 0.41 | **0.4** | **0.29** | 0.17 |
| **HM.2DF.CT** | 0.56 | 0.74 | **0.77** | 0.52 | 0.63 | 0.56 | **0.84** | **0.99** | **1** | 0.42 | 0.45 | 0.37 | **0.4** | 0.22 | 0.13 |
| **MDAC: 3-Corner** | | | | | | | | | | | | | | | |
| **SCORE.NULL.CT** | **0.74** | 0.73 | 0.5 | **0.69** | **0.85** | 0.64 | 0.58 | **0.98** | **0.98** | **0.63** | **0.68** | 0.44 | **0.56** | **0.42** | **0.2** |
| **SCORE.CT** | **0.74** | 0.73 | 0.5 | **0.69** | **0.85** | 0.64 | 0.59 | **0.98** | **0.98** | **0.63** | **0.68** | 0.44 | **0.56** | **0.42** | **0.2** |
| **SCORE.NULL.EV** | 0.64 | 0.68 | 0.48 | 0.6 | 0.81 | 0.61 | 0.23 | 0.85 | 0.91 | 0.53 | 0.63 | 0.41 | 0.48 | 0.37 | 0.16 |
| **SCORE.EV** | 0.66 | 0.69 | 0.48 | 0.62 | 0.81 | 0.62 | 0.25 | 0.86 | 0.92 | 0.56 | 0.64 | 0.41 | 0.5 | 0.38 | 0.17 |
| **SCORE.MERLIN** | **0.74** | 0.73 | 0.5 | **0.68** | **0.85** | 0.65 | 0.58 | **0.98** | **0.99** | **0.63** | **0.68** | 0.44 | **0.57** | **0.42** | **0.19** |
| **SCORE.MERLIN.AV** | **0.73** | 0.74 | 0.5 | **0.68** | **0.85** | 0.64 | 0.58 | **0.98** | **0.97** | **0.63** | **0.69** | 0.44 | **0.56** | **0.42** | **0.19** |
| **HM.MERLIN** | **0.73** | **0.8** | **0.67** | **0.68** | 0.79 | 0.64 | 0.59 | **0.98** | **0.98** | **0.63** | 0.64 | 0.44 | **0.57** | 0.38 | 0.17 |
| **HM.CT** | **0.73** | **0.79** | **0.65** | **0.69** | 0.8 | 0.63 | 0.59 | **0.97** | **0.97** | **0.63** | 0.65 | 0.44 | **0.57** | 0.38 | 0.17 |
| **SCORE.MAX** | **0.74** | 0.75 | 0.56 | **0.7** | **0.86** | **0.72** | 0.59 | **0.98** | **0.99** | **0.63** | **0.7** | **0.5** | **0.57** | **0.44** | **0.22** |
| **SCORE.2DF.CT** | 0.71 | 0.69 | 0.48 | **0.67** | 0.83 | 0.61 | **0.66** | **0.99** | **0.99** | 0.6 | 0.65 | 0.41 | **0.58** | **0.43** | **0.19** |
| **HM.2DF.CT** | 0.71 | 0.75 | 0.62 | **0.67** | 0.76 | 0.6 | **0.66** | **0.97** | **0.97** | 0.6 | 0.6 | 0.4 | **0.58** | 0.38 | 0.17 |

Note: For each model, power values within 3% of the maximum are highlighted in bold.

# BIBLIOGRAPHY

Abecasis, G. R., L. R. Cardon, and W. O. Cookson, 2000 A general test of association for quantitative traits in nuclear families. Am J Hum Genet *66*(1): 279–292.

Abecasis, G. R., L. R. Cardon, W. O. Cookson, P. C. Sham, and S. S. Cherny, 2001 Association analysis in a variance components framework. Genet Epidemiol **21 Suppl 1:** S341–6.

Abecasis, G. R., W. O. Cookson, and L. R. Cardon, 2000 Pedigree tests of transmission disequilibrium. Eur J Hum Genet *8*(7): 545–551.

Abecasis, G. R. and J. E. Wigginton, 2005 Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet *77*(5): 754–767.

Allison, D. B., 1997 Transmission-disequilibrium tests for quantitative traits. Am J Hum Genet *60*(3): 676–690.

Allison, D. B., M. C. Neale, R. Zannolli, N. J. Schork, C. I. Amos, and J. Blangero, 1999 Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Am J Hum Genet *65*(2): 531–544.

Almasy, L. and J. Blangero, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet *62*(5): 1198–1211.

Amos, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet *54*(3): 535–543.

Amos, C. I. and R. C. Elston, 1989 Robust methods for the detection of genetic linkage for quantitative data from pedigrees. Genet Epidemiol *6*(2): 349–360.

Barrett, J. H., 2002 Association studies. Methods Mol Biol **195:** 3–12.

Bhattacharjee, S., C. L. Kuo, N. Mukhopadhyay, G. N. Brock, D. E. Weeks, and E. Feingold, 2008 Robust score statistics for QTL linkage analysis. Am J Hum Genet *82*(3): 567–582.

Bourgain, C., S. Hoffjan, R. Nicolae, D. Newman, L. Steiner, K. Walker, R. Reynolds, C. Ober, and M. S. McPeek, 2003 Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am J Hum Genet *73*(3): 612–626.

Chen, W. M., K. W. Broman, and K. Y. Liang, 2004 Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genet Epidemiol *26*(4): 265–272.

Chen, W. M., K. W. Broman, and K. Y. Liang, 2005 Power and robustness of linkage tests for quantitative traits in general pedigrees. Genet Epidemiol *28*(1): 11–23.

Diao, G. and D. Y. Lin, 2005 A powerful and robust method for mapping quantitative trait loci in general pedigrees. Am J Hum Genet *77*(1): 97–111.

Diao, G. and D. Y. Lin, 2006 Improving the power of association tests for quantitative traits in family studies. Genet Epidemiol *30*(4): 301–313.

Dupuis, J., D. O. Siegmund, and B. Yakir, 2007 A unified framework for linkage and association analysis of quantitative traits. Proc Natl Acad Sci U S A *104*(51): 20210–20215.

Ewens, W. J., M. Li, and R. S. Spielman. A Review of Family-Based Tests for Linkage Disequilibrium between a Quantitative Trait and a Genetic Marker. PLoS.

Falconer, D. W., 1981 *Introduction to Quantitative Genetics*. Harlow, UK: Longman Group.

Forrest, W. F. and E. Feingold, 2000 Composite statistics for QTL mapping with moderately discordant sibling pairs. Am J Hum Genet *66*(5): 1642–1660.

Fulker, D. W., S. S. Cherny, P. C. Sham, and J. K. Hewitt, 1999 Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet *64*(1): 259–267.

Ghosh, S. and G. De, 2007 Association analysis of population-based quantitative trait data: an assessment of ANOVA. Hum Hered *64*(1): 82–88.

Haseman, J. K. and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav Genet *2*(1): 3–19.

Hegele, R. A., S. B. Harris, A. J. Hanley, and B. Zinman, 1999 Association between AGT codon 235 polymorphism and variation in serum concentrations of creatinine and urea in Canadian Oji-Cree. Clin Genet *55*(6): 438–443.

Hopper, J. L. and J. D. Mathews, 1982 Extensions to multivariate normal models for pedigree analysis. Ann Hum Genet *46*(Pt 4): 373–383.

Horvath, S., X. Xu, and N. M. Laird, 2001 The family based association test method: strategies for studying general genotype–phenotype associations. Eur J Hum Genet *9*(4): 301–306.

Laird, N. M., S. Horvath, and X. Xu, 2000 Implementing a unified approach to family-based tests of association. Genet Epidemiol **19 Suppl 1:** S36–42.

Laird, N. M. and C. Lange, 2006 Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet *7*(5): 385–394.

Lake, S. L., D. Blacker, and N. M. Laird, 2000 Family-based tests of association in the presence of linkage. Am J Hum Genet *67*(6): 1515–1525.

Lange, C., D. DeMeo, E. K. Silverman, S. T. Weiss, and N. M. Laird, 2004 PBAT: tools for family-based association studies. Am J Hum Genet *74*(2)**:** 367–369.

Lange, K., 2002 *Mathematical and Statistical Methods for Genetic Analysis* (Second Edition ed.). Springer-Verlag.

Lange, K., J. Westlake, and M. A. Spence, 1976 Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet *39*(4)**:** 485–491.

Lebrec, J., H. Putter, and J. C. van Houwelingen, 2004 Score test for detecting linkage to complex traits in selected samples. Genet Epidemiol *27*(2)**:** 97–108.

Li, C. C. and L. Sacks, 1954 The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics **10:** 347–360.

Martin, E. R., S. A. Monks, L. L. Warren, and N. L. Kaplan, 2000 A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet *67*(1)**:** 146–154.

Monks, S. A. and N. L. Kaplan, 2000 Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. Am J Hum Genet *66*(2)**:** 576–592.

Mukhopadhyay, N., S. Bhattacharjee, C. L. Kuo, B. Reck, D. E. Weeks, and E. Feingold. QTL-ALL: Analysis and linkage library for human QTL mapping. (unpublished data).

O'Donnell, C. J., K. Lindpaintner, M. G. Larson, V. S. Rao, J. M. Ordovas, E. J. Schaefer, R. H. Myers, and D. Levy, 1998 Evidence for association and genetic linkage of the angiotensin-converting enzyme locus with hypertension and blood pressure in men but not women in the Framingham Heart Study. Circulation *97*(18)**:** 1766–1772.

Peng, J. and D. Siegmund, 2006 QTL mapping under ascertainment. Ann Hum Genet *70*(Pt 6)**:** 867–881.

Purcell, S., 2008 PLINK v1.3 URL http://pngu.mgh.harvard.edu/purcell/plink/ (17 July, 2008).

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*(3)**:** 559–575.

Purcell, S., P. Sham, and M. J. Daly, 2005 Parental phenotypes in family-based association analysis. Am J Hum Genet *76*(2)**:** 249–259.

Putter, H., L. A. Sandkuijl, and J. C. van Houwelingen, 2002 Score test for detecting linkage to quantitative traits. Genet Epidemiol *22*(4)**:** 345–355.

R Development Core Team, 2008 R: A Language and Environment for Statistical Computing URL http://www.R-project.org/.

Rabinowitz, D., 1997 A transmission disequilibrium test for quantitative trait loci. Hum Hered *47*(6)**:** 342–350.

Rabinowitz, D. and N. Laird, 2000 A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered *50*(4)**:** 211–223.

Rao, C. R., 1948 Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Proc. Cambridge Philos. Soc. **44:** 50–57.

Rao, C. R., 2002, Second Edition)*Linear Statistical Inference*. New York: Wiley.

Rao, C. R. and J. S. Poti, 1946 On locally most powerful tests when alternatives are one sided. Sankhya **7:** 439–440.

Sengul, H., S. Bhattacharjee, E. Feingold, and D. E. Weeks, 2007 The elusive goal of pedigree weights. Genet Epidemiol *31*(1)**:** 51–65.

Sham, P. C. and S. Purcell, 2001 Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. Am J Hum Genet *68*(6)**:** 1527–1532.

Sham, P. C., S. Purcell, S. S. Cherny, and G. R. Abecasis, 2002 Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet *71*(2)**:** 238–253.

Sham, P. C., J. H. Zhao, S. S. Cherny, and J. K. Hewitt, 2000 Variance-Components QTL linkage analysis of selected and non-normal samples: conditioning on trait values. Genet Epidemiol **19 Suppl 1:** S22–8.

Shih, M. C. and A. S. Whittemore, 2002 Tests for genetic association using family data. Genet Epidemiol *22*(2)**:** 128–145.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens, 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet *52*(3)**:** 506–516.

Szatkiewicz, J. P. and E. Feingold, 2004 A powerful and robust new linkage statistic for discordant sibling pairs. Am J Hum Genet *75*(5)**:** 906–909.

Szatkiewicz, J. P., K. T.Cuenco, and E. Feingold, 2003 Recent advances in human quantitative-trait-locus mapping: comparison of methods for discordant sibling pairs. Am J Hum Genet *73*(4)**:** 874–885.

Tang, H., 2000 Using variance components to map quantitative trait loci in humans. Ph. D. thesis, Stanford University.

Tang, H. and D. Siegmund, 2001 Mapping quantitative trait loci in oligogenic models. Biostatistics *2*(2)**:** 147–162.

T.Cuenco, K., J. P. Szatkiewicz, and E. Feingold, 2003 Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. Am J Hum Genet *73*(4)**:** 863–873.

Terwilliger, J. D. and J. Ott, 1992 A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered *42*(6)**:** 337–346.

Wang, K., 2002 Efficient score statistics for mapping quantitative trait loci with extended pedigrees. Hum Hered *54*(2): 57–68.

Wang, K., 2005 A likelihood approach for quantitative-trait-locus mapping with selected pedigrees. Biometrics *61*(2): 465–473.

Whittemore, A. S. and J. Halpern, 2003 Genetic association tests for family data with missing parental genotypes: a comparison. Genet Epidemiol *25*(1): 80–91.