Stony Brook University

A family-based likelihood ratio test for general pedigree structures that allows for missing data and genotyping errors

A Dissertation Presented

by

Yang Yang

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

December 2007

Stony Brook University

The Graduate School

Yang Yang

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend

acceptance of this dissertation.

Stephen J. Finch – Dissertation Advisor
Professor, Applied Mathematics & Statistics

John Reinitz – Chairperson of Defense
Professor, Applied Mathematics & Statistics

Wei Zhu
Professor, Applied Mathematics & Statistics

Derek Gordon
Associate Professor, Department of Genetics
Rutgers, The State University of New Jersey

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

A family-based likelihood ratio test for general pedigree structures that allows for missing data and genotyping errors

by

Yang Yang

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Statistics)

Stony Brook University

2007

The purpose of this work is to design a likelihood ratio test (LRT) that uses the information of both affected and unaffected individuals from a general pedigree to test association between marker and disease. The null hypothesis is that of equal marker penetrances, and the alternative hypothesis implies the presence of both allelic association and linkage between the disease and marker loci. The test is based on a conditional likelihood, which is a product of two factors: the first factor, $L_{Founder}$, uses founder's genotypes and phenotypes to estimate population frequencies of marker genotypes. The second factor, $L_{Nonfounder}$, evaluates disequilibrium in transmission of marker alleles from parents to offspring. The test statistic built on this conditional likelihood allows for two problems: (1) missing parental genotypes, and (2) random genotyping errors. Derivations of the conditional likelihoods are given for trios (two parents and a child), general nuclear families, multiple-marriage nuclear families, and zero-looped three- and four-generation pedigrees. For example, the following scenarios are considered for a general nuclear family: complete parental genotype data and no genotyping errors; only one genotyped parent and no genotyping errors; no parental genotype data and no genotyping errors; and with genotyping errors in the previous three scenarios. A robust algorithm grid-UOBYQA is used to locate log-likelihood maxima under the null and alternative hypotheses as well as to estimate marker penetrances and population genotype frequencies.

The results of a null simulation study suggest that the test statistic appears to follow a central chi-square distribution with one degree of freedom under the null hypothesis, even in the presence of missing data and genotyping errors. The power comparison based on a 2^3 factorial design shows that this LRT is more powerful than the original TDT, even when 20% genotypes in trios are missing and 1% genotypes are mistyped. Including the information of unaffected children in the likelihood calculation appears to increase the power to test marker-disease association. Finally, the application of this LRT to an idiopathic scoliosis dataset and a psoriasis dataset successfully identifies the significant associations between the markers and the disease that were previously published.

I dedicate this work to my parents Guangming Yang and Jinfeng Wang, and my fiancé, Maohua Lu

Table of Contents

List of Figures

List of Tables

Acknowledgments

I want to express my gratitude to all those who helped and supported me to complete my graduate study in Stony Brook University.

I am deeply indebted to my supervisor Dr. Stephen J. Finch. It is my fortune to have him as a dissertation advisor. He gave me the freedom to choose my research topic. He won me an honored opportunity to present in a biostatistics seminar in Columbia University. He put great effort into revising my dissertation. His patient guidance, insightful ideas, and inspiring encouragement gave me possibilities to finish my PhD study within only two and a half years.

My colleague, Dr. Derek Gordon, who is on top of my dissertation topic, has been always there to listen and give advice. He provided such a wonderful opportunity for me to collaborate on an invited paper from Human Heredity. He helped me set up the dissertation outline and provided two real datasets for the application of my likelihood ratio test. I am also thankful to him for reasoning my reports, commenting on my writings, and helping me understand and enrich my ideas.

I would like to thank the other two committee members, Dr. John Reinitz and Dr. Wei Zhu, who gave me permission to switch my dissertation topic. They helped me refine my dissertation writing. They cast insights and gave constructive criticisms during my preliminary exam and dissertation defense.

I am thankful to Dr. Nancy Mendell, who taught me two important advanced courses in statistics, and helped me a lot in my first paper in biostatistics. I am also thankful to Dr. Melody S. Goodman for her guidance during the infant mortality project. She also gave me full support in my career development.

Another professor I would like express my gratitude is Dr. Xiaolin Li, the graduate director of the Applied Mathematics and Statistics (AMS). He has been helping and supporting me since my graduate application. He made me more determined to further my graduate study in the United State during the interview for the graduate application in China. He encouraged me to be an instructor in my second semester, which helped me improve my oral English. He also gave practical suggestions on how to define my career path.

I also want to thank my friends to make my life full of happiness. Their support and encouragement helped me overcome the difficulties in my life and study in the United States. I cherish their friendships and deeply appreciate their confidence with me.

The last but not the least, I would like to express my heart-felt gratitude to my family. I warmly appreciate the consideration and understanding from them. Without their love, none of this would be possible. Specifically, I would like to thank my fiancé, Maohua Lu, for his love, support and patience in my everyday life in the United State.

Chapter 1 Introduction

1.1 History of transmission disequilibrium test and the other family-based tests

Spielman et al. (1993) proposed the transmission disequilibrium test (TDT), which was designed to test for linkage between a genetic marker and a disease-susceptibility locus (DSL) for a trait of interest, provided that there is allelic association. Allelic association (or linkage disequilibrium) is defined as the excessive co-occurrence of certain combinations of alleles in the same gamete because of tight linkage, or for other reasons (Sham, 1997). The TDT uses data from case-parent trios to evaluate the transmission of the associated marker allele from a heterozygous parent to an affected offspring. Under the null hypothesis of no linkage in the presence of allelic association, the number of alleles that are transmitted to the affected offspring is determined by Mendel's law. If the observed number of the transmitted alleles is significantly different from the number of those expected in Mendelian transmissions, a DSL appears to be associated and closely linked to the marker locus.

Since both linkage and allelic association between the marker locus and the DSL have to be present for the TDT to reject the null hypothesis, the TDT is also valid as a test of allelic association for case-parent trios provided that there is linkage. The linkage analysis typically identifies large candidate regions, while the evidence of allelic association in the presence of linkage may indicate which markers in the region are closest to a disease locus (Martin et al, 1997). This makes the TDT more valuable than linkage studies in pinpointing a narrower region where a DSL might lie. As a test of allelic association, the TDT is particularly suited for markers that may be at a DSL or very close to a DSL (Lander and Kruglyak, 1995; Risch and Merikangas, 1996). The TDT is not sensitive to the allelic association caused by admixture and/or population stratification (Spielman and Ewens 1998). As a nonparametric test, the TDT is robust to misspecification of the disease model or trait distribution (Laird and Lange, 2006).

The family-based design for TDT uses complete and errorless genotype data from the case-parent trios. To extend the family-based test to more general situations in linkage and association studies, numerous methodological extensions, as reviewed in Laird and Lange (2006), have been developed to allow for: specific mode of inheritance, arbitrary pedigree structures, complex phenotypes, missing parental genotypes and genotyping errors.

Schaid and his colleagues (Schaid and Sommer, 1994; Schaid, 1996) examined the power of the association tests under different genetic mechanisms (for example, dominant, recessive, and multiplicative mode of inheritance) leading to disease. These results demonstrate substantial gains in power for statistical tests designed to detect specific genetic mechanisms. The application of these tests was limited to independent case-parent trios. Martin et al. (1997) proposed two test statistics that focus on the set of transmissions from a parent to his/her affected offspring, rather than focusing on the individual transmissions to each offspring. They explored the test statistics for independent nuclear families with two affected offspring. Their tests are valid under the null hypothesis of no allelic association or no linkage, and generally are more powerful than the original TDT. Laird and her colleagues (Rabinowitz and Laird, 2000; Laird et al, 2000) developed a broad class of family-based association tests (FBAT) that adjust for admixture for either dichotomous or complex phenotypes.

The FBAT score statistic is based on the covariance of genotype and phenotype. Although genotypes of unaffected children are used to infer parental genotypes when parental genotypes are incomplete, the genotypes of the unaffected children are not incorporated in the score statistic. Also, under the null hypothesis of no linkage and no allelic association, the FBAT does not provide a valid test for allelic association in the presence of linkage for general nuclear families beyond trios. Therefore, Lake et al. (2000) updated the FBAT by incorporating an empirical variance in the score statistic to provide a valid test for allelic association. Under such a scenario, the null hypothesis of FBAT becomes no allelic association in the presence of linkage. Allison (1997) developed five tests for use with quantitative phenotypes such as body-mass index or blood pressure. These tests are based on the assumption that the residual distribution is normal or the sample size is large, allowing reliance on the central limit theorem. The test for quantitative phenotypes proposed by Rabinowitz (1997) needs no parametric assumptions on the distribution of the traits.

One of the most important issues regarding robustness of the family-based tests is incomplete parental genotype data. When one or both parental genotypes are missing, the resulting trio with incomplete genotype data must be discarded to ensure validity of the TDT, thereby sacrificing information. Curtis and Sham (1995) showed that the computation of the TDT statistic on trios in which one parental genotype is unknown increases the type I error rate of the statistic. Spielman and Ewens (1998) proposed the S-TDT that extends the original TDT to multiplex nuclear families whose parental genotypes are unknown. It compares the marker genotypes in affected and unaffected sibs instead of using marker data from their parents. However, there is a requirement on the sib-ship configuration when using the S-TDT. The smallest sib-ships that can give information for the S-TDT should contain exactly one affected and one unaffected sib, with different marker genotypes. Sun et al. (1999) proposed the 1-TDT that uses genotypes of affected children and only one available parent for each affected child. Weinberg (1999) generalized the work by Schaid and Sommer (1993) and set the missing parental genotype problem in a likelihood framework. Her likelihood ratio test (LRT) based on a log-linear model for genetic data is not sensitive to allelic association that is due to genetic admixture and is robust enough to maintain good power. Under a strict null hypothesis that the allele under study is neither linked to nor associated with the disease, the relative risks associated with inheriting one or more copies of the variant allele equals 1. When used as a test of allelic association, her LRT can be regarded as an alternative to the TDT. However, her LRT only considered case-parent trios.

Another issue for family-based tests is the presence of genotyping errors. Gordon et al. (2001) demonstrated that, when the TDT is applied to data in which Mendelian-inconsistent trios are removed, the detected genotyping errors can significantly increase the type I error rate. Their simulation showed that random genotyping errors that result in Mendelian-consistent genotype data for trios also cause an increase in type I error when their data are analyzed with the TDT. Therefore, they introduced TDTae, a family-based likelihood method allowing for random errors in the genotype data of trios. Considering both the missing parental genotype data and the genotyping errors, Gordon et al. (2004) extended the TDTae to involve general pedigrees. It is valid to test for linkage in the presence of allelic association. More recently, Cheng and Chen (2007) proposed a simple family-based association test that is not only robust against population stratification, but is also robust against genotyping error with error rates varying across families. However, these extensions of the TDT that allow for genotyping errors consider only affected offspring in the

families.

1.2 Brief introduction of my work

Among the more recent developments in family-based test is including estimated penetrance values for general pedigrees (Lange et al., 2005). The concept is to use all phenotype and genotype information in the pedigree rather than just using genotype information on affected children. The purpose of this work is the development of a likelihood-based method that uses information of both affected and unaffected individuals in a general pedigree and allows for random genotyping errors. It tests for marker-disease association by penetrance estimation, and is robust to missing genotype and/or phenotype data and random genotyping errors in general pedigrees. This test can be used for candidate-gene studies or a genome-wide association studies. It is also valid as a test of linkage in the presence of allelic association.

First, I derive the likelihood functions under all possible scenarios for trios, nuclear families, and three- and four-generation pedigrees. Based on these likelihood functions, I apply the grid-UOBYQA algorithm to locate the maximum log-likelihood under each hypothesis. To assess the null distribution of the test statistic and the type I errors, I perform null simulations on different types of families. Then I compare the power of the original TDT and this LRT with a 2^3 factorial design by Monte-Carlo simulation. Finally I apply the LRT to two previously published genetic studies and compare the results with those obtained by other family-based tests.

Chapter 2 includes comprehensive derivations of likelihood functions, null simulation, power calculation, and information of two real datasets. Chapter 3 introduces the grid-UOBYQA algorithm and describes the two-step search procedure. Chapter 4 lists the results of null simulation, power comparison, and the real applications of this LRT. Chapter 5 discusses the likelihood functions, missing parental genotype problem, genotyping errors, and maximization algorithms.

Chapter 2 Methods

This chapter begins with the notation that will be used in the subsequent chapters. This work only considers the bi-allelic situation, so that the disease-susceptibility locus (DSL), with alleles d_i coded as + (low-risk) or *d* (high-risk), has three possible genotypes: ++, +*d*, and *dd*. The bi-allelic marker locus with alleles m_j coded as *a* or *b*, also has three possible genotypes: *aa*, *ab*, and *bb*.

Genetic parameters

g_{DSL} = genotype of one individual at a DSL.

g = genotype of one individual at a marker locus, with the coding 0, 1, or 2 defined as the number of *b* alleles in the marker genotype. If the marker genotype is unknown, $g = Miss$.

$\vec{g}$ = the set of marker genotypes of a family.

$\vec{g}_C$ = the set of children's marker genotypes of a family.

$\vec{g}_{Obs}$ = the set of observed marker genotypes of a family.

$\vec{g}_{True}$ = one possible set of consistent marker genotypes of a family corrected from $\vec{g}_{Obs}$ with one Mendelian inconsistency.

$\vec{G}$ = genotypes of multiple families involved in the LRT.

p = phenotype, or the affection status of one individual, with $p = A$ for an individual being affected, $p = U$ for an individual being unaffected and $p = Miss$ for an individual with missing affection status.

$\vec{p}$ = the set of phenotypes of a family.

$\vec{p}_C$ = the set of children's phenotypes of a family.

$\vec{P}$ = phenotypes of multiple families involved in the LRT.

f_i = disease penetrance, defined as the probability of an individual being affected given that his/her genotype at the DSL is i ($i = 0$ for the ++ genotype, $i = 1$ for the $+d$ genotype, and $i = 2$ for the dd genotype).

R_i = genotype relative risk at the DSL. $R_1 = f_1 / f_0$ and $R_2 = f_2 / f_0$, where f_0 is the reference disease penetrance.

ϕ_i = marker penetrance (or marker effect), defined as the probability of an individual being affected given that his/her genotype at the marker locus is *i* (Nielsen and Weir, 2001) ($i = 0$ for the *aa* genotype, $i = 1$ for the *ab* genotype, and $i = 2$ for the *bb* genotype).

$\overline{R}_i$ = genotype relative risk at the marker locus. $\overline{R}_2 = \phi_1 / \phi_0$ and $\overline{R}_2 = \phi_2 / \phi_0$, where ϕ_0 is the reference marker genotype penetrance.

π_i = population frequency of a marker genotype. $i = 0$ for the *aa* genotype, $i = 1$ for the *ab* genotype, and $i = 2$ for the *bb* genotype. Also, $\pi_0 + \pi_1 + \pi_2 = 1$.

D = a measure of linkage disequilibrium (LD) (Robbins, 1918), defined as $D = \Pr(m_1 = a, d_1 = +) - \Pr(m_1 = a)\Pr(d_1 = +)$. It measures the deviation of the observed haplotype frequencies from the expected frequencies.

D' = the proportion of maximum LD (Lewontin and Kojima, 1960). $D' = D / D_{\max}$ when $D \geq 0$ and $D' = D / D_{\min}$ when $D < 0$, where $D_{\max} = \min\{\Pr(d)\Pr(a), \Pr(+)\Pr(b)\}$ and

$D_{\min} = \max\{-\Pr(d)\Pr(b), -\Pr(+)\Pr(a)\}$. It is a normalized value that lies between 0 and 1.

Notation for family members

$FF, MF, FM, MM, F, M, C (or \vec{C}) =$ paternal grandfather, paternal grandmother, maternal grandfather, maternal grandmother, father, mother and a child (or a set of all children), respectively. This work uses subscripts for a given set to indicate individual members of the set. For example, if $\vec{C}$ represents the set of children in a nuclear family, then C_r refers to the r^{th} child.

$n_{i.C}$ = number of children with marker genotype i in a family. $i = 0$ for the aa genotype, $i = 1$ for the ab genotype, and $i = 2$ for the bb genotype.

Notation for frequently used likelihood functions

L_{Trios} = the likelihood factor from a trio

$L_{Trios.a}$ = the likelihood factor from a trio with complete parental genotypes

$L_{Trios.b}$ = the likelihood factor from a trio with one untyped parent and one typed parent

$L_{Trios.c}$ = the likelihood factor from a trio with two untyped parents

$L_{Nuclear}$ = the likelihood factor from a nuclear family

$L_{Nuclear.a}$ = the likelihood factor from a nuclear family with complete parental genotypes

$L_{Nuclear.b}$ = the likelihood factor from a nuclear family with one untyped parent and one typed parent

$L_{Nuclear.c}$ = the likelihood factor from a nuclear family with two untyped parents

Genotyping error parameters

η = probability for a homozygote incorrectly coded as a heterozygote.

γ = probability for a heterozygote incorrectly coded as a homozygote.

ε = error rate of the simplified DSB error model (Douglas et al., 2002), in which $\varepsilon = \eta = \gamma$.

2.1 Null hypothesis and assumptions

2.1.1 Null hypothesis

This family-based likelihood method is designed to test the association of a candidate gene (or a marker) and disease. Under the null hypothesis of no association, the marker penetrances should be equal ($\phi_0 = \phi_1 = \phi_2$). Rejecting the null hypothesis implies an association of marker with disease, which exists only when the marker is both linked and associated with a DSL affecting the trait. (Schaid and Sommer, 1993).

Because both linkage and allelic association should be present to reject the null hypothesis, this likelihood method can be used to (1) test linkage or allelic association for candidate-gene or genome-wide association studies, or to (2) test linkage in the presence of allelic association for the follow up of case-control association studies. However, like the original TDT, this likelihood method may not be valid as a test of allelic association in the presence of linkage for families with more than one affected

offspring (Martin et al., 2003).

2.1.2 Assumptions

The following assumptions are given in this work:
a. Hardy-Weinberg equilibrium (HWE)

If p is defined as the frequency of allele a and q as the frequency of another allele b for a trait controlled by a pair of alleles, then HWE will give Hardy-Weinberg proportions $p(aa) = p^2$, $p(ab) = 2pq$, $p(bb) = q^2$ (Hardy, 1908; Weinberg, 1908).

The implications of HWE are: (1) the frequencies of alleles in a population will remain constant from generation to generation; (2) the genotype frequencies will remain constant from generation to generation; (3) the Hardy-Weinberg proportions will be reached in a single generation of random mating.

b. Random mating between parental gametes

Let $d_1m_1 : d_2m_2$ denotes a parental gamete pair. Under the assumption of random mating between parental gametes, the joint probability of the gamete pair can be decomposed into the product: $\Pr(d_1m_1 : d_2m_2) = \Pr(d_1m_1)\Pr(d_2m_2)$ (Martin et al, 1998).

c. Multiplicative mode of inheritance

This work assumes multiplicative mode of inheritance at the DSL: $f_1^2 = f_0 f_2$ (or equivalently, $R_2 = R_1^2$). Multiplicative mode of inheritance is also known as the log-additive gene model (Schaid and Sommer, 1994). Fitting a multiplicative model is a reasonable and simple start for this association test, since in general, the change in risk on the 'induced' relative risk is approximately multiplicative regardless of the mode of inheritance at the true disease locus (e.g., dominant or recessive) (Siegmund and Gauderman, 2001). Under assumptions (a) and (b), multiplicative penetrances at the DSL will result in multiplicative penetrances at the marker locus: $\phi_1^2 = \phi_0\phi_2$ (see Appendix I). This reduces the number of parameters to be estimated in this work. Under the multiplicative model, the null test statistic should follow a central χ^2 distribution with one degree of freedom.

Note that this likelihood method allows the flexibility to remove this assumption.

d. No parental imprinting

Parental imprinting describes the phenomenon of differential gene function based on whether the transmission of an allele was from the mother or the father (Chaudhuri and Messing, 1994). This work assumes that there is no parental imprinting. Let d_1d_2 denote the disease genotype of a child, where d_1 is transmitted from father, and d_2 is transmitted from mother. If there is no parental imprinting, disease genotypes $+d$ and $d+$ have the same gene effect on the phenotype of the child.

e. Independence of parental genotypes (i.e. no assortative mating)

This work assumes that paternal genotype is independent of maternal genotype. That is, $\Pr(g_F, g_M) = \Pr(g_F)\Pr(g_M)$. The assumption is made to reduce the number of parameters to be estimated.

f. Independence of marker genotypes and phenotypes

The individuals' phenotypes are conditionally independent given their genotypes. Children's genotypes are conditionally independent given the parental mating type and the children's phenotypes.

This assumption is required to make the likelihood function valid for families with multiple affected sibs (Schaid and Sommer, 1993).

g. Missing at random (MAR)

The missing data for a variable X are "missing at random" if the probability of missing data on X is unrelated to the value of X (Little and Rubin, 2002). This work assumes that phenotypes and marker genotypes are MAR, so that the probability of missing data on phenotype or marker genotype is unrelated to the values of phenotype or marker genotype. The MAR assumption also indicates that an individual's missing phenotype information is independent of missing marker genotype information, and vice versa.

h. Each nuclear family contains at most one Mendelian inconsistency

To simplify the likelihood computation for pedigrees with genotyping errors, Ehm et al. (1996) assumes at most one error per pedigree. Douglas et al. (2002) calculated the error rates in nuclear families by assuming that there is exactly one genotyping error per family. To allow small to moderate mistyping rates, this work assumes that each nuclear family (or a nuclear family decomposed from a general pedigree) contains at most one Mendelian inconsistency. Gordon et al. (1999, 2000) calculated the error detection rates for Mendelian inconsistent pedigrees. They found that the error detection rates are very low.

i. Independent and random genotyping errors, no phenotyping errors

As in Gordon and Ott (2001) and Gordon et al. (2004), this work assumes that genotyping errors are introduced randomly and independently into alleles at a bi-allelic locus and that there are no phenotyping errors.

2.2 The likelihood function of a family possibly with missing data but with no genotyping errors

In this work, the likelihood functions for nuclear families with complete parental genotypes are similar to those previously published (Tu et al., 2000; Whittemore and Tu, 2000) (See Chapter 5 Discussion: 5.1 The likelihood function). I use the complete-data likelihood conditional on the observed data to compute the likelihood factor for a nuclear family with incomplete parental genotypes.

Let θ denote the parameter (or a vector of parameters) of interest, Y_{obs} denote the observed data and Y_{mis} the missing data, and $L(Y_{obs}, Y_{mis}; \theta)$ denote the complete-data likelihood that would have been constructed had there been no missing data. Conditional on Y_{obs}, Y_{mis} takes on J possible values: $y_{mis.1} \ldots y_{mis.j} \ldots y_{mis.J}$, where J depends on the family structure and missing data pattern. The complete-data likelihood conditional on the observed data (Lyles et al., 2001; Schafer and Graham, 2000) for discrete missing data problems can be written as:

$$\sum_{j=1}^{J} L(Y_{obs}, Y_{mis} = y_{mis.j}; \theta) \Pr(Y_{mis} = y_{mis.j} \mid Y_{obs}). \qquad (1)$$

In this work, $\theta = \{\pi_0, \pi_1, \phi_0, \phi_2\}$, Y_{obs} denotes the observed genotype and phenotype data, and Y_{mis} denote the missing parental genotype data. Note that estimates of π_2 and ϕ_1 can be inferred from $\pi_0 + \pi_1 + \pi_2 = 1$ and $\phi_1^2 = \phi_0\phi_2$.

As discussed in Section 5.2 of Chapter 5, the likelihoods as in equation (1) are not restricted to MAR problems (Schafer and Graham, 2000).

2.2.1 Conditional likelihood function of trios

Complete parental genotype data

From assumption (f) that one individual's phenotype is independent of the other individuals' phenotypes conditional on their genotypes, and assumption (e) that there is no assortative mating, the conditional likelihood function for one trio with complete parental marker genotype data is:

$$\begin{aligned} L_{Trio.a} &= \Pr(g_F, g_M \mid p_F, p_M)\Pr(g_C \mid p_C, g_F, g_M) \\ &= \Pr(g_F \mid p_F)\Pr(g_M \mid p_M)\Pr(g_C \mid p_C, g_F, g_M) \\ &= \frac{\Pr(g_F)\Pr(g_M)\Pr(p_F \mid g_F)\Pr(p_M \mid g_M)}{\Pr(p_F)\Pr(p_M)}\Pr(g_C \mid p_C, g_F, g_M). \end{aligned} \tag{2}$$

The probability of an individual being affected is:

$$\Pr(p = A) = \sum_{i=0}^{2}\Pr(A \mid g = i)\Pr(g = i) = \sum_{i=0}^{2}\phi_i\pi_i ,$$

and the probability of an individual being unaffected is:

$$\Pr(p = U) = \sum_{i=0}^{2}\Pr(U \mid g = i)\Pr(g = i) = \sum_{i=0}^{2}(1-\phi_i)\pi_i .$$

For an individual with genotype data but missing affection status, under the assumption of MAR, $\Pr(g_F \mid p_F = Miss) = \Pr(g_F)$ or $\Pr(g_M \mid p_M = Miss) = \Pr(g_M)$ if the individual is a parent, and $\Pr(g_C \mid p_C = Miss, g_F, g_M) = \Pr(g_C \mid g_F, g_M)$ if a child.

There are four possible types of children based on the availability of genotype and phenotype information.

In the first type, the child has both genotype and phenotype data. If the child is affected and $g_C = i$, where $i = 0,1,or\,2$

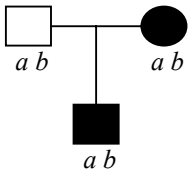
$$\begin{aligned} &\Pr(g_C = i \mid p_C = A, g_F, g_M) \\ &= \frac{\Pr(p_C = A, g_C = i, g_F, g_M)}{\Pr(p_C = A, g_F, g_M)} \\ &= \frac{\Pr(p_C = A \mid g_C = i, g_F, g_M)\Pr(g_C = i \mid g_F, g_M)\Pr(g_F, g_M)}{\sum_{j=0}^{2}\Pr(p_C = A \mid g_C = j, g_F, g_M)\Pr(g_C = j \mid g_F, g_M)\Pr(g_F, g_M)} \\ &= \frac{\Pr(p_C = A \mid g_C = i)\Pr(g_C = i \mid g_F, g_M)}{\sum_{j=0}^{2}\Pr(p_C = A \mid g_C = j)\Pr(g_C = j \mid g_F, g_M)} \\ &= \frac{\phi_i \Pr(g_C = i \mid g_F, g_M)}{\sum_{j=0}^{2}\phi_j \Pr(g_C = j \mid g_F, g_M)}. \end{aligned} \tag{3}$$

In equation (3), $\Pr(p_C = A \mid g_C = i, g_F, g_M) = \Pr(p_C = A \mid g_C = i)$ from the assumption

of independence of marker genotypes and phenotypes. The conditional probabilities $\Pr(g_C \mid g_F, g_M)$ often referred to as transmission probabilities (Demenais and Elston, 1981), are listed in Table 1.

As an example, Figure 1 shows a case-parent trio, with marker genotypes $g_F = 1, g_M = 1$, and $g_C = 1$. The left white square (male) indicates an unaffected father, the right black circle (female) indicates an affected mother, and the middle black square indicates an affected male child. The two letters (in Figure 1, '*a b*') below each square or circle are two marker alleles for each individual.

Figure 1: A case-parent trio with complete parental genotypes



The conditional probability of the child's genotype given his affection status is:

$$\Pr(g_C = 1 \mid p_C = A, g_F = g_M = 1)$$

$$= \frac{\phi_1 \Pr(g_C = 1 \mid g_F = g_M = 1)}{\phi_0 \Pr(g_C = 0 \mid g_F = g_M = 1) + \phi_1 \Pr(g_C = 1 \mid g_F = g_M = 1) + \phi_2 \Pr(g_C = 2 \mid g_F = g_M = 1)}$$

$$= \frac{\phi_1 \frac{1}{2}}{\phi_0 \frac{1}{4} + \phi_1 \frac{1}{2} + \phi_2 \frac{1}{4}} = \frac{2\phi_1}{\phi_0 + 2\phi_1 + \phi_2}.$$

The conditional probabilities $\Pr(g_C \mid p_C = A, g_F, g_M)$ for all parent-child genotype configurations are listed in Table 2.

Similarly, if the child is unaffected and $g_C = i$, for $i = 0,1,or\,2$, the conditional probability is given by:

$$\Pr(g_C = i \mid p_C = U, g_F, g_M) = \frac{(1-\phi_i)\Pr(g_C = i \mid g_F, g_M)}{\sum_{j=0}^{2}(1-\phi_j)\Pr(g_C = j \mid g_F, g_M)}. \quad (4)$$

The values in equation (4) for all parent-child genotype configurations are specified in Table 3.

In the second type, the child has genotype data but no phenotype data. Under the assumption of MAR, the likelihood factor from the child is

$$\Pr(g_C \mid p_C = Miss, g_F, g_M) = \Pr(g_C \mid g_F, g_M).$$

The conditional probabilities $\Pr(g_C \mid g_F, g_M)$ are given in Table 1.

In the third type, the child has phenotype data but no genotype data. Under the assumption of MAR, the marginal probability can be used for the observed data (Little and Rubin, 2002). The marginal likelihood factor contributed by the child without genotype data is:

$$\sum_{g_C} \Pr(g_C \mid p_C, g_F, g_M) = 1,$$

indicating that no information is contributed by a child with only phenotype data to this association test.

In the fourth type, the child has neither genotype nor phenotype data. It is obvious that a child with no genetic information has no contribution to the likelihood.

Incomplete parental genotype data: one parental genotype is missing

Consider a trio with one missing parental genotype, without loss of generality, I specify that the paternal genotype is missing. If the child's genotype is also unknown, the marginal likelihood factor of such trio is

$$\sum_{g_F}\sum_{g_C}\Pr(g_F\,|p_F)\Pr(g_M\mid p_M)\Pr(g_C\mid p_C,g_F,g_M)$$
$$=\Pr(g_M\mid p_M)\sum_{g_F}\Pr(g_F\mid p_F)\sum_{g_C}\Pr(g_C\mid p_C,g_F,g_M)=\Pr(g_M\mid p_M).$$

That is, such trio can only contribute its maternal information to the likelihood. From assumption (f) that the individuals' phenotypes are conditionally independent given their genotypes, I have $\Pr(g_F\mid g_M,g_C,p_F,p_M,p_C)=\Pr(g_F\mid g_M,g_C,p_F)$. If the child's genotype is observed but not the father's, the likelihood factor is

$$L_{Trio.b}=\sum_{Y_{mis}}L_{Trio.a}(\theta\mid Y_{obs},Y_{mis})\Pr(Y_{mis}\mid Y_{obs},\theta)$$
$$=\sum_{g_F}L_{Trio.a}(\theta\mid g_M,g_C,p_F,p_M,p_C,g_F)\Pr(g_F\mid g_M,g_C,p_F,p_M,p_C)$$
$$=\sum_{g_F}L_{Trio.a}(\theta\mid g_F,g_M,g_C,p_F,p_M,p_C)\Pr(g_F\mid g_M,g_C,p_F),$$

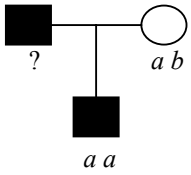
where $Y_{obs}=\{g_M,\vec{g}_{\bar{C}},p_F,p_M,\vec{p}_{\bar{C}}\}$, $Y_{mis}=\{g_F\}$ and

$$\Pr(g_F\mid g_M,g_C,p_F)=\frac{\Pr(g_F,g_M,g_C,p_F)}{\sum_i\Pr(g_F=i,g_M,g_C,p_F)}. \quad (5)$$

Note that the maternal and the child's phenotypes do not enter in equation (5) to compute the conditional probability $\Pr(g_F\mid g_M,g_C,p_F)$.

As an example, Figure 2 shows a case-parent trio with marker genotypes $g_F=Miss$, $g_M=1$, and $g_C=0$. The '?' below the black square indicates that the genotype of the affected father is unknown.

Figure 2: A case-parent trio with only maternal genotype



Based on the maternal and the child's genotypes, the missing paternal genotype must be either 0 or 1. To compute conditional probabilities $\Pr(g_F\mid g_M,g_C,p_F)$ for such a trio, I first compute

$$\Pr(g_F=0,g_M=1,g_C=0,p_F=A)$$
$$=\Pr(g_F=0)\Pr(p_F=A\mid g_F=0)\Pr(g_M=1)\Pr(g_C=0\mid g_F=0,g_M=1)$$
$$=\pi_0\phi_0\pi_1\frac{1}{2}=\frac{\pi_0\pi_1\phi_0}{2},\text{and}$$
$$\Pr(g_F=1,g_M=1,g_C=0,p_F=A)$$
$$=\Pr(g_F=1)\Pr(p_F=A\mid g_F=1)\Pr(g_M=1)\Pr(g_C=0\mid g_F=1,g_M=1)$$
$$=\pi_1\phi_1\pi_1\frac{1}{4}=\frac{\pi_1^2\phi_1}{4}.$$

Then based on equation (5), I have:

$$\Pr(g_F = 0 \mid g_M = 1, g_C = 0, p_F = A) = \frac{\pi_0\pi_1\phi_0 / 2}{\pi_0\pi_1\phi_0 / 2 + \pi_1^2\phi_1 / 4} = \frac{2\pi_0\phi_0}{2\pi_0\phi_0 + \pi_1\phi_1}, \text{and}$$

$$\Pr(g_F = 1 \mid g_M = 1, g_C = 0, p_F = A) = \frac{\pi_1^2\phi_1 / 4}{\pi_0\pi_1\phi_0 / 2 + \pi_1^2\phi_1 / 4} = \frac{\pi_1\phi_1}{2\pi_0\phi_0 + \pi_1\phi_1}.$$

In the event that the father were unaffected in the trio, I would have

$$\Pr(g_F = 0 \mid g_M = 1, g_C = 0, p_F = U) = \frac{2\pi_0(1-\phi_0)}{2\pi_0(1-\phi_0) + \pi_1(1-\phi_1)}, \text{and}$$

$$\Pr(g_F = 1 \mid g_M = 1, g_C = 0, p_F = U) = \frac{\pi_1(1-\phi_1)}{2\pi_0(1-\phi_0) + \pi_1(1-\phi_1)}.$$

In the event that the paternal phenotype were unknown in the trio,

$$\Pr(g_F = 0 \mid g_M = 1, g_C = 0, p_F = Miss) = \frac{2\pi_0}{2\pi_0 + \pi_1}, \text{and}$$

$$\Pr(g_F = 1 \mid g_M = 1, g_C = 0, p_F = Miss) = \frac{\pi_1}{2\pi_0 + \pi_1}$$

For any possible paternal phenotype, the conditional probabilities can be written as

$$\Pr(g_F = 0 \mid g_M = 1, g_C = 0, p_F) = \frac{2\pi_0\eta_0}{2\pi_0\eta_0 + \pi_1\eta_1}, \text{and}$$

$$\Pr(g_F = 1 \mid g_M = 1, g_C = 0, p_F) = \frac{\pi_1\eta_1}{2\pi_0\eta_0 + \pi_1\eta_1},$$

where

$$\eta_i = \begin{cases} \phi_i & if \ \ p_F = A \\ 1-\phi_i & if \ \ p_F = U \\ 1 & if \ \ p_F = Miss \end{cases}. \quad (6)$$

The conditional probabilities $\Pr(g_F \mid g_M, g_C, p_F)$ for an arbitrary trio with untyped father are listed in Table 4.

Incomplete parental genotype data: both parental genotypes are missing

Consider a trio without parental genotypes. If the child's genotype is also unknown, the marginal likelihood factor

$$\sum_{g_F}\sum_{g_M}\sum_{g_C} \Pr(g_F \mid p_F)\Pr(g_M \mid p_M)\Pr(g_C \mid p_C, g_F, g_M) = 1$$

This implies that such a trio cannot contribute any information to the likelihood. It is also the reason why this likelihood method only considers families with at least one genotyped individual. From assumption (f) that the individuals' phenotypes are conditionally independent given their genotypes, I have $\Pr(g_F, g_M \mid g_C, p_F, p_M, p_C) = \Pr(g_F, g_M \mid g_C, p_F, p_M)$. If the child's genotype data is available, the likelihood is

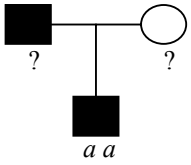
$$\begin{aligned} L_{Trio.c} &= \sum_{Y_{mis}} L_{Trio.a}(\theta \mid Y_{obs}, Y_{mis})\Pr(Y_{mis} \mid Y_{obs}) \\ &= \sum_{g_F}\sum_{g_M} L_{Trio.a}(\theta; g_C, p_F, p_M, p_C, g_F, g_M)\Pr(g_F, g_M \mid g_C, p_F, p_M, p_C) \\ &= \sum_{g_F}\sum_{g_M} L_{Trio.a}(\theta; g_F, g_M, g_C, p_F, p_M, p_C)\Pr(g_F, g_M \mid g_C, p_F, p_M) \end{aligned}$$

where $Y_{obs} = \{g_C, p_F, p_M, p_C\}$, $Y_{mis} = \{g_F, g_M\}$, and

$$\Pr(g_F, g_M \mid g_C, p_F, p_M) = \frac{\Pr(g_F, g_M, g_C, p_F, p_M)}{\sum_i \sum_j \Pr(g_F = i, g_M = j, g_C, p_F, p_M)}. \qquad (7)$$

As an example, Figure 3 shows a case-parent trio with $g_F = Miss$, $g_M = Miss$, and $g_C = 0$.

Figure 3: A case-parent trio with unknown parental genotypes



Based on the child's genotype $g_C = 0$, his parental genotypes $\{g_F, g_M\}$ must be {0, 0}, {0, 1}, {1, 0}, or {1, 1}. First I compute

$$\Pr(g_F = 0, g_M = 0, g_C = 0, p_F = A, p_M = U)$$
$$= P(g_F = 0, g_M = 0) P(p_F = A \mid g_F = 0) P(p_M = U \mid g_M = 0) \Pr(g_C = 0 \mid g_F = 0, g_M = 0)$$
$$= \pi_0^2 \phi_0 (1 - \phi_0) \cdot 1,$$
$$\Pr(g_F = 0, g_M = 1, g_C = 0, p_F = A, p_M = U)$$
$$= P(g_F = 0, g_M = 1) P(p_F = A \mid g_F = 0) P(p_M = U \mid g_M = 1) \Pr(g_C = 0 \mid g_F = 0, g_M = 1)$$
$$= \pi_0 \pi_1 \phi_0 (1 - \phi_1) \frac{1}{2},$$
$$\Pr(g_F = 1, g_M = 0, g_C = 0, p_F = A, p_M = U)$$
$$= P(g_F = 1, g_M = 0) P(p_F = A \mid g_F = 1) P(p_M = U \mid g_M = 0) \Pr(g_C = 0 \mid g_F = 1, g_M = 0)$$
$$= \pi_1 \pi_0 \phi_1 (1 - \phi_0) \frac{1}{2}, \text{ and}$$
$$\Pr(g_F = 1, g_M = 1, g_C = 0, p_F = A, p_M = U)$$
$$= P(g_F = 1, g_M = 1) P(p_F = A \mid g_F = 1) P(p_M = U \mid g_M = 1) \Pr(g_C = 0 \mid g_F = 1, g_M = 1)$$
$$= \pi_1^2 \phi_1 (1 - \phi_1) \frac{1}{4}.$$

Then based on equation (7), I have

$$\Pr(g_F = 0, g_M = 0 \mid g_C = 0, p_F = A, p_M = U)$$
$$= \frac{\pi_0^2 \phi_0 (1 - \phi_0)}{\pi_0^2 \phi_0 (1 - \phi_0) + \pi_0 \pi_1 \phi_0 (1 - \phi_1) \frac{1}{2} + \pi_1 \pi_0 \phi_1 (1 - \phi_0) \frac{1}{2} + \pi_1^2 \phi_1 (1 - \phi_1) \frac{1}{4}}$$
$$= \frac{4\pi_0^2 \phi_0 (1 - \phi_0)}{4\pi_0^2 \phi_0 (1 - \phi_0) + 2\pi_0 \pi_1 \phi_0 (1 - \phi_1) + 2\pi_1 \pi_0 \phi_1 (1 - \phi_0) + \pi_1^2 \phi_1 (1 - \phi_1)},$$
$$\Pr(g_F = 0, g_M = 1 \mid g_C = 0, p_F = A, p_M = U)$$
$$= \frac{2\pi_0 \pi_1 \phi_0 (1 - \phi_1)}{4\pi_0^2 \phi_0 (1 - \phi_0) + 2\pi_0 \pi_1 \phi_0 (1 - \phi_1) + 2\pi_1 \pi_0 \phi_1 (1 - \phi_0) + \pi_1^2 \phi_1 (1 - \phi_1)},$$

$$\Pr(g_F = 1, g_M = 0 \mid g_C = 0, p_F = A, p_M = U)$$
$$= \frac{2\pi_1\pi_0\phi_1(1-\phi_0)}{4\pi_0^2\phi_0(1-\phi_0) + 2\pi_0\pi_1\phi_0(1-\phi_1) + 2\pi_1\pi_0\phi_1(1-\phi_0) + \pi_1^2\phi_1(1-\phi_1)}, \text{ and}$$
$$\Pr(g_F = 1, g_M = 1 \mid g_C = 0, p_F = A, p_M = U)$$
$$= \frac{\pi_1^2\phi_1(1-\phi_1)}{4\pi_0^2\phi_0(1-\phi_0) + 2\pi_0\pi_1\phi_0(1-\phi_1) + 2\pi_1\pi_0\phi_1(1-\phi_0) + \pi_1^2\phi_1(1-\phi_1)}.$$

For any possible combination of parental phenotypes, the conditional probability $\Pr(g_F = 0, g_M = 0 \mid g_C = 0, p_F, p_M)$ for example, can be written as

$$\frac{4\pi_0^2\Theta(g=0,p_F)\Theta(g=0,p_M)}{4\pi_0^2\Theta(0,p_F)\Theta(0,p_M) + 2\pi_0\pi_1\Theta(0,p_F)\Theta(1,p_M) + 2\pi_1\pi_0\Theta(1,p_F)\Theta(0,p_M) + \pi_1^2\Theta(1,p_F)\Theta(1,p_M)}$$

where

$$\Theta(g=i,p) = \begin{cases} \phi_i & if\ p = A \\ 1-\phi_i & if\ p = U \\ 1 & if\ p = Miss \end{cases}, i = 0,1,2. \quad (8)$$

For example, if both parental phenotypes were unknown in the trio, I would have $\Theta(g_F = i, p_F = Miss) = 1$ and $\Theta(g_M = i, p_M = Miss) = 1, i = 0,1,2$. Then

$$\Pr(g_F = 0, g_M = 0 \mid g_C = 0, p_F = Miss, p_M = Miss) = \frac{2\pi_0^2}{2\pi_0^2 + 2\pi_0\pi_1 + 2^{-1}\pi_1^2},$$

$$\Pr(g_F = 0, g_M = 1 \mid g_C = 0, p_F = Miss, p_M = Miss) = \frac{\pi_0\pi_1}{2\pi_0^2 + 2\pi_0\pi_1 + 2^{-1}\pi_1^2},$$

$$\Pr(g_F = 1, g_M = 0 \mid g_C = 0, p_F = Miss, p_M = Miss) = \frac{\pi_1\pi_0}{2\pi_0^2 + 2\pi_0\pi_1 + 2^{-1}\pi_1^2}, \text{ and}$$

$$\Pr(g_F = 1, g_M = 1 \mid g_C = 0, p_F = Miss, p_M = Miss) = \frac{2^{-1}\pi_1^2}{2\pi_0^2 + 2\pi_0\pi_1 + 2^{-1}\pi_1^2}.$$

Table 5 lists the conditional probabilities $\Pr(g_F, g_M \mid g_C, p_F, p_M)$ for an arbitrary trio without parental information. Since there are several conditional probability situations that need to be considered for a full development of this method, I list the conditional probabilities $\Pr(g_F, g_M \mid g_C, p_F, p_M)$ in Table 6.

2.2.2 Conditional likelihood function of general nuclear families

Complete parental genotype data

Similar to equation (2), the conditional likelihood function of a nuclear family with complete parental genotype data is

$$L_{Nuclear.a} = \Pr(g_F \mid p_F)\Pr(g_M \mid p_M)\Pr(\vec{g}_{\vec{C}} \mid \vec{p}_{\vec{C}}, g_F, g_M)$$
$$= \frac{\Pr(g_F)\Pr(g_M)\Pr(p_F \mid g_F)\Pr(p_M \mid g_M)}{\Pr(p_F)\Pr(p_M)} \prod_{\vec{C}=\{C_r\}} \Pr(g_{C_r} \mid p_{C_r}, g_F, g_M). \quad (9)$$

Here $\Pr(\vec{g}_{\vec{C}} \mid \vec{p}_{\vec{C}}, g_F, g_M) = \prod_{\vec{C}=\{C_r\}} \Pr(g_{C_r} \mid p_{C_r}, g_F, g_M)$ is from assumption (f) that the children's genotypes are conditionally independent given the children's phenotypes and the parental mating type.

Based on the availability of genotype and phenotype information, each child in

a nuclear family can be placed into one of four disjoint sets:

$\vec{C}_1$: Children with both genotype and phenotype data, where r is the index for affected children and s for unaffected children;

$\vec{C}_2$: Children with genotype data but no phenotype data, where t is the index;

$\vec{C}_3$: Children with phenotype data but no genotype data;

$\vec{C}_4$: Children with neither genotype nor phenotype data.

The likelihood factor from $\vec{C}_1$ is

$$\Pr(\vec{g}_{\vec{C}_1} \mid \vec{p}_{\vec{C}_1}, g_F, g_M) = \prod_r \Pr(g_{C_{1r}} \mid p_{C_{1r}} = A, g_F, g_M) \cdot \prod_s \Pr(g_{C_{1s}} \mid p_{C_{1s}} = U, g_F, g_M).$$

The likelihood factor from $\vec{C}_2$ is:

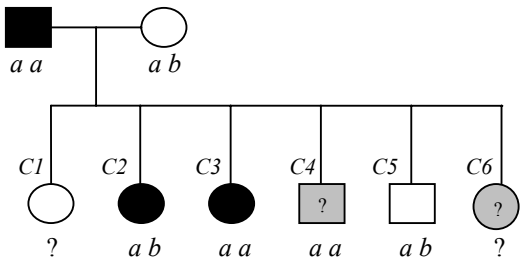
$$\Pr(\vec{g}_{\vec{C}_2} \mid \vec{p}_{\vec{C}_2}, g_F, g_M) = \prod_t \Pr(g_{C_{2t}} \mid p_{C_{2t}} = Miss, g_F, g_M)$$

$$= \prod_t \Pr(g_{C_{2t}} \mid g_F, g_M) \text{ (from MAR assumption).}$$

The likelihood factor from set $\vec{C}_3$ and $\vec{C}_4$ is 1, which implies that untyped children do not enter in the likelihood calculation.

For example, Figure 4 shows a large nuclear family with two genotyped parents and six children, four genotyped and two untyped. The father is affected with $g_F = 0$, and the mother is unaffected with $g_M = 1$. Note that the six children can be partitioned into four sets $\vec{C}_1, \ldots, \vec{C}_4$, where $\vec{C}_1 = \{C_2, C_3, C_5\}$ with $r = 2,3$ and $s = 5$, $\vec{C}_2 = \{C_4\}$ with $t = 4$, $\vec{C}_3 = \{C_1\}$, and $\vec{C}_4 = \{C_6\}$. The grey square and the grey circle with a '?' in the middle indicate that the affection status of these two children is unknown.

Figure 4: A nuclear family of size 8 with complete parental genotypes



The conditional likelihood factor from the parents (founders): $L_{Founder} =$

$$\Pr(g_F \mid p_F)\Pr(g_M \mid p_M) = \frac{\Pr(g_F = 0)\Pr(g_M = 1)\Pr(p_F = A \mid g_F = 0)\Pr(p_M = U \mid g_M = 1)}{\Pr(p_F = A)\Pr(p_M = U)}$$

$$= \frac{\pi_0\pi_1\phi_0(1-\phi_1)}{(\pi_0\phi_0 + \pi_1\phi_1 + \pi_2\phi_2)(1 - \pi_0\phi_0 - \pi_1\phi_1 - \pi_2\phi_2)}.$$

The conditional likelihood factor from the children (nonfounders): $L_{Nonfounder} =$

$$\prod_r \Pr(g_{C_{1r}} \mid p_{C_{1r}} = A, g_F, g_M) \prod_s \Pr(g_{C_{1s}} \mid p_{C_{1s}} = U, g_F, g_M) \cdot \prod_t \Pr(g_{C_{2t}} \mid g_F, g_M) \cdot 1 \cdot 1$$

$$= \left[\Pr(g_{C_2} = 1 \mid p_{C_2} = A, g_F, g_M)\Pr(g_{C_3} = 0 \mid p_{C_3} = A, g_F, g_M)\Pr(g_{C_5} = 1 \mid p_{C_5} = U, g_F, g_M)\right]$$

$$\cdot \Pr(g_{C_4} = 0 \mid g_F, g_M)$$

$$= \left[\frac{\phi_0\phi_1(1-\phi_1)}{(\phi_0 + \phi_1)^2(2 - \phi_0 - \phi_1)}\right] \cdot \frac{1}{2} = \frac{\phi_0\phi_1(1-\phi_1)}{2(\phi_0 + \phi_1)^2(2 - \phi_0 - \phi_1)}.$$

Therefore, the likelihood factor from the nuclear family with complete parental data
$L_{Nuclear.a} = L_{Founder} \cdot L_{Nonfounder}$

$$= \frac{1}{2}\pi_0\pi_1 \frac{\phi_0^2\phi_1(1-\phi_1)^2}{(\pi_0\phi_0+\pi_1\phi_1+\pi_2\phi_2)(1-\pi_0\phi_0-\pi_1\phi_1-\pi_2\phi_2)(\phi_0+\phi_1)^2(2-\phi_0-\phi_1)}.$$

Using the partitioning of children into four disjoint sets $\vec{C}_1,...,\vec{C}_4$, equation (9) can be written as:

$$L_{Nuclear.a} = \Pr(g_F \mid p_F)\Pr(g_M \mid p_M)\cdot\prod_r \Pr(g_{C_{1_r}} \mid p_{C_{1_r}} = A, g_F, g_M)$$
$$\cdot\prod_s \Pr(g_{C_{1_s}} \mid p_{C_{1_s}} = U, g_F, g_M)\cdot\prod_t \Pr(g_{C_{2_t}} \mid g_F, g_M),$$

where the values of $\Pr(g_{C_{1_r}} \mid p_{C_{1_r}} = A, g_F, g_M)$, $\Pr(g_{C_{1_s}} \mid p_{C_{1_s}} = U, g_F, g_M)$ and $\Pr(g_{C_{2_t}} \mid g_F, g_M)$ can be found from Table 2, Table 3 and Table 1, respectively.

Incomplete parental genotypes: one parental genotype is missing

Consider a nuclear family with one untyped parent. As before, I specify that the paternal genotype is missing. If all the children are untyped, one can only use the maternal information to infer the estimates. That is, $L_{Nuclear.b} = \Pr(g_M \mid p_M)$. From assumption (f) that the individuals' phenotypes are conditionally independent given their genotypes, I have $\Pr(g_F \mid g_M, \vec{g}_{\vec{C}}, p_F, p_M, \vec{p}_{\vec{C}}) = \Pr(g_F \mid g_M, \vec{g}_{\vec{C}}, p_F)$. If at least one child is genotyped, the likelihood factor from the nuclear family is

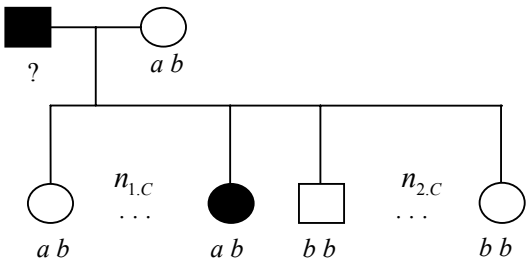
$$L_{Nuclear.b} = \sum_{g_F} L_{Nuclear.a}(\theta; g_F, g_M, \vec{g}_{\vec{C}}, p_F, p_M, \vec{p}_{\vec{C}})\Pr(g_F \mid g_M, \vec{g}_{\vec{C}}, p_F, p_M, \vec{p}_{\vec{C}})$$
$$= \sum_{g_F} L_{Nuclear.a}(\theta; g_F, g_M, \vec{g}_{\vec{C}}, p_F, p_M, \vec{p}_{\vec{C}})\Pr(g_F \mid g_M, \vec{g}_{\vec{C}}, p_F),$$

where

$$\Pr(g_F \mid g_M, \vec{g}_{\vec{C}}, p_F) = \frac{\Pr(g_F, g_M, \vec{g}_{\vec{C}}, p_F)}{\sum_i \Pr(g_F = i, g_M, \vec{g}_{\vec{C}}, p_F)}. \qquad (10)$$

For example, Figure 5 shows a nuclear family with one untyped parent. The nuclear family has $n_{1.C}$ children with genotype 1, and $n_{2.C}$ children with genotype 2. The father is affected and untyped, and the mother is unaffected with $g_M = 1$.

Figure 5: A large nuclear family with one parental genotype missing



Based on the maternal and children's genotypes, the paternal genotype must be either 1 or 2. First I compute:

$$\Pr(g_F = 1, g_M = 1, \vec{g}_{\bar{C}}, p_F = A) =$$
$$\Pr(g_F = 1, g_M = 1)\Pr(p_F = A \mid g_F = 1)$$
$$\times[\Pr(g_C = 1 \mid g_F = 1, g_M = 1)]^{n_{1.C}}[\Pr(g_C = 2 \mid g_F = 1, g_M = 1)]^{n_{2.C}}$$
$$= \pi_1^2\phi_1\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{4}\right)^{n_{2.C}}, \text{and}$$
$$\Pr(g_F = 2, g_M = 1, \vec{g}_{\bar{C}}, p_F = A) =$$
$$\Pr(g_F = 2, g_M = 1)\Pr(p_F = A \mid g_F = 2)$$
$$\times[\Pr(g_C = 1 \mid g_F = 2, g_M = 1)]^{n_{1.C}}[\Pr(g_C = 2 \mid g_F = 2, g_M = 1)]^{n_{2.C}}$$
$$= \pi_2\pi_1\phi_2\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{2}\right)^{n_{2.C}}$$

Then based on equation (10), I have

$$\Pr(g_F = 1 \mid g_M = 1, \vec{g}_{\bar{C}}, p_F = A)$$
$$= \frac{\pi_1^2\phi_1\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{4}\right)^{n_{2.C}}}{\pi_1^2\phi_1\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{4}\right)^{n_{2.C}} + \pi_2\pi_1\phi_2\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{2}\right)^{n_{2.C}}} = \frac{\pi_1\phi_1}{\pi_1\phi_1 + 2^{n_{2.C}}\pi_2\phi_2}, \text{and}$$

$$\Pr(g_F = 2 \mid g_M = 1, \vec{g}_{\bar{C}}, p_F = A) = \frac{2^{n_{2.C}}\pi_2\phi_2}{\pi_1\phi_1 + 2^{n_{2.C}}\pi_2\phi_2}.$$

If the father were unaffected, and Figure 5 were otherwise the same, I would have

$$\Pr(g_F = 1 \mid g_M = 1, \vec{g}_{\bar{C}}, p_F = U) = \frac{\pi_1(1-\phi_1)}{\pi_1(1-\phi_1) + 2^{n_{2.C}}\pi_2(1-\phi_2)}, \text{and}$$

$$\Pr(g_F = 2 \mid g_M = 1, \vec{g}_{\bar{C}}, p_F = U) = \frac{2^{n_{2.C}}\pi_2(1-\phi_2)}{\pi_1(1-\phi_1) + 2^{n_{2.C}}\pi_2(1-\phi_2)}.$$

If the paternal phenotype were unknown in Figure 5,

$$\Pr(g_F = 1 \mid g_M = 1, \vec{g}_{\bar{C}}, p_F = Miss) = \frac{\pi_1}{\pi_1 + 2^{n_{2.C}}\pi_2}, \text{and}$$

$$\Pr(g_F = 2 \mid g_M = 1, \vec{g}_{\bar{C}}, p_F = Miss) = \frac{2^{n_{2.C}}\pi_2}{\pi_1 + 2^{n_{2.C}}\pi_2}.$$

For any possible paternal phenotype, the conditional probabilities are

$$\Pr(g_F = 1 \mid g_M = 1, \vec{g}_{\bar{C}}, p_F) = \frac{\pi_1\eta_1}{\pi_1\eta_1 + 2^{n_{2.C}}\pi_2\eta_2}, \text{and}$$

$$\Pr(g_F = 2 \mid g_M = 1, \vec{g}_{\bar{C}}, p_F) = \frac{2^{n_{2.C}}\pi_2\eta_2}{\pi_1\eta_1 + 2^{n_{2.C}}\pi_2\eta_2},$$

where η_i is defined as in equation (6).

The conditional probabilities $\Pr(g_F \mid g_M, \vec{g}_{\bar{C}}, p_F)$ for an arbitrary nuclear family with untyped father are listed in Table 7.

Incomplete parental genotypes: both parental genotype data are missing

Consider a nuclear family without parental genotypes. When at least one child is genotyped, the likelihood factor from the nuclear family is

$$L_{Nuclear.c} = \sum_{g_F}\sum_{g_M} L_{Nuclear.a}(\theta; g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F, g_M \mid \vec{g}_{\bar{C}}, p_F, p_M),$$

where

$$\Pr(g_F, g_M \mid \vec{g}_{\bar{C}}, p_F, p_M) = \frac{\Pr(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M)}{\sum_i \sum_j \Pr(g_F = i, g_M = j, \vec{g}_{\bar{C}}, p_F, p_M)}. \qquad (11)$$

Suppose both parental genotypes in the nuclear family in Figure 5 were unknown. Based on the children's genotypes, the mating type of the parents $\{g_F, g_M\}$ must be {1, 1}, {1, 2} or {2, 1}. First I compute

$$\begin{aligned}
&\Pr(g_F = 1, g_M = 1, \vec{g}_{\bar{C}}, p_F = A, p_M = U) \\
&= \Pr(g_F = 1, g_M = 1) \Pr(p_F = A \mid g_F = 1) \Pr(p_M = U \mid g_M = 1) \\
&\qquad \cdot [\Pr(g_C = 1 \mid g_F = 1, g_M = 1)]^{n_{1.C}} [\Pr(g_C = 2 \mid g_F = 1, g_M = 1)]^{n_{2.C}} \\
&= \pi_1^2 \phi_1 (1 - \phi_1) \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{4}\right)^{n_{2.C}}, \\
&\Pr(g_F = 1, g_M = 2, \vec{g}_{\bar{C}}, p_F = A, p_M = U) \\
&= \Pr(g_F = 1, g_M = 2) \Pr(p_F = A \mid g_F = 1) \Pr(p_M = U \mid g_M = 2) \\
&\qquad \cdot [\Pr(g_C = 1 \mid g_F = 1, g_M = 2)]^{n_{1.C}} [\Pr(g_C = 2 \mid g_F = 1, g_M = 2)]^{n_{2.C}} \\
&= \pi_1 \pi_2 \phi_1 (1 - \phi_2) \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{2}\right)^{n_{2.C}}, \text{and} \\
&\Pr(g_F = 2, g_M = 1, \vec{g}_{\bar{C}}, p_F = A, p_M = U) = \pi_1 \pi_2 \phi_2 (1 - \phi_1) \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{2}\right)^{n_{2.C}}.
\end{aligned}$$

Then based on equation (11), I have

$$\Pr(g_F = 1, g_M = 1 \mid \vec{g}_{\bar{C}}, p_F = A, p_M = U) = \frac{\pi_1 \phi_1 (1 - \phi_1)}{\pi_1 \phi_1 (1 - \phi_1) + 2^{n_{2.C}} \pi_2 \phi_1 (1 - \phi_2) + 2^{n_{2.C}} \pi_2 \phi_2 (1 - \phi_1)},$$

$$\Pr(g_F = 1, g_M = 2 \mid \vec{g}_{\bar{C}}, p_F = A, p_M = U) = \frac{2^{n_{2.C}} \pi_2 \phi_1 (1 - \phi_2)}{\pi_1 \phi_1 (1 - \phi_1) + 2^{n_{2.C}} \pi_2 \phi_1 (1 - \phi_2) + 2^{n_{2.C}} \pi_2 \phi_2 (1 - \phi_1)},$$

and

$$\Pr(g_F = 2, g_M = 1 \mid \vec{g}_{\bar{C}}, p_F = A, p_M = U) = \frac{2^{n_{2.C}} \pi_2 \phi_2 (1 - \phi_1)}{\pi_1 \phi_1 (1 - \phi_1) + 2^{n_{2.C}} \pi_2 \phi_1 (1 - \phi_2) + 2^{n_{2.C}} \pi_2 \phi_2 (1 - \phi_1)}.$$

For any possible combination of parental phenotypes, the conditional probability $\Pr(g_F = 1, g_M = 1 \mid \vec{g}_{\bar{C}}, p_F, p_M)$ can be written as

$$\frac{\pi_1 \Theta(g = 1, p_F) \Theta(g = 1, p_M)}{\pi_1 \Theta(1, p_F) \Theta(1, p_M) + 2^{n_{2.C}} \pi_2 \Theta(1, p_F) \Theta(2, p_M) + 2^{n_{2.C}} \pi_2 \Theta(2, p_F) \Theta(1, p_M)},$$

where $\Theta(g, p)$ can be computed by equation (8). For example, if the parental phenotypes and genotypes were unknown in Figure 5, I would have

$$\begin{aligned}
&\Pr(g_F = 1, g_M = 1 \mid \vec{g}_{\bar{C}}, p_F = Miss, p_M = Miss) = \frac{\pi_1}{\pi_1 + 2^{n_{2.C}} \pi_2 + 2^{n_{2.C}} \pi_2} = \frac{\pi_1}{\pi_1 + 2^{n_{2.C}+1} \pi_2}, \\
&\Pr(g_F = 1, g_M = 2 \mid \vec{g}_{\bar{C}}, p_F = Miss, p_M = Miss) = \frac{2^{n_{2.C}} \pi_2}{\pi_1 + 2^{n_{2.C}+1} \pi_2}, \text{and} \\
&\Pr(g_F = 2, g_M = 1 \mid \vec{g}_{\bar{C}}, p_F = Miss, p_M = Miss) = \frac{2^{n_{2.C}} \pi_2}{\pi_1 + 2^{n_{2.C}+1} \pi_2}.
\end{aligned}$$

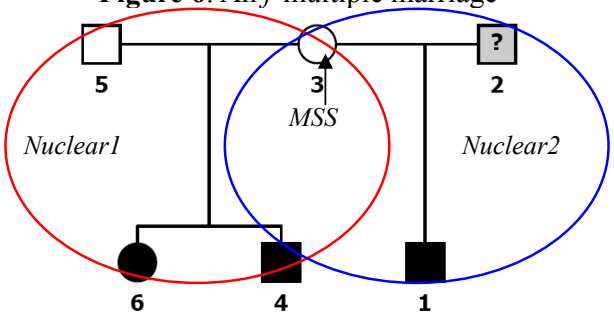
Table 8 lists the conditional probabilities $\Pr(g_F, g_M \mid \vec{g}_{\bar{C}}, p_F, p_M)$ for an arbitrary nuclear family with two untyped parents.

2.2.3 Conditional likelihood function of multiple-marriage nuclear families

A "multiple marriage" is a nuclear family in which one parent parented offspring with multiple distinct mates, both of whom are in the pedigree (Schäffer, 2000). An *f*-multiple marriage is one in which the multiply married spouse (MMS) is a founder. An *n*-multiple marriage is one in which the MMS is not a founder in the pedigree (Schäffer, 2000). This work only considers *f*-multiple marriages. An example is shown in Figure 6.

The *f*-multiple marriage in Figure 6 is a nuclear family from a previously published psoriasis dataset (Helms et al., 2003). The figure is drawn by HaploPainter developed by Thiele and Nürnberg (2005). Unlike earlier figures, this one does not contain any genotype information. The number below each square or circle represents each individual. Individual 3 (ID3 for short) is the MMS. She is a founder and has one affected child (ID1) with a male (ID2) whose affection status is unknown. She has two affected children (ID4 and ID6) fathered by an unaffected male (ID5).

Figure 6: An *f*-multiple marriage



I start with the *f*-multiple marriage where the MMS has 2 mates (as the one in Figure 6). The *f*-multiple marriage is decomposed into two nuclear families ($Nuclear_1$ and $Nuclear_2$) at the MMS. $L_{Nuclear_1}$ denotes the likelihood factor from $Nuclear_1$ (the MMS, one mate and their children), and $L_{Nuclear_2}$ denotes the likelihood factor from $Nuclear_2$ (the MMS, the other mate and their children). Let g_{MMS} and p_{MMS} denote the genotype and phenotype of the MMS, respectively. I consider the following scenarios to derive the likelihood factor from a nuclear family with two multiple marriages.

The genotype of the multiply married spouse is available

The likelihood factor from such *f*-multiple marriage with genotyped MMS is

$$L_{MM.a} = L_{Nuclear_1} L_{Nuclear_2} / \Pr(g_{MMS} \mid p_{MMS}). \qquad (12)$$

Notice that since the information of the MMS is used both in $L_{Nuclear_1}$ and $L_{Nuclear_2}$, equation (12) removes the duplicated factor by dividing by $\Pr(g_{MMS} \mid p_{MMS})$.

The genotype of the multiply married spouse is unknown

If $g_{MMS} = Miss$, I use the available genotypes in the *f*-multiple marriage to infer the possible genotype of the MMS. Let $\{g_{MMS.1}\}$ and $\{g_{MMS.2}\}$ denote two sets of

possible genotypes of the MMS determined by the observed genotypes from $Nuclear_1$ and $Nuclear_2$, respectively. The set of possible genotypes of the MMS inferred by the observed genotypes from the *f*-multiple marriage is $\{g_{MMS}\} = \{g_{MMS.1}\} \cap \{g_{MMS.2}\}$. Then the likelihood factor from the *f*-multiple marriage can be approximated by

$$L_{MM.b} = L_{Nuclear_1} L_{Nuclear_2}, \qquad (13)$$

where $L_{Nuclear_i} = L_{Nuclear_i|\{g_{MMS}\}}, i = 1,2$ is the complete-data likelihood of $Nuclear_i$, with $\{g_{MMS}\}$ the set of possible genotypes of the MMS. When $\{g_{MMS.1}\} = \{g_{MMS.2}\}$, the computation of $L_{Nuclear_i}, i = 1,2$ in equation (13) follows the procedures described in Section 2.2.2.

The likelihood functions (12) and (13) can be extended to nuclear families with more than two multiple marriages. Suppose the MMS has *k* mates. Let $L_{Nuclear_i}, i \in \{1,2,...,k\}$ denotes the likelihood factor from the *i*-th nuclear family decomposed from the *f*-multiple marriage. The likelihood factor from the *f*-multiple marriage is

$$L_{MM.a} = \prod_{i=1}^{k} L_{Nuclear_i} \Big/ \Pr(g_{MMS} \mid p_{MMS})^{k-1} \text{ if } g_{MMS} \text{ is available;}$$

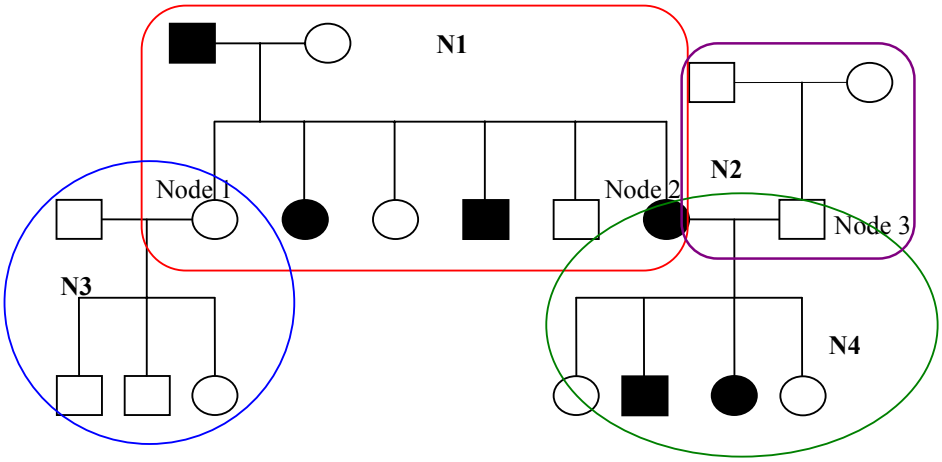
$$L_{MM.b} = \prod_{i=1}^{k} L_{Nuclear_i} \text{ if } g_{MMS} \text{ is unknown.}$$

In $L_{MM.b}$, $L_{Nuclear_i} = L_{Nuclear_i|\{g_{MMS}\}}, i \in \{1,2,...,k\}$ is the complete-data likelihood of the *i*-th decomposed nuclear family, with $\{g_{MMS}\} = \{g_{MMS.1}\} \cap \cdots \cap \{g_{MMS.k}\}$ the set of possible genotypes of the MMS. $\{g_{MMS.i}\}, i \in \{1,2,...,k\}$ denotes a set of possible genotypes of the MMS determined by the observed genotypes from the *i*-th decomposed nuclear family. The likelihood $L_{MM.b}$ here is also an approximation.

2.2.4 Conditional likelihood function of zero-looped three- and four-generation pedigrees

A pedigree will be termed *looped*, or *zero-looped*, according to whether it has, or has not, any cycles (Berge, 1962). Cannings et al. (1978) defined a *zero-looped* pedigree to be a *tree* of individuals and marriages, such as the one shown in Figure 7.

Figure 7: A zero-looped three-generation pedigree



Four related nuclear families, denoted as $N1, N2, N3$ and $N4$, are decomposed from the three-generation pedigree in Figure 7. I define a decomposed nuclear family from a three-generation pedigree as a *high-level* nuclear family if it contains the information about the first generation and as a *low-level* nuclear family if it contains information about the third generation. In one decomposed nuclear family, a nonfounder that has offspring in the pedigree is defined as a *node*. In Figure 7, $N1$ and $N2$ are high-level nuclear families, and $N3$ and $N4$ are low-level families. There are three nodes denoted as *Node1*, *Node2* and *Node3* in this three-generation family. *Node1* is the mother in $N3$ and a child in $N1$. *Node2* is the mother in $N4$ and a child in $N1$. *Node3* is the father in $N4$ and the child in $N2$.

When a large proportion of grandparental and the parental genotypes are missing in a large three- or four-generation pedigree, the likelihood function will be very complicated. The LRT will take prohibitively long. Under these circumstances, pedigree splitting is often used to approximate the likelihood (Blanton et al., 1991; Hasstedt, 1993; Lake et al., 2000). One method to compute the likelihood for three- and four- generation pedigrees is to decompose them into multiple nuclear families. However, the likelihood on the decomposed nuclear families involves duplicated information from the nodes. When applied to a small sample of large pedigrees, the likelihood will lead to a substantial loss of information and power, and may risk inflation of type I errors (Allen-Brady et al., 2006).

For example, I calculate the likelihood of the three- and four-generation pedigrees that occurred in two datasets: (1) Psoriasis data (Helms et al., 2003) (2) Idiopathic scoliosis (IS) data for CHD7 gene (Gao et al., 2007). Both datasets contain many zero-looped three- and four- generation pedigrees such as those in Figure 8 and Figure 9.

Figure 8: Pedigree A from psoriasis data set: a zero-looped three-generation pedigree

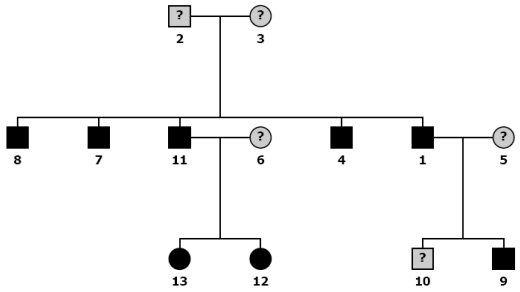
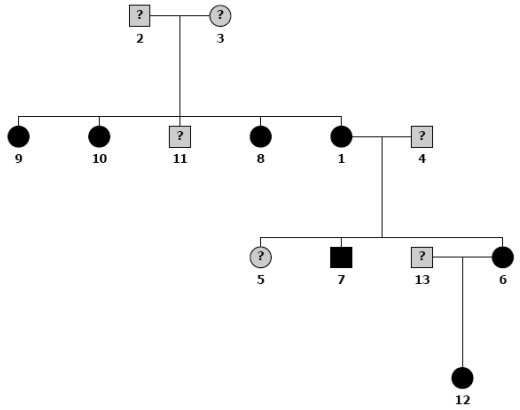


Figure 9: Pedigree B from IS data set: a zero-looped four-generation pedigree



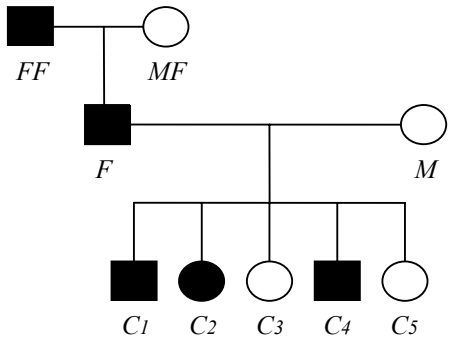
The zero-looped three-generation pedigree (Pedigree A) in Figure 8 is from the

psoriasis data set. The grandparents (ID2 and ID3) have 5 affected children, two, ID11 and ID1, are nodes. There are one high-level family and two low-level families in this pedigree.

The zero-looped four-generation pedigree (Pedigree B) in Figure 9 is from the IS data set. Since it contains four generations, I call the nuclear family composed of individuals from the middle two generations the *middle-level* nuclear family. Grandparents (ID2 and ID3) have 4 affected children, with ID1 a node. ID1 and ID4 have two affected children, with ID6 being the second node of the pedigree. With mate ID13, ID6 has an affected child (ID12), which is the fourth generation of this pedigree.

To derive the likelihood functions of three- and four-generation pedigrees, I start with a simple zero-looped three-generation pedigree with one high-level family such as the one in Figure 10. There, *FF*, *MF*, *F* and *M* denote paternal grandfather, paternal grandmother, father and mother, respectively. The five children of the second generation *F* and *M* are denoted as C_1, … , C_5.

Figure 10: A simple zero-looped three-generation pedigree



The pedigree can be decomposed into a high-level trio (*FF*, *MF* and *F*), and a low-level nuclear family of size 7 with *F*, *M* and their five children. Let L_{Trio} and $L_{Nuclear}$ denote the likelihood factors from the high-level trio and the low-level nuclear family, respectively. There are seven possible scenarios describing the availability of parental genotypes and grandparental genotypes.

The paternal genotype and at least one grandparental genotype are available

Since the information of the father is duplicated, I divide the product of L_{Trio} and $L_{Nuclear}$ by $\Pr(g_F \mid p_F)$. The pedigree likelihood factor is

8.c

$$L_{Pedigree.a} = L_{Trio} \cdot L_{Nuclear} / \Pr(g_F \mid p_F),$$

where $L_{Trio} = L_{Trio.a}(g_{FF}, g_{MF}, g_F, p_{FF}, p_{FM}, p_F)$ when both grandparental genotypes are available, $L_{Trio} = L_{Trio.b}(g_{FF}, g_F, p_{FF}, p_{FM}, p_F)$ when only one grandparental genotype is available (without loss of generality, I specify that g_{FF} is available), and $L_{Nuclear} = L_{Nuclear.a}(g_F, g_M, \vec{g}_{\vec{C}}, p_F, p_M, \vec{p}_{\vec{C}})$ when the maternal genotype is available, $L_{Nuclear} = L_{Nuclear.b}(g_F, \vec{g}_{\vec{C}}, p_F, p_M, \vec{p}_{\vec{C}})$ when the maternal genotype is unknown.

The paternal genotype is unknown, but the maternal genotype and both grandparental genotypes are available

The likelihood factor under this scenario is approximated by the product of $L_{Trio} = \Pr(g_{FF} \mid p_{FF}) \Pr(g_{MF} \mid p_{MF})$ and $L_{Nuclear}$, where $L_{Nuclear}$ is not $L_{Nuclear.b}$ in Section

2.2.2. In $L_{Nuclear.b}$, the possible paternal genotypes are only determined by the maternal and children's genotypes. The grandparental genotypes, however, should also be considered to determine the possible paternal genotypes. Therefore, the likelihood factor is

$$L_{Pedigree.b} = \Pr(g_{FF} \mid p_{FF}) \Pr(g_{MF} \mid p_{MF}) \times \sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F),$$

where

$$\Pr(g_F \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F) = \frac{\Pr(g_{FF}, g_{MF}, g_F, g_M, \vec{g}_{\bar{C}}, p_F)}{\sum_{i=0}^{2} \Pr(g_{FF}, g_{MF}, g_F = i, g_M, \vec{g}_{\bar{C}}, p_F)}. \quad (14)$$

Consider the three-generation pedigree in Figure 10 with $p_F = A$, $g_{FF} = g_{MF} = 1$, $g_F = Miss$, $g_M = 1$, $g_{C_1} = g_{C_2} = g_{C_3} = 1$ and $g_{C_4} = g_{C_5} = 2$ ($n_{1.C} = 3$ and $n_{2.C} = 2$). From the grandparental genotypes, g_F must be 0, 1 or 2; from the maternal and children's genotypes, g_F must be 1 or 2. Therefore, from all the observed genotypes in the pedigree, g_F must be 1 or 2. I first compute

$$\begin{aligned}
&\Pr(g_{FF} = 1, g_{MF} = 1, g_F = 1, g_M = 1, \vec{g}_{\bar{C}}, p_F = A) \\
&= \Pr(g_{FF} = 1, g_{MF} = 1) \Pr(g_F = 1 \mid g_{FF} = 1, g_{MF} = 1) \Pr(g_M = 1) \\
&\quad \times \Pr(\vec{g}_C \mid g_F = 1, g_M = 1) \Pr(p_F = A \mid g_F = 1) \\
&= \pi_1^2 \frac{1}{2} \pi_1 \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{4}\right)^{n_{2.C}} \phi_1, \text{and} \\
&\Pr(g_{FF} = 1, g_{MF} = 1, g_F = 2, g_M = 1, \vec{g}_{\bar{C}}, p_F = A) \\
&= \Pr(g_{FF} = 1, g_{MF} = 1) \Pr(g_F = 2 \mid g_{FF} = 1, g_{MF} = 1) \Pr(g_M = 1) \\
&\quad \times \Pr(\vec{g}_C \mid g_F = 2, g_M = 1) \Pr(p_F = A \mid g_F = 2) \\
&= \pi_1^2 \frac{1}{4} \pi_1 \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{2}\right)^{n_{2.C}} \phi_2.
\end{aligned}$$

Then based on equation (14), I have

$$\begin{aligned}
&\Pr(g_F = 1 \mid g_{FF} = 1, g_{MF} = 1, g_M = 1, \vec{g}_{\bar{C}}, p_F = A) \\
&= \frac{\pi_1^2 \frac{1}{2} \pi_1 \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{4}\right)^{n_{2.C}} \phi_1}{0 + \pi_1^2 \frac{1}{2} \pi_1 \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{4}\right)^{n_{2.C}} \phi_1 + \pi_1^2 \frac{1}{4} \pi_1 \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{2}\right)^{n_{2.C}} \phi_2} = \frac{2\phi_1}{2\phi_1 + 2^{n_{2.C}} \phi_2}, \text{and} \\
&\Pr(g_F = 2 \mid g_{FF} = 1, g_{MF} = 1, g_M = 1, \vec{g}_{\bar{C}}, p_F = A) = \frac{2^{n_{2.C}} \phi_2}{2\phi_1 + 2^{n_{2.C}} \phi_2}.
\end{aligned}$$

8.c There are some special cases under which $\Pr(g_F = i \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F) = 1$. For example, when $\{g_{FF}, g_{MF}\} = \{0,0\}, \{0,2\}, \{2,0\}$ and $\{2,2\}$, g_F must be 0, 1, 1, and 2, respectively. When $\{g_{FF}, g_{MF}\} = \{1,2\}$ or $\{2,1\}$, $g_M = 0$ and $\{\vec{g}_{\bar{C}}\} = \{0\}$ (all the children are genotyped 0), g_F must be 1. Table 9 and Table 10 list the conditional probabilities $\Pr(g_F \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F)$ for all possible combinations of $g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}$ and p_F.

The parental genotypes are unknown, but both grandparental genotypes are available

Similar to the calculation of $L_{Pedigree.b}$, I first compute the likelihood factors for the decomposed nuclear families $L_{Trio} = \Pr(g_{FF} \mid p_{FF}) \Pr(g_{MF} \mid p_{MF})$ and $L_{Nuclear}$. Then I multiply these two values as an approximation to the likelihood factor from the three-generation pedigree. The possible parental genotypes are determined not only by the children's genotype but also by the grandparental genotypes. The likelihood factor from such three-generation pedigree is

$$L_{Pedigree.c} = \Pr(g_{FF} \mid p_{FF}) \Pr(g_{MF} \mid p_{MF}) \times \sum_{g_F} \sum_{g_M} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F, g_M \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F, p_M),$$

where

$$\Pr(g_F, g_M \mid g_{FF}, g_{MF}, \vec{g}_{\bar{C}}, p_F, p_M) = \frac{\Pr(g_{FF}, g_{MF}, g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M)}{\sum_{i=0}^{2} \sum_{j=0}^{2} \Pr(g_{FF}, g_{MF}, g_F = i, g_M = j, \vec{g}_{\bar{C}}, p_F, p_M)}. \quad (15)$$

Table 11 lists $\Pr(g_F, g_M \mid g_{FF}, g_{MF}, \vec{g}_{\bar{C}}, p_F, p_M)$ for all possible combinations of $g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F$ and p_M.

The paternal genotype is unknown, but the maternal genotype and one grandparental genotype are available

If the paternal genotype is unknown, and without loss of generality, g_{FF} is available, I have $L_{Trio} = \Pr(g_{FF} \mid p_{FF})$. The likelihood factor from the pedigree can be approximated by

$$L_{Pedigree.d} = \Pr(g_{FF} \mid p_{FF}) \times \sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}, p_F),$$

where

$$\Pr(g_F \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}, p_F) = \frac{\sum_{g_{MF}} \Pr(g_{FF}, g_{MF}, g_F, g_M, \vec{g}_{\bar{C}}, p_{MF}, p_F)}{\sum_{i=0}^{2} \sum_{g_{MF}} \Pr(g_{FF}, g_{MF}, g_F = i, g_M, \vec{g}_{\bar{C}}, p_{MF}, p_F)}. \quad (16)$$

Consider the three-generation pedigree in Figure 10 with $p_{MF} = U$ and $p_F = A$. In the event that $g_{FF} = 1$, $g_{MF} = Miss$, $g_F = Miss$, $g_M = 1$, $n_{0.C} = 0$, $n_{1.C} > 0$ and $n_{2.C} > 0$ (at least one child is genotyped 1, at least one child is genotyped 2, but no child is genotyped 0), g_F must be either 1 or 2. First I compute

$$\sum_{g_{MF}} \Pr(g_{FF} = 1, g_{MF}, g_F = 1, g_M = 1, \vec{g}_{\bar{C}}, p_{MF} = U, p_F = A)$$

$$= \Pr(g_{FF} = 1) \sum_{j=0}^{2} [\Pr(g_{MF} = j) \Pr(p_{MF} = U \mid g_{MF} = j) \Pr(g_F = 1 \mid g_{FF} = 1, g_{MF} = j)] \times \Pr(g_M = 1) \Pr(\vec{g}_C \mid g_F = 1, g_M = 1) \Pr(p_F = A \mid g_F = 1)$$

$$= \pi_1 [\pi_0 (1 - \phi_0) \frac{1}{2} + \pi_1 (1 - \phi_1) \frac{1}{2} + \pi_2 (1 - \phi_2) \frac{1}{2}] \cdot \pi_1 \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{4}\right)^{n_{2.C}} \phi_1$$

$$= \frac{1}{2} (1 - \pi_0 \phi_0 - \pi_1 \phi_1 - \pi_2 \phi_2) \pi_1^2 \left(\frac{1}{2}\right)^{n_{1.C}} \left(\frac{1}{4}\right)^{n_{2.C}} \phi_1, \text{ and}$$

$$\sum_{g_{MF}} \Pr(g_{FF}=1, g_{MF}, g_F=2, g_M=1, \vec{g}_{\bar{C}}, p_{MF}=U, p_F=A)$$

$$= \Pr(g_{FF}=1)\sum_{j=0}^{2} \Pr(g_{MF}=j)\Pr(p_{MF}=U \mid g_{MF}=j)\Pr(g_F=2 \mid g_{FF}=1, g_{MF}=j)$$
$$\times \Pr(g_M=1)\Pr(\vec{g}_C \mid g_F=2, g_M=1)\Pr(p_F=A \mid g_F=2)$$

$$= \pi_1[\pi_0(1-\phi_0)\cdot 0 + \pi_1(1-\phi_1)\frac{1}{4} + \pi_2(1-\phi_2)\frac{1}{2}]\cdot \pi_1\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{2}\right)^{n_{2.C}}\phi_2$$

$$= \frac{1}{4}[\pi_1(1-\phi_1)+2\pi_2(1-\phi_2)]\pi_1^2\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{2}\right)^{n_{2.C}}\phi_2.$$

Then based on equation (16), I have

$$\Pr(g_F=1 \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}=U, p_F=A)$$
$$= \frac{\frac{1}{2}(1-\pi_0\phi_0-\pi_1\phi_1-\pi_2\phi_2)\pi_1^2\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{4}\right)^{n_{2.C}}\phi_1}{\frac{1}{2}(1-\pi_0\phi_0-\pi_1\phi_1-\pi_2\phi_2)\pi_1^2\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{4}\right)^{n_{2.C}}\phi_1+\frac{1}{4}[\pi_1(1-\phi_1)+2\pi_2(1-\phi_2)]\pi_1^2\left(\frac{1}{2}\right)^{n_{1.C}}\left(\frac{1}{2}\right)^{n_{2.C}}\phi_2}$$
$$= \frac{2[(1-\phi_0)\pi_0+(1-\phi_1)\pi_1+(1-\phi_2)\pi_2]\phi_1}{2[(1-\phi_0)\pi_0+(1-\phi_1)\pi_1+(1-\phi_2)\pi_2]\phi_1+[\pi_1(1-\phi_1)+2\pi_2(1-\phi_2)]2^{n_{2.C}}\phi_2}, \text{and}$$

$$\Pr(g_F=2 \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}=U, p_F=A)$$
$$= \frac{[\pi_1(1-\phi_1)+2\pi_2(1-\phi_2)]2^{n_{2.C}}\phi_2}{2[(1-\phi_0)\pi_0+(1-\phi_1)\pi_1+(1-\phi_2)\pi_2]\phi_1+[\pi_1(1-\phi_1)+2\pi_2(1-\phi_2)]2^{n_{2.C}}\phi_2}.$$

If the paternal grandmother were affected,

$$\Pr(g_F=1 \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}=A, p_F=A) = \frac{2(\phi_0\pi_0+\phi_1\pi_1+\phi_2\pi_2)\phi_1}{2(\phi_0\pi_0+\phi_1\pi_1+\phi_2\pi_2)\phi_1+(\pi_1\phi_1+2\pi_2\phi_2)2^{n_{2.C}}\phi_2},$$

and

$$\Pr(g_F=2 \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}=A, p_F=A) = \frac{(\pi_1\phi_1+2\pi_2\phi_2)2^{n_{2.C}}\phi_2}{2(\phi_0\pi_0+\phi_1\pi_1+\phi_2\pi_2)\phi_1+(\pi_1\phi_1+2\pi_2\phi_2)2^{n_{2.C}}\phi_2}.$$

If the phenotype of the paternal grandmother were unknown,

$$\Pr(g_F=1 \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}=Miss, p_F=A) = \frac{2\phi_1}{2\phi_1+(\pi_1+2\pi_2)2^{n_{2.C}}\phi_2}, \text{and}$$

$$\Pr(g_F=2 \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}=Miss, p_F=A) = \frac{(\pi_1+2\pi_2)2^{n_{2.C}}\phi_2}{2\phi_1+(\pi_1+2\pi_2)2^{n_{2.C}}\phi_2}.$$

Table 12 and Table 13 only list $\Pr(g_F \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_F, p_{MF}=Miss)$ due to space limitations. Similar probabilities can be easily derived for $p_{MF}=A$ and $p_{MF}=U$.

The parental genotypes are unknown, but one grandparental genotype is available

Without loss of generality, I specify that g_{FF} is available. First I have $L_{Trio}=\Pr(g_{FF} \mid p_{FF})$. Similar to equation (15), the possible parental genotypes are determined not only by the children's genotypes but also by g_{FF}. Then the likelihood factor from such a three-generation pedigree is approximated as

$$L_{Pedigree.e} = \Pr(g_{FF} \mid p_{FF})$$
$$\times \sum_{g_F}\sum_{g_M} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}})\Pr(g_F, g_M \mid g_{FF}, \vec{g}_{\bar{C}}, p_{MF}, p_F, p_M),$$

where

$$\Pr(g_F, g_M \mid g_{FF}, \vec{g}_{\bar{C}}, p_{MF}, p_F, p_M) = \frac{\sum_{g_{MF}} \Pr(g_{FF}, g_{MF}, g_F, g_M, \vec{g}_{\bar{C}}, p_{MF}, p_F, p_M)}{\sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{g_{MF}} \Pr(g_{FF}, g_{MF}, g_F = i, g_M = j, \vec{g}_{\bar{C}}, p_{MF}, p_F, p_M)}.$$

The conditional probability $\Pr(g_F, g_M \mid g_{FF}, \vec{g}_{\bar{C}}, p_{MF} = Miss, p_F, p_M)$ can be obtained from Table 11 by setting $g_{MF} = Miss$.

Both grandparental genotypes are unknown

In this case, the three-generation pedigree is reduced to a nuclear family. The likelihood factor is

$$L_{Pedigree.f} = L_{Nuclear},$$

where $L_{Nuclear} = L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}})$ when both parents are genotyped; $L_{Nuclear} = L_{Nuclear.b}(g_F, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}})$ when father is genotyped but mother is untyped; $L_{Nuclear} = L_{Nuclear.b}(g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}})$ when father is untyped but mother is genotyped; $L_{Nuclear} = L_{Nuclear.c}(\vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}})$ when both parents are untyped.

Now consider a complex three-generation pedigree such as Pedigree A in Figure 8. I denote the likelihood factor from Pedigree A by L_{PedA}. The high-level nuclear family decomposed from this pedigree is a large nuclear family of size 7, consisting of ID2, ID3, ID8, ID7, ID11, ID4, and ID1. It contains five affected children (ID8, ID7, ID11, ID4, and ID1), with nodes ID1 and ID11 being the fathers of two low-level families, respectively. $L_{Nuclear.High}$, $L_{Nuclear.Low_1}$ and $L_{Nuclear.Low_2}$ denote the likelihood factors from the high-level and the low-level families, respectively.

If the genotypes of ID1 and ID11 are available, the likelihood factor contributed by Pedigree A is

$$L_{PedA} = \frac{L_{Nuclear.High}(\vec{g}_{High}, \vec{p}_{High}) L_{Nuclear.Low_1}(\vec{g}_{Low_1}, \vec{p}_{Low_1}) L_{Nuclear.Low_2}(\vec{g}_{Low_2}, \vec{p}_{Low_2})}{\Pr(g_1 \mid p_1)\Pr(g_{11} \mid p_{11})},$$

where $\vec{g}_{High} = \{g_2, g_3, g_8, g_7, g_{11}, g_4, g_1\}$, $\vec{p}_{High} = \{p_2, p_3, p_8, p_7, p_{11}, p_4, p_1\}$,

$\vec{g}_{Low_1} = \{g_1, g_5, g_{10}, g_9\}$, $\vec{p}_{Low_1} = \{p_1, p_5, p_{10}, p_9\}$,

$\vec{g}_{Low_2} = \{g_{11}, g_6, g_{13}, g_{12}\}$, and $\vec{p}_{Low_2} = \{p_{11}, p_6, p_{13}, p_{12}\}$.

If one node's genotype is unknown (without loss of generality, $g_{11} = Miss$), and the genotype of his mate (g_6) and the grandparental genotypes (g_2, g_3) are available, the approximate likelihood factor contributed by Pedigree A is

$$L_{PedA} = \frac{L_{Nuclear.High}(\vec{g}_{High}, \vec{p}_{High}) L_{Nuclear.Low_1}(\vec{g}_{Low_1}, \vec{p}_{Low_1}) L_{Nuclear.Low_2}(\vec{g}_{Low_2}, \vec{p}_{Low_2})}{\Pr(g_1 \mid p_1)},$$

where

$$L_{Nuclear.Low_2} = \sum_{g_{11}} L_{Nuclear.Low_2}(g_{11}, g_6, g_{13}, g_{12}, \vec{p}_{Low_2}) \Pr(g_{11} \mid g_2, g_3, g_6, g_{13}, g_{12}, p_{11}). \quad (17)$$

For other possible scenarios considering the availability of g_2, g_3 and g_6, formulas similar to equation (17) can be derived.

If both genotypes of ID1 and ID11 are unknown, and g_2, g_3, g_5 and g_6 are available, the approximate likelihood factor contributed by Pedigree A is

$$L_{PedA} = L_{Nuclear.High}(\vec{g}_{High}, \vec{p}_{High}) L_{Nuclear.Low_1}(\vec{g}_{Low_1}, \vec{p}_{Low_1}) L_{Nuclear.Low_2}(\vec{g}_{Low_2}, \vec{p}_{Low_2}),$$

where

$$L_{Nuclear.Low_1} = \sum_{g_1} L_{Nuclear.Low_1}(g_1, g_5, g_{10}, g_9, \vec{p}_{Low_1}) \Pr(g_1 \mid g_2, g_3, g_5, g_{10}, g_9, p_1), \quad (18)$$ and

$$L_{Nuclear.Low_2} = \sum_{g_{11}} L_{Nuclear.Low_2}(g_{11}, g_6, g_{13}, g_{12}, \vec{p}_{Low_2}) \Pr(g_{11} \mid g_2, g_3, g_6, g_{13}, g_{12}, p_{11}). \quad (19)$$

For other possible scenarios considering the availability of g_2, g_3, g_5 and g_6, formulas similar to equations (18) and (19) can also be derived.

For a three-generation pedigree with one high-level family, if an untyped node has siblings and at least one sibling's genotype is available, one should include the available siblings' genotypes to infer the possible genotypes of the node and the relevant conditional probabilities. To simplify the likelihood calculations, this work does not consider the sibling's genotypes.

The likelihood computation of Pedigree B in Figure 9 is similar to that of Pedigree A, except that the fourth generation is included in the likelihood calculation. L_{PedB} denotes the likelihood factor from this pedigree. $L_{Nuclear.High}$, $L_{Nuclear.Middle}$ and $L_{Nuclear.Low}$ denote the likelihood factors from the high-level nuclear family (consists of ID1, ID2, ID3, ID9, ID10, ID11, ID8, and ID1), the middle-level nuclear family(consists of ID4, ID1, ID5, ID7, and ID6) and the low-level nuclear family (consists of ID13, ID6, and ID12), respectively. Node ID1 is the mother of the middle-level nuclear family, and node ID6 is the mother of the low-level nuclear family.

If the genotypes of ID1 and ID6 are available, the likelihood factor of Pedigree B is

$$L_{PedB} = \frac{L_{Nuclear.High}(\vec{g}_{High}, \vec{p}_{High}) L_{Nuclear.Middle}(\vec{g}_{Middle}, \vec{p}_{Middle})}{\Pr(g_1 \mid p_1)} \cdot \frac{L_{Nuclear.Low}(\vec{g}_{Low}, \vec{p}_{Low})}{\Pr(g_6 \mid p_6)},$$

where $\vec{g}_{High} = \{g_2, g_3, g_9, g_{10}, g_{11}, g_8, g_1\}$, $\vec{p}_{High} = \{p_2, p_3, p_9, p_{10}, p_{11}, p_8, p_1\}$,

$\vec{g}_{Middle} = \{g_4, g_1, g_5, g_7, g_6\}$, $\vec{p}_{Middle} = \{p_4, p_1, p_5, p_7, p_6\}$,

$\vec{g}_{Low} = \{g_{13}, g_6, g_{12}\}$, and $\vec{p}_{Low} = \{p_{13}, p_6, p_{12}\}$.

In the event that the genotype of ID1 is available, the genotype of ID6 is unknown, and the genotypes of ID6's mate (ID13) and parents (ID1 and ID4) are available, the approximate likelihood factor of Pedigree B is

$$L_{PedB} = \frac{L_{Nuclear.High}(\vec{g}_{High}, \vec{p}_{High}) L_{Nuclear.Middle}(\vec{g}_{Middle}, \vec{p}_{Middle})}{\Pr(g_1 \mid p_1)} \cdot L_{Nuclear.Low}(\vec{g}_{Low}, \vec{p}_{Low}),$$

where

$$L_{Nuclear.Low} = \sum_{g_6} L_{Nuclear.Low}(g_{13}, g_6, g_{12}, \vec{p}_{Low}) \Pr(g_6 \mid g_1, g_4, g_{13}, g_{12}, p_6). \quad (20)$$

For other possible scenarios considering the availability of g_{13}, g_1 and g_4, formulas similar to equation (20) can be derived.

In the event that the genotype of ID6 is available, the genotype of ID1 is unknown, and the genotypes of ID1's mate (ID4) and parents (ID2 and ID3) are available, the approximate likelihood factor of Pedigree B is then

$$L_{PedB} = L_{Nuclear.High}(\vec{g}_{High}, \vec{p}_{High}) L_{Nuclear.Middle}(\vec{g}_{Middle}, \vec{p}_{Middle}) \cdot \frac{L_{Nuclear.Low}(\vec{g}_{Low}, \vec{p}_{Low})}{\Pr(g_6 \mid p_6)},$$

where

$$L_{Nuclear.Middle} = \sum_{g_1} L_{Nuclear.Middle}(g_4, g_1, g_5, g_7, g_6, \vec{p}_{Low}) \Pr(g_1 \mid g_2, g_3, g_4, g_5, g_7, g_6, p_1). \quad (21)$$

For other possible scenarios considering the availability of g_4, g_2 and g_3, formulas similar to equation (21) can be derived.

In the event that the genotypes of ID1 and ID6 are unknown, and the genotypes of ID1's parents (ID2 and ID3), ID1's mate (ID4), and ID6's mate (ID13) are available, the approximate likelihood factor of Family 14 is then

$$L_{PedB} = L_{Nuclear.High}(\vec{g}_{High}, \vec{p}_{High}) L_{Nuclear.Middle}(\vec{g}_{Middle}, \vec{p}_{Middle}) L_{Nuclear.Low}(\vec{g}_{Low}, \vec{p}_{Low}),$$

where

$$L_{Nuclear.Middle} = \sum_{g_1} L_{Nuclear.Middle}(g_4, g_1, g_5, g_7, g_6, \vec{p}_{Low}) \Pr(g_1 \mid g_2, g_3, g_4, g_5, g_7, g_6, p_1), \text{and}$$

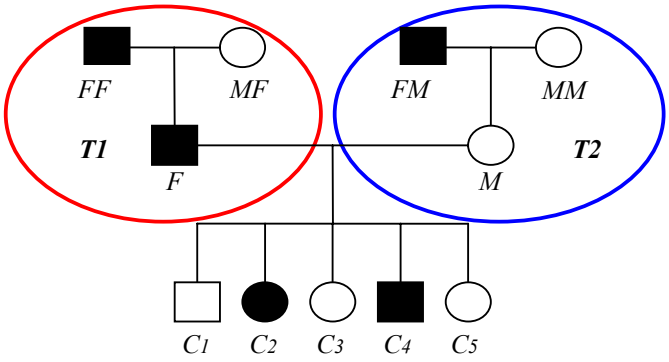
$$L_{Nuclear.Low} = \sum_{g_6} L_{Nuclear.Low}(g_{13}, g_6, g_{12}, \vec{p}_{Low}) \Pr(g_6 \mid g_4, g_{13}, g_{12}, p_6). \quad (22)$$

For other possible scenarios considering the availability of g_2, g_3, g_4 and g_{13}, formulas similar to equation (22) can be derived.

2.2.5 Conditional likelihood function for CEPH pedigrees

Another typical pedigree structure is shown in Figure 11, which consists of four grandparents (*FF*, *MF*, *FM*, and *MM*), two parents (*F* and *M*), and multiple offspring ($\vec{C} = \{C_1, C_2, ..., C_5\}$) (Chakravarti, 1991). Such a pedigree is an example of a *Centre d'Etude du Polymorphisme Humain* (CEPH) *pedigree*. Since a CEPH pedigree is not commonly used for dichotomous trait genetic studies, I could not find a data set containing CEPH pedigrees for an application of this likelihood method. The pedigree in Figure 11 (Pedigree C) generates the additional complexity of grandparental information from both father and mother.

Figure 11: Pedigree C: a CEPH pedigree



The parents *F* and *M* are the nodes in Pedigree C. I denote the left high-level trio in Figure 11 by *T1*, and the right high-level trio by *T2*. Let L_{T1} and L_{T2} denote the likelihood factor from *T1* and *T2*, respectively.

In the event that g_F is available and two grandparents in *T1* are genotyped,

$$L_{T1} = L_{Trio.a}(g_{FF}, g_{MF}, g_F, p_{FF}, p_{MF}, p_F);$$

In the event that g_F is available and only one grandparent in *T1* is genotyped (without loss of generality, *FF* is genotyped),

$$L_{T1} = L_{Trio.b}(g_{FF}, g_F, p_{FF}, p_{MF}, p_F).$$

In the event that g_F is unknown and two grandparents in *T1* are genotyped,

$$L_{T1} = \Pr(g_{FF} \mid p_{FF}) \Pr(g_{MF} \mid p_{MF});$$

In the event that g_F is unknown and only one grandparent in *T1* is genotyped (without loss of generality, *FF* is genotyped),

$$L_{T1} = \Pr(g_{FF} \mid p_{FF}).$$

The calculation of L_{T2} is similar to that of L_{T1} under the four scenarios.

There are three possible scenarios considering the availability of g_F and g_M.

Both parental genotypes are available

For a CEPH pedigree with two genotyped parents and at least one genotyped grandparent each high-level trio, the likelihood factor is

$$\begin{aligned} L_{CEPH.a} &= L_{T1} L_{T2} \frac{L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}})}{\Pr(g_F \mid p_F) \Pr(g_M \mid p_M)} \\ &= L_{T1} L_{T2} \frac{\Pr(g_F \mid p_F) \Pr(g_M \mid p_M) \Pr(\vec{g}_{\bar{C}} \mid \vec{p}_{\bar{C}}, g_F, g_M)}{\Pr(g_F \mid p_F) \Pr(g_M \mid p_M)} \\ &= L_{T1} L_{T2} \Pr(\vec{g}_{\bar{C}} \mid \vec{p}_{\bar{C}}, g_F, g_M). \end{aligned}$$

In the event that the grandparents from a high-level trio are untyped (without loss of generality, $g_{FF} = g_{MF} = Miss$) and at least one grandparental genotype in the other high-level trio is available, the CEPH pedigree is reduced to a three-generation pedigree with one high-level trio, as in Figure 10. The likelihood factor is then

$$L_{CEPH.a} = L_{T2} \cdot P(g_F \mid p_F) \Pr(\vec{g}_{\bar{C}} \mid \vec{p}_{\bar{C}}, g_F, g_M).$$

If the event that all four grandparental genotypes are unknown, the CEPH pedigree is reduced to a general nuclear family with complete parental genotype information. The likelihood factor is

$$L_{CEPH.a} = P(g_F \mid p_F) P(g_M \mid p_M) \Pr(\vec{g}_{\bar{C}} \mid \vec{p}_{\bar{C}}, g_F, g_M).$$

One parental genotype is unknown, and the other parental genotype is available

Consider a CEPH pedigree with one genotyped parent (without loss of generality, $g_F = Miss$). If the genotypes of the paternal grandparents are available, the possible genotypes of the father are determined by g_{FF}, g_{MF}, g_M and $\vec{g}_{\bar{C}}$. In the event that at least one maternal grandparent is genotyped, the likelihood factor is

$$L_{CEPH.b} = L_{T1} L_{T2} \frac{\sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F)}{\Pr(g_M \mid p_M)}.$$

The division by $\Pr(g_M \mid p_M)$ is necessary to remove the duplicated information of the mother in the likelihood calculation. The conditional probabilities $\Pr(g_F \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F)$ are listed in Table 9 and Table 10. In the event that no maternal grandparent is genotyped, the likelihood factor is

$$L_{CEPH.b} = L_{T1} \sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F).$$

Since the information of the mother is only used in $L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}})$, the division by $\Pr(g_M \mid p_M)$ is not included in this likelihood function.

If only one paternal grandparent is genotyped (without loss of generality, g_{FF} is available), possible genotypes of the father are determined by g_{FF}, g_M and $\vec{g}_{\bar{C}}$. In the

event that at least one maternal grandparent if genotyped, the likelihood factor is

$$L_{CEPH.b} = L_{T1}L_{T2}\frac{\sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}, p_F)}{\Pr(g_M \mid p_M)},$$

where $\Pr(g_F \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}, p_F)$ is given by equation (16). The division by $\Pr(g_M \mid p_M)$ removes the duplicated information of the mother in the likelihood calculation. The conditional probabilities $\Pr(g_F \mid g_{MF}, g_M, \vec{g}_{\bar{C}}, p_{MF} = Miss, p_F)$ are listed in Table 12 and Table 13. In the event that no maternal grandparent is genotyped, the likelihood factor is

$$L_{CEPH.b} = L_{T1}\sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF}, p_F).$$

If both genotypes of the paternal grandparents are unknown, the CEPH pedigree is reduced to a three-generation pedigree with one high-level trio, like the one in Figure 10. In the event that at least one maternal grandparent is genotyped, the likelihood factor is

$$L_{CEPH.b} = L_{T1}L_{T2}\frac{\sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F \mid g_M, \vec{g}_{\bar{C}}, p_F)}{\Pr(g_M \mid p_M)}.$$

In the event that no maternal grandparent is genotyped, the likelihood factor is

$$L_{CEPH.b} = L_{T1}\sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F \mid g_M, \vec{g}_{\bar{C}}, p_F).$$

The values of $\Pr(g_F \mid g_M, \vec{g}_{\bar{C}}, p_F)$ are listed in Table 7.

Both parental genotypes are unknown

In the event that four grandparental genotypes are available and both parental genotypes are unknown, the possible parental genotypes are determined by the grandparental and children's genotypes. The likelihood function is given as

$$\begin{aligned} L_{CEPH.c} &= \Pr(g_{FF}, g_{MF} \mid p_{FF}, p_{MF}) \Pr(g_{FM}, g_{MM} \mid p_{FM}, p_{MM}) \\ &\cdot \sum_{g_F}\sum_{g_M} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F, g_M \mid g_{FF}, g_{MF}, g_{FM}, g_{MM}, \vec{g}_{\bar{C}}, p_F, p_M), \end{aligned} \quad (23)$$

where

$$= \frac{\Pr(g_F, g_M, g_{FF}, g_{MF}, g_{FM}, g_{MM}, \vec{g}_{\bar{C}}, p_F, p_M)}{\sum_i \sum_j \Pr(g_F = i, g_M = j, g_{FF}, g_{MF}, g_{FM}, g_{MM}, \vec{g}_{\bar{C}}, p_F, p_M)}.$$

This work uses an approximation to the likelihood in equation (23):

$$\begin{aligned} L_{CEPH.c} &= \Pr(g_{FF}, g_{MF} \mid p_{FF}, p_{MF}) \Pr(g_{FM}, g_{MM} \mid p_{FM}, p_{MM}) \\ &\times L_{Nuclear.c}(\vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}} \mid \{g_F\}, \{g_M\}), \end{aligned} \quad (24)$$

where $L_{Nuclear.c}(\vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}} \mid \{g_F\}, \{g_M\}, \{g_F, g_M\})$ denotes the likelihood of the low-level nuclear family with two untyped parents. The calculation of $L_{Nuclear.c}(\vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}} \mid \{g_F\}, \{g_M\})$ is similar to that of $L_{Nuclear.c}(\vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}})$, except that the set of possible parental genotypes in the former likelihood is $\{\{g_F\} \times \{g_M\}\} \cap \{g_F, g_M\}$, where $\{g_F\}$ denotes a set of the possible paternal genotypes inferred g_{FF} and g_{MF}, $\{g_M\}$ a set of the possible maternal genotypes inferred from g_{FM} and g_{MM}, and $\{g_F, g_M\}$ a set of the possible parental genotypes inferred from $\vec{g}_{\bar{C}}$.

In the event that one grandparental genotype is unknown (without loss of generality, $g_{MF} = Miss$) and the other three grandparental genotypes are available, and both parental genotypes are missing, the approximate likelihood is similar to equation (24):

$$L_{CEPH.c} = \Pr(g_{FF} \mid p_{FF}) \Pr(g_{FM}, g_{MM} \mid p_{FM}, p_{MM}) L_{Nuclear.c}(\vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}} \mid \{g_F\}, \{g_M\}). \quad (25)$$

The possible paternal genotypes $\{g_F\}$ in equation (25) are determined by g_{FF}, and the possible maternal genotypes $\{g_M\}$ are determined by g_{FM} and g_{MM}.

In the event that the grandparental genotypes from a high-level trio are unknown (without loss of generality, $g_{FM} = g_{MM} = Miss$) and the other two grandparental genotypes are available, and both parental genotypes are missing, the CEPH pedigree is reduced to the small three-generation pedigree in Figure 10, with the likelihood

$$L_{CEPH.c} = \Pr(g_{FF}, g_{MF} \mid p_{FF}, p_{MF}) \cdot \sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F, g_M \mid g_{FF}, g_{MF}, \vec{g}_{\bar{C}}, p_F, p_M),$$

where the values of $\Pr(g_F, g_M \mid g_{FF}, g_{MF}, \vec{g}_{\bar{C}}, p_F, p_M)$ are listed in Table 11.

In the event that two grandparental genotypes from the different high-level trios are unknown (without loss of generality, $g_{MF} = g_{MM} = Miss$) and the other two grandparental genotypes are available, and both parental genotypes are missing, similar to equations (24) and (25), the likelihood factor is

$$L_{CEPH.c} = \Pr(g_{FF} \mid p_{FF}) \Pr(g_{FM} \mid p_{FM}) L_{Nuclear.c}(\vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}} \mid \{g_F\}, \{g_M\}).$$

The possible paternal genotypes $\{g_F\}$ are determined by g_{FF}, and the possible maternal genotypes $\{g_M\}$ are determined by g_{FM}.

In the event that only one grandparental genotype is available (without loss of generality, g_{FF} is available), and both parental genotypes are missing, the CEPH pedigree is reduced to the small three-generation pedigree in Figure 10, with the likelihood

$$L_{CEPH.c} = \Pr(g_{FF} \mid p_{FF}) \cdot \sum_{g_F} L_{Nuclear.a}(g_F, g_M, \vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}) \Pr(g_F, g_M \mid g_{FF}, \vec{g}_{\bar{C}}, p_{MF}, p_F, p_M),$$

where $\Pr(g_F, g_M \mid g_{FF}, \vec{g}_{\bar{C}}, p_{MF} = Miss, p_F, p_M)$ can be computed from Table 11 by setting $g_{MF} = Miss$.

In the event that all the four grandparental genotypes are unknown, and both parental genotypes are missing, the CEPH pedigree is reduced to a general nuclear family with unknown parental genotypes, with likelihood

$$L_{CEPH.c} = L_{Nuclear.c}(\vec{g}_{\bar{C}}, p_F, p_M, \vec{p}_{\bar{C}}).$$

2.3 Incorporating the Mendelian inconsistencies into the likelihood function

Mendelian consistency is arguably the most important and common criterion for identifying genotyping errors (Zou et al., 2003). Families that are not Mendelian-consistent are often checked for genotyping errors. Three assumptions (see assumptions (h) and (i)) are given in this section: (1) there is at most one inconsistency in a nuclear family, (2) the genotyping errors are independent and

random and (3) there are no phenotyping errors. This work does not identify and adjust the errors in families displaying Mendelian consistency.

The error model (Table 14) used here is based on Douglas et al. (2002). I set $\eta = \gamma = \varepsilon$ in the R program to avoid failure of identifiability. Tests that only consider Mendelian-inconsistencies cannot give correct estimates of error rates since families displaying Mendelian consistency may also have genotyping errors (Gordon et al., 2001).

2.3.1 The likelihood function of a nuclear family with at most one genotyping inconsistency

Similar to the complete-data likelihood conditional on the observed data in equation (1), the likelihood function for one nuclear family (including the one with missing parental genotype data) with at most one inconsistency is:

$$L_{error} = \sum_{\vec{g}_{True}} L(\vec{g}_{True}, \vec{p}; \theta) \Pr_{error}(\vec{g}_{True} \mid \vec{g}_{Obs}, M) \quad (26)$$

where $\vec{g}_{Obs}$ refers to the observed genotype data in a nuclear family with at most one inconsistency, $\vec{g}_{True}$ refers to any possible set of genotypes corrected from $\vec{g}_{Obs}$, with $\sum_{\vec{g}_{True}} \Pr_{error}(\vec{g}_{True} \mid \vec{g}_{Obs}, M) = 1$, and M is an indicator for Mendelian consistency. If $M = 1$, the nuclear family is Mendelian-consistent. Then $\vec{g}_{True} = \vec{g}_{Obs}$, and $\Pr_{error}(\vec{g}_{True} \mid \vec{g}_{Obs}, M = 1) = \Pr_{error}(\vec{g}_{Obs} \mid \vec{g}_{Obs}) = 1$. Equation (26) is written as $L_{error} = L(\vec{g}_{Obs}, \vec{p}; \theta)$, which is the likelihood function of a nuclear family without genotyping errors (see Section 2.2.2).

If $M = 0$, the nuclear family has exactly one Mendelian inconsistency. Let m be the number of Mendelian consistent genotype sets for the family in which exactly one genotype has been corrected. I reorder the observed genotypes for the n family members so that correcting the first observed genotype can make the family Mendelian-consistent. That is, the observed and the corrected genotypes are reordered as $\vec{g}_{Obs.i} = \{g^{i.1}_{error}, g^{i.2}_{obs}, ..., g^{i.n}_{obs}\}$ and $\vec{g}_{True.i} = \{g^{i.1}_{true}, g^{i.2}_{true}, ..., g^{i.n}_{true}\}$, $i = 1,2,...,m$. In each $\vec{g}_{True.i}$, $g^{i.1}_{true}$ is the genotype corrected from $g^{i.1}_{error}$ ($g^{i.1}_{true} \neq g^{i.1}_{error}$) but $g^{i.k}_{true} = g^{i.k}_{obs}$, $k = 2,...,n$. The conditional probability

$$\Pr_{error}(\vec{g}_{True.i} \mid \vec{g}_{Obs.i}, M = 0) = \frac{\Pr(\vec{g}_{True.i} \mid \vec{g}_{Obs.i})}{\sum_{j=1}^{m} \Pr(\vec{g}_{True.j} \mid \vec{g}_{Obs.j})}, \quad i = 1,...,m, \quad (27)$$

where $\Pr(\vec{g}_{True.i} \mid \vec{g}_{Obs.i}) = \Pr(g^{i.1}_{true} \mid g^{i.1}_{error}) \Pr(g^{i.2}_{true} \mid g^{i.2}_{obs}) \cdots \Pr(g^{i.n}_{true} \mid g^{i.n}_{obs})$. Table 15 lists the values of

$$\Pr(g_{true} \mid g_{obs}) = \frac{\Pr(g_{obs} \mid g_{true}) \Pr(g_{true})}{\sum_{k=0}^{2} \Pr(g_{obs} \mid g_{true} = k) \Pr(g_{true} = k)}. \quad (28)$$

Recall that this work does not consider phenotyping errors.

In the following example and discussions, $g_{Obs.x}$ and $g_{True.x}$ denote the observed and the corrected genotype of x respectively, where x is the specified family member (F, M or C_{INC} whose genotypes are inconsistent) . For example, consider a nuclear family of size n. The observed parental genotypes $g_{Obs.F} = 0$ and $g_{Obs.M} = Miss$. The

second child with observed genotype 2 is denoted as C_{INC}. The observed genotypes of the remaining $n-3$ children are 1's. There are two possible consistent sets of genotypes for the nuclear family, with (1) $g_{True.F}=1$ corrected from $g_{Obs.F}=0$, or (2) $g_{True.C_{INC}}=1$ corrected from $g_{Obs.C_{INC}}=2$. Based on equations (27) and (28), I first compute

$$\Pr_{error}(\vec{g}_{True.1}\mid\vec{g}_{Obs.1},M=0)$$
$$=\frac{\Pr(g_{True.F}=1\mid g_{Obs.F}=0)\Pr(g_{True.C_{INC}}=2\mid g_{Obs.C_{INC}}=2)}{\Pr(g_{True.F}=1\mid g_{Obs.F}=0)\Pr(g_{True.C_{INC}}=2\mid g_{Obs.C_{INC}}=2)+\Pr(g_{True.F}=0\mid g_{Obs.F}=0)\Pr(g_{True.C_{INC}}=1\mid g_{Obs.C_{INC}}=2)}$$
$$=\frac{0.5\gamma\pi_1\cdot(1-\eta)\pi_2}{0.5\gamma\pi_1\cdot(1-\eta)\pi_2+(1-\eta)\pi_0\cdot0.5\gamma\pi_1}=\frac{\pi_2}{\pi_0+\pi_2}$$

$$\Pr_{error}(\vec{g}_{True.2}\mid\vec{g}_{Obs.2},M=0)=1-\Pr_{error}(\vec{g}_{True.1}\mid\vec{g}_{Obs.1},M=0)=\frac{\pi_0}{\pi_0+\pi_2}.$$

Then the likelihood factor from this family is

$$L_{error}=\sum_{i=1}^{2}L(\vec{g}_{True.i},\vec{p};\theta)\Pr_{error}(\vec{g}_{True.i}\mid\vec{g}_{Obs.i},M=0)$$
$$=\frac{\pi_2}{\pi_0+\pi_2}L(\{g_{True.F}=1,g_{True.M}=Miss,g_{True.C_1}=1,g_{True.C_{INC}}=2,g_{True.C_3}=1,\ldots g_{True.C_{n-2}}=1\},\vec{p};\theta)$$
$$+\frac{\pi_0}{\pi_0+\pi_2}L(\{g_{True.C_{INC}}=1,g_{True.F}=0,g_{True.M}=Miss,g_{True.C_1}=1,g_{True.C_3}=1,\ldots g_{True.C_{n-2}}=1\},\vec{p};\theta)$$

Consider all possible scenarios for one nuclear family with at most one inconsistency:

a. $\{g_{Obs.F},g_{Obs.M}\}=\{0,0\}$

If $n_{1.C}\geq2$, then $m=2$: (1) $g_{True.F}=1$, or (2) $g_{True.M}=1$.

If $n_{1.C}=1$, then $m=3$: (1) $g_{True.F}=1$, (2) $g_{True.M}=1$, or (3) $g_{True.C_{INC}}=0$ corrected from $g_{Obs.C_{INC}}=1$.

b. $\{g_{Obs.F},g_{Obs.M}\}=\{0,1\}$ or $\{g_{Obs.F},g_{Obs.M}\}=\{1,0\}$

If $n_{2.C}\geq2$, then $m=1$: $g_{True.F}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{0,1\}$, or $g_{True.M}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{1,0\}$.

If $n_{2.C}=1$, then $m=2$: (1) $g_{True.F}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{0,1\}$, or $g_{True.M}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{1,0\}$, or (2) $g_{True.C_{INC}}=1$ corrected from $g_{Obs.C_{INC}}=2$.

c. $\{g_{Obs.F},g_{Obs.M}\}=\{0,2\}$ or $\{g_{Obs.F},g_{Obs.M}\}=\{2,0\}$

If $n_{0.C}\geq1$ and $n_{1.C}\geq1$, then $m=1$: $g_{True.M}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{0,2\}$, or $g_{True.F}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{2,0\}$.

If $n_{0.C}\geq2$ and $n_{1.C}=0$, then $m=1$: $g_{True.M}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{0,2\}$, or $g_{True.F}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{2,0\}$.

If $n_{1.C}\geq1$ and $n_{2.C}\geq1$, then $m=1$: $g_{True.F}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{0,2\}$, or $g_{True.M}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{2,0\}$.

If $n_{1.C}=0$ and $n_{2.C}\geq2$, then $m=1$: $g_{True.F}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{0,2\}$, or $g_{True.M}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{2,0\}$.

If $n_{0.C}=1$ and $n_{1.C}=0$, then $m=2$: (1) $g_{True.M}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{0,2\}$, or $g_{True.F}=1$ when $\{g_{Obs.F},g_{Obs.M}\}=\{2,0\}$, or (2) $g_{True.C_{INC}}=1$ corrected from

$g_{Obs.C_{INC}} = 0$.

If $n_{1.C} = 0$ and $n_{2.C} = 1$, then $m = 2$: (1) $g_{True.F} = 1$ when $\{g_{Obs.F}, g_{Obs.M}\} = \{0, 2\}$, or $g_{True.M} = 1$ when $\{g_{Obs.F}, g_{Obs.M}\} = \{2, 0\}$, or (2) $g_{True.C_{INC}} = 1$ corrected from $g_{Obs.C_{INC}} = 2$.

d. $\{g_{Obs.F}, g_{Obs.M}\} = \{1, 2\}$ or $\{g_{Obs.F}, g_{Obs.M}\} = \{2, 1\}$

If $n_{0.C} \geq 2$, then $m = 1$: $g_{True.M} = 1$ when $\{g_{Obs.F}, g_{Obs.M}\} = \{1, 2\}$, or $g_{True.F} = 1$ when $\{g_{Obs.F}, g_{Obs.M}\} = \{2, 1\}$.

If $n_{0.C} = 1$, then $m = 2$: (1) $g_{True.M} = 1$ when $\{g_{Obs.F}, g_{Obs.M}\} = \{1, 2\}$, or $g_{True.F} = 1$ when $\{g_{Obs.F}, g_{Obs.M}\} = \{2, 1\}$, or (2) $g_{True.C_{INC}} = 1$ corrected from $g_{Obs.C_{INC}} = 0$.

e. $\{g_{Obs.F}, g_{Obs.M}\} = \{2, 2\}$

If $n_{1.C} \geq 2$, then $m = 2$: (1) $g_{True.F} = 1$, or (2) $g_{True.M} = 1$.

If $n_{1.C} = 1$, then $m = 3$: (1) $g_{True.F} = 1$, (2) $g_{True.M} = 1$, or (3) $g_{True.C_{INC}} = 2$ corrected from $g_{Obs.C_{INC}} = 1$.

f. $g_{Obs.F} = 0$ and $g_{Obs.M} = Miss$, or $g_{Obs.F} = Miss$ and $g_{Obs.M} = 0$

If $n_{2.C} \geq 2$, then $m = 1$: $g_{True.F} = 1$ when $g_{Obs.F} = 0$, or $g_{True.M} = 1$ when $g_{Obs.M} = 0$.

If $n_{2.C} = 1$, then $m = 2$: (1) $g_{True.F} = 1$ when $g_{Obs.F} = 0$, or $g_{True.M} = 1$ when $g_{Obs.M} = 0$, or (2) $g_{True.C_{INC}} = 1$ corrected from $g_{Obs.C_{INC}} = 2$.

g. $g_{Obs.F} = 2$ and $g_{Obs.M} = Miss$, or $g_{Obs.F} = Miss$ and $g_{Obs.M} = 2$

If $n_{0.C} \geq 2$, then $m = 1$: $g_{True.F} = 1$ when $g_{Obs.F} = 2$, or $g_{True.M} = 1$ when $g_{Obs.M} = 2$.

If $n_{0.C} = 1$, then $m = 2$: (1) $g_{True.F} = 1$ when $g_{Obs.F} = 2$, or $g_{True.M} = 1$ when $g_{Obs.M} = 2$, or (2) $g_{True.C_{INC}} = 1$ corrected from $g_{Obs.C_{INC}} = 0$.

Note that a nuclear family with one parent untyped and the other parent genotyped 1 is always Mendelian consistent, so is a nuclear family without parental genotypes.

Table 16 lists the conditional probabilities $\Pr_{error}(\vec{g}_{True.i} \mid \vec{g}_{Obs.i}, M = 0)$ for an arbitrary nuclear family with one inconsistency.

2.3.2 The likelihood function of a three- or four-generation pedigree with one or more genotyping inconsistencies

Some association tests (such as the FBAT) remove the pedigree with one or more genotyping inconsistencies from the analysis. Instead of sacrificing all the information from the pedigree, I first check the consistency of the high-level nuclear family. If it contains one inconsistency, I remove or adjust the genotypes of the subjects that cause the inconsistency using the Mendelian protocol given next. If one node' genotype in the high-level nuclear family is removed, I consider the node as untyped when checking the consistency and/or correcting the genotypes of a lower-level nuclear family where the node is one parent. The procedure is repeated until I correct all the inconsistencies from the pedigree.

Consider such a pedigree: it contains one or more inconsistencies, but each nuclear family decomposed from the pedigree contains at most one inconsistency. F, M and C_{INC} denote the father, the mother and the child whose genotypes are inconsistent. These three subjects are from a high-level, a middle-level or a low-level

nuclear family that contains exactly one inconsistency. Since I use equations in section 2.2.4 and 2.2.5 to compute the likelihood of the pedigree with the corrected genotypes, I use g_F, g_M and $g_{C_{err}}$ instead of $g_{True.F}$, $g_{True.M}$ and $g_{True.C_{err}}$.

a. $\{g_{Obs.F}, g_{Obs.M}\} = \{0,0\}$

If $n_{1.C} \geq 2$, I set $g_F = Miss$ and $g_M = Miss$.

If $n_{1.C} = 1$, I set $g_F = Miss$ and $g_M = Miss$. For C_{INC} with $g_{Obs.C_{INC}} = 1$, I set $g_{C_{INC}} = Miss$.

b. $\{g_{Obs.F}, g_{Obs.M}\} = \{0,1\}$ or $\{g_{Obs.F}, g_{Obs.M}\} = \{1,0\}$

If $n_{2.C} \geq 2$, I set $g_F = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{0,1\}$, or $g_M = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{1,0\}$.

If $n_{2.C} = 1$, I set $g_F = Miss$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{0,1\}$, or $g_M = Miss$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{1,0\}$. For C_{INC} with $g_{Obs.C_{INC}} = 2$, I set $g_{C_{INC}} = Miss$.

c. $\{g_{Obs.F}, g_{Obs.M}\} = \{0,2\}$ or $\{g_{Obs.F}, g_{Obs.M}\} = \{2,0\}$

If $n_{0.C} \geq 1$ and $n_{1.C} \geq 1$, I set $g_M = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{0,2\}$, or $g_F = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{2,0\}$.

If $n_{0.C} \geq 2$ and $n_{1.C} = 0$, I set $g_M = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{0,2\}$, or $g_F = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{2,0\}$.

If $n_{1.C} \geq 1$ and $n_{2.C} \geq 1$, I set $g_F = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{0,2\}$, or $g_M = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{2,0\}$.

If $n_{1.C} = 0$ and $n_{2.C} \geq 2$, I set $g_F = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{0,2\}$, or $g_M = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{2,0\}$.

If $n_{0.C} = 1$ and $n_{1.C} = 0$, I set $g_M = Miss$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{0,2\}$, or $g_F = Miss$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{2,0\}$. For C_{INC} with $g_{Obs.C_{INC}} = 0$, I set $g_{C_{INC}} = Miss$.

If $n_{1.C} = 0$ and $n_{2.C} = 1$, I set $g_F = Miss$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{0,2\}$, or $g_M = Miss$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{2,0\}$. For C_{INC} with $g_{Obs.C_{INC}} = 2$, I set $g_{C_{INC}} = Miss$.

d. $\{g_{Obs.F}, g_{Obs.M}\} = \{1,2\}$ or $\{g_{Obs.F}, g_{Obs.M}\} = \{2,1\}$

If $n_{0.C} \geq 2$, I set $g_M = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{1,2\}$, or $g_F = 1$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{2,1\}$.

If $n_{0.C} = 1$, I set $g_M = Miss$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{1,2\}$, or $g_F = Miss$ if $\{g_{Obs.F}, g_{Obs.M}\} = \{2,1\}$. For C_{INC} with $g_{Obs.C_{INC}} = 0$, I set $g_{C_{INC}} = Miss$

e. $\{g_{Obs.F}, g_{Obs.M}\} = \{2,2\}$

If $n_{1.C} \geq 2$, I set $g_F = Miss$ and $g_M = Miss$.

If $n_{1.C} = 1$, I set $g_F = Miss$ and $g_M = Miss$. For C_{INC} with $g_{Obs.C_{INC}} = 1$, I set $g_{C_{INC}} = Miss$.

f. $g_{Obs.F} = 0$ and $g_{Obs.M} = Miss$, or $g_{Obs.F} = Miss$ and $g_{Obs.M} = 0$

If $n_{2.C} \geq 2$, I set $g_F = 1$ if $g_{Obs.F} = 0$, or $g_M = 1$ if $g_{Obs.M} = 0$.

If $n_{2.C} = 1$, I set $g_F = Miss$ if $g_{Obs.F} = 0$, or $g_M = Miss$ if $g_{Obs.M} = 0$. For C_{INC} with $g_{Obs.C_{INC}} = 2$, I set $g_{C_{INC}} = Miss$.

g. $g_{Obs.F} = 2$ and $g_{Obs.M} = Miss$, or $g_{Obs.F} = Miss$ and $g_{Obs.M} = 2$

If $n_{0.C} \geq 2$, I set $g_F = 1$ if $g_{Obs.F} = 2$, or $g_M = 1$ if $g_{Obs.M} = 2$.

If $n_{0.C} = 1$, I set $g_F = Miss$ if $g_{Obs.F} = 2$, or $g_M = Miss$ if $g_{Obs.M} = 2$. For C_{INC} with $g_{Obs.C_{INC}} = 0$, I set $g_{C_{INC}} = Miss$.

A more exact analysis would generate m consistent genotype sets that have only one genotyping correction from the observed genotypes and then weight these consistent genotype sets in the likelihood function as in section 2.3.1. This will be a subject of my future work.

2.4 The likelihood ratio test statistic

The likelihood ratio test statistic for N families with observed marker genotypes $\vec{G}$ and phenotypes $\vec{P}$ is:

$$LRT = 2[\ln L_{error}(\vec{G}, \vec{P}; \hat{\pi}_0, \hat{\pi}_1, \hat{\phi}_0, \hat{\phi}_2) - \ln L_{error}(\vec{G}, \vec{P}; \hat{\hat{\pi}}_0, \hat{\hat{\pi}}_1, \hat{\hat{\phi}}_0)] \quad (29)$$

where $L_{error}(\vec{G}, \vec{P}; *) = \prod_{l=1}^{N} L_{error}(\vec{g}_l, \vec{p}_l; *)$ and the likelihood factor from one family $L_{error}(\vec{g}_l, \vec{p}_l; *)$ may be found in equation (26). In equation (29), $\hat{\pi}_0, \hat{\pi}_1, \hat{\phi}_0, \hat{\phi}_2$ are the estimates of $\pi_0, \pi_1, \phi_0, \phi_2$ under the alternative hypothesis, and $\hat{\hat{\pi}}_0, \hat{\hat{\pi}}_1, \hat{\hat{\phi}}_0$ are the estimates under the null hypothesis.

2.5 Null simulation

I simulate nine sets of 500 null replicates with full or missing genotypes and phenotypes. Seven sets have no genotyping error, and two contain genotyping errors.

Each replicate in the first set contains 200 case-parent trios, for a total of 600 individuals. The replicate in the second set contains 200 quartets (two parents, one affected child and one unaffected child), for a total of 800 individuals. All the replicates in these two sets are 100% genotyped and do not contain genotyping errors. I set marker allele frequencies to $\Pr(a) = 0.9$ and $\Pr(b) = 0.1$, disease allele frequencies $\Pr(+) = 0.88$ and $\Pr(d) = 0.12$, and the proportion of maximum linkage disequilibrium $D' = 0.95$. I specify equal disease penetrances $f_0 = f_1 = f_2 = 0.1$. Recombination fraction is set at 0.5, indicating that the marker is unlinked to the DSL. The simulation program SLINK program (Weeks et al, 1990) is used to simulate the parental phenotypes and marker genotypes when fixing the affection status of the children.

Each replicate in the third set contains 100 quartets, with 75% affected sib pairs and 25% sib pairs with discordant affection status. I set marker allele frequencies $\Pr(a) = 0.6$ and $\Pr(b) = 0.4$. The SIMULATE program (Terwilliger and Ott, 1994) is used to simulate the null data. Note that only the marker allele frequencies are used to simulate the genotype data since the SIMULATE program ignores the DSL specifications in simulation. All the replicates in the third set are 80% genotyped and do not contain genotyping errors. An R function is written to remove randomly 20% phenotypes and 20% genotypes from the simulated data.

The fourth to the sixth sets use 92 fixed nuclear family structures with a total of 366 individuals in each replicate. 46 nuclear families derived from a previously published IS study (Gao et al., 2007) are replicated twice in each null data replicate. The largest nuclear family contains 7 individuals, while the smallest has 3. The

median size of these nuclear families is 4. I set marker allele frequencies $\Pr(a) = 0.6$ and $\Pr(b) = 0.4$. The fourth set uses 100% genotyped data, and the fifth and the sixth sets use 80% genotyped data. I also randomly insert genotyping errors with error rate $\varepsilon = 0.01$ in the sixth set. The SIMULATE program is used to simulate the genotypes, and an R function is written to randomly insert genotyping errors and then remove genotypes.

In the last three sets, 53 fixed multiplex IS families with a total of 313 individuals are used for each replicate. The largest family contains 18 individuals, while the smallest contains 3. The median family size is 6. I set marker allele frequencies $\Pr(a) = 0.6$ and $\Pr(b) = 0.4$. The SIMULATE program is used to simulate the null data. The seventh set uses 100% genotyped data, and the eighth and the ninth sets use 80% genotyped data. I also randomly insert genotyping errors with error rate $\varepsilon = 0.01$ in the ninth set.

I use the Kolmogorov-Smirnov (KS) goodness of fit test (Kolmogoroff, 1941; Smirnov, 1939) to determine whether the null distribution of the LRT for each setting fits well to a central χ^2 with one degree of freedom. The decision rule is that a p-value > 0.05 of KS test statistic indicates that the data comes from a central χ^2 distribution with one degree of freedom.

2.6 Power comparison

To compare the power of the TDT and this LRT, I perform an unrepeated 2^3 factorial design with three factors: disease genotype relative risk $R_1 = f_1 / f_0$ (1.75 or 2), marker allele frequency $p(b)$ (0.1 or 0.2), and number of trios or quartets (125 or 175). Note that this work assumes multiplicative mode of inheritance, so that $R_2 = R_1^2$. The disease allele frequency $\Pr(d)$ is 0.12 or 0.24 when marker allele frequency $\Pr(b)$ is 0.1 or 0.2, with $D' = 0.95$. Recombination fraction is set at 0. For each setting, I first compare the power of the TDT and the power of this LRT on the same number of case-parent trios. To test the statistical significance of main and two-way interaction effects on the power difference (power of the LRT – power of the TDT), ANOVA is used for the parsimonious model after removing the non-significant effects at 10% significant level. Then I compare the power of the LRT on the same number of trios and quartets (two parents, one affected child and the other unaffected). I also compare the power of the LRT on the same number of trios or quartets when the simulated parental phenotype data is used or not. Finally, I perform power calculation of the LRT on trios with 80% available genotypes, and trios with 1% genotyping errors and 80% available genotypes. The power of the TDT is calculated analytically using the R package powerpkg developed by Weeks (2005), based on the asymptotic power formula of Abel and Muller-Myhsok (1998). The power of the LRT is calculated via Monte Carlo computer simulation with 500 replicates. The data for each replicate is simulated by the SLINK program.

2.7 Applications

Idiopathic scoliosis data for CHD7 gene

I apply the likelihood method to a previously published genetic study for idiopathic scoliosis (IS), a common disease of children displaying a complex inheritance pattern but lacking known causative genes. In that study, a follow-up

analysis of genome-wide linkage scans provided supporting evidence of linkage to human chromosome 8q12 in a total cohort of 53 multiplex families in which 130 individuals were affected. A fine-mapping study of the CHD7 candidate gene encoded in the 8q12 candidate region was subsequently performed by genotyping 25 single nucleotide polymorphic (SNP) loci evenly spaced throughout the ~93 kb region. Two of the 25 SNPs were not sufficiently polymorphic and were dropped. Application of TDT methods produced significant results for ~19 of the 23 SNPs. Re-sequencing conserved regions underlying the peak of association identified a potential functional SNP, rs4738824, which was also significant in tests of transmission disequilibrium. These data identified CHD7 as the first candidate gene for IS (Gao et al., 2007).

I first study the 92 nuclear families derived from 53 IS families. The median family size is 4, and there are 352 individuals (including duplicated individuals after decomposition), of whom 145 were affected. While there are no genotype inconsistencies in this data, there were 58.8% individuals with missing phenotype information, and approximately 30% individuals with missing genotype data on each SNP locus. I apply the likelihood method to test association between each of the 23 SNPs and IS. I compare the results with those obtained by TDTae under the multiplicative mode of inheritance.

Then I apply this LRT to the cohort of 53 IS families with 313 individuals. Among these individuals, 133 were affected with IS while the phenotypes of the remaining 180 were unknown. The largest family size is 16. The median family size is 5. The data contains 3 four-generation pedigrees, 18 three-generation pedigrees, and 32 nuclear families, among which 3 are *f*-multiple marriages. I apply the likelihood method to test association between each of the 23 SNPs and IS. I compare the results with those reported by Gao et al. (2007).

Psoriasis data on chromosome 17q25

The second application is to a psoriasis study. The data contains 79 SNPs and 29 polymorphic microsatellites from chromosome 17q25 at an average resolution of 80kb genotyped in 242 psoriasis families with multiple affected and unaffected individuals each family. There are 1056 individuals, of whom 596 (56.4%) were affected and 221 (20.9%) were unaffected. The largest family size is 13. The median family size is 4. The data contains 6 three-generation pedigrees, and 236 nuclear families, among which 4 are *f*-multiple marriages. The previously published study identifies significant linkage for multiple SNPs and two peaks of strong association with psoriasis in 17q25 region on Chromosome 17 (Helms et al., 2003). Gordon et al. (2004) further studied 16 SNPs in this region. They found that two SNPs displayed significant evidence of linkage at the 5% significance level after correction for multiple testing via the false discovery rate method (Benjamini and Hochberg, 1995). Gordon et al. (2004) also detected inconsistent genotypes at each of these SNPs.

I apply this LRT for 13 of the SNPs to test association between each SNP and psoriasis. Approximately 30% individuals were untyped on each of the 13 SNPs. I compare the results with those by the TDTae under multiplicative mode of inheritance and the FBAT using additive coding (Laird, 2006). Note that the additive coding in FBAT reflects an underlying additive or multiplicative mode of inheritance (Laird, 2006). I also compare the genotype relative risks estimated by the TDTae and this LRT. To detect the genotyping errors, I write an R function that can identify families with genotype inconsistencies. The FBAT removes the inconsistent families from the analysis, while the LRT and the TDTae incorporate the inconsistencies into the likelihood functions.

Chapter 3 Grid-UOBYQA algorithm

Gordon et al. (2004) applied a two-stage optimization procedure to locate the maximum log-likelihoods under each hypothesis. They used Powell's quadratically convergent algorithm as implemented in the 'Numerical Recipes in C' text (Press et al., 2002). Their results suggest that the grid search and Powell algorithm, both of which are direct search algorithms, work efficiently for their likelihood-based method for general pedigrees with missing parental genotypes and genotyping errors.

Many direct search algorithms, including line search methods, the restriction of vectors of variables to discrete grids, the use of geometric simplexes, conjugate direction procedures, and true region algorithms that form linear or quadratic approximations to the objective function, have been proposed for optimization calculations that do not require the calculation of derivatives (Powell 1998). Among these algorithms, Powell's quadratically convergent method (Powell, 1964), denoted as Powell in Gordon et al. (2004), was widely used and extended with 927 citations to date. However, the problem of linear dependence in this algorithm may make the search procedure end with the maximum/minimum of the objective function only over a subspace of the full n-dimensional case (Press et al., 2002). The linear dependence problem was fixed by a singular value decomposition algorithm (Press et al., 2002).

The UOBYQA (Unconstrained Optimization BY Quadratic Approximation), another derivative-free method developed by Powell (2000) for general unconstrained optimization, uses multivariate quadratic Lagrange interpolations to approximate the objective function and uses the trust region technique (Celis et al., 1985) to ensure convergence. It uses two trust region radii. The first radius is similar to the trust region radius in the standard trust region method, while the second radius is used as a stopping criterion to control the goodness of the quadratic model. Numerical results and theoretical analyses show that the UOBYQA algorithm is globally convergent for general objective functions when the second trust region radius converges to zero. It has also displayed quadratic convergence in numerical experiments (Powell, 2000; Han and Liu, 2004). I implement the fixed Powell algorithm in R and compare its convergence rate with that of UOBYQA by R package powell (Powell, 2000). Results of numerical experiments show that the UOBYQA appears to have a faster convergence rate.

Since UOBYQA is specifically designed for unconstrained optimization calculation in multi-dimensions, one strategy to apply it to constrained optimization problems is to constrain the search region by setting an infinite value to the objective function when the search reaches beyond the parameter space bounded by the lower and upper limits for each parameter. Since the discrete grid method was specifically proposed for variables bounded by constraints (Torczon, 1997), on the constrained search region, the discrete grid method can be used to identify several starting points around which the optimal point may lie. Motivated by the simplicity of the discrete grid search and the advantage in convergence of the UOBYQA, this work uses a grid search in the first stage and the UOBYQA in the second stage similar to the two-stage maximization procedure applied in the TDTae (Gordon et al., 2004). The composite algorithm is called grid-UOBYQA.

The grid-UOBYQA algorithm parameters are a superset of those for the grid selection and UOBYQA. At each grid point, the values of the objective function are computed and then compared. Those grid points corresponding to the first few largest/lowest values of the objective function are selected as the starting points for the UOBYQA optimization. For each starting point, UOBYQA determines the local

fit-statistic maximum/minimum. The largest/smallest of all observed maxima/minima is then considered as the global fit-statistic maximum/minimum. The advantage of Grid-UOBYQA is that it can provide a thorough sampling of the parameter space. For a continuously differentiable objective function like the log-likelihood in this work, this combined algorithm can make full use of the advantages of both grid search and UOBYQA. It is good for situations where the best-fit parameter values are not easily determined a priori, and where there is a high probability that false maxima/minima would be found if one-shot techniques such as UOBYQA are used instead (Freeman et al., 2001). However, the biggest disadvantage is that it can be very slow, especially when the number of grid points is large.

In this work, grid-UOBYQA is applied to maximize the log-likelihoods under each hypothesis. First I identify the K best starting points for the parameters $\{\pi_0, \pi_1, \phi_0\}$ under the null hypothesis $\phi_0 = \phi_1 = \phi_2$, selecting from G^3 grid points on a 3-dimensional rectangle. From each of these K starting points, UOBYQA optimization will end with a local maximum of log-likelihood together with the corresponding estimates $\{\hat{\pi}_{0k}, \hat{\pi}_{1k}\}, k = 1,..., K$. Then I use these estimates as the starting values for the genotype frequency estimation. I start with KG^2 grid points to find the optimal value for $\{\pi_0, \pi_1, \phi_0, \phi_2\}$. That is, for each of the K estimates $\{\hat{\pi}_{0k}, \hat{\pi}_{1k}\}, k = 1,..., K$, I examine G^2 rectangular grid points for $\{\phi_0, \phi_2\}$. I select the K best starting points from the KG^2 grid points. As in the null hypothesis likelihood maximization procedure, the UOBYQA algorithm will find the local maximum of log-likelihood under the alternative hypothesis starting from each of the K best points. If the K searches under each hypothesis locate the same local maximum, I consider this local maximum to be the global maximum under that hypothesis. If this condition is not met, I try a larger G and a larger K. In the event of failure of a common maximum, I report the largest observed log-likelihood and the corresponding parameter estimates under each hypothesis.

For example, suppose that I use a grid search for the $K = 6$ best starting points for the parameters $\{\pi_0, \pi_1, \phi_0\}$ using $G = 5$ values starting from 0 and ending at 1, with an increment of 0.25. The log-likelihood will be computed $5^3 = 125$ times and the grid search will end with 6 starting points corresponding to the six largest log-likelihood values under the null. If, for each of the six searches, the UOBYQA finds the same local maximum log-likelihood, I denote the local maximum as the global maximum, and record the corresponding estimate $\{\hat{\pi}_{0k}, \hat{\pi}_{1k}, \hat{\phi}_{0k}\}$. Then for each of the six estimated pairs $\{\hat{\pi}_{0k}, \hat{\pi}_{1k}\}$, each of the parameters $\{\phi_0, \phi_2\}$ is tested at 5 values: [0, 0.25, 0.5, 0.75, 1]. Starting from the six best starting points out of $6 \times 5^2 = 150$, the UOBYQA search will give the local maxima under the alternative. If the six searches under the null or the alternative hypothesis do not converge to the same maximum, I use $K = 8$ and $G = 10$. If new search fails to locate the same maximum, I denote the largest among the eight local maximums as the maximum under the null and the alternative hypotheses.

This work uses the Powell package (Powell, 2000) for the UOBYQA search procedures. The R program to implement the grid-UOBYQA maximization is available on http://www.ams.sunysb.edu/~yayang.

Chapter 4 Results

4.1 Null simulations

Table 17 lists the *p*-values from the KS goodness of fit test for each simulation and the empirical type I error rates and their 95% confidence intervals. These results suggest that, for these simulated data sets, the empirical distribution of LRT appears to fit to a central χ^2 distribution with one degree of freedom at the 5% significance level. Specifically, this LRT is valid as a test of linkage in the presence of allelic association, or as a test of linkage or allelic association.

4.2 Power comparison

Table 18 lists the powers of the TDT and this LRT based on the 2^3 factorial design. The results suggest that the LRT is better than the original TDT (compare columns labeled TDT and LRT^b), even in the presence of 20% missing genotype data and 1% genotyping errors (compare columns labeled TDT, LRT^c, and LRT^d). Since the parental phenotypes are simulated in the SLINK, removing the parental phenotype information will decrease the power to test association of marker with disease. That is the reason why the values in columns labeled LRT^a and LRT^e are smaller than those in any other column. The values in columns labeled LRT^e and LRT^f are larger than those in columns labeled LRT^a and LRT^b, respectively, indicating that including information of the unaffected children in the LRT appears to increase power of the family-based association test.

Table 19 displays the ANOVA table of the unrepeated three-level design for the power difference of the TDT and the LRT (see values in columns labeled TDT and LRT^b in Table 18). The three main effects, genotype relative risk at the disease locus (GRRD), marker allele frequency (MAF), and number of trios (NT), and the interaction of GRRD and NT are significant at the 10% significance level. Only the main effects are significant at the 5% significance level.

4.3 Application to real datasets

Idiopathic scoliosis data for CHD7 gene

The results of this LRT and the TDTae for the 92 IS nuclear families are compared in Figure 12, which shows $-\log_{10}(p\text{-value})$ for each test. The results of the LRT are consistent with those by TDTae. Specifically, $p\text{-value} < 0.05$ is equivalent to $-\log_{10}(p\text{-value}) > 1.3$. Seven of the 23 SNPs have *p*-values that are less than 0.001 (or equivalently, $-\log_{10}(p\text{-value}) > 3$). They are, from with the smallest *p*-value to the largest, rs7843033, rs7000766, hcv509504, rs1038851, hcv509505, hcv148921, and rs7842389. The marker with the largest LRT of 12.96 ($p\text{-value} = 0.000319$; $-\log_{10}(p\text{-value}) = 3.50$) is rs7843033. This marker is the most significant marker in the TDTae analysis. For marker rs7843033, the estimated genotype relative risks at the marker locus are $\hat{\bar{R}}_1 = \hat{\phi}_1 / \hat{\phi}_0 = 2.40$ and $\hat{\bar{R}}_2 = \hat{\phi}_2 / \hat{\phi}_0 = 5.77$, consistent with the strong genetic effect observed for this marker locus in the previous study (Gao et al., 2007).

Figure 13 compares the results of this LRT and other family-based tests in Gao et al. (2007) for the 53 multiplex IS families. Table 20 lists more detailed results,

including the marker genotype relative risks estimated by the TDTae and the LRT. The results of the LRT are similar to those of the haplotype-based haplotype relative risk (HHRR) association test (Terwilliger and Ott, 1992), the affected sib pair (ASP) linkage test (Terwilliger, 1995) and the TDTae. Eight of the 23 SNPs have LRT p-values that are less than 0.001 (or equivalently, $-\log_{10}(p\text{-value}) > 3.0$). From with the smallest p-value to the largest, these eight SNPs are rs1483207, rs7843033, rs4392940, rs7000766, rs4237036, rs1038351, hcv148921, and rs7017676. The most significant marker identified by the HHRR, the TDTae and the ASP are rs7017676, rs7843033, and rs7000766, respectively. All of them are among the eight most significant SNPs identified by this LRT.

Figure 12: The TDTae and LRT p-values ($-\log_{10}$ transformed) on 23 SNPs in CHD7 gene for the 92 IS nuclear families

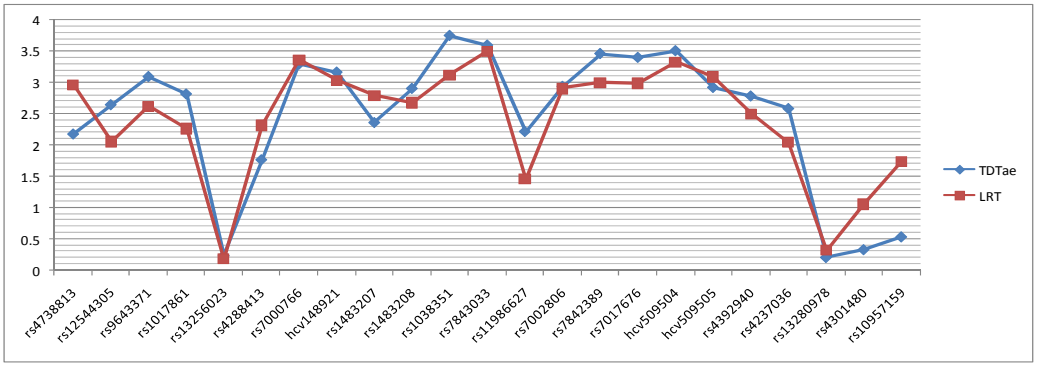
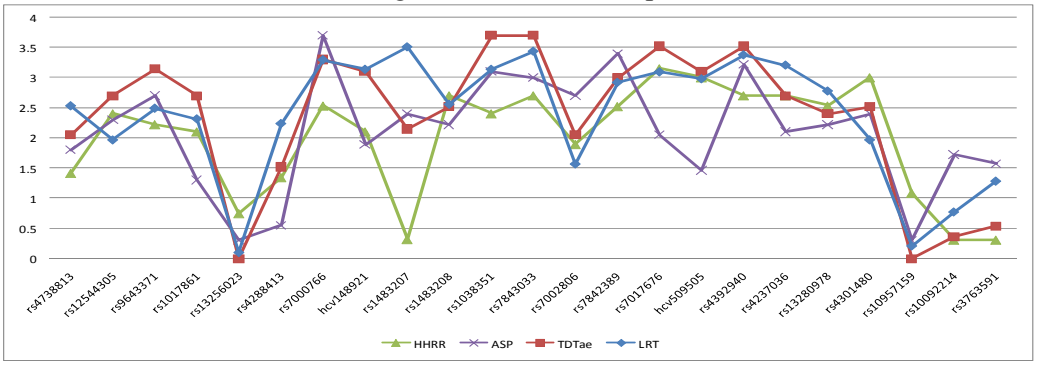


Figure 13: The HHRR, ASP, TDTae and LRT p-values ($-\log_{10}$ transformed) for 23 SNPs in CHD7 gene for the 53 multiplex IS families

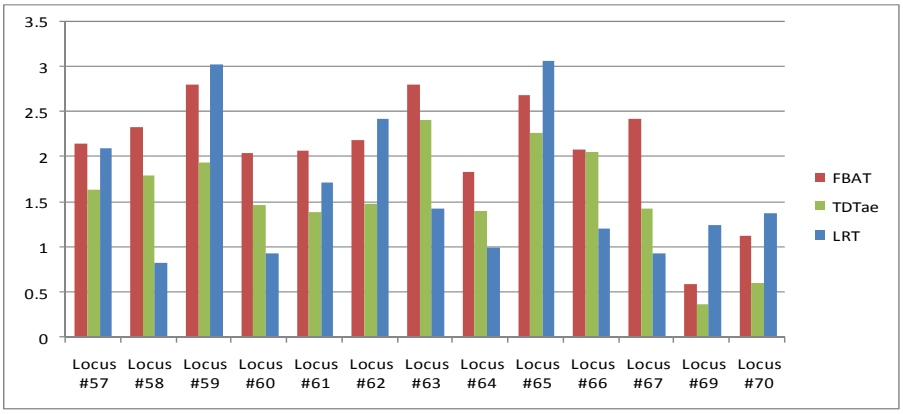


Psoriasis data on chromosome 17q25

Figure 14 compares the results of the FBAT, the TDTae and this LRT on 13 SNPs, as indicated by $-\log_{10}$(p-value). More detailed results are listed in Table 21, including the genotype relative risks estimated by the TDTae and the LRT. At the 5% significance level, the FBAT and the TDTae identifies 11 significant markers while the LRT finds 7 significant markers. The three most significant markers by the FBAT, locus#59, locus#62, and locus#65, also display significant association with the disease using the TDTae and the LRT. Specifically, locus#65 is the second most significant marker identified by the TDTae. Locus#65 is the most significant marker, and locus#59 is the second most significant marker by the LRT. For locus#59, the FBAT p-value is 0.0015, the TDTae p-value is 0.0116, and the LRT p-value is 0.0009, with

estimated relative risk of the marker genotypes $\hat{\bar{R}}_1 = 1.47$ and $\hat{\bar{R}}_2 = 2.16$. For locus#65, the FBAT *p*-value is 0.0020, the TDTae *p*-value is 0.0053, and the LRT *p*-value is 0.00086, with $\hat{\bar{R}}_1 = 1.47$ and $\hat{\bar{R}}_2 = 2.17$. The LRT gives quite different results from the other two methods on locus#58, locus#60, locus#64 and locus#67. The FBAT and the TDTae have significant results on the four SNPs, while the LRT *p*-values are greater than 0.05. I conjecture that one reason for the different results may be that the genotypes of the unaffected individuals are incorporated into the test statistic of the LRT. As reviewed in Chapter 1, the TDTae considers only affected offspring. Although the FBAT uses genotypes of unaffected offspring to infer the incomplete parental genotypes, the test statistic does not contain the genotypes of the unaffected children. Another possible reason is that the FBAT does not use the information from the inconsistent families while the LRT and the TDTae incorporate the inconsistencies into the likelihood functions. The Mendelian check on genotypes finds that there are two or more families (6 at most) with inconsistent genotypes at each of these SNP loci.

Figure 14: The FBAT, TDTae and LRT *p*-values ($-\log_{10}$ transformed) for 13 SNPs on chromosome 17q25 for 242 psoriasis families



Chapter 5 Discussion

5.1 The likelihood function

The overall likelihood for a pedigree is

$$L = \sum_{g_1} \cdots \sum_{g_n} \prod_{founder} \Pr(g_{founder}) \prod_{\{c,f,m\}} \Pr(g_c \mid g_f, g_m) \prod_i \Pr(p_i \mid g_i),$$

where $g_{founder}$ denotes a founder's genotype, g_f, g_m and g_c denote genotypes for father, mother and a child, and p_i and g_i the phenotype and genotype for the *i*th individual in the pedigree of size n. The likelihood function consists of three factors from left to right: (1) probability of founder genotypes, (2) probability of children's genotypes given parental genotypes, and (3) probability of phenotypes given genotypes for all individuals in a pedigree (Sham, 1997). Since the likelihood requires intensive computation, many algorithms have been proposed to speed the calculation (Elston and Stewart, 1971; Lander and Green, 1987; Kruglyak and Lander, 1998).

One of the most widely used algorithms for likelihood computation is the Elston-Stewart algorithm (Elston and Stewart, 1971). Their algorithm works efficiently for larger but simple pedigrees and a small number of markers. It computes the likelihood as a function of the recombination fraction between a disease and marker locus. That is, their likelihood algorithm is designed to test for linkage whether or not there is an association (Gordon et al., 2004). Weeks suggests that likelihood method based on the overall pedigree likelihood conditional on the parental marker genotypes can be used as an alternative to the TDT (personal communication through email).

There are other conditional likelihood methods that are proposed to test association as extensions of the TDT (Spielman et al., 1993; Schaid and Sommer, 1993; Whittemore and Tu, 2004). Schaid and Sommer presented two likelihood methods to test association between marker and disease for trios of two parents and one affected child: (1) a likelihood method appropriate when HWE holds and (2) a likelihood method conditional on parental genotypes when HWE does not hold. The first Schaid-Sommer likelihood for n trios is

$$L_{SS.1} = \prod_{i=1}^{n} \frac{\Pr(g_{C_i}, p_{C_i} = A, g_F, g_M)}{\Pr(p_{C_i} = A)} = \prod_{i=1}^{n} \Pr(g_F, g_M \mid p_{C_i} = A) \Pr(g_{C_i} \mid p_{C_i} = A, g_F, g_M)$$

where

$$\frac{\Pr(g_{C_i} = k, p_{C_i} = A, g_F, g_M)}{\Pr(p_{C_i} = A)} = \frac{\phi_k \Pr(g_{C_i} = k \mid g_F, g_M) \Pr(g_F, g_M)}{\phi_0\pi_0 + \phi_1\pi_1 + \phi_2\pi_2}, k = 0,1,or 2.$$

The second Schaid-Sommer likelihood is

$$\begin{aligned} L_{SS.2}(\overline{R}_1, \overline{R}_2) = C & \left(\frac{\overline{R}_2}{\overline{R}_1 + \overline{R}_2}\right)^{x_{22}} \left(\frac{\overline{R}_1}{\overline{R}_1 + \overline{R}_1}\right)^{x_{21}} \\ & \times \left(\frac{\overline{R}_2}{\overline{R}_2 + 2\overline{R}_1 + 1}\right)^{x_{42}} \left(\frac{2\overline{R}_1}{\overline{R}_2 + 2\overline{R}_1 + 1}\right)^{x_{41}} \left(\frac{1}{\overline{R}_2 + 2\overline{R}_1 + 1}\right)^{x_{40}}, \\ & \times \left(\frac{\overline{R}_1}{\overline{R}_1 + 1}\right)^{x_{51}} \left(\frac{1}{\overline{R}_1 + 1}\right)^{x_{50}} \end{aligned}$$

where the constant $C = \binom{n_2}{x_{22}}\binom{n_4}{x_{42}x_{41}x_{40}}\binom{n_5}{x_{51}}$ with n_i the number of affected children from mating type i (see Table 1), and x_{ij} the number of affected children from mating type i with j mutant alleles b .Their results for the relative efficiency of these two likelihood methods suggest that their second likelihood method may at times be preferable, even when HWE holds.

Based on Schaid and Sommer's work, Whittemore and Tu (2000) use the likelihood function for a family at locus *t*,

$$L_{WT} = \Pr(\vec{g} \mid R, \vec{p}, t) = \frac{\Pr(\vec{g})\Pr(\vec{p} \mid \vec{g}, R, t)}{\Pr(\vec{p} \mid R, t)}.$$

Their likelihood is defined as the probability of the family's observed marker genotypes $\vec{g}$, given the family's genealogical structure *R*, the vector of phenotypes $\vec{p}$, and that *t* is a DSL. Let $\vec{g}_{DSL}$ denote the vector of genotypes at a DSL. They have

$$\Pr(\vec{p} \mid \vec{g}) = \sum_{\vec{g}_{DSL}} \Pr(\vec{p} \mid \vec{g}_{DSL}) \Pr(\vec{g}_{DSL} \mid \vec{g})$$

after suppressing the dependence of the probabilities both on the family structure R and on the particular locus t. Here $\sum_{\vec{g}_{DSL}}$ denotes summation over all possible genotype vector $\vec{g}_{DSL}$. Substitution of the equation above into L_{WT} gives the conditional likelihood as

$$\Pr(\vec{g} \mid \vec{p}) = \Pr(\vec{g}) \sum_{\vec{g}_{DSL}} \frac{\Pr(\vec{p} \mid \vec{g}_{DSL})}{\Pr(\vec{p})} \Pr(\vec{g}_{DSL} \mid \vec{g}).$$

When the marker locus is at the DSL ($\vec{g}_{DSL} = \vec{g}$) or near to the DSL, $\Pr(\vec{g}_{DSL} \mid \vec{g})$ equals or approaches 1. The likelihood for a family

$$\Pr(\vec{g} \mid \vec{p}) = \Pr(\vec{g}) \sum_{\vec{g}_{DSL}} \frac{\Pr(\vec{p} \mid \vec{g}_{DSL})}{\Pr(\vec{p})} \cdot 1 = \Pr(\vec{g}_{DSL}) \frac{\Pr(\vec{p} \mid \vec{g}_{DSL})}{\Pr(\vec{p})} = \Pr(\vec{g}_{DSL} \mid \vec{p}).$$

That is, the following likelihood of marker genotypes given the phenotypes and the family structure *R* can be used to test association for general nuclear families:

$$L_{WT*} = \Pr(\vec{g} \mid \vec{p}, R) = \Pr(g_F, g_M, \vec{g}_{\bar{C}} \mid p_F, p_M, \vec{p}_{\bar{C}})$$
$$= \Pr(g_F, g_M \mid p_F, p_M, \vec{p}_{\bar{C}}) \Pr(\vec{g}_{\bar{C}} \mid g_F, g_M, p_F, p_M, \vec{p}_{\bar{C}}).$$

The likelihood function in this work has two factors. The first factor, $L_{Founder} = \Pr(g_F, g_M \mid p_F, p_M)$, uses founder's genotypes and phenotypes to estimate population frequencies of parental marker genotypes (under HWE, it can be used to estimate population frequencies of marker genotypes for all generations). The second factor, $L_{Nonfounder} = \Pr(\vec{g}_{\bar{C}} \mid g_F, g_M, \vec{p}_{\bar{C}})$, evaluates disequilibrium in transmission of marker alleles from parents to offspring. It follows the approach of the second likelihood function in Schaid and Sommer (1993). The product of these two factors gives the likelihood factor from a general nuclear family with complete parental genotypes:

$$L_{YY} = L_{Founder} L_{Nonfounder} = \Pr(g_F, g_M \mid p_F, p_M) \Pr(\vec{g}_{\bar{C}} \mid g_F, g_M, \vec{p}_{\bar{C}}).$$

I use a different notation L_{YY} here to denote this likelihood for the convenience of comparison of these likelihood functions.

Based on the review of the likelihood methods to test association (Schaid, 1996; Clayton, 1999; Whittemore and Tu, 2000), Laird and Lange (2006) concluded that

with parental data, all information on association is contained in $L_{Nonfounder}$ and likelihood-ratio tests based on $L_{Nonfounder}$ are optimal. Therefore, the LRT based on L_{YY} also contains the information needed to test association.

Note that L_{YY} will be equivalent to L_{WT*} if the following two equations hold:

$$\Pr(\vec{g}_{\bar{C}} \mid g_F, g_M, \vec{p}_{\bar{C}}) = \Pr(\vec{g}_{\bar{C}} \mid g_F, g_M, p_F, p_M, \vec{p}_{\bar{C}}) \quad (30)$$

$$\Pr(g_M, g_F \mid p_M, p_F) = \Pr(g_M, g_F \mid p_M, p_F, \vec{p}_{\bar{C}}) \quad (31)$$

For a child, one can prove that

$$\Pr(g_C \mid g_M, g_F, p_M, p_F, p_C) = \frac{\Pr(g_C, g_M, g_F, p_M, p_F, p_C)}{\Pr(g_M, g_F, p_M, p_F, p_C)}$$

$$= \frac{\Pr(g_M, g_F)\Pr(p_F \mid g_F)\Pr(p_M \mid g_M)\Pr(g_C \mid g_M, g_F)\Pr(p_C \mid g_C)}{\sum_{g_C} \Pr(g_M, g_F)\Pr(p_F \mid g_F)\Pr(p_M \mid g_M)\Pr(g_C \mid g_M, g_F)\Pr(p_C \mid g_C)}$$

$$= \frac{\Pr(g_M, g_F)\Pr(g_C \mid g_M, g_F)\Pr(p_C \mid g_C)}{\sum_{g_C} \Pr(g_M, g_F)\Pr(g_C \mid g_M, g_F)\Pr(p_C \mid g_C)} = \frac{\Pr(g_C, g_M, g_F, p_C)}{\Pr(g_M, g_F, p_C)} = \Pr(g_C \mid g_M, g_F, p_C).$$

Also, given the mating type, the children's genotypes are assumed to be conditionally independent. That is,

$$\Pr(\vec{g}_{\bar{C}} \mid g_M, g_F, p_M, p_F, \vec{p}_{\bar{C}}) = \prod_i \Pr(g_{C_i} \mid g_M, g_F, p_M, p_F, p_{C_i}), \text{ and}$$

$$\Pr(\vec{g}_{\bar{C}} \mid g_M, g_F, \vec{p}_{\bar{C}}) = \prod_i \Pr(g_{C_i} \mid g_M, g_F, p_{C_i}).$$

Therefore, equation (30) holds.

Equation (31) holds if one individual's phenotype is dependent solely on his/her marker genotype so that the parental phenotypes are sufficient to determine the probability of the parental genotypes. However, the assumption will be violated under the following scenarios:

The parental phenotypes are unavailable

Since this work assumes that the phenotypes are MAR, I have $\Pr(g_M, g_F \mid p_M = Miss, p_F = Miss, p_C) = \Pr(g_M, g_F \mid p_C)$ and $\Pr(g_M, g_F \mid p_M = Miss, p_F = Miss) = \Pr(g_M, g_F)$. If equation (31) is right, I have $\Pr(g_F, g_M \mid p_C) = \Pr(g_F, g_M)$ when parental phenotypes are unavailable. However, it is not necessarily true. Actually, Schaid and Sommer (1993) have derived that the conditional likelihood

$$\Pr(g_F, g_M \mid p_C = D) = \Pr(g_F, g_M)\frac{\Pr(p_C = D \mid g_F, g_M)}{\Pr(p_C = D)}$$

$$= \Pr(g_F, g_M)\frac{\sum_{i=0}^{2} \phi_i \Pr(g_C = i \mid g_F, g_M)}{\phi_0\pi_0 + \phi_1\pi_1 + \phi_2\pi_2}.$$

When $\phi_0 = \phi_1 = \phi_2$, equation (31) holds. When the marker penetrances are unequal, $\Pr(g_F, g_M \mid p_C = D)$ is not necessarily equal to $\Pr(g_F, g_M)$. The implication is that the mating type is not independent of the child's phenotype.

The parental phenotypes are available

Weeks shows that children's phenotypes can be used to infer the mating type

of their parents, even when the parental phenotypes are available (personal communication by email). Consider two trios with unaffected parents, one with an unaffected child, and the other with an affected child. Suppose the marker allele *a* is quite frequent. If there is linkage disequilibrium between the marker and a DSL so that the *db* and *+a* haplotypes are the most frequent, then the unaffected child is more likely to be *+a*/*+a*; and so that the mating type of the first trio is more likely to be $aa \times aa$ at the marker. In contrast, the affected child is more likely to be *+a/db*, and so that the mating type of the second trio is more likely to be $aa \times ab$ at the marker.

The violation of the assumption that supports equation (31) indicates that L_{YY} is not necessarily equal to L_{WT^*} especially under the alternative hypothesis. But since L_{YY} contains information on association, it can be used as an approximation of L_{WT^*}. Also, the results of null simulation and power calculation suggest that LRT based on L_{YY} is appropriate for the association test. Another advantage of L_{YY} is that the LRT based on L_{YY} can be readily extended for large pedigrees with missing data and genotyping errors with affordable computation complexity while L_{WT^*} will be computationally inefficient when missing data and genotyping errors appear in large pedigrees.

5.2 Missing parental genotype data

When the Elston-Stewart algorithm is applied to compute likelihoods for pedigrees with missing parental genotype data, it always ends up including a summation over all underlying complete phenotype and marker genotype vectors that are consistent with the observed phenotype and genotype data (personal communication with Weeks). That is, the derivation of the conditional likelihood for pedigrees with missing parental genotype data is

$$\Pr(\vec{g}_{obs} \mid \vec{p}) = \sum_{\vec{g}_{mis}} \Pr(\vec{g}_{obs}, \vec{g}_{mis} \mid \vec{p}).$$

Although it may be feasible to apply this marginal likelihood conditional on the observed phenotypes to this work, the calculation of likelihood for large pedigrees with substantial missing data will be computationally expensive (Nyholt, 2002). Another disadvantage of the marginal likelihood is that it cannot be used to infer the estimates when the genotypes are missing not at random (MNAR) (Little and Rubin, 2002). Purcell et al. (2007) proposed a test for nonrandom genotyping failure with respect to genotype. They find that if the assumption of MAR is violated, one would often expect to see an association between missingness and flanking haplotypes (Purcell et al., 2007). That is, when marker genotypes are MNAR, the likelihood built on the assumption of MAR (such as the marginal likelihood mentioned by Weeks) may result in a biased test. Although this work assumes MAR for genotypes and phenotypes for easy start, the conditional expectation of the complete-data likelihood for nuclear families with missing parental genotype data can also be used if the MAR is violated.

5.3 Genotyping errors

Optimal performance of genetic analyses relies on accurate and efficient genotypes as genotyping errors reduce power to detect and map genetic effects. Even at low error rates (< 2%), genotyping errors, of which 25% are Mendelian consistent,

can result in biased test results (Douglas et al., 2000). Even worse, the Mendelian-consistent errors are difficult to detect, Badzioch et al. (2003) found that even when the error rate is assumed to be as high as 10%, only 50% of the Mendelian-consistent genotyping errors can be found.

Another problem is the computational complexity to perform the LRT for families with potential Mendelian-inconsistent and -consistent errors. It is difficult to infer the estimate of error rate for large and complex pedigrees since the computation time is exponential in the number of individuals when considering Mendelian-consistent errors throughout the family members. I have written an R function to perform the LRT for trios with potential genotyping errors (both consistent and inconsistent), without any constraints on the number of errors per trio. The computation is quite slow, and the likelihood tends to increase when the grid-UOBYQA algorithm searches around the parameter limits (0 or 1). This results in unreasonable parameter estimates. Therefore, I only consider Mendelian inconsistencies and assume at most one inconsistency per nuclear family. Actually, most studies for genotyping errors considered only Mendelian-inconsistent errors in their models (Gordon et al., 2004; Ehm et al., 1996) or assumed that there was exactly one genotyping error per family (Douglas et al., 2002) due to the computational complexity.

In the real application of this LRT, I found that in the psoriasis data, there were a couple of families with a strong evidence of paternal inconsistency. I check the consistency of the genotypes for each family on each of the 13 SNP being tested and detect inconsistencies of the paternal genotype with the genotype of one specific child in at least 5 nuclear families. For example, the paternal genotypes of one psoriasis family were inconsistent with those of the child with ID 1 on all 13 SNPs. When removing these nuclear families with paternal inconsistency from the analysis, there is no inconsistency or very few inconsistencies in the remaining families. To make use of the information from families with paternal inconsistency, one strategy is to remove the paternal genotypes from those families for all the markers, and then use the likelihood function for families with missing parental genotypes to compute the likelihood factors contributed by other individuals. That suggests the use of equation (26), which is similar to the complete-likelihood function (see equation (1)), to compute the likelihood factor for the family with one inconsistency.

5.4 Maximization algorithms

For simple family structures such as case-parent trios with complete or missing parental genotype data, the expectation and maximization (EM) algorithm (Dempster et al., 1977) is normally used to maximize the log-likelihood under the hypotheses of association (Schaid DJ, 1996; Weinberg, 1999). However, it is not easy to extend the EM algorithm to allow for larger families due to the difficulty in deriving the expectation and maximization functions for arbitrary pedigree structures, especially when there are missing data and genotyping errors in the pedigree. Besides the difficulty in implementing the EM algorithm, this work applies grid-UOBYQA for the following reasons. First, the theoretical convergence rate for EM algorithm is linear (Dempster et al., 1977). When the associated model is complicated, it converges more slowly than the UOBYQA since the later has a faster quadratic convergence rate while being globally convergent (Powell, 2000). Second, the EM algorithm may converge to a local optimum, while the grid-UOBYQA can provide a thorough sampling of the parameter space, which makes it easy to locate the global maximum of the

log-likelihood.

The biggest disadvantage of grid-UOBYQA is that the computation complexity increases exponentially in the number of grids. This makes the computation very slow when G is large. Therefore, we set $G = 5$ to search for $K = 6$ best starting points under the two hypotheses.

The experience in performing the simulations and applying the likelihood method to the real data sets is that the maximization is reasonably fast when the data are from nuclear families with little/no missing information. The computational effort increases when considering more general pedigrees and/or pedigrees with more missing data and inconsistencies. These results are similar to those observed with the TDTae.

Table 1: $\Pr(g_C \mid g_F, g_M)$, the probability of a child's genotype conditional on the parental genotypes

Mating Type[a]	$\{g_F, g_M\}$	$\Pr(g_F, g_M)$	$g_C = 0$	$g_C = 1$	$g_C = 2$
6	{0, 0}	$\pi_0^{\,2}$	1	0	0
5	{0, 1}/{1, 0}	$2\pi_0\pi_1$	1/2	1/2	0
4	{0, 2}/{2, 0}	$2\pi_0\pi_2$	0	1	0
3	{1, 1}	π_1^2	1/4	1/2	1/4
2	{1, 2}/{2, 1}	$2\pi_1\pi_2$	0	1/2	1/2
1	{2, 2}	$\pi_2^{\,2}$	0	0	1

[a]The mating types are consistent with those in Schaid and Sommer (1993).

Table 2: $\Pr(g_C \mid g_F, g_M, p_C = A)$, the probability of a child's genotype conditional on the parental genotypes and child being affected

$\{g_F, g_M\}$	$g_C = 0$	$g_C = 1$	$g_C = 2$
{0, 0}	1	0	0
{0, 1}/{1, 0}	$\frac{\phi_0}{\phi_0 + \phi_1}$	$\frac{\phi_1}{\phi_0 + \phi_1}$	0
{0, 2}/{2, 0}	0	1	0
{1, 1}	$\frac{\phi_0}{\phi_0 + 2\phi_1 + \phi_2}$	$\frac{2\phi_1}{\phi_0 + 2\phi_1 + \phi_2}$	$\frac{\phi_2}{\phi_0 + 2\phi_1 + \phi_2}$
{1, 2}/{2, 1}	0	$\frac{\phi_1}{\phi_1 + \phi_2}$	$\frac{\phi_2}{\phi_1 + \phi_2}$
{2, 2}	0	0	1

Table 3: $\Pr(g_C \mid g_F, g_M, p_C = U)$, the probability of a child's genotype conditional on parental genotypes and child being unaffected

$\{g_F, g_M\}$	$g_C = 0$	$g_C = 1$	$g_C = 2$
{0, 0}	1	0	0
{0, 1}/{1, 0}	$\frac{1-\phi_0}{2-\phi_0-\phi_1}$	$\frac{1-\phi_1}{2-\phi_0-\phi_1}$	0
{0, 2}/{2, 0}	0	1	0
{1, 1}	$\frac{1-\phi_0}{4-\phi_0-2\phi_1-\phi_2}$	$\frac{2(1-\phi_1)}{4-\phi_0-2\phi_1-\phi_2}$	$\frac{1-\phi_2}{4-\phi_0-2\phi_1-\phi_2}$
{1, 2}/{2, 1}	0	$\frac{1-\phi_1}{2-\phi_1-\phi_2}$	$\frac{1-\phi_2}{2-\phi_1-\phi_2}$
{2, 2}	0	0	1

Table 4: $\Pr(g_F \mid g_M, g_C, p_F)$ for a trio with untyped father

$\{g_M, g_C\}$	$\Pr(g_F = 0 \mid g_M, g_C, p_F)$	$\Pr(g_F = 1 \mid g_M, g_C, p_F)$	$\Pr(g_F = 2 \mid g_M, g_C, p_F)$
{0, 0}	$\frac{2\pi_0\eta_0}{2\pi_0\eta_0 + \pi_1\eta_1}$	$\frac{\pi_1\eta_1}{2\pi_0\eta_0 + \pi_1\eta_1}$	0
{0, 1}	0	$\frac{\pi_1\eta_1}{\pi_1\eta_1 + 2\pi_2\eta_2}$	$\frac{2\pi_2\eta_2}{\pi_1\eta_1 + 2\pi_2\eta_2}$
{1, 0}	$\frac{2\pi_0\eta_0}{2\pi_0\eta_0 + \pi_1\eta_1}$	$\frac{\pi_1\eta_1}{2\pi_0\eta_0 + \pi_1\eta_1}$	0
{1, 1}	$\frac{\pi_0\eta_0}{\pi_0\eta_0 + \pi_1\eta_1 + \pi_2\eta_2}$	$\frac{\pi_1\eta_1}{\pi_0\eta_0 + \pi_1\eta_1 + \pi_2\eta_2}$	$\frac{\pi_2\eta_2}{\pi_0\eta_0 + \pi_1\eta_1 + \pi_2\eta_2}$
{1, 2}	0	$\frac{\pi_1\eta_1}{\pi_1\eta_1 + 2\pi_2\eta_2}$	$\frac{2\pi_2\eta_2}{\pi_1\eta_1 + 2\pi_2\eta_2}$
{2, 1}	$\frac{2\pi_0\eta_0}{2\pi_0\eta_0 + \pi_1\eta_1}$	$\frac{\pi_1\eta_1}{2\pi_0\eta_0 + \pi_1\eta_1}$	0
{2, 2}	0	$\frac{\pi_1\eta_1}{\pi_1\eta_1 + 2\pi_2\eta_2}$	$\frac{2\pi_2\eta_2}{\pi_1\eta_1 + 2\pi_2\eta_2}$

In this table, $\eta_i = \begin{cases} \phi_i & if \ \ p_F = A \\ 1-\phi_i & if \ \ p_F = U \\ 1 & if \ \ p_F = Miss \end{cases}$

Table 5: $\Pr(g_F, g_M \mid g_C, p_F = Miss, p_M = Miss)$ for a trio without parental data

$\{g_F, g_M\}$	$g_C = 0$	$g_C = 1$	$g_C = 2$
{0, 0}	$\frac{4\pi_0^2}{4\pi_0^2 + 4\pi_0\pi_1 + \pi_1^2}$	0	0
{0, 1}/{1, 0}	$\frac{4\pi_0\pi_1}{4\pi_0^2 + 4\pi_0\pi_1 + \pi_1^2}$	$\frac{\pi_0\pi_1}{2\pi_0\pi_1 + 4\pi_0\pi_2 + \pi_1^2 + 2\pi_1\pi_2}$	0
{0, 2}/{2, 0}	0	$\frac{2\pi_0\pi_2}{2\pi_0\pi_1 + 4\pi_0\pi_2 + \pi_1^2 + 2\pi_1\pi_2}$	0
{1, 1}	$\frac{\pi_1^2}{4\pi_0^2 + 4\pi_0\pi_1 + \pi_1^2}$	$\frac{\pi_1^2}{2\pi_0\pi_1 + 4\pi_0\pi_2 + \pi_1^2 + 2\pi_1\pi_2}$	$\frac{\pi_1^2}{\pi_1^2 + 4\pi_1\pi_2 + 4\pi_2^2}$
{1, 2}/ {2, 1}	0	$\frac{\pi_1\pi_2}{2\pi_0\pi_1 + 4\pi_0\pi_2 + \pi_1^2 + 2\pi_1\pi_2}$	$\frac{4\pi_1\pi_2}{\pi_1^2 + 4\pi_1\pi_2 + 4\pi_2^2}$
{2, 2}	0	0	$\frac{4\pi_2^2}{\pi_1^2 + 4\pi_1\pi_2 + 4\pi_2^2}$

Table 6: $\Pr(g_F, g_M \mid g_C, p_F, p_M)$ for a trio with two untyped parents

$\{g_F, g_M\}$	$g_C = 0$	$g_C = 1$	$g_C = 2$
{0, 0}	$\frac{2\pi_0^2\Theta(0,p_F)\Theta(0,p_M)}{D_0}$	0	0
{0, 1}	$\frac{\pi_0\pi_1\Theta(0,p_F)\Theta(1,p_M)}{D_0}$	$\frac{\pi_0\pi_1\Theta(0,p_F)\Theta(1,p_M)}{D_1}$	0
{0, 2}	0	$\frac{2\pi_0\pi_2\Theta(0,p_F)\Theta(2,p_M)}{D_1}$	0
{1, 0}	$\frac{\pi_1\pi_0\Theta(1,p_F)\Theta(0,p_M)}{D_0}$	$\frac{\pi_1\pi_0\Theta(1,p_F)\Theta(0,p_M)}{D_1}$	0
{1, 1}	$\frac{2^{-1}\pi_1^2\Theta(1,p_F)\Theta(1,p_M)}{D_0}$	$\frac{\pi_1^2\Theta(1,p_F)\Theta(1,p_M)}{D_1}$	$\frac{2^{-1}\pi_1^2\Theta(1,p_F)\Theta(1,p_M)}{D_2}$
{1, 2}	0	$\frac{\pi_1\pi_2\Theta(1,p_F)\Theta(2,p_M)}{D_1}$	$\frac{\pi_1\pi_2\Theta(1,p_F)\Theta(2,p_M)}{D_2}$
{2, 0}	0	$\frac{2\pi_2\pi_0\Theta(2,p_F)\Theta(0,p_M)}{D_1}$	0
{2, 1}	0	$\frac{\pi_2\pi_1\Theta(2,p_F)\Theta(1,p_M)}{D_1}$	$\frac{\pi_2\pi_1\Theta(2,p_F)\Theta(1,p_M)}{D_2}$
{2, 2}	0	0	$\frac{2\pi_2^2\Theta(2,p_F)\Theta(2,p_M)}{D_2}$

In this table, $\Theta(i,p) = \begin{cases} \phi_i & if\ p = A \\ 1-\phi_i & if\ p = U \\ 1 & if\ p = Miss \end{cases}, i = 0, 1, 2.$

In the event that $p_F = p_M = Miss$, $\Theta(i, p_F) = \Theta(i, p_M) = 1$ and Table 6 is reduced to Table 5.

Table 7: $\Pr(g_F \mid g_M, \vec{g}_{\bar{C}}, p_F)$ for a nuclear family with untyped father

$\{g_M, \{\vec{g}_{\bar{C}}\}\}$	$\Pr(g_F=0 \mid g_M, \vec{g}_{\bar{C}}, p_F)$	$\Pr(g_F=1 \mid g_M, \vec{g}_{\bar{C}}, p_F)$	$\Pr(g_F=2 \mid g_M, \vec{g}_{\bar{C}}, p_F)$
{0, {0}}	$\frac{2^{n_{0.C}}\pi_0\eta_0}{2^{n_{0.C}}\pi_0\eta_0+\pi_1\eta_1}$	$\frac{\pi_1\eta_1}{2^{n_{0.C}}\pi_0\eta_0+\pi_1\eta_1}$	0
{0, {1}}	0	$\frac{\pi_1\eta_1}{\pi_1\eta_1+2^{n_{1.C}}\pi_2\eta_2}$	$\frac{2^{n_{1.C}}\pi_2\eta_2}{\pi_1\eta_1+2^{n_{1.C}}\pi_2\eta_2}$
{0, {0,1}}	0	1	0
{1, {0}}	$\frac{2^{n_{0.C}}\pi_0\eta_0}{2^{n_{0.C}}\pi_0\eta_0+\pi_1\eta_1}$	$\frac{\pi_1\eta_1}{2^{n_{0.C}}\pi_0\eta_0+\pi_1\eta_1}$	0
{1, {1}}	$\frac{\pi_0\eta_0}{\pi_0\eta_0+\pi_1\eta_1+\pi_2\eta_2}$	$\frac{\pi_1\eta_1}{\pi_0\eta_0+\pi_1\eta_1+\pi_2\eta_2}$	$\frac{\pi_2\eta_2}{\pi_0\eta_0+\pi_1\eta_1+\pi_2\eta_2}$
{1, {2}}	0	$\frac{\pi_1\eta_1}{\pi_1\eta_1+2^{n_{2.C}}\pi_2\eta_2}$	$\frac{2^{n_{2.C}}\pi_2\eta_2}{\pi_1\eta_1+2^{n_{2.C}}\pi_2\eta_2}$
{1, {0,1}}	$\frac{2^{n_{0.C}}\pi_0\eta_0}{2^{n_{0.C}}\pi_0\eta_0+\pi_1\eta_1}$	$\frac{\pi_1\eta_1}{2^{n_{0.C}}\pi_0\eta_0+\pi_1\eta_1}$	0
{1, {0,2}}	0	1	0
{1, {1,2}}	0	$\frac{\pi_1\eta_1}{\pi_1\eta_1+2^{n_{2.C}}\pi_2\eta_2}$	$\frac{2^{n_{2.C}}\pi_2\eta_2}{\pi_1\eta_1+2^{n_{2.C}}\pi_2\eta_2}$
{1, {0,1,2}}	0	1	0
{2, {1}}	$\frac{2^{n_{1.C}}\pi_0\eta_0}{2^{n_{1.C}}\pi_0\eta_0+\pi_1\eta_1}$	$\frac{\pi_1\eta_1}{2^{n_{1.C}}\pi_0\eta_0+\pi_1\eta_1}$	0
{2, {2}}	0	$\frac{\pi_1\eta_1}{\pi_1\eta_1+2^{n_{2.C}}\pi_2\eta_2}$	$\frac{2^{n_{2.C}}\pi_2\eta_2}{\pi_1\eta_1+2^{n_{2.C}}\pi_2\eta_2}$
{2, {1,2}}	0	1	0

In this table, $\eta_i = \begin{cases} \phi_i & if\ p_F = A \\ 1-\phi_i & if\ p_F = U \\ 1 & if\ p_F = Miss \end{cases}$. The first column labeled $\{g_M, \{\vec{g}_{\bar{C}}\}\}$ lists the maternal genotypes and the set of children's genotypes.

Table 8: $\Pr(g_F, g_M \mid \vec{g}_{\bar{C}}, p_F, p_M)$ for a nuclear family with untyped parents

$\{g_F, g_M\}$	$\{\vec{g}_{\bar{C}}\} = \{0\}$	$\{\vec{g}_{\bar{C}}\} = \{1\}$	$\{\vec{g}_{\bar{C}}\} = \{2\}$
{0, 0}	$\frac{2^{n_{0.C}} \pi_0^2 \Theta(0, p_F)\Theta(0, p_M)}{D_0}$	0	0
{0, 1}	$\frac{\pi_0 \pi_1 \Theta(0, p_F)\Theta(1, p_M)}{D_0}$	$\frac{\pi_0 \pi_1 \Theta(0, p_F)\Theta(1, p_M)}{D_1}$	0
{0, 2}	0	$\frac{2^{n_{1.C}} \pi_0 \pi_2 \Theta(0, p_F)\Theta(2, p_M)}{D_1}$	0
{1, 0}	$\frac{\pi_1 \pi_0 \Theta(1, p_F)\Theta(0, p_M)}{D_0}$	$\frac{\pi_1 \pi_0 \Theta(1, p_F)\Theta(0, p_M)}{D_1}$	0
{1, 1}	$\frac{2^{-n_{0.C}} \pi_1^2 \Theta(1, p_F)\Theta(1, p_M)}{D_0}$	$\frac{\pi_1^2 \Theta(1, p_F)\Theta(1, p_M)}{D_1}$	$\frac{2^{-n_{2.C}} \pi_1^2 \Theta(1, p_F)\Theta(1, p_M)}{D_2}$
{1, 2}	0	$\frac{\pi_1 \pi_2 \Theta(1, p_F)\Theta(2, p_M)}{D_1}$	$\frac{\pi_1 \pi_2 \Theta(1, p_F)\Theta(2, p_M)}{D_2}$
{2, 0}	0	$\frac{2^{n_{1.C}} \pi_2 \pi_0 \Theta(2, p_F)\Theta(0, p_M)}{D_1}$	0
{2, 1}	0	$\frac{\pi_2 \pi_1 \Theta(2, p_F)\Theta(1, p_M)}{D_1}$	$\frac{\pi_2 \pi_1 \Theta(2, p_F)\Theta(1, p_M)}{D_2}$
{2, 2}	0	0	$\frac{2^{n_{2.C}} \pi_2^2 \Theta(2, p_F)\Theta(2, p_M)}{D_2}$
$\{g_F, g_M\}$	$\{\vec{g}_{\bar{C}}\} = \{0,1\}$	$\{\vec{g}_{\bar{C}}\} = \{1,2\}$	$\{\vec{g}_{\bar{C}}\} = \{0,2\} / \{0,1,2\}$
{0, 0}	0	0	0
{0, 1}	$\frac{2^{n_{0.C}} \pi_0 \Theta(0, p_F)\Theta(1, p_M)}{D_{01}}$	0	0
{0, 2}	0	0	0
{1, 0}	$\frac{2^{n_{0.C}} \pi_0 \Theta(1, p_F)\Theta(0, p_M)}{D_{01}}$	0	0
{1, 1}	$\frac{\pi_1 \Theta(1, p_F)\Theta(1, p_M)}{D_{01}}$	$\frac{\pi_1 \Theta(1, p_F)\Theta(1, p_M)}{D_{12}}$	1
{1, 2}	0	$\frac{2^{n_{2.C}} \pi_2 \Theta(1, p_F)\Theta(2, p_M)}{D_{12}}$	0
{2, 0}	0	0	0
{2, 1}	0	$\frac{2^{n_{2.C}} \pi_2 \Theta(2, p_F)\Theta(1, p_M)}{D_{12}}$	0
{2, 2}	0	0	0

Legend of Table 8

In this table, $\Theta(i,p)=\begin{cases}\phi_i & if\ \ p=A\\ 1-\phi_i & if\ \ p=U\\ 1 & if\ \ p=Miss\end{cases}, i=0,1,2.$

$\{\vec{g}_{\vec{C}}\}=\{0\}$ indicates that all the children in the family are genotyped 0. Similarly, $\{\vec{g}_{\vec{C}}\}=\{1\}$ and $\{\vec{g}_{\vec{C}}\}=\{2\}$ indicate that all the children are genotyped 1 and 2. Also,

$$D_0=2^{n_{0.C}}\pi_0^2\Theta(0,p_F)\Theta(0,p_M)+\pi_0\pi_1\Theta(0,p_F)\Theta(1,p_M)+\pi_1\pi_0\Theta(1,p_F)\Theta(0,p_M)$$
$$+2^{-n_{0.C}}\pi_1^2\Theta(1,p_F)\Theta(1,p_M);$$

$$D_1=\pi_0\pi_1\Theta(0,p_F)\Theta(1,p_M)+2^{n_{1.C}}\pi_0\pi_2\Theta(0,p_F)\Theta(2,p_M)+\pi_1\pi_0\Theta(1,p_F)\Theta(0,p_M)$$
$$+\pi_1^2\Theta(1,p_F)\Theta(1,p_M)+\pi_1\pi_2\Theta(1,p_F)\Theta(2,p_M)$$
$$+2^{n_{1.C}}\pi_2\pi_0\Theta(2,p_F)\Theta(0,p_M)+\pi_2\pi_1\Theta(2,p_F)\Theta(1,p_M);$$

$$D_2=2^{-n_{2.C}}\pi_1^2\Theta(1,p_F)\Theta(1,p_M)+\pi_1\pi_2\Theta(1,p_F)\Theta(2,p_M)+\pi_2\pi_1\Theta(2,p_F)\Theta(1,p_M)$$
$$+2^{n_{2.C}}\pi_2^2\Theta(2,p_F)\Theta(2,p_M);$$

$$D_{01}=2^{n_{0.C}}\pi_0[\Theta(0,p_F)\Theta(1,p_M)+\Theta(1,p_F)\Theta(0,p_M)]+\pi_1\Theta(1,p_F)\Theta(1,p_M);$$

$$D_{12}=\pi_1\Theta(1,p_F)\Theta(1,p_M)+2^{n_{2.C}}\pi_0[\Theta(1,p_F)\Theta(2,p_M)+\Theta(2,p_F)\Theta(1,p_M)].$$

Table 9: $\Pr(g_F \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F)$ for a pedigree with untyped father and genotyped paternal grandparents: Part A

$\{g_{FF}, g_{MF}\}$	$g_F = 0$	$g_F = 1$	$g_F = 2$
	$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{0, \{0\}\}$		
{0, 0}	1	0	0
{0, 1}/{1, 0}	$\frac{2^{n_{0.C}}\Theta(0,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+\Theta(1,p_F)}$	$\frac{\Theta(1,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+\Theta(1,p_F)}$	0
{1, 1}	$\frac{2^{n_{0.C}}\Theta(0,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+2\Theta(1,p_F)}$	$\frac{2\Theta(1,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+2\Theta(1,p_F)}$	0
{1, 2}/{2, 1}	0	1	0
	$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{0, \{1\}\}$		
{0, 1}/{1, 0}	0	1	0
{1, 1}	0	$\frac{2\Theta(1,p_F)}{2\Theta(1,p_F)+2^{n_{1.C}}\Theta(2,p_F)}$	$\frac{2^{n_{1.C}}\Theta(2,p_F)}{2\Theta(1,p_F)+2^{n_{1.C}}\Theta(2,p_F)}$
{1, 2}/{2, 1}	0	$\frac{\Theta(1,p_F)}{\Theta(1,p_F)+2^{n_{1.C}}\Theta(2,p_F)}$	$\frac{2^{n_{1.C}}\Theta(2,p_F)}{\Theta(1,p_F)+2^{n_{1.C}}\Theta(2,p_F)}$
{2, 2}	0	0	1
	$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{0, \{0,1\}\}$		
{*, *}	0	1	0
	$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{0\}\}$		
{0, 1}/{1, 0}	$\frac{2^{n_{0.C}}\Theta(0,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+\Theta(1,p_F)}$	$\frac{\Theta(1,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+\Theta(1,p_F)}$	0
{1, 1}	$\frac{2^{n_{0.C}}\Theta(0,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+2\Theta(1,p_F)}$	$\frac{2\Theta(1,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+2\Theta(1,p_F)}$	0
{1, 2}/{2, 1}	0	1	0
	$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{1\}\}$		
{0, 1}/{1, 0}	$\frac{\Theta(0,p_F)}{\Theta(0,p_F)+\Theta(1,p_F)}$	$\frac{\Theta(1,p_F)}{\Theta(0,p_F)+\Theta(1,p_F)}$	0
{1, 1}	$\frac{\Theta(0,p_F)}{\Theta(0,p_F)+2\Theta(1,p_F)+\Theta(2,p_F)}$	$\frac{2\Theta(1,p_F)}{\Theta(0,p_F)+2\Theta(1,p_F)+\Theta(2,p_F)}$	$\frac{\Theta(2,p_F)}{\Theta(0,p_F)+2\Theta(1,p_F)+\Theta(2,p_F)}$
{1, 2}/{2, 1}	0	$\frac{\Theta(1,p_F)}{\Theta(1,p_F)+\Theta(2,p_F)}$	$\frac{\Theta(2,p_F)}{\Theta(1,p_F)+\Theta(2,p_F)}$
	$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{2\}\}$		
{0, 1}/{1, 0}	0	1	0
{1, 1}	0	$\frac{2\Theta(1,p_F)}{2\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$	$\frac{2^{n_{2.C}}\Theta(2,p_F)}{2\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$
{1, 2}/{2, 1}	0	$\frac{\Theta(1,p_F)}{\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$	$\frac{2^{n_{2.C}}\Theta(2,p_F)}{\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$

In this table, $\Theta(i,p) = \begin{cases} \phi_i & if\ p = A \\ 1-\phi_i & if\ p = U \\ 1 & if\ p = Miss \end{cases}, i = 0, 1, 2.$

{*, *} in the first column labeled $\{g_{FF}, g_{MF}\}$ denotes all possible mating types of the paternal grandparents. $\{g_M, \{\vec{g}_{\bar{C}}\}\}$ indicates the maternal genotypes and the set of children's genotypes.

Table 10: $\Pr(g_F \mid g_{FF}, g_{MF}, g_M, \vec{g}_{\bar{C}}, p_F)$ for a pedigree with untyped father and genotyped paternal grandparents: Part B

$\{g_{FF}, g_{MF}\}$	$g_F = 0$	$g_F = 1$	$g_F = 2$
$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{0,1\}\}$			
{0, 1}/{1, 0}	$\frac{2^{n_{0.C}}\Theta(0,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+\Theta(1,p_F)}$	$\frac{\Theta(1,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+\Theta(1,p_F)}$	0
{1, 1}	$\frac{2^{n_{0.C}}\Theta(0,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+2\Theta(1,p_F)}$	$\frac{2\Theta(1,p_F)}{2^{n_{0.C}}\Theta(0,p_F)+2\Theta(1,p_F)}$	0
{1, 2}/{2, 1}	0	1	0
$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{1,2\}\}$			
{0, 1}/{1, 0}	0	1	0
{1, 1}	0	$\frac{2\Theta(1,p_F)}{2\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$	$\frac{2^{n_{2.C}}\Theta(2,p_F)}{2\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$
{1, 2}/{2, 1}	0	$\frac{\Theta(1,p_F)}{\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$	$\frac{2^{n_{2.C}}\Theta(2,p_F)}{\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$
$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{0,2\}/\{1, \{0,1,2\}\}$			
{*, *}	0	1	0
$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{2, \{1\}\}$			
{0, 1}/{1, 0}	$\frac{2^{n_{1.C}}\Theta(0,p_F)}{2^{n_{1.C}}\Theta(0,p_F)+\Theta(1,p_F)}$	$\frac{\Theta(1,p_F)}{2^{n_{1.C}}\Theta(0,p_F)+\Theta(1,p_F)}$	0
{1, 1}	$\frac{2^{n_{1.C}}\Theta(0,p_F)}{2^{n_{1.C}}\Theta(0,p_F)+2\Theta(1,p_F)}$	$\frac{2\Theta(1,p_F)}{2^{n_{1.C}}\Theta(0,p_F)+2\Theta(1,p_F)}$	0
{1, 2}/{2, 1}	0	1	0
$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{2, \{2\}\}$			
{0, 1}/{1, 0}	0	1	0
{1, 1}	0	$\frac{2\Theta(1,p_F)}{2\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$	$\frac{2^{n_{2.C}}\Theta(2,p_F)}{2\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$
{1, 2}/{2, 1}	0	$\frac{\Theta(1,p_F)}{\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$	$\frac{2^{n_{2.C}}\Theta(2,p_F)}{\Theta(1,p_F)+2^{n_{2.C}}\Theta(2,p_F)}$
$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{2, \{1,2\}\}$			
{*, *}	0	1	0

In this table, $\Theta(i,p) = \begin{cases} \phi_i & if\ p = A \\ 1-\phi_i & if\ p = U \\ 1 & if\ p = Miss \end{cases}, i = 0, 1, 2.$

{*, *} in the first column labeled $\{g_{FF}, g_{MF}\}$ denotes all possible mating types of the paternal grandparents. $\{g_M, \{\vec{g}_{\bar{C}}\}\}$ indicates the maternal genotypes and the set of children's genotypes.

Table 11: $\Pr(g_F, g_M \mid g_{FF}, g_{MF}, \vec{g}_{\tilde{C}}, p_F, p_M)$ for a pedigree with untyped parents and genotyped paternal grandparents

$\{g_F, g_M\}$	$\{\vec{g}_{\tilde{C}}\} = \{0\}$	$\{\vec{g}_{\tilde{C}}\} = \{1\}$	$\{\vec{g}_{\tilde{C}}\} = \{2\}$
{0, 0}	$\frac{2^{n_{0.C}}\omega_0\pi_0\Theta(0,p_F)\Theta(0,p_M)}{D_0}$	0	0
{0, 1}	$\frac{\omega_0\pi_1\Theta(0,p_F)\Theta(1,p_M)}{D_0}$	$\frac{\omega_0\pi_1\Theta(0,p_F)\Theta(1,p_M)}{D_1}$	0
{0, 2}	0	$\frac{2^{n_{1.C}}\omega_0\pi_2\Theta(0,p_F)\Theta(2,p_M)}{D_1}$	0
{1, 0}	$\frac{\omega_1\pi_0\Theta(1,p_F)\Theta(0,p_M)}{D_0}$	$\frac{\omega_1\pi_0\Theta(1,p_F)\Theta(0,p_M)}{D_1}$	0
{1, 1}	$\frac{2^{-n_{0.C}}\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)}{D_0}$	$\frac{\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)}{D_1}$	$\frac{2^{-n_{2.C}}\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)}{D_2}$
{1, 2}	0	$\frac{\omega_1\pi_2\Theta(1,p_F)\Theta(2,p_M)}{D_1}$	$\frac{\omega_1\pi_2\Theta(1,p_F)\Theta(2,p_M)}{D_2}$
{2, 0}	0	$\frac{2^{n_{1.C}}\omega_2\pi_0\Theta(2,p_F)\Theta(0,p_M)}{D_1}$	0
{2, 1}	0	$\frac{\omega_2\pi_1\Theta(2,p_F)\Theta(1,p_M)}{D_1}$	$\frac{\omega_2\pi_1\Theta(2,p_F)\Theta(1,p_M)}{D_2}$
{2, 2}	0	0	$\frac{2^{n_{2.C}}\omega_1\pi_2\Theta(2,p_F)\Theta(2,p_M)}{D_2}$
$\{g_F, g_M\}$	$\{\vec{g}_{\tilde{C}}\} = \{0,1\}$	$\{\vec{g}_{\tilde{C}}\} = \{1,2\}$	$\{\vec{g}_{\tilde{C}}\} = \{0,2\} / \{0,1,2\}$
{0, 0}	0	0	0
{0, 1}	$\frac{2^{n_{0.C}}\omega_0\pi_1\Theta(0,p_F)\Theta(1,p_M)}{D_{01}}$	0	0
{0, 2}	0	0	0
{1, 0}	$\frac{2^{n_{0.C}}\omega_1\pi_0\Theta(1,p_F)\Theta(0,p_M)}{D_{01}}$	0	0
{1, 1}	$\frac{\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)}{D_{01}}$	$\frac{\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)}{D_{12}}$	1
{1, 2}	0	$\frac{2^{n_{2.C}}\omega_1\pi_2\Theta(1,p_F)\Theta(2,p_M)}{D_{12}}$	0
{2, 0}	0	0	0
{2, 1}	0	$\frac{2^{n_{2.C}}\omega_2\pi_1\Theta(2,p_F)\Theta(1,p_M)}{D_{12}}$	0
{2, 2}	0	0	0

Legend of Table 11

In this table, $\Theta(i,p)=\begin{cases}\phi_i & if\ p=A\\ 1-\phi_i & if\ p=U\\ 1 & if\ p=Miss\end{cases}, i=0,1,2.$

$\{\vec{g}_{\vec{C}}\}=\{0\}$ indicates that all the children in the family are genotyped 0. Similarly, $\{\vec{g}_{\vec{C}}\}=\{1\}$ and $\{\vec{g}_{\vec{C}}\}=\{2\}$ indicate that all the children are genotyped 1 and 2, respectively. Also,

$$D_0 = 2^{n_{0.C}}\omega_0\pi_0\Theta(0,p_F)\Theta(0,p_M)+\omega_0\pi_1\Theta(0,p_F)\Theta(1,p_M)+\omega_1\pi_0\Theta(1,p_F)\Theta(0,p_M)$$
$$+2^{-n_{0.C}}\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)$$

$$D_1 = \omega_0\pi_1\Theta(0,p_F)\Theta(1,p_M)+2^{n_{1.C}}\omega_0\pi_2\Theta(0,p_F)\Theta(2,p_M)+\omega_1\pi_0\Theta(1,p_F)\Theta(0,p_M)$$
$$+\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)+\omega_1\pi_2\Theta(1,p_F)\Theta(2,p_M)$$
$$+2^{n_{1.C}}\omega_2\pi_0\Theta(2,p_F)\Theta(0,p_M)+\omega_2\pi_1\Theta(2,p_F)\Theta(1,p_M)$$

$$D_2 = 2^{-n_{2.C}}\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)+\omega_1\pi_2\Theta(1,p_F)\Theta(2,p_M)+\omega_2\pi_1\Theta(2,p_F)\Theta(1,p_M)$$
$$+2^{n_{2.C}}\omega_2\pi_2\Theta(2,p_F)\Theta(2,p_M)$$

$$D_{01} = 2^{n_{0.C}}[\omega_0\pi_1\Theta(0,p_F)\Theta(1,p_M)+\omega_1\pi_0\Theta(1,p_F)\Theta(0,p_M)]+\omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)$$

$$D_{12} = \omega_1\pi_1\Theta(1,p_F)\Theta(1,p_M)+2^{n_{2.C}}[\omega_1\pi_2\Theta(1,p_F)\Theta(2,p_M)+\omega_2\pi_1\Theta(2,p_F)\Theta(1,p_M)]$$

The coefficient $\omega_i, i\in\{0,1,2\}$ varies with genotypes of paternal grandparents as follows:

$\{g_{FF},g_{MF}\}$	ω_0	ω_1	ω_2
{0, 0}	1	0	0
{0, 1}/{1, 0}	0.5	0.5	0
{0, 2}/{2, 0}	0	1	0
{1, 2}/{2, 1}	0	0.5	0.5
{2, 2}	0	0	1
{0, ?}/{?, 0}	$\pi_0+0.5\pi_1$	$0.5\pi_1+\pi_2$	0
{1, ?}/{?, 1}	$0.5\pi_0+0.25\pi_1$	0.5	$0.25\pi_1+0.5\pi_2$
{2, ?}/{?, 2}	0	$\pi_0+0.5\pi_1$	$0.5\pi_1+\pi_2$

"?" in the first column labeled $\{g_{FF},g_{MF}\}$ indicates that the parental genotype is missing.

Table 12: $\Pr(g_F \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF} = Miss, p_F)$ for a pedigree with untyped father and paternal grandmother: Part A

g_{FF}	$g_F = 0$	$g_F = 1$	$g_F = 2$
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{0, \{0\}\}$	
0	$\frac{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	$\frac{(\frac{\pi_1}{2} + \pi_2)\eta_1}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	0
1	$\frac{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1}$	$\frac{\eta_1}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1}$	0
2	0	1	0
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{0, \{1\}\}$	
0	0	1	0
1	0	$\frac{\eta_1}{\eta_1 + 2^{n_{1.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{2^{n_{1.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}{\eta_1 + 2^{n_{1.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$
2	0	$\frac{(\pi_0 + \frac{\pi_1}{2})\eta_1}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + 2^{n_{1.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{2^{n_{1.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}{\eta_1 + 2^{n_{1.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{0, \{0,1\}\}$	
*	0	1	0
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{0\}\}$	
0	$\frac{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	$\frac{(\frac{\pi_1}{2} + \pi_2)\eta_1}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	0
1	$\frac{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1}$	$\frac{\eta_1}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1}$	
2	0	1	0
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{1\}\}$	
0	$\frac{(\pi_0 + \frac{\pi_1}{2})\eta_0}{(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	$\frac{(\frac{\pi_1}{2} + \pi_2)\eta_1}{(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	0
1	$\frac{(\pi_0 + \frac{\pi_1}{2})\eta_0}{(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1 + (\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{\eta_1}{(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1 + (\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{(\frac{\pi_1}{2} + \pi_2)\eta_2}{(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1 + (\frac{\pi_1}{2} + \pi_2)\eta_2}$
2	0	$\frac{(\pi_0 + \frac{\pi_1}{2})\eta_1}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + (\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{(\frac{\pi_1}{2} + \pi_2)\eta_2}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + (\frac{\pi_1}{2} + \pi_2)\eta_2}$
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{2\}\}$	
0	0	1	0
1	0	$\frac{\eta_1}{\eta_1 + 2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}{\eta_1 + 2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$
2	0	$\frac{(\pi_0 + \frac{\pi_1}{2})\eta_1}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + 2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + 2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$

Legend of Table 12

In this table, $\eta_i = \begin{cases} \phi_i & if \ \ p_F = A \\ 1-\phi_i & if \ \ p_F = U \\ 1 & if \ \ p_F = Miss \end{cases}$. '*' in the first column labeled g_{FF} denotes all possible mating types of the paternal grandparents. $\{g_M, \{\vec{g}_{\vec{C}}\}\}$ indicates the maternal genotypes and the set of children's genotypes.

Table 13: $\Pr(g_F \mid g_{FF}, g_M, \vec{g}_{\bar{C}}, p_{MF} = Miss, p_F)$ for a pedigree with untyped father and paternal grandmother: Part B

g_{FF}	$g_F = 0$	$g_F = 1$	$g_F = 2$
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{0,1\}\}$	
0	$\frac{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	$\frac{(\frac{\pi_1}{2} + \pi_2)\eta_1}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	0
1	$\frac{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1}$	$\frac{\eta_1}{2^{n_{0.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1}$	
2	0	1	0
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{1,2\}\}$	
0	0	1	0
1	0	$\frac{\eta_1}{\eta_1 + 2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}{\eta_1 + 2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$
2	0	$\frac{(\pi_0 + \frac{\pi_1}{2})\eta_1}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + 2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + 2^{n_{2.C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{1, \{0,2\}\} / \{1, \{0,1,2\}\}$	
*	0	1	0
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{2, \{1\}\}$	
0	$\frac{2^{n_{1.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0}{2^{n_{1.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	$\frac{(\frac{\pi_1}{2} + \pi_2)\eta_1}{2^{n_{1.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + (\frac{\pi_1}{2} + \pi_2)\eta_1}$	0
1	$\frac{2^{n_{1.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0}{2^{n_{1.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1}$	$\frac{\eta_1}{2^{n_{1.C}}(\pi_0 + \frac{\pi_1}{2})\eta_0 + \eta_1}$	0
2	0	1	0
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{2, \{2\}\}$	
0	0	1	0
1	0	$\frac{\eta_1}{\eta_1 + 2^{n_{2C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{2^{n_{2C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}{\eta_1 + 2^{n_{2C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$
2	0	$\frac{(\pi_0 + \frac{\pi_1}{2})\eta_1}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + 2^{n_{2C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$	$\frac{2^{n_{2C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}{(\pi_0 + \frac{\pi_1}{2})\eta_1 + 2^{n_{2C}}(\frac{\pi_1}{2} + \pi_2)\eta_2}$
		$\{g_M, \{\vec{g}_{\bar{C}}\}\} = \{2, \{1,2\}\}$	
*	0	1	0

In this table, $\eta_i = \begin{cases} \phi_i & if\ \ p_F = A \\ 1 - \phi_i & if\ \ p_F = U \\ 1 & if\ \ p_F = Miss \end{cases}$. '*' in the first column labeled g_{FF} denotes all possible mating types of the paternal grandparents. $\{g_M, \{\vec{g}_{\bar{C}}\}\}$ indicates the maternal genotypes and the set of children's genotypes.

Table 14: Error model $\Pr(g_{Obs} \mid g_{True})$

	$g_{True}=0$	$g_{True}=1$	$g_{True}=2$
$g_{Obs}=0$	$1-\eta$	0.5γ	0
$g_{Obs}=1$	η	$1-\gamma$	η
$g_{Obs}=2$	0	0.5γ	$1-\eta$

Table 15: $\Pr(g_{True} \mid g_{Obs})$

	$g_{Obs}=0$	$g_{Obs}=1$	$g_{Obs}=2$
$g_{True}=0$	$\frac{(1-\eta)\pi_0}{(1-\eta)\pi_0+0.5\gamma\pi_1}$	$\frac{\eta\pi_0}{\eta\pi_0+(1-\gamma)\pi_1+\eta\pi_2}$	0
$g_{True}=1$	$\frac{0.5\gamma\pi_1}{(1-\eta)\pi_0+0.5\gamma\pi_1}$	$\frac{(1-\gamma)\pi_1}{\eta\pi_0+(1-\gamma)\pi_1+\eta\pi_2}$	$\frac{0.5\gamma\pi_1}{0.5\gamma\pi_1+(1-\eta)\pi_2}$
$g_{True}=2$	0	$\frac{\eta\pi_2}{\eta\pi_0+(1-\gamma)\pi_1+\eta\pi_2}$	$\frac{(1-\eta)\pi_2}{0.5\gamma\pi_1+(1-\eta)\pi_2}$

Table 16: $\Pr_{error}(\vec{G}_{True.i} \mid \vec{G}_{Obs.i}, M = 0)$ for inconsistent nuclear families

$\{g_{Obs.F}, g_{Obs.M}\} = \{0, 0\}$			
$g_{Obs.\vec{C}}$	$g_{True.F} = 1, g_{Obs.F} = 0$	$g_{True.F} = 1, g_{Obs.F} = 0$	$g_{True.C} = 0, g_{Obs.C} = 1$
$n_{1.C} \geq 2$	1/2	½	---
$n_{1.C} = 1$	$\frac{0.5\gamma(1-\gamma)\pi_1^2}{\gamma(1-\gamma)\pi_1^2 + \eta(1-\eta)\pi_0^2}$	$\frac{0.5\gamma(1-\gamma)\pi_1^2}{\gamma(1-\gamma)\pi_1^2 + \eta(1-\eta)\pi_0^2}$	$\frac{\eta(1-\eta)\pi_0^2}{\gamma(1-\gamma)\pi_1^2 + \eta(1-\eta)\pi_0^2}$
$\{g_{Obs.F}, g_{Obs.M}\} = \{0, 1\} / \{1, 0\} / \{0, ?\} / \{?, 0\}$			
$g_{Obs.\vec{C}}$	$g_{True.F} = 1, g_{Obs.F} = 0$ / $g_{True.M} = 1, g_{Obs.M} = 0$	$g_{True.C} = 1, g_{Obs.C} = 2$	---
$n_{2.C} \geq 2$	1	---	---
$n_{2.C} = 1$	$\frac{\pi_2}{\pi_2 + \pi_0}$	$\frac{\pi_0}{\pi_0 + \pi_2}$	---
$\{g_{Obs.F}, g_{Obs.M}\} = \{0, 2\} / \{2, 0\}$			
$g_{Obs.\vec{C}}$	$g_{True.F} = 1, g_{Obs.F} = 2$ / $g_{True.M} = 1, g_{Obs.M} = 2$	$g_{True.C} = 1, g_{Obs.C} = 0$ / $g_{True.C} = 1, g_{Obs.C} = 2$	$g_{True.F} = 1, g_{Obs.F} = 0$ / $g_{True.M} = 1, g_{Obs.M} = 0$
$n_{0.C} \geq 1,$ $n_{1.C} \geq 1$	1	---	---
$n_{0.C} \geq 2,$ $n_{1.C} = 0$	1	---	---
$n_{1.C} \geq 1,$ $n_{2.C} \geq 1$	0	---	1
$n_{1.C} = 0,$ $n_{2.C} \geq 2$	0	---	1
$n_{0.C} = 1,$ $n_{1.C} = 0$	$\frac{\pi_0}{\pi_0 + \pi_2}$	$\frac{\pi_2}{\pi_0 + \pi_2}$	---
$n_{1.C} = 0,$ $n_{2.C} = 1$	---	$\frac{\pi_0}{\pi_0 + \pi_2}$	$\frac{\pi_2}{\pi_0 + \pi_2}$
$\{g_{Obs.F}, g_{Obs.M}\} = \{1, 2\} / \{2, 1\} / \{?, 2\} / \{2, ?\}$			
$g_{Obs.\vec{C}}$	$g_{True.F} = 1, g_{Obs.F} = 2$ / $g_{True.M} = 1, g_{Obs.M} = 2$	$g_{True.C} = 1, g_{Obs.C} = 0$	---
$n_{0.C} \geq 2$	1	---	---
$n_{0.C} = 1$	$\frac{\pi_0}{\pi_0 + \pi_2}$	$\frac{\pi_2}{\pi_0 + \pi_2}$	---
$\{g_{Obs.F}, g_{Obs.M}\} = \{2, 2\}$			
$g_{Obs.\vec{C}}$	$g_{True.F} = 1, g_{Obs.F} = 2$	$g_{True.M} = 1, g_{Obs.M} = 2$	$g_{True.C} = 2, g_{Obs.C} = 1$
$n_{1.C} \geq 2$	1/2	½	---
$n_{1.C} = 1$	$\frac{0.5\gamma(1-\gamma)\pi_1^2}{\gamma(1-\gamma)\pi_1^2 + \eta(1-\eta)\pi_2^2}$	$\frac{0.5\gamma(1-\gamma)\pi_1^2}{\gamma(1-\gamma)\pi_1^2 + \eta(1-\eta)\pi_2^2}$	$\frac{\eta(1-\eta)\pi_2^2}{\gamma(1-\gamma)\pi_1^2 + \eta(1-\eta)\pi_2^2}$

Legend of Table 16

In this table, '?' indicates that the paternal genotype or the maternal genotype is missing in the observed parental genotypes $\{g_{Obs.F}, g_{Obs.M}\}$. In the last two columns, '---' indicates that the probability $\Pr_{error}(\vec{g}_{True.i} \mid \vec{g}_{Obs.i}, M = 0)$ is not available.

When $\eta = \gamma$,

$\frac{0.5\gamma(1-\gamma)\pi_1^2}{\gamma(1-\gamma)\pi_1^2 + \eta(1-\eta)\pi_i^2} = \frac{0.5\pi_1^2}{\pi_1^2 + \pi_i^2}$ and $\frac{\eta(1-\eta)\pi_i^2}{\gamma(1-\gamma)\pi_1^2 + \eta(1-\eta)\pi_i^2} = \frac{\pi_i^2}{\pi_1^2 + \pi_i^2}$, $i = 0$ *or* 2.

Table 17: Results of null simulation

Set	10% Signif. level[j]	5% Signif. level	1% Signif. level	KS (*P*-value)
1[a]	0.110 [0.085, 0.140]	0.058 [0.041, 0.082]	0.010 [0.004, 0.023]	0.050 (0.1646)
2[b]	0.112 [0.087, 0.143]	0.048 [0.032, 0.070]	0.010 [0.004, 0.023]	0.047 (0.2179)
3[c]	0.114 [0.089, 0.145]	0.058 [0.041, 0.082]	0.010 [0.004, 0.023]	0.035 (0.5704)
4[d]	0.118 [0.093, 0.149]	0.058 [0.041, 0.082]	0.024 [0.014, 0.041]	0.043 (0.3145)
5[e]	0.112 [0.087, 0.143]	0.076 [0.056, 0.103]	0.018 [0.009, 0.034]	0.031 (0.7218)
6[f]	0.122 [0.096, 0.154]	0.068 [0.049, 0.094]	0.022 [0.012, 0.039]	0.059 (0.0606)
7[g]	0.110 [0.085, 0.140]	0.052 [0.036, 0.075]	0.006 [0.002, 0.017]	0.057 (0.0761)
8[h]	0.114 [0.089, 0.145]	0.064 [0.046, 0.089]	0.010 [0.004, 0.023]	0.055 (0.0929)
9[i]	0.114 [0.089, 0.145]	0.048 [0.032, 0.070]	0.008 [0.003, 0.020]	0.042 (0.3395)

[a]200 trios, 600 individuals, 100% genotyped, no genotyping errors;
[b]200 quartets, 800 individuals, 100% genotyped, no genotyping errors;
[c]100 mixed quartets, among which 25% contain affected sib pairs while 75% contain sib pairs with discordant affection status, 80% genotyped, no genotyping errors;
[d]92 nuclear families, 366 individuals, 100% genotyped, no genotyping errors;
[e]92 nuclear families, 366 individuals, 80% genotyped, no genotyping errors;
[f]92 nuclear families, 366 individuals, 80% genotyped, 1% genotyping errors;
[g]53 multiplex families, 313 individuals, 100% genotyped, no genotyping errors;
[h]53 multiplex families, 313 individuals, 80% genotyped, no genotyping errors;
[i]53 multiplex families, 313 individuals, 80% genotyped, 1% genotyping errors;
[j]95% Wilson confidence intervals in brackets determined using binomial distribution as implemented by function *binconf* in R (written by Rollin Brant).

Table 18: Power comparison of the TDT and this LRT at $\alpha = 0.05$

R_1	Pr(d), Pr(b)	N	TDT	LRT[a]	LRT[b]	LRT[c]	LRT[d]	LRT[e]	LRT[f]
1.75	0.12, 0.10	125	0.612	0.518	0.726	0.704	0.664	0.552	0.736
		175	0.740	0.654	0.842	0.826	0.788	0.696	0.862
	0.24, 0.20	125	0.774	0.700	0.854	0.857	0.820	0.778	0.892
		175	0.885	0.838	0.958	0.952	0.930	0.856	0.960
2.00	0.12, 0.10	125	0.792	0.754	0.896	0.890	0.870	0.810	0.914
		175	0.899	0.842	0.968	0.954	0.930	0.882	0.970
	0.24, 0.20	125	0.915	0.888	0.974	0.954	0.932	0.924	0.978
		175	0.974	0.946	0.996	0.988	0.992	0.974	0.998

[a]Trios, without using information of parental phenotypes, 100% genotyped
[b]Trios, using information of parental phenotypes, 100% genotyped
[c]Trios, using information of parental phenotypes, 80% genotyped
[d]Trios, using information of parental phenotypes, 80% genotyped, 1% genotyping errors
[e]Quartets, without using information of parental phenotypes, 100% genotyped
[f]Quartets, using information of parental phenotypes, 100% genotyped

The label of the first column, ' R_1 ', stands for the genotype relative risk at the disease locus defined as f_1/f_0 . Under the multiplicative mode of inheritance, $R_2 = f_2/f_0 = R_1^2$. The label of the second column, ' Pr(d), Pr(b) ', stands for the population frequency of the disease allele *d* and the population frequency of marker allele *b*. The label of the third column, '*N*', stands for the number of families in the power calculation.

Table 19: ANOVA table of the unrepeated 2^3 factorial design on the power difference of the TDT and the LRT

Response: Difference in Power						
Factor	DF	Sum of Square	Mean Sum of Square	*F*-value	Pr(>*F*)	Signif. Codes[a]
GRRD	1	0.00165312	0.00165312	44.1324	0.006950	**
MAF	1	0.00300312	0.00300312	80.1724	0.002939	**
NT	1	0.00103512	0.00103512	27.6340	0.013410	*
GRRD:NT	1	0.00035112	0.00035112	9.3737	0.054921	.
Residuals	3	0.00011237	0.00003746			

[a]Significance Codes: [0, 0.001]: '***', (0.001, 0.01]: '**', (0.01, 0.05]: '*' , (0.05, 0.1]: '.' , (0.1, 1]: ' '.

Table 20: Results of four family-based tests (ASP, HHRR, TDTae and LRT) for 23 SNPs in CHD7 gene

ID	SNP	*P*-value				GRRM[a] (TDTae)		GRRM (LRT)	
		ASP	HHRR	TDTae	LRT	$\hat{\overline{R}}_1$	$\hat{\overline{R}}_2$	$\hat{\overline{R}}_1$	$\hat{\overline{R}}_2$
1	rs4738813	0.016	0.39	0.009	0.003	1.900	3.609	1.989	3.955
2	rs1254430	0.005	0.004	0.002	0.011	2.373	5.627	1.843	3.396
3	rs9643371	0.002	0.006	0.0007	0.003	2.478	6.139	1.962	3.848
4	rs1017861	0.05	0.008	0.002	0.005	2.084	4.342	1.894	3.586
5	rs1325602	0.500	0.184	1.000	0.795	1.000	1.000	1.154	1.332
6	rs4288413	0.284	0.046	0.030	0.006	1.755	3.079	1.969	3.877
7	rs7000766	0.0002	0.003	0.0005	0.0005	2.701	7.294	2.334	5.448
8	hcv148921	0.013	0.008	0.0008	0.0007	2.196	4.820	2.092	4.378
9	rs1483207	0.004	0.486	0.007	0.0003	2.222	4.933	2.442	5.965
10	rs1483208	0.006	0.002	0.003	0.003	2.284	5.216	2.133	4.551
11	rs1038351	0.0008	0.004	0.0002	0.0007	3.059	9.355	2.385	5.689
12	rs7843033	0.001	0.002	0.0002	0.0004	2.994	8.961	2.469	6.095
13	rs7002806	0.002	0.013	0.009	0.027	2.049	4.200	1.693	2.867
14	rs7842389	0.0004	0.003	0.001	0.001	2.518	6.341	2.194	4.812
15	rs7017676	0.009	0.0007	0.0003	0.0008	2.860	8.182	2.332	5.440
16	hcv509505	0.035	0.001	0.0008	0.001	2.455	6.028	2.233	4.986
17	rs4392940	0.0006	0.002	0.0003	0.0004	2.909	8.460	2.410	5.810
18	rs4237036	0.008	0.002	0.002	0.0006	2.340	5.476	2.344	5.494
19	rs13280978	0.006	0.003	0.004	0.002	2.105	4.431	2.151	4.626
20	rs4301480	0.004	0.001	0.003	0.011	2.498	6.240	1.907	3.638
21	rs10957159	0.5	0.084	1.000	0.620	1.000	1.000	1.154	1.332
22	rs10092214	0.019	0.50	0.434	0.169	1.181	1.395	1.332	1.774
23	rs3763591	0.027	0.50	0.288	0.052	1.289	1.660	1.531	2.344

[a]Genotype relative risks at the marker locus (GRRM) are estimated under the multiplicative mode of inheritance

Table 21: Results of three family-based tests (FBAT, TDTae, and LRT) for 13 SNPs on chromosome 17q25 for 242 psoriasis families

Locus	FBAT *p*-value	TDTae *p*-value	LRT *p*-value	$\hat{\overline{R}}_1$	$\hat{\overline{R}}_2$	#INC[a]
#57	0.0071	0.0228	0.0082	1.359	1.848	4
#58	0.0047	0.0162	0.1793	1.172	1.374	6
#59	0.0015	0.0116	0.0011	1.468	2.156	3
#60	0.0089	0.0342	0.1327	1.190	1.417	3
#61	0.0085	0.0408	0.0201	1.308	1.710	3
#62	0.0065	0.0328	0.0038	1.392	1.937	4
#63	0.0016	0.0038	0.0371	1.172	1.373	5
#64	0.0146	0.0391	0.1114	1.203	1.448	3
#65	0.0020	0.0053	0.0009	1.474	2.173	3
#66	0.0082	0.0087	0.0590	1.001	1.001	5
#67	0.0037	0.0376	0.1247	1.192	1.420	5
#69	0.2560	0.4270	0.0574	1.553	2.412	3
#70	0.0737	0.2464	0.0443	1.314	1.726	2

[a]Number of inconsistencies in the genotype data

References

Abel L, Muller-Myhsok B. 1998. Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. American Journal of Human Genetics 63(2):664-667.

Allen-Brady K, Wong J, Camp NJ. 2006. PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. BMC Bioinformatics 7:209.

Allison DB. 1997. Transmission-disequilibrium tests for quantitative traits. American Journal of Human Genetics 60:676-690.

Badzioch MD, DeFrance HB, Jarvik GP. 2003. An examination of the genotyping error detection function of SIMWALK2. BMC Genetics 4(Suppl 1):S40.

Berge C. 1962. The theory of graphs. Methuen, London.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of Royal Statistical Society: Series B 57:289-300.

Blanton SH, Heckenlively JR, Cottingham AW, Friedman J, Sadler LA, Wagner M, Friedman LH, Daiger SP. 1991. Linkage mapping of autosomal dominant retinitis pigmentosa (RP1) to the pericentric region of human chromosome 8. Genomics 11(4):857-69.

Cannings C, Thompson EA, Skolnick MH. 1978. Probability functions on complex pedigrees. Advances in Applied Probability 10(1):22-61.

Celis M, Dennis JE, Tapia RA. 1985. A trust region strategy for nonlinear equality constrained optimization. In Proceedings of the SIAM Conference on Numerical Optimization: 71-82.

Chakravarti A. 1991. Information content of the Centre d'Eude du Polymorphisme Humain (CEPH) family structures for linkage studies. Human Genetics 87(6):721-724.

Chaudhuri S, Messing J. 1994. Allele-specific parental imprinting of dzr1, a posttranscriptional regulator of zein accumulation. Proceedings of the National Academy of Sciences of the United States of America 91(11):4867-4871.

Cheng KF, Chen JH. 2007. A simple and robust TDT-type test against genotyping error with error rates varying across families. Human Heredity 64:114-112.

Clayton D. 1999. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. American Journal of Human Genetics 65:1170-1177.

Curtis D, Sham PC. 1995. A note on the application of the transmission disequilibrium test when a parent is missing. American Journal of Genetics 56:811-812.

Demenais FM, Elston RC. 1981. A general transmission probability model for pedigree data. Human Heredity 31:93-99.

Dempster AP, Larid NM, Rubin DB. 1997. Maximum likelihood estimation from imcomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society Series B 39: 1-38.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for find-scale mapping. Genomics 29:311-322.

Douglas JA, Boehnke M, Lange K. 2000. A multipoint method for detecting genotyping errors and mutatins in sibling-pair linkage data. American Journal of Human Genetics 66:1287-1297.

Douglas JA, Skol AD, Boehnke M. 2002. Probability of detection of genotyping

errors and mutations as inheritance inconsistencies in nuclear-family data. American Journal of Human Genetics 70:487-495.
Ehm MG, Kimmel M, Cottingham RW Jr. (1996). Error detection for genetic data, using likelihood methods. American Journal of Human Genetics 58:255-234.
Elston RC, Stewart J. 1971. A general model for the analysis of pedigree data. Human Heredity 21:523-542.
Freeman PE, Doe S, Siemiginowska A. 2001. Sherpa: a mission-independent data analysis application. SPIE Proceedings 4477:76.
Gao X, Gordon D, Zhang D, Browne R, Helms C, Gillum J, Weber S, Devroy S, Swaney S, Dobbs M, Morcuende J, Sheffield V, Lovett M, Bowcock A, Herring J, Wise C. 2007. CHD7 gene polymorphisms are associated with susceptibility to idiopathic scoliosis. American Journal of Human Genetics 80:957-965.
Gordon D, Heath SC, Ott J. 1999. True pedigree errors more frequent than apparent errors for single nucleotide polymorphism. Human Heredity 49:65-70.
Gordon D, Leal SM, Heath SC, Ott J. 2000. An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. Pacific Symposium on Biocomputing 5:663-674.
Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Ott J. 2004. A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. European Journal of Human Genetics 12:752-761.
Gordon D, Health SC, Liu X, Ott J. 2001. A transmission disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. American Journal of Human Genetics 69:371-380.
Gordon D, Ott J. 2001. Assessment and management of single nucleotide polymorpohism genotype errors in genetic association analysis. Pacific Symposium Biocomputing: 18-19.
Han L, Liu G. 2004. On the convergence of the UOBYQA method. Journal of Applied Mathematics and Computing 16(1-2):125-142.
Hardy GH. 1908. Mendelian proportions in a mixed population. Science 28:49-50.
Hasstedt DJ. 1993. Variance components/major locus likelihood approximation for quantitative, polychotomous, and multivariate data. Genetic Epidemiology 10(3):145-158.
Helms C, Cao L, Krueger JG *et al.* 2003. A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. Nature Genetics 35:349-356.
Kolmogoroff A. 1941. Confidence limits for an unknown distribution function. Annals of Mathematical Statistics 12:461-463.
Kruglyak L, Lander ES. 1998. Faster multipoint linkage analysis using Fourier transforms. Journal of Computational Biology 5(1):1-7.
Lake SL, Blacker D, Laird NM. 2000. Family-Based tests of association in the presence of linkage. American Journal of Human Genetics 67(6):1515-1525.
Laird NM. 2006. Family-based association tests and the FBAT-toolkit. http://www.biostat.harvard.edu/~fbat/fbat.htm.
Laird NM, Horvath S, Xu X. 2000. Implementing a unified approach to family based tests of association. Genetic Epidemiology 19(Suppl 1):S26-S42.
Laird NM, Lange C. 2006. Family-based designs in the age of large-scale gene-association studies. Nature Reviews. Genetics 7: 385-394.
Lander ES, Green P. 1987. Construction of multilocus genetic maps in humans.

Proceedings of the National Academy of Sciences of USA 84:2363-2367.
Lander ES, Kruglyak L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature Genetics 11: 241-247.
Lange K, Sinsheimer JS, Sobel E. 2005. Association testing with Mendel. Genetic Epidemiology 29:36-50.
Lewontin R, Kojima K. 1960. The evolutionary dynamics of complex polymorphisms. Evolution 14:458-472.
Little RJA, Rubin DB. 2002. Statistical analysis with missing data, second edition. Wiley-Interscience. Chapter 5 Section 3: 88-92.
Lyles RH, Taylor DJ, Hanfelt JJ, Kupper LL. 2001. An alternative parametric approach for discrete missing data problems. Communications in Statitisics - Theory and Methods 30:1969-1988.
Martin ER, Kaplan NL, Weir BS. 1997. Tests for linkage and association in nuclear families. American Journal of Human Genetics 62:450-458.
Martin ER, Bass MP, Hauser ER, Kaplan NL. 2003. Accounting for linkage in family-based tests of association with missing parental genotypes. American Journal of Human Genetics 73:1016-1026.
Martin RB, Alda M, Maclean CJ. 1998. Parental genotype reconstruction: applications of haplotype relative risk to imcomplete parental data. Genetic Epidemiology 15:471-490.
Morris RW. 2003. Likelihood ratio tests for association with multiple disease susceptibility alleles, genotyping errors, or missing parental data. PhD dissertation, North Carolina State University: 38.
Nielsen DM, Weir BS. 2001. Association studies under general disease models. Theoretical Population Biology 60:253-263.
Nyholt DR. 2002. GENEHUNTER: Your ‘one-stop shop’ for statistical genetic analysis? Human Heredity 53:2-7.
Powell MJD. 1964. An efficient method for finding the minimum of a function of several variables without calculating derivatives. The Computer Journal 7:155-162.
Powell MJD. 1998. Direct search algorithms for optimization calculations. Acta Numerica 7:287-336.
Powell MJD. 2000. UOBYQA: unconstrained optimization by quadratic approximation. Mathematical Programming 92(3):555-582.
Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 2002. Numerical recipes in C. The art of scientific computing. Cambridge: Cambridge University Press.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, DeBakker PIW, Daly MJ, Sham PC. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses.
Rabinowitz D. 1997. A transmission disequilibrium test for quantitative trait loci. Human Heredity 47:342-350.
Rabinowitz D, Laird N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Human Heredity 50:211-223.
Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516-1517.
Rollin B. http://cran.r-project.org/doc/packages/Hmisc.pdf.
Robbins RB. 1918. Some applications of mathematics to breeding problems III. Genetics 3:375-389.
Schafer JL, Graham JW. 2002. Missing data: our view of the state of the art.

Psychological Methods 7(2): 147-177.
Schaid DJ. 1996. General score tests for associations of genetic marker with disease using cases and their parents. Genetic Epidemiology 13:423-449.
Schaid DJ, Sommer SS. 1993. Genotype relative risks: methods for design and analysis of candidate-gene association studies. American Journal of Human Genetics 53(5):1114-1125.
Schaid DJ, Sommer SS. 1994. Comparison of statistics for candidate-gene association studies using cases and parents. American Journal of Human Genetics 55:402-409.
Schäffer AA. 2000. Pedigree Travesal in FASTLINK. Rice University.
Sham P. 1997. Statistics in Human Genetics. First edition. Arnold Publishing.
Siegmund KD and Gauderman WJ. 2001. Association tests in nuclear families. Human Heredity 52:66-76.
Smirnov N. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin de 1 Universite de Moscou, Serie internationale (Mathematiques) 2:3-14.
Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics 52:506-516.
Spielman RS, Ewens WJ. 1996. The TDT and other family-based tests for linkage disequilibrium and association. American Journal of Human Genetics 59:983-989.
Spielman RS, Ewens WJ. 1998. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. American Journal of Human Genetics 62: 450-458.
Sun F, Flanders WD, Yang Q, Khoury MJ. 1999. Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT. American Journal of Epidemiology 150:97-104.
Terwilliger JD. 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. American Journal of Human Genetics 56:777-787.
Terwilliger JD, Ott J. 1992. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Human Heredity 42(6):337-346.
Terwilliger JD, Ott J. 1994. Handbook of Human Genetic Linkage. Baltimore, Johns Hopkins.
Thiele H and Nürnberg P. 2005. HaploPainter: a tool for drawing pedigrees with complex haplotypes. Bioinformatics 21(8):1730-1732.
Torczon V (1997). On the convergence of pattern search algorithms. SIAM Journal of Optimization 7:1-25.
Tu IP, Balise RR, Whittemore AS. 2000. Detection of disease genes by use of family data. II. Application to nuclear families. American Journal of Human Genetics 66:1341-1350.
Weeks DE. 2005. http://cran.r-project.org/doc/packages/powerpkg.pdf.
Weeks DE, Ott J, Lathrop GM. 1990. SLINK: a general simulation program for linkage analysis. American Journal of Human Genetics 47:A204.
Weinberg CR. 1999. Allowing for missing parents in genetic studies of case-parent triads. American Journal of Human Genetics 64:1189-1193.
Weinberg W. 1908. Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg 64: 368–382.
Whittemore AS, Tu IP. 2000. Detection of disease genes by use of family data. I.

Likelihood-based theory. American Journal of Human Genetics 66: 1328-1340.
Zou G, Pan D, Zhao H. 2003. Genotyping error detection through tightly linked markers. Genetics 164:1161-1173.

Appendix

Under the following assumptions:

(1) HWE on the marker locus: $\Pr(m_1 m_2) = \Pr(m_1)\Pr(m_2)$,

(2) random mating between the parental gamete:
$\Pr(d_1 m_1 : d_2 m_2) = \Pr(d_1 m_1)\Pr(d_2 m_2)$, and

(3) multiplicative penetrances at the DSL: $f_{d_1 d_2}{}^2 = f_{d_1 d_1} f_{d_2 d_2}$,

the marker penetrance is given by

$$\phi_{m_1 m_2} = \sum_{d_1}\sum_{d_2} f_{d_1 d_2} \Pr(d_1 m_1 : d_2 m_2 \mid m_1 m_2) = \sum_{d_1}\sum_{d_2} f_{d_1 d_2} \Pr(d_1 m_1)\Pr(d_2 m_2)/\Pr(m_1 m_2)$$

$$= \sum_{d_1}\sum_{d_2} f_{d_1 d_2}\left[\Pr(d_1 m_1)/\Pr(m_1)\right]\left[\Pr(d_2 m_2)/\Pr(m_2)\right]$$

$$= \sum_{d_1}\sum_{d_2} \sqrt{f_{d_1 d_1} f_{d_2 d_2}}\left[\Pr(d_1 m_1)/\Pr(m_1)\right]\left[\Pr(d_2 m_2)/\Pr(m_2)\right]$$

$$= \sum_{d_1} \sqrt{f_{d_1 d_1}}\,\Pr(d_1 \mid m_1)\sum_{d_2} \sqrt{f_{d_2 d_2}}\,\Pr(d_2 \mid m_2) \quad \text{(Morris, 2003)}$$

For a di-allelic marker locus, the marker penetrances

$$\phi_0 = \phi_{aa} = [\sqrt{f_{++}}\,\Pr(+\mid a) + \sqrt{f_{dd}}\,\Pr(d \mid a)]^2 = [\sqrt{f_0}\,\Pr(+\mid a) + \sqrt{f_2}\,\Pr(d \mid a)]^2$$

$$\phi_1 = \phi_{ab} = [\sqrt{f_{++}}\,\Pr(+\mid a) + \sqrt{f_{dd}}\,\Pr(d \mid a)][\sqrt{f_{++}}\,\Pr(+\mid b) + \sqrt{f_{dd}}\,\Pr(d \mid b)]$$

$$= [\sqrt{f_0}\,\Pr(+\mid a) + \sqrt{f_2}\,\Pr(d \mid a)][\sqrt{f_0}\,\Pr(+\mid b) + \sqrt{f_2}\,\Pr(d \mid b)]$$

$$\phi_2 = \phi_{bb} = [\sqrt{f_{++}}\,\Pr(+\mid b) + \sqrt{f_{dd}}\,\Pr(d \mid b)]^2 = [\sqrt{f_0}\,\Pr(+\mid b) + \sqrt{f_2}\,\Pr(d \mid b)]^2,$$

where $\Pr(d \mid \cdot) = 1 - \Pr(+ \mid \cdot)$.

Note that the conditional probabilities above are the respective probabilities of disease allele d_i given the marker allele m_j. For example, $\Pr(+ \mid a)$ is the probability of the low risk disease allele + given that the marker's allele is *a*. It is obvious that these marker penetrances are also multiplicative: $\phi_1^2 = \phi_0 \phi_2$.