ABSTRACT

CHUNG, REN-HUA. Statistical Methods for Family-Based Association Studies for Complex Human Diseases: Single-Locus and Haplotype Methods. (Under the direction of Dr. Eden Martin.)

Disease-gene fine-mapping is an important task in human genetics. Linkage and association analyses are the two main approaches for exploring disease susceptibility genes. In Chapter 1, we introduce the development of methods for disease-gene mapping in the past decades and present the rationale behind our new method development. Family-based association analyses have provided powerful tools for disease-gene mapping. The Association in the Presence of Linkage test (APL), a family-based association method, can use nuclear families with multiple affected siblings and infer missing parental genotypes properly in the linkage region. In Chapter 2, we generalized and extended APL so that it can be applied to general nuclear family structures using a bootstrap variance estimator. Unlike the original APL that can handle at most two affected siblings, the new APL can handle up to three affected siblings. We also extended APL from a single-marker test to a multiple-marker haplotype analysis. According to our simulations, the new APL has a correct type I error rate and more power than other family-based association methods such as PDT, FBAT/HBAT, and PDTPHASE in nuclear families with missing parents. The robustness of APL when there are rare alleles or haplotypes and when there is population substructure such that the allele frequencies in the population deviate from the Hardy-Weinberg Equilibrium (HWE) assumption was also examined in Chapter 2. Genes on the X chromosome play a role in many common diseases. Linkage analyses have identified regions on the X chromosome with high linkage peaks for several diseases. Currently there are few family-based association

methods available for X-chromosome markers. In order to fill in this gap, we proposed a novel family-based association method, X-APL, in Chapter 3. X-APL is a modification of APL and shares some important properties with APL. X-APL can also perform haplotype analyses, which is the only family-based test of association we are aware of for testing haplotypes for the X-chromosome markers. Our simulation results showed that X-APL has a correct type I error rate and has more power than other family-based association methods for X chromosome such as XS-TDT, XPDT and XMCPDT for single-marker analysis in nuclear families. The robustness of X-APL when there are deviations of genotype frequencies from HWE was also examined in Chapter 3. Linkage and family-based association analyses are often applied simultaneously in the same data in order to maximize use of family data sets. However, it is not intuitively clear under what conditions association and linkage tests performed in the same data set may be correlated. In Chapter 4, we used computer simulations and theoretical statements to estimate the correlation between linkage statistics (affected sib pair maximum LOD scores) and family-based association statistics (PDT and APL) under various hypotheses. Different types of pedigrees were studied: nuclear families with affected sib pairs, extended pedigrees and incomplete pedigrees. Both simulation and theoretical results showed that when there is either no linkage or no association, the linkage and association statistics are not correlated. When there is linkage and association in the data, the two tests have a positive correlation.

STATISTICAL METHODS FOR FAMILY-BASED ASSOCIATION STUDIES FOR COMPLEX HUMAN DISEASES: SINGLE-LOCUS AND HAPLOTYPE METHODS

by

REN-HUA CHUNG

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

BIOINFORMATICS

Raleigh, North Carolina

2006

APPROVED BY:

Bruce S. Weir Co-chair of Advisory Committee	Eden R. Martin Co-chair of Advisory Committee
Trudy F.C. Mackay	Dahlia M. Nielsen
Jung-Ying Tzeng	

To my parents,
Mr. Hai-Jung Chung and Mrs. Mei-Chih Li

僅以此獻給我的父母
鍾海榮先生以及李美枝女士

Biography

Ren-Hua Chung was born in Hualien County in Taiwan on September 21, 1978. He received his Bachelor of Science degree in computer science from National Chiao-Tung University in 2000. After working for about one year as a programmer for the Stark Company in Hsinchu in Taiwan, he went to the University of California at Davis for his Master of Science degree in computer science. He was working with Dr. Dan Gusfield, who guided him into the exciting and interesting field of bioinformatics for haplotype inference based on the perfect phylogenetic tree model. He completed his M.S. degree in 2003, and started his doctoral work in bioinformatics at North Carolina State University. While working toward his doctoral degree, he interned at the Center for Human Genetics at Duke University under the direction of Dr. Eden Martin. His research interests focus on method development for complex disease gene-mapping, particularly for family-based association analysis.

Acknowledgement

I am very grateful for the advice of my advisor, Eden Martin, during my doctoral studies. She is always patient and provides very helpful suggestions for the problems I encounter during my research. I always feel lucky and am proud of working with her. I also thank Dr Bruce Weir for his comments on my research and advice about course selection. I also would like to thank my other committee, Trudy Mackay, Dahlia Nielsen, Jung-Ying Tzeng, and also the graduate school representative, David McAllister.

I would like to thank the staff at the Bioinformatics Research Center, especially Juliebeth Briseno, who answered so many questions I had about graduate school regulations. I also thank Elizabeth Hauser, Richard Morris and Yi-Ju Li at Center for Human Genetics at Duke University for their advice and comments on my research. I also appreciate the support from members at CHG for testing and suggestions for the computer programs I have written.

I would like to express my special thanks to Ying-Erh Chen, who supports me with love and sincerity.

Needless to say, none of this would be possible without the support of my family, including my parents and sister and brother-in-law.

Table of Contents

List of Tables

List of Figures

Chapter 1

Review

1.1 Introduction to disease-gene mapping

In 1865, the Bohemian monk Gregor Mendel published "Experiments of Plant Hybridyzation," which later became Mendel's law of inheritance, and this law turned into an essential chapter in today's genetics textbooks. Mendel studied traits that were mainly caused by segregation of a single gene. Thus, diseases caused by mutations in one gene are referred to as Mendelian diseases. Finding disease susceptibility genes is one of the major tasks in human genetics studies. Disease-gene mapping has been fairly successful for Mendelian disorders, mainly by the process of positional cloning [Risch, 2000]. Traditionally, genes were isolated based on the amino acid sequences of known proteins. Positional cloning has the property that genes are identified and mapped solely based on the inherited traits and no biological knowledge regarding the traits is required [Botstein and Risch, 2003]. A total of 1822 genes have been reported that cause monogenic Mendelian diseases in "the online version of Mendelian Inheritance in Man" database (OIMM) [Antonarakis and McKusick, 2000; Antonarakis and Beckmann, 2006].

For complex diseases that multiple genes, as well as environmental risk factors, may directly or interactively cause, more factors should be considered for disease-gene mapping. Some examples of complex diseases are Alzheimer's disease, schizophrenia, type-2 diabetes and cancer. Disease-gene mapping for complex diseases is more challenging than mapping genes for Mendelian disorders due to genetic heterogeneity in which mutations in different genes can cause the same disease phenotype [Lander and Schork, 1994; Risch, 2000]. Other factors

such as incomplete penetrances, phenocopies and late age at disease onset also limit the progress of complex disease gene mapping [Gillanders et al., 2006]. Hence, disease-gene mapping efforts for complex diseases have not been as successful as those for Mendelian disorders [Weiss and Terwilliger, 2000; Todd, 2001; Tabor et al., 2002]. For example, the number of genes and environmental factors involved in schizophrenia is not clear. The genes encoding dysbindin (DTNBP1) and neuregulin 1 (NRG1) are considered to have strong evidence of association with schizophrenia [Owen et al., 2005]. Other genes such as "disrupted in schizophrenia 1" (DISC1), "D-amino-acid oxidase" (DAO), "D-amino-acid oxidase activator" (DAOA) and "regulator of G-protein signaling 4" (RGS4) still do not have convincing results for schizophrenia [Owen et al., 2005].

Although 99.9% of the human genomes are identical between people, there are still millions of differences among the 3.2 billion base pairs [Kruglyak and Nickerson, 2001]. These genetic variations can cause phenotypic variations among people and are potentially associated with traits or diseases. Genetic markers, which are nucleotide variants with known positions, are often used for human disease analyses. Several types of markers exist, such as Restriction Fragment Length Polymorphisms (RFLP's), microsatellites, and single nucleotide polymorphisms (SNPs). Markers can be used to construct a genetic map, which can be used as a reference for disease-gene mapping [Dib et al., 1996]. Researchers can genotype markers for studying their relationship with diseases according to a genetic map. Botstein et al. [1980] proposed the concept using RFLP's as the markers to construct a genetic map. Later, genetic

maps were constructed using denser microsatellites [Cooperative Human Linkage Center, 1994; Dib et al., 1996]. SNPs, which usually contain two alleles, have drawn significant attention as markers for genetic disease-mapping studies due to their high abundance across the human genome [Kruglyak, 1997; The International SNP Map Working Group, 2001]. It was estimated that there are around 7.1 million SNPs with a minimal allele frequency of at least 0.05 in the human population [Kruglyak and Nickerson, 2001]. With the completion of PHASE I of the HapMap Project, the number of SNPs in the public database (dbSNP) increased from 2.6 million to 9.2 million [The International HapMap Consortium, 2005].

As genotyping cost has become cheaper and the process has become faster, genotyping for markers can be performed on a genome-wide scale, which produces a large amount of marker data for analysis [Gunderson et al., 2005; Syvanen, 2006]. Hence, statistical methods are required after numerous markers are genotyped from collected samples. Two commonly used statistical methods are linkage and association (linkage disequilibrium) analyses [Lander and Schork, 1995]. Theoretical methods for linkage tests were proposed around 1930 [Fisher, 1935a; 1935b; Penrose, 1935]. Association analyses can be performed based on case-control samples or samples collected from families [Falk and Rubinstein, 1987; Spielman et al., 1993]. These two methods will be introduced in the following sections.

1.2 Linkage analysis

The first step of positional cloning is linkage analysis. Then genes are cloned according to the mapped positions from linkage analyses to study their functions. Linkage analyses are used to find chromosome regions that do not recombine with a proposed disease locus. Linkage is often evaluated by the logarithm of the odds (LOD) score [Morton et al., 1955], which is the logarithm of odds of the recombination rate equal to θ estimated from the observed data with respect to the assumption that the recombination rate is 0.5. A traditional LOD score assumes that a single locus contributes to the disease with a specific model of inheritance (e.g., dominant or recessive model). Hence, this type of method that requires a genetic model assumption is called parametric and may not have high power for complex diseases, since an obvious genetic segregation of markers cannot be observed in the polygenic disorders [Weeks and Lathrop, 1995].

Another category of methods utilizes the observation of allele sharing between affected sib pairs [Kruglyak et al., 1996; Kong and Cox, 1997; Whittemore and Tu, 1998]. The LOD score is defined based on the allele-sharing probabilities for identity-by-descent (IBD) between affected siblings. If there is no linkage between the disease locus and marker, the transmissions of alleles from parents to siblings are independent and the probabilities of an affected sib pair sharing 0, 1, and 2 alleles IBD are 0.25, 0.5, and 0.25, respectively. If there is linkage between the disease locus and marker, the transmissions of alleles from parents to affected siblings are not independent and the IBD probabilities may vary. Therefore, one can

compare the observed estimates of IBD parameters in the data with the expected IBD parameters (1/4, 1/2, 1/4) derived under the null assumption that there is no linkage. A significant departure of the observed IBD parameters from the expected IBD parameters implies the presence of linkage. A major advantage of the allele-sharing method is that it does not require the information of a genetic model. Hence, it is referred as a non-parametric model and is robust under different genetic models.

Linkage analysis can be performed with either two-point or multipoint estimates [Kruglyak et al., 1996]. For two-point linkage analysis, only one marker and the disease locus are considered when calculating the statistic. For multipoint linkage analysis, several markers are considered simultaneously with the disease locus. Hence, we can define the most likely position of the disease locus on the marker map. A map of the markers with distances between them is required for multipoint linkage analysis.

1.3 Association analysis

Linkage analysis can lack power for common diseases caused by multiple genes and environmental factors [Cardon and Bell, 2001]. Since the linkage test uses LOD scores to measure the degree of linkage in data, it is not intuitively obvious how large a LOD score should be in order to be claimed a significant finding [Lander and Kruglyak, 1995; Curtis, 1996; Risch and Botstein, 1996; Morton, 1998]. The identified linkage regions also rarely reached a resolution less than a few megabases [Cardon and Bell, 2001]. Association analysis

can be used as a complementary method to linkage analysis. The association test can be more powerful than the linkage test, and it requires fewer samples than linkage analysis to achieve the same power for common complex diseases [Risch and Merikangas, 1996].

Association analysis tests whether the disease and marker alleles are in linkage disequilibrium (LD). Disease phenotypes are used for association analyses instead of disease loci since, in general, the disease loci are unknown [Weiss and Terwillinger, 2000]. LD generally spans only small distances, and the markers used for association analysis are often very tightly spaced. Therefore, association analysis provides a higher resolution for locating disease genes than linkage analysis. A common strategy for identifying complex disease genes is to conduct linkage analyses first and then follow significant results with tests for association at a denser panel of markers in an attempt to further localize the disease gene [Cardon and Bell, 2001].

1.3.1 Population-based association analysis

Two main categories of statistical methods, population-based (case-control and case-cohort studies) and family-based studies, are often used for association analysis [Laird and Lange, 2006]. Population-based analysis requires samples to be independently collected. It compares the differences of distributions of allele frequencies between the affected individuals (cases) and unaffected individuals (controls) [Risch, 2000]. A contingency table can be created and the Pearson chi-squared statistic or Fisher's exact test can be used to test for association.

Regression-based analyses such as logistic regression can also be used in the case-control test [Agresti, 2002]. The main limitation of the case-control analysis is that the presence of confounding effects in the samples could cause a high false positive rate in the analysis [Risch, 2000; Devlin et al., 2001]. For example, population admixture and population substructure can cause confounding, which can produce association between unlinked loci [Ewens and Spielman, 1995].

Two major types of approaches were proposed to solve this problem: genomic control (GC) [Devlin et al., 1999; 2001] and structured analysis (SA) [Prichard et al., 2000]. In GC, Devlin and Roeder [1999] demonstrated that the effect of confounding is constant across the genome, which potentially allows for correction on the test statistic. A set of null markers across the genome was used to estimate the effect of confounding. The confounding effect is then removed from the test statistic for association to achieve a reasonable type I error rate. SA analysis assumed the population was derived from several subpopulations and the allele frequencies were different between subpopulations. A Markov Chain Monte Carlo (MCMC) algorithm was applied to infer the origin of each individual in the sample using a set of loci unlinked to the candidate gene, given a specific number of origins. Individuals from the same origin were clustered into a group. Then association analysis was performed conditionally on each inferred group.

1.3.2 Family-based association analysis

Another approach for the association test uses family data. A widely used family-based method, the TDT [Spielman et al. 1993], compares the differences of alleles transmitted and untransmitted from parents to affected siblings in triad families (one affected offspring and both parents). A McNemar's chi-squared test is used for the paired transmitted and untransmitted statistics. The TDT was originally proposed to test for linkage in the presence of association, but it is also a valid test for association in the presence of linkage [Ewens and Spielman, 2005]. In terms of statistical power, the TDT has similar power compared with case-control studies for association tests when the number of triad families is equal to the number of cases and the number of cases is equal to the number of controls for case-control studies [McGinnis et al., 2002]. Hence, performing case-control studies for association can cost less, since collecting family data generally requires more resources in terms of time and money [Laird and Lange, 2006]. However, the TDT test has the advantage that it is valid even when population stratification is present in the data [Ewens and Spielman, 1995], since the test is conditional on parental data.

In the TDT, each pair of transmitted/untransmitted alleles from a parent to an affected sibling is treated as independent to construct the McNemar's test. However, as a test for association in a linkage region, this assumption does not hold for transmissions between affected siblings. Hence, the TDT is not a valid test for association when more than one affected sibling is used and there is linkage between marker and disease loci [Martin et al., 1997].

One solution is to randomly select one affected sibling from each family and perform the TDT [Wang et al., 1996]. However, affected sibling pairs can significantly increase the power and efficiency of the family-based association test [Risch 2000]. It was estimated that less than half of the number of families with one affected sib are required for families with two affected sibs to achieve the same power as families with one affected sib [McGinnis et al., 2002]. Hence, it is not an optimal solution for the TDT to use only one affected sibling in the family when other affected siblings' information is available.

Several modifications of the TDT for association test were proposed to account for linkage in families with multiple affected siblings. Martin et al. [1997] proposed the Pedigree Disequilibrium Test (PDT) that treats the transmissions from a parent to the affected sib pair as a unit, and the unit can be shown to be independent between parents. The PDT statistic and its variance were constructed based on the unit of transmissions and can avoid the independence assumption between affected siblings used in TDT. Rabinowitz and Laird [2000] compared the difference between the transmissions from parents to the affected siblings and the expected value conditional on the minimum sufficient statistics for the null distribution. The distribution for the statistic can be generated by the Monte-Carlo method [Kaplan et al., 1997], approximated by asymptotic normal distribution, or computed by the exact distribution when the number of pedigrees is small [Rabinowitz and Laird, 2000].

TDT was also extended to large pedigrees (extended pedigrees). In Martin et al. [2000a], the extended pedigrees are partitioned into several related nuclear families, and the transmissions in each related nuclear family sums to a statistic. The variance for the statistic was estimated based on independent transmissions between each extended pedigree. Abecasis et al. [2000] also used a similar strategy to Martin et al. [2000a] that generalized TDT to extended pedigrees.

For late-onset diseases, parental genotypes for the affected siblings are often missing. In order to accommodate the loss of information, one approach for a family-based association test is to compare the difference of allele frequencies between affected and unaffected siblings without using parental data, such as the S-TDT test proposed in Spielman and Ewens [1998]. However, S-TDT still has the requirement that only one affected sibling with one unaffected sibling should be used in each family for a valid test for association in the presence of linkage. S-TDT was generalized to multiple affected sibs in Horvath and Laird [1998] as the SDT test. The difference of the averaged numbers of a certain allele between the affected and unaffected siblings was used to form a sign test in SDT. An exact distribution for the SDT statistic was calculated based on the distribution of the sign test. When the genotypes of some parents as well as some unaffected siblings are available, the statistic for the transmissions from parents to affected siblings such as the TDT statistic can be combined with the statistic for the difference of the number of a certain allele between affected and unaffected siblings [Spielman and Ewens, 1998; Martin et al., 2000a].

Another approach to deal with missing parental genotypes uses siblings' genotypes to infer the missing parental genotypes and then compares the number of alleles transmitted and untransmitted from parents to affected siblings [Weinberg, 1999]. Knapp [1999] proposed "reconstruction combined TDT" (RC-TDT), which reconstructs missing parental genotypes first and then performs the combined TDT and S-TDT test. The missing parental mating-types are reconstructed based on siblings' genotypes. However, because of the same property inherited from TDT and S-TDT, RC-TDT is not a valid test for association in the presence of linkage when multiple affected siblings are used in the data. Clayton [1999] proposed a score test derived from the likelihood of parental genotypes and offspring genotypes conditional on disease in the offspring. When there are missing parents in the data, the likelihood for possible parental genotypes was used for in the likelihood to derive the score test. The variance was estimated by taking the variability for inferring possible parental genotypes into consideration.

Linkage between disease and marker loci should be considered when inferring the missing parental genotypes based on siblings' genotypes with multiple affected sibs present in the data [Martin et al., 2003]. The score test proposed by Clayton [1999] is implemented in the software package TRANSMIT. The inference of missing parental genotype in TRANSMIT is based on Mendelian probabilities for the siblings, which assumes the transmissions to the affected sibs are independent. The inference is appropriate when there is no linkage between

disease and marker loci. However, in a linkage region, ignoring linkage when inferring the missing parental genotypes with multiple affected sibs in the data can inflate the type I error rate in TRANSMIT [Martin et al., 2003].

1.3.3 Association in the Presence of Linkage method

The Association in the Presence of Linkage (APL) method, proposed by Martin et al. [2003], is a powerful family-based association tool. The APL uses nuclear families with at least one affected offspring. The APL compares the difference between the number of copies of a specific allele in affected offspring and the expected number under the null hypothesis of no association conditional on parental genotypes. The APL can infer missing parental genotypes properly in the linkage region by taking the IBD parameters for affected siblings into consideration. Hence, APL does not have the problem of possible inflation of the type I error rate, as in TRANSMIT, if linkage is present and multiple affected siblings' data are used [Martin et al., 2003]. Martin et al. [2003] demonstrated that APL can have more power than PDT and FBAT [Rabinowitz and Laird, 2000] for nuclear family data with missing parents. Hence, APL provides a useful family-based association tool for late-onset diseases in which parental data are usually not available.

The original APL proposed in Martin et al. [2003] is not flexible for mixed nuclear family structures (including the mixture of singleton and multiplex families with an arbitrary number of unaffected sibs) due to the fact that different parameters in the variance estimator

should be considered for each type of family structure separately. Moreover, the original APL can use nuclear families with up to two affected sibs. However, in real data analysis, the data can contain a mixture of different nuclear family structures. For disease with higher prevalence, more than two affected sibs can be present in a family. Taking more affected sibs into account in the statistic may increase the power for a family-based association test. IBD status between each pair of affected sibs should be considered in the linkage region to infer missing parental genotypes when including multiple affected sibs in the test. In order to resolve these problems, a novel variance estimator based on the bootstrap approach [Efron and Tibshirani, 1993], which allows APL for different nuclear family structures and extension to use of three affected sibs, will be introduced in Chapter 2. A strategy of inference for missing parental genotypes based on IBD status between multiple affected sibs will also be introduced in Chapter 2.

Haplotype analyses can show more power than single-marker analyses if the joint LD between markers and the disease locus is stronger than the pairwise LD between a single marker and the disease locus [Morris and Kaplan, 2003; Nielsen et al., 2004]. A global test for all haplotypes jointly considers the effects of all haplotypes on the disease under the global null hypothesis that none of the haplotypes are associated with the disease. A global test for all haplotypes can be more informative than individual haplotype tests since it can capture multiple haplotype effects [Horvath et al., 2004]. The global test can also have more power than individual haplotype tests since it does not have the multiple-testing issue faced

when analyzing haplotypes individually [Morris et al., 1997]. Hence, a tool for haplotype association tests is desirable. A single-marker test was proposed in the original APL method. An extension of the single-locus APL test to a multiple-locus haplotype test will also be introduced in Chapter 2.

The TDT has the advantage over a population-based association test in that it is robust to population stratification. Thus, keeping this feature in developing a family-based association test is important. The presence of population admixture can cause the allele frequencies to deviate from the Hardy-Weinberg Equilibrium (HWE). For the APL statistic, HWE for allele or haplotype frequencies may be required to reduce the number of parameters required to be estimated by APL. Hence, there was a need for the robustness of APL toward the deviations from HWE assumption. We used simulations to generate data sets with allele and haplotype frequencies that are deviated from HWE and evaluated the robustness of APL in Chapter 2.

An informative family for APL is the family in which the difference of the number of a certain allele between parents and affected siblings is not equal to 0. Families with both homozygous parents are not informative for APL. Although these "uninformative" families can still help in the estimation of allele frequency and IBD parameters, they do not contribute to the APL statistic. The APL statistic is asymptotically a normal distribution if the number of informative families is large [Martin et al., 2003], based on the central limit theorem [Feller, 1971]. When the number of informative families is small, this normality assumption

may not hold. The violation of the normal approximation under the null hypothesis may inflate the type I error rate for the APL test. Hence, it is very important to investigate the type I error rate for the APL test when the tested markers have rare alleles. In Chapter 2, a guideline of deciding when the APL test is a valid test for rare alleles or haplotypes will be provided and discussed.

1.4 X-linked analysis

The mammalian sex chromosomes (the X and Y chromosomes) derived from a pair of autosomes around 300 million years ago, and the Y chromosome then lost almost all genes shared with the X chromosome [Ross et al., 2005]. The X chromosome has the property that females have two copies of the chromosomes – one is inherited from the mother and the other from the father – while males only inherit one X chromosome from the mother. One copy of the X chromosomes in females undergoes X inactivation in early development and remains inactivated in somatic tissues [Gartler, 1983]. This process achieves dosage compensation, which equalizes gene expression between males and females [Lyon, 1961]. The inactivated X chromosome later enters a reactivation step in meiosis [Gartler, 1983]. The mechanism of choosing which X chromosome will undergo inactivation in females is still not fully understood [Vallender et al., 2005].

Sex-linked traits were first discovered in the fruit fly (*Drosophila*) in 1910 [Morgan, 1910]. Morgan observed that the mutated white-eyed flies did not appear randomly between sexes

but were sex-limited. X-linked diseases are diseases in which genes responsible for the diseases are located on the X chromosome. X-linked inheritance has several specific properties [Dobyns, 1996]. For example, male-to-male transmission of X-linked disease can never happen. Female siblings are always heterozygous for the X-linked disease when the father is affected but the mother is not. In general, X-linked disease genes affect a greater proportion of males than females due to the fact that the hemizygous males can express recessive traits [Dobyns, 2006].

To study X-linked diseases, linkage analyses have identified regions on the X-chromosome with high linkage peaks for several complex diseases. For Parkinson disease, Pankratz et al. (2003) identified the position 109 cM on the X chromosome that has a nonparametric linkage LOD score 3.1 after removing families with *parkin* mutations from their sample. Shao et al. (2002) found locus DXS6789 on the X-chromosome position 62.5 cM that shows linkage with autism (with maximum nonparametric LOD score 1.81). In order to further localize the disease susceptibility genes, association analyses should be applied in these linkage regions. However, association methods were designed primarily for autosomal markers and cannot be applied directly to the markers on the X chromosome. Fine-mapping in these regions has been slow in part due to the lack of appropriate statistical methods for family-based association analysis on the X chromosome.

The X-linked sibling TDT (XS-TDT) and reconstruction-combined transmission/disequilibrium test for X-chromosome markers (XRC-TDT), proposed in Horvath et al. [2000], are the first association methods specifically for X-chromosome markers. XRC-TDT was modified from RC-TDT proposed by Knapp [1999], which can reconstruct missing parental genotypes and combine the transmissions from parents to affected siblings and the difference of the number of a specific allele between affected and unaffected siblings. XS-TDT was modified from S-TDT proposed by Spielman and Ewens [1998], which compares the difference of the number of a specific allele between affected and unaffected siblings. XRC-TDT and XS-TDT were originally proposed for linkage tests, but theoretically they are also valid tests for association in the presence of linkage for families with a single proband. However, XRC-TDT and XS-TDT, which assume independent transmissions between affected siblings, are not valid tests for association when linkage is present and there are affected sib pairs in the data. This is the same problem faced by RC-TDT and S-TDT, as discussed in the previous sections.

More recently, the PDT, proposed for autosomal markers by Martin et al. [1997; 2000a], was extended to X-chromosomal PDT (XPDT) and X-chromosomal MC PDT (XMCPDT) [Ding et al., 2006]. As demonstrated in Martin et al. [2003] and Chung et al. [2006], PDT can have lower power than the APL test if parental genotypes are not available. Like PDT, XPDT can have low power in families with missing parents. XMCPDT can infer missing genotypes conditional on estimated or true allele frequency based on the Monte Carlo approach. In real

data analyses, true allele frequency is always unknown. As discussed in Ding et al. [2006], the allele frequency can be estimated from the parents (founders), but the statistic does not account for the variability in this estimate. For late-onset diseases the parental genotypes are often missing. It is not clear if a small portion of founders for the estimate of allele frequency can affect the validity of the XMCPDT test. Though the examples simulated by Ding et al. [2006] show no inflation of type I error, the validity of the test with varying amounts of missing parental data has not been thoroughly examined.

The APL test accounts for linkage when inferring missing parental genotypes based on affected sib pair data [Martin et al., 2003]. The APL can estimate allele frequency based on the siblings' data even when parental data are not available. The extended APL uses the bootstrap approach to account for the variability in the parameter estimation [Chung et al., 2006]. The APL also remains a valid test when multiple affected siblings are used in the linkage region. The same strategies can be applied to X-chromosome markers as well. In Chapter 3, we present X-APL, a novel test for association in the presence of linkage on the X chromosome.

Disease loci can have different effects on males and females. For example, *BRCA1* and *BRCA2* genes associated with breast cancer result in different prevalences for females and males [Cornelisse et al., 1996; Ormiston, 1996; Frank et al., 2002; Fentiman et al., 2006], and the *Cox-2* gene is associated with prostate cancer only for males [Panguluri et al., 2004;

Shahedi et al., 2006]. Knowing that the effects of genes vary according to sex helps in follow-up studies. For example, resequencing may be performed only in males or females if there is a sex-specific effect. Therefore, in addition to the test using all data from both sexes, a strategy for testing sex-specific effects will also be introduced in Chapter 3.

1.5 Correlation between linkage and association analyses

Linkage analysis can be powerful for finding rare variants associated with the disease. However, the regions identified by linkage tests are often very large, often as large as 40cM. Association analysis provides a higher resolution of finding disease genes than linkage analysis, since dense markers can be genotyped to test LD between the markers and disease loci. To take advantage of their complementary properties, a common strategy for identifying complex disease genes is to conduct linkage analyses first and then follow significant results with family-based tests for association at a denser panel of markers in an attempt to further localize the disease gene [Cardon et al., 2001]. Using this strategy, many studies have found significant association results from regions that showed high linkage peaks. For example, Martin et al. [2000b] identified several SNPs significantly associated with late-onset Alzheimer's disease (AD) in the APOE region, which was a well-established susceptibility gene for AD by linkage analyses [Pericak-Vance et al., 1991]. Van der Walt et al. [2004] found three SNPs located in the fibroblast growth factor 20 (FGF20) gene significantly associated with Parkinson disease (PD) in the linkage region 8p identified in Scott et al. [2001].

For family-based association analysis design, the same data are often tested for linkage and association analyses. For example, in the study of linkage and association for schizophrenia in Schwab et al. [2002], microsatellite markers in the region on chromosome 6q were genotyped from 69 families with at least two affected siblings per family. Nonparametric multipoint linkage analysis and TDT for association were both applied on the same microsatellite markers. In the study of linkage and association for alcoholism in McQueen et al. [2005], a total of 11555 SNPs, released by the Genetic Analysis Workshop 14 (GAW 14), were genotyped from 143 families. Multipoint linkage analysis and quantitative trait association analysis were both performed on the same SNP markers. As discussed in McQueen et al. [2005], this strategy can provide more information than just performing linkage or association analysis alone.

Recently, advanced technology and reduced genotyping costs have made genome-wide association (GWA) analyses of hundreds of thousands of single nucleotide polymorphism (SNP) markers possible. With the completion of PHASE I of the HAPMAP project [International HapMap Consortium, 2003; Altshuler et al., 2005], about 6 million new SNPs were genotyped to promote the discovery of high-quality SNPs and to define LD structures in the human genome as a framework for whole-genome association analyses. Whole-genome association analyses can be performed without information from linkage analyses. However, a large sample size is required to compensate for the power lost from multiple comparison

corrections required for the huge number of hypothesis tests. This multiple-testing issue is a challenging problem for whole-genome association analysis [Carlson et al., 2004]. Recently, a novel approach for GWA analyses uses linkage test results to weight the p-values of association tests, and this approach shows more power than association tests alone if the linkage tests are informative [Roeder et al., 2006]. If the linkage tests are not informative, the loss of power for association is small. Hence, even in the era of genome-wide association analysis, linkage analysis can still play an important role. Furthermore, we must keep in mind that due to the limitation of association analyses for finding rare variants associated with the diseases, linkage analyses will still remain essential [Wang et al., 2005].

However, it is not intuitively clear under what conditions association and linkage tests performed in the same data set may be correlated. For association tests such as APL, which includes IBD parameters for estimating parental mating-type, it is also not clear whether the IBD-based test statistic may be correlated with the linkage statistic. If there is correlation between linkage and family-based association test statistics, the results obtained from the tests may not be appropriate, particularly when one test is performed based on the results from the other test. For example, if there is correlation between linkage and family-based association test statistics when there is linkage and no association, then association tests performed based on significant results from linkage tests may tend to be liberal.

To help interpret the results from linkage and association tests conducted on the same data, it is desirable to know when the tests are correlated. In Chapter 4, we will introduce our theoretical and simulation studies to estimate the correlation between the linkage and association statistics. General pedigree structures such as extended pedigrees and incomplete pedigrees (families with missing parents) were used in the simulations to estimate the correlation between the linkage and association statistics. Commonly used methods for linkage and association implemented in software packages were used. For linkage statistics, the Kong and Cox's LOD score [Kong and Cox, 1997], extended from the allele-sharing method in Kruglyak et al. [1996] and implemented in the software package MERLIN [Abecasis et al., 2002], were used. For association statistics, the PDT software package [Martin et al., 2000a], which can handle extended pedigrees, and APL [Martin et al., 2003], which is implemented in the APL software package and can handle missing parents in nuclear families [Chung et al., 2006], were used.

1.6 Conclusion

Linkage and association methods have played very important roles in human complex disease gene mapping. They can be applied complementarily to obtain the maximum information from the data. The APL test proposed in Martin et al. [2003] provides a powerful family-based test for association in the linkage region. The extension of the APL method and examination of its robustness will be described in Chapter 2. Due to the lack of family-based association methods for the X-chromosome markers, we have developed the X-APL, which

was extended from APL and will be introduced in Chapter 3. Since linkage and association approaches are often applied on the same data, it is important to know when the two statistics may be correlated. An examination and interpretation of the correlation between the linkage and association statistics will be described in Chapter 4. Finally, some future work such as extension of the discussed methods will be discussed in Chapter 5.

Chapter 2

The APL Test: Extension to General Nuclear Families and Haplotypes and Examination of Its Robustness

Ren-Hua Chung, Elizabeth R. Hauser, Eden R. Martin

(2006) Human Heredity 61:189-199

2.1 Abstract

Objective: The Association in the Presence of Linkage test (APL) is a powerful statistical method that allows for missing parental genotypes in nuclear families. However, in its original form, the statistic does not easily extend to mixed nuclear family structures nor to multiple-marker haplotypes. Furthermore, the robustness of APL in practice has not been examined. Here we present a generalization of the APL model and an examination of its robustness under a variety of non-standard scenarios. Methods: The generalization is made possible by incorporating a bootstrap variance estimator instead of the original robust variance estimator. This allows for use of more than two affected siblings. Haplotype analysis was accomplished by combining estimation of haplotype phase into the EM algorithm. Computer simulation was used to examine robustness of the APL to departures from test assumptions. Results: The extended APL tests both single-marker and multiple-marker haplotypes and shows more power than other association methods. Simulation results showed that the single-marker APL test is robust to the departure from HWE. For the haplotype test, violation of the HWE assumption can inflate type I error. We also evaluated general guidelines for the validity of APL with rare alleles and rare haplotypes. Software for the APL test is available from http://www.chg.duke.edu/research/apl.html.

2.2 Introduction

Family-based association tests provide powerful tools for finding genes associated with complex diseases. The classic transmission/disequilibrium test (TDT) [Spielman et al., 1993],

for example, can be used to test for association in the presence of linkage (i.e. linkage disequilibrium) in family triads (one affected offspring and both parents). However, the TDT is not a valid test of association when more than one affected sibling is used and there is linkage between disease loci and markers. Modifications of the TDT have been proposed to take linkage into consideration for families with multiple affected siblings [Martin et al., 1997, 2000a; Abecasis et al., 2000; Rabinowitz and Laird, 2000].

In some cases, such as the late-onset diseases, parental genotypes are often missing. One approach to deal with this problem is to compare the allele frequencies between affected siblings and unaffected siblings only [Martin et al., 2000a; Horvath and Laird, 1998; Monks et al., 1998]. These methods all properly allow for correlation due to linkage in their statistics. Another approach is to use genotypes of the siblings to infer the genotypes of their parents [Clayton, 1999; Knapp, 1999; Weinberg, 1999; Martin et al., 2003]. A widely used program TRANSMIT was implemented based on the method in Clayton [1999]. As indicated in Martin et al. [2003], when linkage is present between marker and disease loci, TRANSMIT, which assumes independent transmission between siblings from their parents, has an inflated type I error rate when parental genotypes are missing. The Association in the Presence of Linkage (APL) method, proposed in Martin et al. [2003], correctly infers missing parental genotypes in regions of linkage by simultaneously estimating identity-by-descent (IBD) parameters. However, the original APL is not flexible for mixed nuclear family structures (including the mixture of singletons and multiplexes with arbitrary number of unaffected sibs)

due to the fact that the different parameters in the variance estimator should be considered for each type of family structure separately. Furthermore, only the single-marker test was proposed in the original APL method and up to two affected siblings were considered. The robustness of APL toward the rare alleles or rare haplotypes and the deviations from the Hardy-Weinberg Equilibrium (HWE) assumption were not examined either.

In order to generalize APL to be flexible in real data applications, we modified and extended the APL method. A bootstrap variance estimator, instead of the original robust variance estimator, is used. The bootstrap variance estimator has the advantage that mixed family structures can be easily incorporated when estimating the variance. Two affected siblings in families were considered for estimating IBD parameters for inferring missing parental genotypes in Martin et al. [2003]. We extended APL to utilize three affected siblings by considering IBD between every two affected siblings in the three siblings and we compared its power to APL using only two affected siblings.

APL was also extended from single-marker analysis to multiple-marker haplotype analysis. Haplotypes can be more informative than genotypes if multiple markers contribute to the trait [Martin et al., 2000a]. Hence, developing a powerful haplotype analysis tool for association is desirable. Our modified APL model allows for unknown phase, and the missing parental haplotypes are correctly inferred by taking IBD parameters into consideration. A global test measuring the overall effect of all possible haplotypes is also calculated. We then compared

the power of the modified APL to other association analysis methods. For the single-marker test, we compared the power of APL to two alternative methods in nuclear families: the pedigree disequilibrium test (PDT) [Martin et al., 2000a] and the family-based association test (FBAT) [Lake et al., 2000]. For the haplotype test, we compared the power of APL to PDTPHASE [Dudbridge, 2003] and the haplotype FBAT (HBAT) [Horvath et al., 2000].

To examine the robustness of the APL statistic, we use computer simulation to investigate two issues that frequently occur in real data analyses: rare alleles or rare haplotypes and the deviations from the HWE assumption. Since APL assumes HWE for the haplotype test, we simulated several data sets that have deviations from HWE and observed the effect on APL statistic. Finally, a powerful software package is provided based on the implementation of the generalized APL model, which will be very useful for family-based disease association studies.

2.3 Methods

2.3.1 Review of the APL model

We followed the statistical development proposed in Martin et al. [Martin et al., 2003] and show the modifications and extensions that have been made. The APL uses nuclear families with at least one affected offspring. The APL is based on the statistic T which is the difference between the number of copies of a specific allele in affected offspring and the expected number under the null hypothesis of no association conditional on parental

genotypes. T_s is the sum of T's over all families in the sample. If the parental genotypes are missing, probabilities of consistent parental mating types are used to estimate the expected copies of the allele in parents. When affected siblings are used, mating types are correctly inferred by taking linkage into consideration. Linkage between disease loci and markers was accommodated by including IBD parameters for affected siblings when estimating parental mating-type probabilities. For an affected sib-pair family, the probability of parental mating-type G_p was estimated based on the siblings' genotypes G and their affection status A in Martin et al. [2003] equation (2):

$$P(G_p \mid G, A) = \frac{\mu_{G_p} \sum_{k=0}^{2} z_k P(G \mid G_p, IBD = k)}{P(G \mid A)} \qquad (1)$$

where μ_{G_p} is the unconditional mating-type probability and z_k is the IBD parameter which denotes the probability that the affected siblings share k alleles IBD. The parameters μ_{G_p} and z_k (k = 0, 1, 2) can be estimated by EM algorithm. The probability $P(G \mid G_p, IBD = k)$ in the numerator is simply a function of Mendelian segregation probabilities. The probability $P(G \mid A)$ in the denominator can be calculated by summing all terms in the numerator over all possible parents, G_p, for a given G.

If there is only one affected sibling in a family, the probabilities $P(G \mid G_p, IBD = k)$ reduce to $P(G \mid G_p)$, which are simply Mendelian transmission probabilities. As shown in Martin et al. [2003], unaffected siblings can be used to improve the estimation of the parental mating-type probabilities. When the disease penetrances are low, which is common for complex diseases,

transmissions to unaffected siblings are independent conditional on parental genotypes and do not depend on disease status. Hence, if unaffected siblings are present, the probabilities $P(G \mid G_p, IBD = k)$ are multiplied by the Mendelian transmission probabilities for the unaffected siblings for a given parental genotypes. Partial parental genotypes can also help APL estimate the parental mating-type probabilities. If one parental genotype, P_1, is present and the other parent, P_2, is missing, equation (1) can be modified by conditioning on P_1 as well: $P(P_2 \mid P_1, G, A)$. The calculation procedures are similar to equation (1) with a restriction that P_1 is known.

Under the null hypothesis that there is no association (with or without linkage), the expected value of T_s is 0. T_s can be standardized to have an asymptotic normal distribution with mean 0 and variance 1. The statistic, called the APL statistic, takes the following form Martin et al. [2003]:

$$T_s \Big/ \sqrt{\hat{V}ar(T_s)} \tag{2}$$

where $\hat{V}ar(T_s)$ is an estimate of the variance of T_s.

The hypothesis test based on the APL statistic will be referred as the APL test. Specifically the null hypothesis is that there is no linkage or no association between marker and disease loci.

2.3.2 Variance estimation

In Martin et al. [2003], a robust variance estimator which takes into account the variance associated with estimation of IBD parameters was used to estimate the variance of T_s. However, the estimator is practically difficult to implement when various family structures exist in a data set. To offer more flexibility, we implement a different approach for variance estimation based on the bootstrap method. Assume there are n families in the sample. We perform k bootstrap resamplings. Each family is treated as an independent unit for resampling. For each bootstrap sample, a new set of n families are resampled with replacement from the original n families. We measure the T_s statistic from the ith set of families and we can obtain T_i, where $i = 1,2,...k$. The estimation of the variance of T_s is the sample variance of the k T_i's [Efron and Tibshirani, 1993]:

$$\hat{V}ar(T_s) = \sum_{i=1}^{k} \left(T_i - \overline{T}\right)^2 \Big/ (k-1) \qquad (3)$$

where $$\overline{T} = \sum_{i=1}^{k} T_i / k$$

When k is large, the sample variance of the k T_i's is close to the variance of T_s.

Even when there are mixed nuclear family structures in the data, the bootstrap variance estimator still works properly. Although the numbers of different types of nuclear families may change during each bootstrap resampling, the bootstrap procedure simulates the

sampling scheme based on current data. Hence, the bootstrap variance estimator estimates the variance caused by sampling errors and parameter estimation in the APL model. We verified the validity of the bootstrap variance estimator by simulations.

2.3.3 Extension to three affected siblings

In Martin et al. [2003], no more than two affected siblings were considered when inferring the missing parental genotypes. This requires three IBD parameters between the two affected siblings (z_0: IBD=0, z_1: IBD=1, z_2: IBD=2). Here we extend the algorithm to three affected siblings by considering IBD status between every pair of the three affected siblings. We follow the IBD configurations for IBD sharing among three siblings in Whittemore et al. [1998] table 3. Four IBD parameters, k_0, k_1, k_2 and k_3, denote IBD sharing of (2,1,1), (2,2,2), (1,0,1), and (2,0,0) alleles among three sibling pairs, respectively. For example, for three siblings A, B and C, (2,1,1) means that A and B share 2 alleles IBD, B and C share 1 allele IBD and A and C also share 1 allele IBD. Hence, when three affected siblings are present in a pedigree, the probability of a missing mating-type G_p in Martin et al. [2003] equation (2) is modified as:

$$P(G_p \mid G, A) = \frac{\mu_{G_P} \sum_{i=0}^{3} k_i P(G \mid G_p, IBD\,status = k_i)}{P(G \mid A)} \qquad (4)$$

where G is the set of genotypes of the three affected siblings, A is the affection status, and μ_{G_P} is the mating-type probability for parents with genotypes G_p.

The IBD parameters k_0, k_1, k_2 and k_3 are estimated by the EM algorithm jointly with μ_{G_P} and z_0, z_1 and z_2 from sibpair families. The IBD parameters between two individuals can be obtained from the IBD parameters between three individuals by simply considering the IBD of the first pair of the three individuals. The relationships between z_0, z_1, z_2 and k_0, k_1, k_2, k_3 are indicated in equation (5). These relationships are included as one additional step in the "M-Step" of EM algorithm.

$$z_0 = (k_2 / 3) + 2 \times (k_3 / 3)$$
$$z_1 = 2 \times (k_0 / 3) + 2 \times (k_2 / 3)$$
$$z_2 = (k_0 / 3) + k_1 + (k_3 / 3) \qquad (5)$$

2.3.4 Extension to multiple-marker haplotype analysis

We extended the APL test to a multiple-marker haplotype test suggested by Martin et al. [2003]. For simplicity, no recombination is assumed to occur between the markers within the families. The number of parental mating types increases exponentially with the number of haplotypes that are used for testing. Hence, the APL program assumes the Hardy-Weinberg Equilibrium (HWE) for haplotype frequencies in order to reduce the number of parameters that will be estimated. Under HWE, only haplotype frequencies and IBD parameters need to be estimated to obtain the mating-type probabilities.

The strategy of haplotype testing is analogous to a multiple-allele analysis. Probabilities for consistent haplotype phases within each family are estimated through the EM algorithm. The

probabilities of the haplotype phases for each family are calculated by taking IBD parameters into consideration. Therefore, the modified APL test correctly infers the phase probabilities under the null even when linkage is present. $\boldsymbol{T_j}$ is calculated as an expected value of the statistic $\boldsymbol{T}$ from all possible phases for a family and $\boldsymbol{T_s}$ is the sum of all $\boldsymbol{T_j}$'s over all families. Here $\boldsymbol{T}$ is a vector where each element in $\boldsymbol{T}$ corresponds to a specific haplotype.

We calculate the global test statistic X to evaluate the haplotype effect. A global test of all haplotypes can be used to capture the multiple haplotype effect [Horvath et al., 2004]. The statistic X is a quadratic form that asymptotically follows a chi-squared distribution under the null hypothesis that none of the haplotypes is in linkage disequilibrium with the disease allele:

$$X = \boldsymbol{T_s}' \Sigma^{-1} \boldsymbol{T_s} \sim \chi^2_{p-1} \qquad (6)$$

where Σ is the variance-covariance matrix of $\boldsymbol{T_s}$ and p is the number of haplotypes tested.

The variance-covariance matrix Σ of $\boldsymbol{T_s}$ can be estimated easily from the bootstrap samples. Due to the property that the elements in $\boldsymbol{T_s}$ sum to 0, the variance-covariance matrix of $\boldsymbol{T_s}$ is not full rank and is not invertible. We substitute the generalized inverse of Σ in the quadratic form. The statistic still has a chi-squared distribution, where the degree of freedom is the rank of Σ [Rao, 1971].

Though it has been suggested that separate tests can be conducted for individual haplotypes, that is to test whether a specific haplotype is in linkage disequilibrium with a disease allele,

this procedure may not be valid. In the APL test, IBD and mating-type parameters are estimated under the global null hypothesis. Estimation under a haplotype-specific null hypothesis is not straightforward in this context. Consequently, we implement only the global test for haplotype analysis.

2.3.5 Rare alleles and rare haplotypes

The APL statistic (equation (2)) is normally distributed when the number of transmissions of alleles or haplotypes is large. However, when alleles or haplotypes are rare, this assumption of normality may not be appropriate. As a guideline for whether the approximation is valid or not, the online manual of TRANSMIT suggests considering the size of the variance estimate. The size 2.5 provides a general guideline of deciding whether the statistic in TRANSMIT is valid or not. This rule can be applied to the APL statistic as well since the square of APL statistic has an asymptotically chi-squared distribution that is analogous to the TRANSMIT statistic. As suggested in the online manual of TRANSMIT, raising the value of the variance estimate to 5 or higher would provide more confidence in validity of the statistic. The same guideline can also be applied to rare haplotypes. We can compute the variance of the T_s statistic of a specific haplotype to decide whether it should be included for global haplotype statistic calculation or not. We evaluate these guidelines in simulations. In order to investigate the guidelines for a broad range of disease models, we simulated data sets under dominant, recessive, additive and multiplicative models. We also generated data sets with a

mixture of singleton and multiplex families to examine if the guidelines are appropriate for mixed nuclear family structures.

2.3.6 HWE assumption

The default version of the single-marker APL test assumes HWE. A version of single-marker APL test without the HWE assumption is also implemented. The HWE assumption for haplotype frequencies is used solely in the multiple-marker haplotype test in order to reduce parameters that are estimated by the EM algorithm. In the real data, genotyping errors or population admixtures may cause the deviation from HWE. To examine the effect of the deviation from HWE on the APL test, we generated a data set by combing two simulated data sets from random-mating populations with different allele frequencies into one data set. This population admixture generates deviation from HWE. Several data sets with different degrees of deviation were generated. Data with two markers were generated to evaluate the effect for haplotype tests. We measure the degree of HWE deviation using the HWE goodness-of-fit test statistic, which has a chi-squared distribution with 1 degree of freedom. We evaluate the robustness of the APL test when the HWE assumption is violated in simulations.

2.3.7 Computer simulations

Computer simulations were used to evaluate the type I error and power for the modified APL statistic. We used the SIMLA computer program [Bass et al., 2004] to generate replicate

samples of families with different disease models and three types of family structures AA, AAU and AAAU. An AA family has one affected sibling pair, an AAU family has one affected sibling pair and one unaffected sibling and an AAAU family has three affected siblings and one unaffected sibling. Families with both parental genotypes missing and only one parental genotype missing were simulated. We used the same parameter values for SIMLA as Martin et al. [2003] table 1 to generate different disease models including four recessive models (RecA, RecB, RecC, RecD) and four multiplicative models (MultA, MultB, MultC, MultD) and marker loci, where the recurrence-risk ratios for siblings range from 1.26 to 1.02. Genetic markers were simulated under the assumption of complete linkage to the disease locus.

For type I error simulations, there was no association between the disease and marker alleles. For single-marker tests, five different samples composed of different types of family structures were used: (1) 300 AAU families with all parental genotypes missing, (2) 300 AAAU families with all parental genotypes missing (3) 250 AA plus 250 AAU families with all parental genotypes missing, and (4) 200 AAAU families plus 100 AAU families with all parental genotypes missing (5) 250 AAU families with all parental genotypes missing plus 250 AAU families with one of the parents' genotypes missing. For multiple-marker haplotype tests for type I error, three markers were simulated. There are eight possible haplotypes (111), (112), (121), (122), (211), (212), (221), (222) simulated for the three

markers with frequencies 0.512, 0.128, 0.128, 0.032, 0.128, 0.032, 0.032, 0.008, respectively. None of them is associated with the disease allele.

For single-marker tests, power simulations assumed that the marker and disease alleles were in perfect association. Therefore, the marker locus is in fact equivalent to the disease locus. For multiple-marker haplotype tests, two markers were simulated having four possible haplotypes (11), (12), (21), (22) with frequencies 0.3, 0.3, 0.2, 0.2, respectively. Table 2.1 shows the association configurations between these haplotypes and the disease alleles used in SIMLA.

2.4 Results

2.4.1 Type I errors and power

In Martin et al. [2003], the APL statistic using the robust variance estimator was shown to have a correct type I error rate. Here we verify proper implementation of the modified APL statistic with the novel bootstrap variance estimator by testing the type I error for both single- and multiple-marker haplotype analysis. Tables 2.2 and 2.3 show the results of the type I error for single-marker and multiple-marker haplotype analysis with the APL program, respectively. Table 2.2 shows that under different disease models and different types of family structures, the mean of the APL statistic is close to 0, the variance is close to 1, and the type I error is close to the nominal level of 0.05. Table 2.3 shows that the type I error rate for the global test of all haplotypes is also close to 0.05.

We next compared the power of the modified APL test and other family based tests of association. PDT [Martin et al., 2000a], FBAT/HBAT [Lake et al., 2000; Horvath et al., 2004] and PDTPHASE [Dudbridge, 2003] were selected for comparison since they remain valid tests for association when linkage is present. TRANSMIT was not included in the comparison since it has an inflated type I error when linkage is present [Martin et al., 2003]. Figure 1 shows the results for single-marker analysis. Compared with figure 1 in Martin et al. [2003], APL using bootstrap variance estimator obtains more power than the APL test using the original variance estimator. For example, APL using the bootstrap estimator has estimated power 0.44 at the 0.05 significance level for the RecD model for 250AAU families and the power estimate for APL using the original variance estimator under the same model is 0.32.

It is not surprising to see that when families with three affected siblings are present in the data, the extension of APL which takes all three affected siblings into account obtains more power than APL using only two affected siblings (Figure 1). For example, for the RecD model for 250 AAAU families, APL using three affected siblings has power 0.725 at the 0.05 significance level and APL using only two affected siblings has power 0.553 under the same model.

We also see that, for different types of family structures, the modified APL has more power than the other two methods under all genetic models considered. APL typically has an outstanding power for the RecD and MultD models. In the combined data sets with both AAU and AA families, the PDT and FBAT do not use the families without unaffected siblings. The APL test will use information from the entire collection of families. The results in figure 1 show that these AA families add a little power to the APL test for all models while the PDT and FBAT maintain the same power.

Figure 2 shows the comparison of the power of global haplotype analysis between APL, HBAT, and PDTPHASE. The pattern is similar to the single-marker results. APL continues to have more power than HBAT and PDTPHASE in most of the examples tested, and in some cases can have considerably more power.

2.4.2 Rare alleles and haplotypes

Rare alleles or rare haplotypes may cause an inflated or conservative type I error rate for the APL test. Table 2.4 shows that, for single-marker APL test, the type I error tends to be inflated for both rare alleles with expected frequencies 0.1 and 0.01 for the smaller samples for different disease models and family structures. The type I error is close to 0.08 if the rare allele frequency is approximately 0.01 in 100 AAU families. It decreases to close to 0.05 if 600 AAU families were used. Hence, collecting more families helps the accuracy of APL statistic when rare alleles are present.

Table 2.4 also shows the adjusted type I error rate after all replicates that have estimated variances of T_s less than 2.5 and 5 were eliminated. For allele frequency of 0.01, the adjusted type I error rate becomes smaller and is often conservative, relative to the inflated rate before adjustment. Table 2.4 shows the guideline using variance 2.5 may not work well in some cases. For example, for a sample that has 100 AAU families and a rare allele with frequency 0.1, the adjusted type I error rates using a cutoff of 2.5 are inflated (0.061 and 0.058) for the RecA and MultA models, respectively. The guideline of variance 5 generally avoids inflated type I error rate shown in Table 2.4. The results of simulated data sets for disease models RecA and MultA generally show the same pattern in Table 2.4. Moreover, we can also observe the same pattern in the mixed nuclear family structures of singleton and multiplex families in Table 2.4. Note that we did not show the adjusted type I error rate using variance of 5 for the mixed family structures of 50 singleton and 50 multiplex families since only a few data sets have variance greater than 5. Hence, it is not suggested that such small sample should be tested with APL for a rare allele frequency 0.01 in real data analysis. The same pattern was observed in dominant and additive models as well using the same family structures in Table 2.4 (results not shown).

Table 2.5 shows estimates of the type I error rate of the global haplotype test when several rare haplotypes exist. We can see that the type I error rate of the global test tends to be conservative. The adjusted type I error rates that are calculated by excluding rare haplotypes

with corresponding variances of T_s less than 5 are close to the 0.05 level. We also observed the same pattern in dominant and additive models using the same family structures. The simulation results show that requiring a variance estimate > 5 serves as a good guideline of deciding whether the APL test is valid or not when rare alleles or rare haplotypes are present.

2.4.3 HWE effect

Table 2.6 shows the effect of deviations from HWE for APL single-marker analysis. It shows that even with deviations from HWE (Goodness-of-fit statistics from 0.036 to 5.055), the APL test remains valid so that the type I error under different models is close to the nominal 0.05 level. The version of APL single-marker test without the HWE assumption was also tested and found to have correct type I error rate by simulations (results not shown). The APL single-marker test with the HWE assumption is preferred since it has more power than the version without the HWE assumption, according to our simulation results (results not shown).

Table 2.6 also shows the effect of deviations from HWE for multiple-marker haplotype analysis. It is no surprise to see that the haplotype APL test is more sensitive to the HWE deviation than the single-marker test since more parameters need to be estimated. We conclude that APL statistic is not very sensitive to the deviation of HWE of allele frequencies in single-marker analysis. However, in haplotype analysis, haplotype frequencies in HWE

would be crucial for APL. Therefore, it may not be prudent to conduct haplotype analysis with the APL if there is evidence of deviation from HWE.

2.4.4 Performance

The APL program is written in C++ and available for Linux, Sun and Windows platforms. Since APL needs to perform a certain amount of bootstrap iterations, it is not as efficient as Transmit without the bootstrap option, which has the same time complexity for calculating T_s as APL. Generally APL can finish a single-marker analysis within one minute for 300 families with all parental genotypes missing. Haplotype analysis causes higher density of calculations and APL usually takes 30 minutes for a data set that has 300 families and 3 markers with all parental genotypes missing running on a Sun workstation equipped with a 1.2GHz CPU.

2.5 Discussion

In this paper, we present the generalization and examinations of the APL test that will be useful for the association analysis in family data. Both single-marker and multiple-marker haplotype tests are provided. We replaced the robust variance estimator proposed in Martin et al. [2003], which is difficult to implement in practice when different types of family structures exist, with the bootstrap variance estimator. The bootstrap variance estimator differs from the robust variance estimator in that it estimates the variance of T_s under the observed model while the robust variance estimator estimates the variance under the null

model. Our simulation results showed that by using the bootstrap variance estimator, the modified APL obtains more power than the APL statistic using the robust variance estimator. We demonstrated the proper implementation of the modified APL by testing its type I error rate from different family structures and disease models.

Our simulation results showed that under different family structures, the bootstrap variance estimator correctly estimated the variance of T_s. Even when there was a mixture of different nuclear family structures, the variances were correctly estimated. In addition to the results for mixtures of multiplex families presented in Table 2.2, we also simulated a mixture of singleton and multiplex families and the type I error rate was as expected (results not shown). Hence, the bootstrap variance estimator is robust to mixed nuclear family structures.

APL was extended to consider three affected siblings if they are present in the data. Simulation results showed the extended APL obtains more power than APL using only two affected siblings. We also compared the power of the APL test, the PDT, FBAT/HBAT and PDTPHASE. The APL test consistently had the highest power for both single-marker and multiple-marker haplotype analysis for nuclear family data. Hence, APL may be preferred in analyzing nuclear family data sets. Although APL may have greater power in nuclear families, the test, unlike PDT, is not valid in extended pedigrees, nor does it offer the flexibility to handle quantitative traits as the FBAT does. So each of these tests offers advantages in different situations.

We examined the robustness of APL for the deviations of HWE and rare alleles or rare haplotypes. HWE is assumed in the haplotype analysis to reduce the number of parameters. Though we found deviations from HWE had little impact on single-marker tests, they do affect validity of haplotype analyses. We also considered the situation when rare alleles or haplotypes are present in the data. They could affect the validity of APL statistic leading to an inflated or conservative type I error rate. When there is extensive LD between markers, rare haplotypes are more likely to exist. In this case, the global haplotype statistic may not be valid. As the TRANSMIT online manual suggests, we confirmed with simulations that variance greater or equal to 5 provides a general guideline of deciding whether to accept APL statistic or not. For global haplotype analysis, haplotypes with variance less than 5 are ignored. Another possible approach to reduce the effect of extensive LD on global haplotype analysis is to collapse rare haplotypes into one haplotype. However, the strategy of choosing which haplotypes to collapse is not trivial. The simulations presented here suggest that collapsing rare haplotypes until they exceed a variance of 5 may be a good strategy.

Since APL considers only nuclear families, the next challenging problem is to extend APL to larger families. In order to do so, the inheritance vectors should be considered when estimating parental mating-type probabilities and IBD parameters since these parameters are correlated within an extended pedigree. An alternative test, the PDT, can handle extended

pedigrees and can be used as a complementary tool for APL for family-based association studies.

In conclusion, we have included several useful extensions to the APL algorithm and provided a comprehensive software package for single marker and haplotype analysis. The APL software provides a useful approach for fine-mapping complex disease genes in sibships or nuclear families that can dramatically outperform existing methods. APL can be publicly accessed at http://www.chg.duke.edu/research/apl.html.

2.6 Acknowledgements

The work was supported in part by generous funding from the National Institute of Health, NS51355 and MH59528. We are grateful to several members of the Duke Center for Human Genetics for testing and providing useful feedback on the APL program. We are also thankful for the generous support of the Beowulf Cluster from PAMS at NC State University, which made our intensive simulations possible.

2.7 Tables

Table 2.1 Association configurations in SIMLA for the power simulations for haplotype tests.

Haplotypes	E(freq | d)[a]	E(freq | D)[b]
11	0.30	1.00
12	0.30	0.00
21	0.20	0.00
22	0.20	0.00

[a] The expected haplotype frequencies on chromosomes with wild-type allele d.

[b] The expected haplotype frequencies on chromosomes with disease allele D.

Table 2.2 Mean, Variance and Type I error of the single-marker APL test across 5000 replicate data sets.

Data	APL test		
N and model of inheritance	Mean	Variance	Type I error[a]
N = 300 AAU:			
RecA	-.011	.994	.048
RecB	.051	.988	.045
MultA	-.013	1.018	.051
MultB	-.014	1.004	.049
N = 300 AAAU:			
RecA	-.047	1.029	.052
RecB	-.018	.991	.049
MultA	-.023	.989	.047
MultB	-.013	1.028	.054
N = 250 AA + 250 AAU			
RecA	-.028	.995	.050
RecB	-.051	1.101	.051
MultA	-.056	.994	.051
MultB	-.035	1.010	.048
N = 200 AAAU + 100 AAU			
RecA	.004	.978	.046
RecB	-.017	.989	.050
MultA	-.014	1.025	.052
MultB	-.058	.972	.047
N = 250 AAU* + 250 AAU			
RecA	.007	.963	.049
RecB	-.021	1.009	.049
MultA	-.013	1.000	.053
MultB	-.023	1.047	.054

[a] Proportion of data sets with p-value ≤ 0.05.

* Only one parental genotype in every family is missing. Other families without * were generated without parental genotypes.

Table 2.3 Type I error of the multiple-marker haplotype APL global test across 5000 replicate data sets.

Data	APL test
N and haplotypes	Type I error[a]
N = 300 AAU:	
RecA	.046
MultA	.050
N = 250 AA + 250 AAU:	
RecA	.041
MultA	.047
N = 250 AAU* + 250 AAU:	
RecA	.052
MultA	.053

[a] Proportion of data sets with p-value ≤ 0.05.

* Only one parental genotype in every family is missing. Other families without * were generated without parental genotypes.

Table 2.4 Type I error of the single-marker APL test before and after the adjustment.

Data		APL test		
N and model of inheritance	E(freq)[a]	Type I error	Type I error[b]	Type I error[c]
N = 100 AAU				
RecA	0.10	.065	.061 (99%[d])	.052 (97%)
RecA[e]	0.01	.069	.005 (18%)	.041 (2%)
MultA	0.10	.060	.058 (99%)	.047 (96%)
MultA[e]	0.01	.076	.006 (18%)	.052 (2%)
N = 300 AAU				
RecA	0.10	.053	.053 (100%)	.053 (100%)
RecA	0.01	.076	.033 (77%)	.020 (34%)
MultA	0.10	.051	.051 (100%)	.051 (100%)
MultA	0.01	.078	.042 (76%)	.023 (34%)
N = 600 AAU				
RecA	0.10	.050	.050 (100%)	.050 (100%)
RecA	0.01	.055	.054 (99%)	.039 (33%)
MultA	0.10	.053	.053 (100%)	.053 (100%)
MultA	0.01	.063	.069 (99%)	.047 (88%)
N = 50 AU + 50 AAU				
RecA	0.10	.066	.064 (99%)	.040 (88%)
RecA[e]	0.01	.071	.010 (9%)	N/A
MultA	0.10	.048	.048 (98%)	.043 (83%)
MultA[e]	0.01	.078	.018 (6%)	N/A
N = 150 AU + 150 AAU				
RecA	0.10	.051	.051 (100%)	.051 (100%)
RecA[e]	0.01	.061	.024 (60%)	.032 (15%)
MultA	0.10	.049	.049 (100%)	.049 (100%)
MultA[e]	0.01	.058	.022 (58%)	.027 (15%)

Table 2.4 (*Continued*)
Three types of family structures, 100 AAU, 300 AAU and 600 AAU, were simulated with two types of allele frequencies, 0.01 and 0.1, for the RecA and MultA models across 5000 replicate data sets. Two types of mixed family structures, 50AU (one affected child and one unaffected child) plus 50AAU and 150 AU plus 150 AAU, were also simulated. All parental genotypes are assumed to be missing.
[a] E(freq) is the expected frequency of the rare allele.
[b] The adjusted type I error rate calculated by the proportion of data sets with p-value ≤ 0.05 where alleles with variance < 2.5 were excluded.
[c] The adjusted type I error rate calculated by the proportion of data sets with p-value ≤ 0.05 where alleles with variance < 5 were excluded.
[d] The percentage shows the percentage of the replicates remaining for type I error calculation.
[e] 100,000 replicate data sets were generated for 100 families and 20000 replicate data sets were generated for 300 families to better approximate the type I error.

Table 2.5 Type I error of the multiple-marker haplotype APL test before and after the adjustment.

Data	APL test	
Haplotypes	Set1	Set2
	E(freq)[c]	E(freq)
111	0.84645	0.612
211	0.09405	0.108
121	0.04455	0.153
221	0.00855	0.027
112	0.00495	0.068
122	0.00045	0.017
212	0.00095	0.012
222	0.00005	0.003
Models	Type I error	
N = 300 AAU RecA:		
Global test	.027	.035
Global test[a]	.044	.038
Global test[b]	.048	.049
N = 300 AAU MultA:		
Global test	.021	.036
Global test[a]	.037	.039
Global test[b]	.047	.053

Two sets, set1 and set2, were simulated across 5000 replicate data sets with different haplotype frequencies, including rare haplotypes, for the RecA and MultA models. All parental genotypes are assumed to be missing.

[a] The adjusted type I error rate calculated by the proportion of data sets with p-value ≤ 0.05 where haplotypes with variance < 2.5 were excluded.

[b] The adjusted type I error rate calculated by the proportion of data sets with p-value ≤ 0.05 where haplotypes with variance < 5 were excluded.

[c] E(freq) are the expected frequencies of the haplotypes used in SIMLA configurations.

Table 2.6 Type I error of APL tests with HWE deviations.

Data	APL test (type I error)[a]					
	Single-marker test				Haplotype test	
HWE statistic	RecA	RecB	MultA	MultB	RecA	MultA
0.036	.038	.049	.039	.046	.045	.051
2.479	.045	.039	.045	.054	.067	.119
5.055	.047	.050	.051	.046	.225	.213

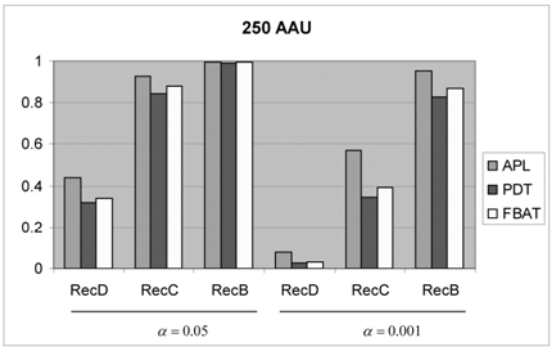
300 AAU families were simulated with HWE deviations for RecA, RecB, MultA and MultB models without parental genotypes.

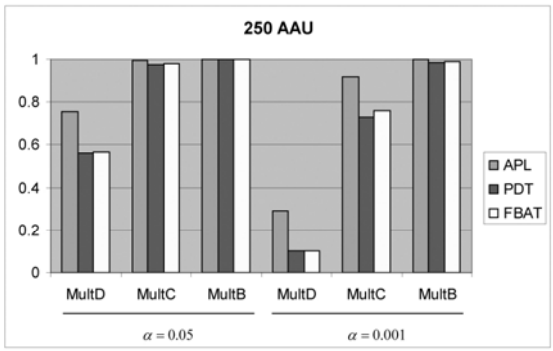
[a] Proportion of data sets with p-value ≤ 0.05.

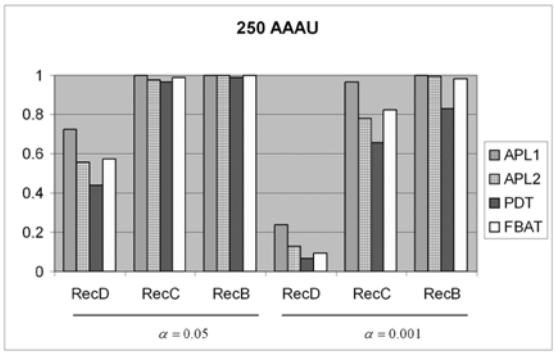
2.8 Figures

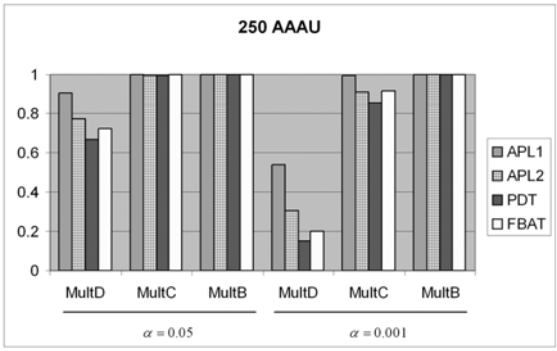
Figure 2.1 Power comparison for single-marker analysis

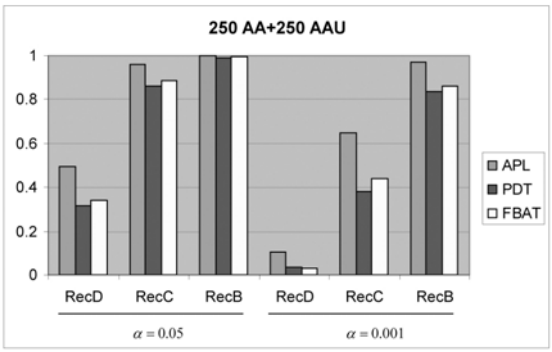
Power of the single-marker APL test, PDT, and FBAT over 2000 replicates. All families were simulated without parental data except AAU* families which have only one parental genotype missing. In the figures for 250AAAU families, APL1 and APL2 are the APL tests using three and two affected siblings, respectively. Power is calculated for significance level (α) of 0.05 and 0.001 for recessive and multiplicative models.

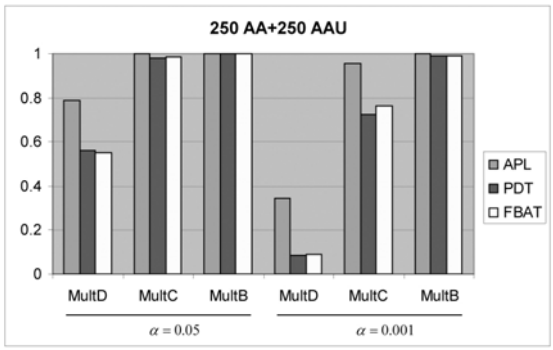

250 AAU
1
0.8
0.6
0.4
0.2
0
RecD RecC RecB RecD RecC RecB
α = 0.05
α = 0.001
APL
PDT
FBAT

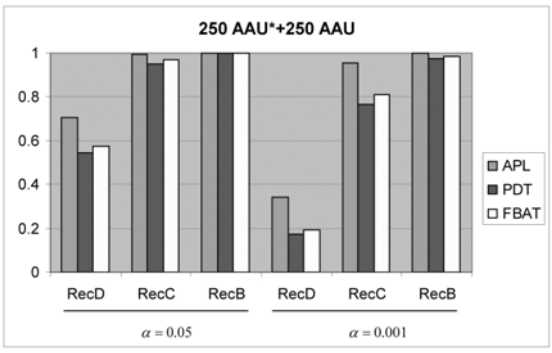

250 AAU
1
0.8
0.6
0.4
0.2
0
MultD MultC MultB MultD MultC MultB
α = 0.05
α = 0.001
APL
PDT
FBAT

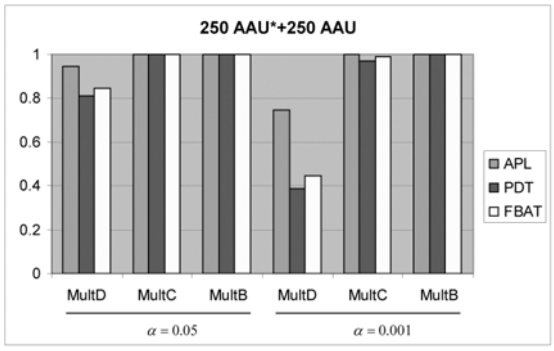

250 AAAU
1
0.8
0.6
0.4
0.2
0
RecD RecC RecB RecD RecC RecB
α = 0.05
α = 0.001
APL1
APL2
PDT
FBAT


250 AAAU
1
0.8
0.6
0.4
0.2
0
MultD MultC MultB MultD MultC MultB
α = 0.05
α = 0.001
APL1
APL2
PDT
FBAT


250 AA+250 AAU
1
0.8
0.6
0.4
0.2
0
RecD RecC RecB RecD RecC RecB
α = 0.05
α = 0.001
APL
PDT
FBAT


250 AA+250 AAU
1
0.8
0.6
0.4
0.2
0
MultD MultC MultB MultD MultC MultB
α = 0.05
α = 0.001
APL
PDT
FBAT


250 AAU*+250 AAU
1
0.8
0.6
0.4
0.2
0
RecD RecC RecB RecD RecC RecB
α = 0.05
α = 0.001
APL
PDT
FBAT


250 AAU*+250 AAU
1
0.8
0.6
0.4
0.2
0
MultD MultC MultB MultD MultC MultB
α = 0.05
α = 0.001
APL
PDT
FBAT

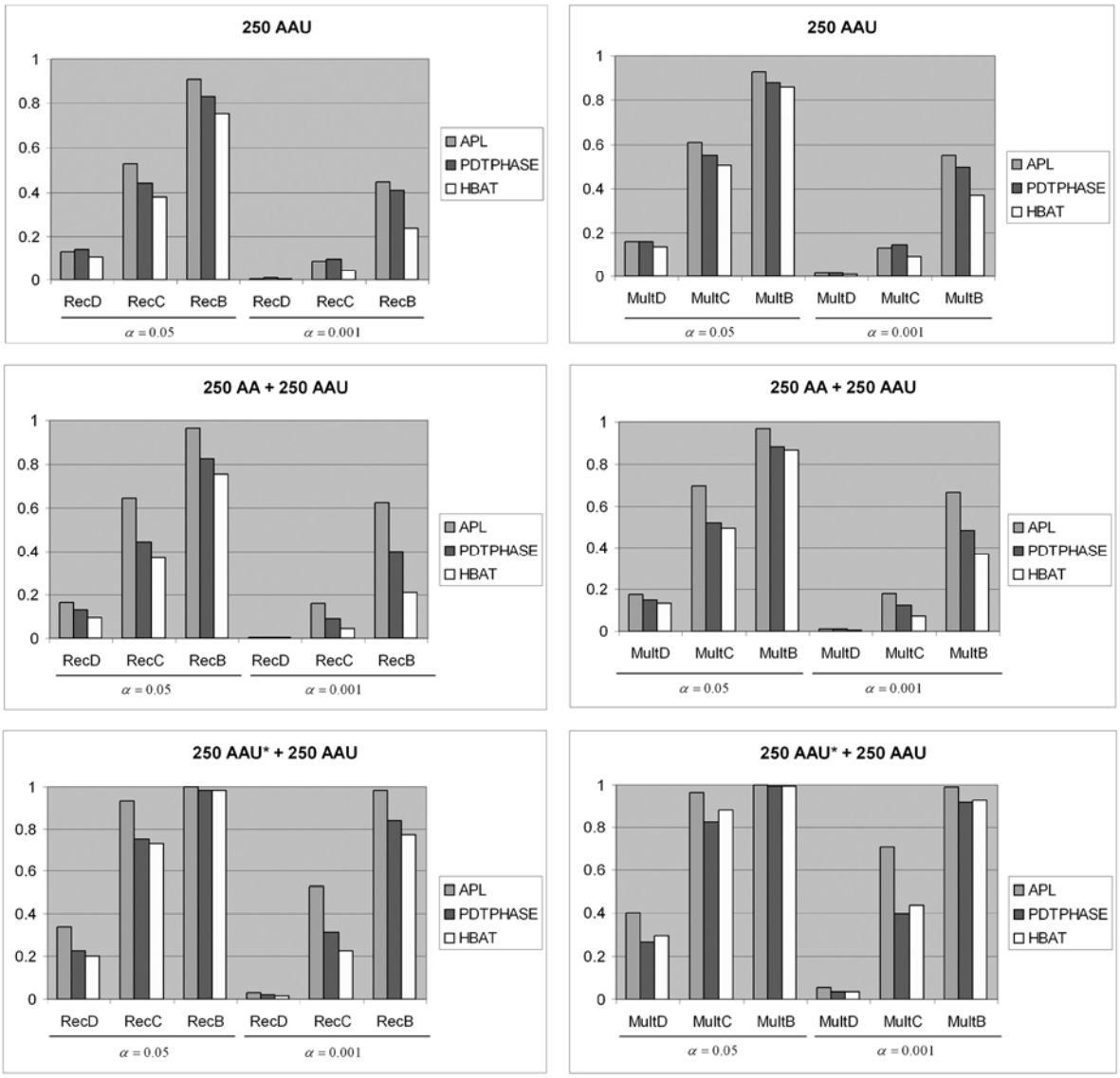


Figure 2.2 Power comparison for haplotype analysis

Power of the haplotype APL test, PDTPHASE, and HBAT over 2000 replicates. All families were simulated without parental data except AAU* families which have only one parental genotype missing. Power is calculated for significance level (α) of 0.05 and 0.001 for recessive and multiplicative models.

Chapter 3

X-APL: An Improved Family-Based Test of Association for the X Chromosome

Ren-Hua Chung, Richard W. Morris, Li Zhang, Yi-Ju Li, Eden R. Martin

(2007) The American Journal of Human Genetics 80:59-68

3.1 Abstract

Family-based association methods have been developed primarily for autosomal markers. XS-TDT and XRC-TDT are the first association-based methods for testing markers on the X chromosome in family data sets. These are valid tests of association in family triads or discordant sib pairs, but are not theoretically valid in multiplex families when linkage is present. Recently, XPDT and XMCPDT, modified versions of the pedigree disequilibrium test (PDT), were proposed. Like the PDT, XPDT compares genotype transmissions from parents to affected offspring or genotypes of discordant siblings; however the XPDT can have low power if there are many missing parental genotypes. XMCPDT uses a Monte Carlo sampling approach to infer missing parental genotypes based on true or estimated population allele frequencies. Though the XMCPDT was shown to be more powerful than the XPDT, variability in the statistic due to using an estimate of allele frequency is not properly accounted for. Here we present a novel family-based test for association, X-APL, a modification of the APL test for Association in the Presence of Linkage. Like the APL, X-APL can use singleton or multiplex families and properly infers missing parental genotypes in linkage regions by considering identity-by-descent parameters for affected sibs. Sampling variability of parameter estimates is accounted for through a bootstrap procedure. X-APL can test marker loci individually or X-chromosome haplotypes. To allow for different penetrances in males and females, separate sex-specific tests are provided. Using simulated data, we demonstrated validity and showed that the X-APL is more powerful than alternative

tests. To show its utility and discuss interpretation in real data analysis, we also applied the X-APL to candidate gene data in a Parkinson disease family sample.

3.2 Introduction

Family-based association methods are often used for localizing genes in complex diseases when family data are available; however, methodological developments have focused primarily on analysis of autosomal markers [Spielman et al., 1993; Martin et al., 1997; 2000; Abecasis et al., 2000; Rabinowitz and Laird, 2000]. Linkage analyses have identified regions on the X chromosome for several diseases, such as Parkinson disease [Scott et al., 2001; Pankratz et al., 2003], autism [Shao et al., 2002; Vincent et al., 2005] and early-onset cardiovascular disease [Hauser et al., 2003]. Although association analysis is often applied to further localize disease susceptibility genes in linkage regions, fine-mapping of such regions on the X chromosome has been slow, in part due to the lack of appropriate statistical methods for family-based association analysis on the X chromosome.

The "X-linked sibling TDT" (XS-TDT) and "reconstruction-combined transmission/disequilibrium test for X-chromosome markers" (XRC-TDT), proposed in Horvath et al. [2000], are the first family-based association methods that test specifically X-chromosome markers. These are valid tests of association in family designs which include a single proband, such as triads or discordant sib pairs. For families with multiple affected offspring, these tests, which assume independent gamete transmissions from parents to

affected siblings, can have an inflated type I error rate when linkage is present between a marker and the disease locus. This is the same problem faced by the original TDT and S-TDT [Martin et al., 1997]. Because association analyses are often conducted in regions showing evidence of linkage, it is critical that family-based association tests allow for the presence of linkage under a null hypothesis of no association when multiple affected offspring are available.

More recently, the pedigree disequilibrium test (PDT), originally proposed for autosomal markers by Martin et al. [1997; 2000], was extended to markers on the X chromosome [Ding et al., 2006]. This approach maintains the properties of the PDT. Namely, the XPDT is a test of association in the presence of linkage in general pedigrees, is valid in stratified populations and does not require specification of model parameters. However, in families with missing parental data the XPDT uses only same-sex discordant sibships and thus may not have optimal power. Recognizing this Ding et al. [2006] suggest a Monte Carlo approach to inferring missing parental data, the XMCPDT. They show that this approach generally has more power than the XPDT, and the power difference increases with increasing amount of missing parental data. A limitation of the XMCPDT is that allele frequencies must be provided. Unknown allele frequencies are estimated from known parental (founder) genotypes, but the statistic does not account for variability in this estimate. Though the examples simulated by Ding et al. [2006] show no inflation of type I error, validity of the test with varying amounts of missing parental data has not been thoroughly examined.

Here we extend the Association in the Presence of Linkage (APL) method [Martin et al., 2003] developed for autosomal markers to the analysis of X-chromosome markers in nuclear families. Like the APL, our proposed procedure, which we refer to as X-APL, properly infers missing parental genotypes in regions of linkage by considering identity-by-descent (IBD) parameters for affected siblings. We use a bootstrap procedure to adjust for the variation in parameter estimates, which does not assume allele frequencies are given. X-APL can perform both single-locus and haplotype association tests. Recognizing the existence of sex-limited traits, we introduce into the X-APL separate tests for males and females, which allow inference about different effects in the sexes.

We used computer simulations to demonstrate the validity of the X-APL statistic and examine robustness and power. We compared the power of the X-APL to the power of the XS-TDT, XPDT and XMCPDT under a range of models and sampling schemes. We compared the power of the X-APL test using all data with separate tests for males and females under sex-specific penetrance models. We then applied X-APL to a real data set containing families with Parkinson disease. We tested markers in two X-linked genes, monoamine oxidase A and B (MAOA and MAOB), which have been examined previously as candidate genes for Parkinson disease [Kurth et al., 1993; Ho et al., 1995; Costa et al., 1997; Wu et al., 2001; Kang et al., 2006].

3.3 Methods

The X-APL statistic is a modification of the APL statistic [Martin et al., 2003]. The APL statistic is based on the difference between the observed number of copies of a specific allele in affected siblings and the expected number of copies conditional on parental genotypes under the null hypothesis that there is no association or no linkage in nuclear families. When parental genotypes are missing, APL infers missing parental genotypes using siblings' genotypes and accounts for linkage by taking the IBD parameters into consideration, see Martin et al. [2003] for details. The APL software can analyze families with up to three affected siblings and arbitrary numbers of unaffected siblings [Chung et al., 2006].

3.3.1 X-APL statistic

The X-APL is designed for nuclear families with one or more affected siblings. First, consider a sample of n families, each with two affected siblings. Markers are assumed to be biallelic with alleles 1 and 2 on the X chromosome. We denote I_j as the number of copies of allele 1 in the affected siblings in the jth family. In Martin et al. [2003], the APL statistic T_j is the difference between I_j and the conditional expected value of I_j, $E(I_j \mid \boldsymbol{G}_{pj})$, where the parental genotypes are represented by $\boldsymbol{G}_{pj}$ in the jth family. To extend the APL statistic to X-chromosome markers, we consider the sexes of the siblings when calculating the expected value of I_j. We define m as an affected male sibling, f as an affected female sibling and sex as the combination of sexes of the affected siblings:

$sex = (m,m)$ if both affected siblings are male

= (m,f) if one affected sibling is male and the other is female

= (f,f) if both affected siblings are female

Under the null hypothesis, the expected value of I_j can be estimated conditional on the parental genotypes and sexes of the affected siblings:

$$\mathrm{E}(I_j \mid \boldsymbol{G}_{pj}, sex) = \begin{cases} N_{fj} & \text{if } sex = (m,m) \\ N_{fj} + N_{mj} & \text{if } sex = (m,f) \\ N_{fj} + 2N_{mj} & \text{if } sex = (f,f) \end{cases}$$

where N_{fj} , the number of allele 1 in the female parent, takes values 0, 1 or 2 and N_{mj}, the number of allele 1 in the male parent, takes values 0 or 1 in the jth family. The expected value of I_j for a singleton family has a simpler form. $\mathrm{E}(I_j \mid \boldsymbol{G}_{pj}, sex = m)$ is $(1/2)\times N_{fj}$ and $\mathrm{E}(I_j \mid \boldsymbol{G}_{pj}, sex = f)$ is $(1/2)\times N_{fj} + N_{mj}$. We define the statistic T_j to be $T_j = I_j - \mathrm{E}(I_j \mid \boldsymbol{G}_{pj}, sex)$ in the jth family. In complete pedigrees, N_{fj} and N_{mj} can be counted directly from the parental data and the transmissions from male parent to affected siblings cancel in T_j; therefore, male parents provide no information for the X-APL statistic in complete pedigrees.

Although calculation of the statistic T_j is straightforward if the parental genotypes are available, for late-onset diseases, parental genotypes are often missing. In this case, we must infer missing parental genotypes based on the siblings' genotypes. In Martin et al. [2003] equation (2), the probability of a parental mating type $\Pr(\boldsymbol{G}_p \mid \boldsymbol{G}, A)$ was estimated based on siblings' genotypes $\boldsymbol{G}$ and their affection status A. We modified this probability for X-APL by also conditioning on the sexes of the siblings. Like the APL for autosomes, the IBD parameters for an affected sibling pair are used to account for linkage between marker and

disease locus when inferring the missing parental genotypes. The IBD for the alleles transmitted from the male parent is fully determined by the sexes of the affected sibling pair. That is, when we consider IBD sharing for alleles transmitted from the male parent, the affected siblings share 0 allele IBD when *sex* = (*m*,*m*) or (*m*,*f*) and 1 allele IBD when *sex* = (*f*,*f*). Thus, only IBD status for alleles transmitted from the female parent needs to be estimated. The affected siblings share either 0 or 1 allele IBD from the female parent.

When there is no association and tight linkage, the probability $\Pr(\boldsymbol{G}_p \mid \boldsymbol{G}, A, sex)$ is similar to Martin et al. [2003] equation (2) and can be written as:

$$\Pr(\boldsymbol{G}_p \mid \boldsymbol{G}, A, sex) = \frac{\mu_{G_p} \sum_{k=0}^{1} z_k \Pr(\boldsymbol{G} \mid \boldsymbol{G}_p, IBD = k, sex)}{\Pr(\boldsymbol{G} \mid A, sex)} \qquad (1)$$

where μ_{G_p} is the unconditional probability of parental mating type $\boldsymbol{G}_p$ and z_k is the probability that the affected siblings share k alleles IBD from the female parent. Since disease penetrances are expected to be low for any particular locus for complex diseases, transmissions to the unaffected siblings are assumed to be independent of disease status. Then IBD parameters for an unaffected sibling pair, or a pair with one affected and one unaffected sibling, can be approximated by (z_0, z_1) = (1/2, 1/2). Therefore, when there are unaffected siblings in a family, the probabilities $\Pr(\boldsymbol{G} \mid \boldsymbol{G}_p, IBD{=}k, sex)$ are multiplied by the Mendelian transmission probabilities for the unaffected siblings for given parental genotypes. The EM algorithm is used to estimate the parameters μ_{G_p} and z_k. The procedures of the EM algorithm for estimating the mating-type and IBD parameters are similar to those used by

Martin et al. [2003] For singleton families, the probability Pr($\boldsymbol{G} \mid \boldsymbol{G}_p$, $IBD = k$, sex) reduces to Pr($\boldsymbol{G} \mid \boldsymbol{G}_p$, sex), which depends only on Mendelian transmission probabilities. When parental genotypes are missing, the expectation of I_j is taken over missing parental genotypes as well as transmissions from (female) parents, as follows:

$$T_j = I_j - \sum_{i \in \boldsymbol{\Omega}} \Pr(\boldsymbol{G}_{pi} \mid \boldsymbol{G}_j, A, sex) E(I_j \mid \boldsymbol{G}_{pi}, sex) \quad (2)$$

where $\boldsymbol{\Omega}$ is a set of all possible parental mating types. Partial parental genotypes can be used to improve estimation of the parental mating-type parameters, using the same methods discussed in Martin et al. [2003]. Let T_s be the sum of T_j over families, then under the null hypothesis that there is no association or no linkage, the expected value of the statistic T_s is 0. The X-APL test is based on this summary statistic T_s.

3.3.2 Variance estimation

Martin et al. [2003] used a robust estimator to estimate the variance of the APL statistic. However, the estimator is difficult to implement in practice when various nuclear family structures exist in a data set. In Chung et al. [2006], a bootstrap variance estimator, which offers more flexibility for analysis of different family structures, was proposed to replace the original variance estimator. The bootstrap approach can be applied to the X-APL model as well. Families are resampled with replacement to form each bootstrap replicate, and a new T_s is calculated for each replicate. If B bootstrap replicates are performed, then the sample

variance $\hat{Var}(T_s)$ can be obtained from the B T_s's. When B is large, the sample variance will be asymptotically close to the variance of T_s.

Finally, the X-APL statistic takes the following form:

$$\frac{T_s}{\sqrt{\hat{Var}(T_s)}} \tag{3}$$

Under the null hypothesis of no linkage or no association, this statistic is asymptotically normal, with a mean of 0 and a variance of 1.

3.3.3 Separate tests for males and females

Disease loci can have different effects on males and females. Although a test using combined data could still find association between disease and markers, separate tests for males and females may be more powerful and informative for sex-specific effects. A straightforward approach to test association for males and females separately is to divide the transmissions from parents to affected siblings into separate transmissions to affected male and female siblings. They can be calculated using the parental mating-type and IBD parameters estimated by all data from both sexes. However, when a marker is in linkage disequilibrium (LD) with a disease locus, estimating parameters using all data may not be appropriate for separate tests, particularly when the disease locus has an effect in only one sex. The parameters may not be estimated properly since they are estimated under the null hypothesis that the marker alleles are not associated with the disease in either sex. Our simulation results

showed that type I errors for the X-APL statistic for the sex with no disease-locus effect can be inflated when all data are used to estimate the parameters. A solution is to divide the data into two sets: one set that has only male affected siblings and another set that has only female affected siblings. The two sets may have overlapping families if some families have both affected male and female siblings. All unaffected individuals are retained in both sets. Then X-APL tests can be applied separately on the two sets using parameters estimated in their respective sets. We refer to the test using only male affected sibs and the test using only female affected sibs as X-APL male and female test, respectively. Since both male and female tests may be performed simultaneously, multiple testing should be considered when we interpret the p-values from the two tests. An adjustment for the p-values may be required such as Bonferroni correction in order to interpret the p-values properly.

3.3.4 Extension to multiple-marker haplotype analysis

To extend the X-APL test to a multiple-marker haplotype test, we assume no recombination occurs between the markers within the families in the sample. The strategy of haplotype testing is analogous to a multiple-allele analysis for a single marker, but with haplotype phase not always known. Probabilities of consistent haplotype phases within each family are estimated jointly with the estimation of IBD parameters and haplotype frequencies using the EM algorithm. Only phase probabilities for females need to be estimated since phase for the male is always known. A global test statistic G, which follows an asymptotic chi-squared distribution under the null hypothesis that none of the haplotypes are associated with the

disease locus, is calculated using the method in Chung et al. [2006] to measure the overall haplotype effect:

$$G = \boldsymbol{T}_s' \Sigma^{-1} \boldsymbol{T}_s \qquad (4)$$

where the vector $\boldsymbol{T}_s$ contains the X-APL statistics for each possible haplotype, Σ is the variance-covariance matrix of $\boldsymbol{T}_s$. If h is the number of haplotypes tested, then the statistic G is asymptotically distributed as chi-square with h-1 degrees of freedom.

A global test for all haplotypes can be more informative than individual haplotype tests since it can capture multiple haplotype effects [Horvath et al., 2004]. The global test also can have more power than individual haplotype tests since it does not have the multiple-testing issue faced when analyzing haplotypes individually [Morris et al., 1997]. Moreover, in the X-APL test, IBD and parental mating-type parameters are estimated under the global null hypothesis that none of the haplotypes are in LD with a disease allele. It is not straightforward to estimate those parameters under a haplotype-specific null hypothesis. For these reasons, we base inference solely on the global test for haplotype analysis.

3.3.5 Hardy-Weinberg equilibrium assumption

Hardy-Weinberg equilibrium (HWE) for haplotype frequencies is assumed in X-APL for the haplotype test to reduce the number of parameters estimated by EM algorithm. The same assumption is also applied to the separate male and female tests. For single-marker analyses, we implement two versions of X-APL, one with and one without a HWE assumption. In real

data analysis, genotyping errors, mutations, and population stratification may cause genotype frequencies to deviate from HWE. We used computer simulations to generate data sets with different degrees of deviation from HWE by combining samples from two random-mating populations with different allele frequencies into one data set. To evaluate deviation from random mating, we generated data with two markers to evaluate the effect for haplotype tests. The HWE goodness-of-fit (GOF) test statistic, which has an asymptotic chi-squared distribution, was used to measure the degree of HWE deviation. Note that we can only measure this deviation in females since males are haploid for the X chromosome. The HWE GOF statistic was calculated as $(n \sum_{i,j \in \Psi} P(h_i h_j) - P(h_i)P(h_j))^2 / nP(h_i)P(h_j)$, where n is the number of female parents, $P(h_i)$ and $P(h_j)$ are the estimated allele frequencies from the mixed population for alleles h_i and h_j. For single marker, $P(h_ih_j)$ are the estimated genotype frequencies in the mixed population and ψ is a set of all alleles for the marker. For haplotypes with two markers, $P(h_ih_j)$ are the estimated haplotype frequencies and ψ is a set of all haplotypes between the two markers. Hence, the deviations from HWE in the data set were simulated with respect to haplotype frequencies for the two markers. We investigated the effects of deviations of allele or haplotype frequencies from HWE expectations for X-APL test using all data and the separate male and female tests.

3.3.6 Computer simulations

Computer simulations were used to evaluate the type I error and power of X-APL. We used the SIMLA computer program [Schmidt et al., 2005] to generate replicate samples of

families based on different disease models. Family ascertainment included the following family structures: single affected offspring with one unaffected sibling (AU), one affected sibling pair (AA), and one affected sibling pair plus one unaffected sibling (AAU). No parental genotypes were available in these families except as noted in the tables.

The SIMLA parameters used in our simulations are shown in Table 1, which contains six recessive models (RecA, RecB, RecC, RecD, RecE, RecF) and six multiplicative models (MultA, MultB, MultC, MultD, MultE, MultF) with different prevalences and genotypic relative risks (GRR) [Martin et al., 2003]. The NullModel in Table 1, with GRR = 1, was used to simulate disease loci that have no effect on a specific sex. The GRR for females is the penetrance function for homozygous disease alleles (f_{DD}) divided by the penetrance function for homozygous normal alleles (f_{dd}). The GRR for males, assumed to be hemizygous, is the penetrance of the disease allele (f_D) divided by the penetrance of the normal allele (f_d). Hence, when GRR is 1 for each sex, the disease loci do not contribute to the disease phenotype. When GRR is greater than 1, the disease alleles increase risk of developing the disease phenotype. By default we simulated samples in which males and females have the same prevalence and GRR, e.g., GRR for females f_{DD}/f_{dd} is equal to GRR for males f_D/f_d. The sex ratio of males to females is 1:1 due to the equal disease prevalence. Samples with different GRR and prevalences for males and females were also generated to reflect disease loci with different effects in the different sexes. In cases of unequal prevalences in males and females, the sex ratio is determined by the prevalence in males to the prevalence in females.

For type I error simulations, we assumed that disease and marker loci were tightly linked (there was no recombination between them), but there was no association between the disease and marker alleles, except as noted below (*Scenario 4*). Four scenarios were simulated (*Scenarios* 1, 2, 3 and 4), each has different family structures and disease models as described in Table 2. In (*Scenario 4*), the marker and disease locus were in strong LD but the GRR for one sex was 1 and for the other sex was greater than 1. To evaluate type I error for multiple-marker haplotype tests, three markers were simulated with eight possible haplotypes. The haplotype frequencies were 0.512, 0.128, 0.128, 0.032, 0.128, 0.032, 0.032, 0.008 for haplotypes (111), (112), (121), (122), (211), (212), (221), (222), respectively. As indicated in Chung et al. [2006], rare haplotypes may affect the validity of the global haplotype APL statistic. Rare haplotypes may also have effects on the X-APL statistic, particularly when samples are stratified into male and female tests. Hence, we increased the number of families to 1000 in *Scenarios 1*, *3*, and *4* for haplotype analyses.

Power simulations assumed that the marker and disease alleles were in perfect LD for single-locus tests, so that the marker locus was in fact equivalent to the disease locus. The AU, AA and AAU family structures with all parents missing were simulated. For multiple-marker haplotype tests, two markers were simulated having four possible haplotypes (11), (12), (21), (22) with frequencies 0.3, 0.3, 0.2, 0.2, respectively. Haplotype (11) was set to be the risk haplotype and was the only haplotype positively associated with the disease allele.

In order to compare the power between the X-APL test using all data to the separate male and female tests, we simulated two additional scenarios (*Scenarios* 5 and 6) as described in Table 2. The two scenarios were simulated with 250 AAU families with all parental genotypes missing under the recessive model.

For comparison with X-APL, we used a SAS macro downloaded from the author's website to conduct the XS-TDT and XRC-TDT [Horvath et al., 2000]. An R program provided by Ding et al was used to conduct the XPDT and XMCPDT [Ding et al., 2006]. Asymptotic p-values provided by the software were used to evaluate significance.

3.4 Results

3.4.1 Type I error and power

We first considered the effect of linkage on the XRC-TDT and XS-TDT as tests for association using computer simulations. Table 3 shows estimates of type I error for disease models RecA and MultA from Table 1 based on 5000 replicate data sets with all parents missing. We found that when multiple affected siblings are present both XS-TDT and XRC-TDT have inflated type I error rates. For example, with significance level 0.05, XS-TDT has a type I error rate 0.062 and XRC-TDT has a type I error rate 0.088 for the 300 AAU families simulated under the disease model MultA. Unlike XS-TDT and XRC-TDT, the XPDT and XMCPDT allow for correlation among multiple affected siblings. In our

simulations, we found that the type I error of the XPDT and the XMCPDT using the true allele frequencies are close to the nominal level (type I error estimates range from 0.045 to 0.054 at the nominal level of 0.05).

We examined the impact of varying the sample size and the proportion of families with parents on the XMCPDT using allele frequencies estimated from the observed parental genotypes (Table 4). When the proportion of families with parents is small, we found that the type I error for XMCPDT can be inflated. For example, when there are 50 AAU families with parents and 300 AAU families with no parents, XMCPDT has a type I error rate 0.077 with a nominal significance level 0.05. This inflation of type I error was seen in both singleton and multiplex families. When the proportion of families with parents increases, XMCPDT can have a reasonable type I error rate, though an upward bias is still evident (Table 4).

Table 5 shows estimates of type I error for X-APL for single-marker and haplotype tests. Type I error estimates for male and female tests are also shown. Under different disease models and family structures, we found that the type I error rate is close to the nominal level of 0.05 for both single-marker and global haplotype tests. In Scenario 4 where the marker and disease loci were in LD and the disease locus only has an effect in males, the female tests show correct type I error rates. In the reverse case that disease locus has an effect only in females, the type I error rates for the male tests were also correct (data not shown). Type I

errors for a nominal level of 0.005 were also estimated and they were also close the nominal level (data not shown).

Even though as a test for association XS-TDT does not account for linkage when multiple affected siblings are present, its type I error rate is not severely inflated for the significance level (α=0.05 and 0.005) in our simulations (Table 3). Hence, we compared the power of X-APL with XS-TDT as well as XPDT under different disease models and nuclear family structures. Since the type I error rate is reasonable for XMCPDT when the proportion of families with parents is large, we included power comparisons with XMCPDT in such cases. We considered two significance levels for power calculations: 0.05 and 0.005. Figure 1 shows that the X-APL outperforms XS-TDT, XPDT and XMCPDT in the six disease models considered. We did not show the power for RecA, RecB, RecC, MultA, MultB and MultC models in figure 1 since X-APL has power 1 for these models. For the data sets that have 250 AAU families without parents, X-APL typically has substantially more power than XS-TDT and XPDT at significance level 0.005. As mentioned in Martin et al. [2003], even families with no unaffecteds or parents can add some power to the APL. We also observed that adding AA families to AAU families gives more information for the X-APL, while XS-TDT and XPDT maintain the same power since they do not use AA families (Figure 1). With a total sample size of 250 AAU families, we also simulated 100 AAU families with parents and 150 AAU families with no parents. Compared to the power for 250 AAU families with no parents, the power for X-APL, XS-TDT and XPDT is higher when some parental

information is available. In the case of larger sample size (250 AAU with one parent and 250 AAU with missing parents), we can see that all of the tests have increasing power. In examples with parental data, XMCPDT can have comparable power with X-APL. However, when the significance level is reduced to 0.005, X-APL shows notably more power than XMCPDT. We also simulated 100 AU families with parents and 150 AU families with no parents to evaluate the power for families with only a single proband. The same pattern as in figure 1 was also observed, namely that X-APL still shows more power than other tests (Data not shown).

We also compared the power of X-APL single-marker test using all data with the X-APL male and female tests for three cases; both sexes together and each sex separately. We applied a Bonferroni correction for the p-values from male and female tests. We compared the power for the X-APL test using all data at significance level 0.05 and the power for male and female tests at significance levels 0.025. Table 6 shows that the X-APL test using data for both sexes has more power than the separate tests for each sex when disease loci have effects on both males and females (*Scenario 5*). However, when disease loci affect only one sex (*Scenario 6*), separate tests can be more powerful. From Table 6 we can see that separate tests can have similar or more power than the test using all data even using a conservative multiple-testing correction. We can also see that the type I error for male or female test for the sex not affected by the disease locus is close to the 0.025 nominal level, as we expected.

In Table 6 we see that the power for the separate test in females is consistently lower than the separate tests in males even when the same GRRs are specified in the two sexes. This is a consequence of the models selected and constraints on model parameters. For example, in scenario 5 males and females have the same disease prevalence and GRR. This forces the phenocopy rate under a recessive model to be higher in females than in males, which as a consequence reduces power for the female test relative to the male test. Varying the relative disease prevalence in males and females also influences power because we have fixed the total sample size. For scenario 6 where the sex ratios are 7:3 and 3:7 and the genetic effect is present in the sex with the higher disease prevalence, the sex-specific test has more power than in scenario 5 where the sex ratio is 1:1.

3.4.2 HWE effect

Table 7 shows type I error estimates for the single-marker X-APL test for data containing deviations from HWE. The version of single-marker X-APL test that assumes HWE for genotype frequencies was tested. For different degrees of deviations from HWE in Table 7, type I error estimates are all close to the 0.05 nominal level. Thus, even in extreme cases where all parents are missing and there are severe deviations from HWE in the data, the single-marker X-APL test is still valid at a significance level of 0.05. Another version of X-APL test that does not assume HWE was also tested for the same data sets and it has correct type I error as well (data not shown). Our power simulations showed that X-APL assuming HWE has more power than X-APL without the HWE assumption regardless of whether

HWE really exists (data not shown). Table 7 also shows estimates of the type I error for X-APL test for global haplotype test. The type I error is inflated when deviations from HWE are present. More severe departures from HWE cause more liberal global haplotype tests.

3.4.3 MAO genes for Parkinson disease

We applied X-APL, XS-TDT and XPDT on the data set used by Kang et al. [2006]. A total of 774 families including 558 singleton families and 216 multiplex families were used for the overall X-APL test. Since 615 families in these 774 families have no parents, XMCPDT was not included for analysis. A total of 530 families including 437 singleton families and 93 multiplex families were used for the male test. A total of 329 families including 288 singleton families and 41 multiplex families were used for the female test. Table 8 shows the results for X-APL using all data and for separate male and female tests, as well as the results obtained by XS-TDT and XPDT. X-APL found that marker RS3027452, located in intron 5 in MAOB, was significant (p-value = 0.036) for the female test at significance level 0.05 while XS-TDT and XPDT did not show significant results. However, with a Bonferroni correction for multiple testing for male and female test, the p-value 0.036 may not be considered a significant result. We also applied the X-APL global haplotype test to the markers in MAOA and MAOB but did not observe a significant haplotype association with Parkinson disease

3.5 Discussion

We have developed the X-APL for testing association in family-based designs for markers on the X chromosome. Our simulation analyses show that X-APL has the correct type I error rate. This is not generally true for XS-TDT and XRC-TDT, which have inflated type I error when linkage is present and there are multiple affected siblings in the data. Inheriting the properties of PDT, XPDT does have the correct type I error rate in the linkage region for families with multiple affected siblings. XMCPDT, which can infer missing genotypes conditional on population allele frequency, relies on availability of at least some parental genotypes to estimate population allele frequency. As demonstrated in our simulations, when the proportion of genotyped parents is low, XMCPDT may not be a valid test. It is also worth noting that in Table 3, the type I error rate of XS-TDT is not substantially inflated, and is much closer to the nominal level than the type I error rate of XMCPDT.

Our simulation results showed that X-APL is more powerful than XS-TDT, XPDT and XMCPDT for the six disease models used for single-marker analyses in our simulations. As mentioned in Horvath et al. [2000], the partition of siblings into same-sex groups can result in reduced power for XS-TDT. XPDT also requires that the discordant sib pairs be of the same sex [Ding et al., 2006]. X-APL does not require this partitioning, which contributes to its increased power. An unexpected observation was that XS-TDT consistently has more power than XPDT. This difference may be because the hypergeometric distribution assumed

in XS-TDT better approximates the variance under the alternative than the variance estimate used in the XPDT.

Our simulation results show that the X-APL test using data from both sexes can have more power than separate X-APL male and female tests when the X-linked disease locus contributes to disease risk for both sexes. When the X-linked disease locus affects only one sex, separate tests can have more power than the test using all data. Hence, separate tests to determine the effects of disease loci on sex can be useful. However, some information may be lost when all data are divided into separate smaller data sets. Moreover, separate male and female tests give rise to a multiple testing issue and a correction may be required in assessing significance of the tests. We found that even if we used a conservative Bonferroni correction for the p-values from separate tests, the separate tests can still have more power than the test using all data when disease loci have effect on only one sex. Therefore, we suggest that in practice tests using all data and separate male and female tests should be performed to capture the most information in the data. The same strategy for male and female tests can be applied to XPDT as well. Our expectation is that it would have less power, and is in fact consistent with the real data analysis for the MAOB gene.

We are not aware of any haplotype test other than X-APL for the X-chromosome markers. Consequently, we estimated power for the global haplotype test of X-APL using computer simulation and the results were reasonable. X-APL fills the gap for family-based haplotype

association analysis on the X chromosome and will be very useful for haplotype analyses in real data applications. When compared to APL for autosomal markers, X-APL shows considerably more power for the global haplotype test using the same disease models and family structures (data not shown). One reason that we expect to find more power for haplotype analysis on the X chromosome compared to autosomes is that the phases for male haplotypes are determined explicitly for X-chromosome markers. Therefore, haplotype phases can be inferred more precisely in X-APL than APL. When both parental data are present, the haplotype phases can be exactly determined. When both parents are missing in a family, the haplotype of a male sibling determines one haplotype in the female parent. Moreover, the male parent is known to carry one Y chromosome and the other haplotype in the male parent can be determined by a female sibling. Hence, there are fewer possible parental mating-types that need to be considered compared to the autosomal data, which can reduce the variance of estimates of missing parental mating-types. Although these observations are merely of academic interest since a disease locus is either on the X chromosome or is not, they suggest that the X-APL haplotype test performance is consistent with expectations.

We also examined the robustness of X-APL for data containing deviations from HWE. Our simulation results showed that the X-APL test for single-marker analysis is robust to deviations from HWE in the data. Since the X-APL test with HWE assumption has more power than the X-APL test without HWE assumption for single-marker analysis, the version

of X-APL with HWE assumption will be preferred for most real data analyses. For global haplotype analyses, violations of HWE for haplotype frequencies can significantly inflate type I errors for X-APL. Therefore, haplotype analyses are not reliable in data sets deviating from HWE.

The monoamine oxidase genes MAOA and MAOB are located on the X-chromosome and have been found associated with Parkinson disease (PD) [Kurth et al., 1993; Ho et al., 1995; Costa et al., 1997; Wu et al., 2001; Kang et al., 2006]. In Kang et al. [2006], a total of 644 families with PD, consisting both of singleton and multiplex families, were used for association studies. The PDT [Martin et al., 1997; 2000], developed for autosomal markers, was used in Kang et al. [2006] dividing the whole data set into female and male siblings. The marker RS1799836 located in intron 13 in MAOB showed significant association with PD in the data set containing only female siblings (p-value = 0.022). The X-APL female test showed significant association for the marker RS3027452 located in intron 5, but again the effect was restricted to females. This suggests that the MAOB gene might have an effect in females but not in males for PD. Our results are not entirely consistent with the results obtained by Kang et al. [2006] since they are different markers showing significance in the different analyses. The inconsistency may be due to the fact that Kang et al. [2006] restricted analysis to same-sex discordant sib pairs from extended families while X-APL calculated transmissions from parents to affected children only in nuclear families. Moreover, unaffected male and female siblings were used in X-APL sex-specific tests to estimate

parameters but again only same-sex unaffecteds were considered in Kang's analysis. Nevertheless, the associated SNPs show some LD (r^2=0.23 between RS1799836 and RS3027452 in affected females) [Kang et al., 2006] and both show effects limited to female subset. Therefore, it may indicate that there is a yet untested female risk variant in LD with the two SNPs.

In conclusion, we have developed X-APL as a powerful, robust and versatile tool for family-based association analysis on the X chromosome. We demonstrated validity where other tests failed and showed that X-APL is more powerful than alternative tests, particularly when parental genotypes are unavailable as in late-onset diseases. X-APL provides single-marker and global haplotype tests as well as separate tests for males and females allowing for evaluation of a variety of hypotheses. As presented here X-APL uses nuclear families allowing for missing parental genotypes. XPDT and XMCPDT have the advantage of performing analyses for extended pedigrees. Similar approaches can be used to modify the APL and X-APL for extended pedigrees, and the bootstrap procedure to estimate the variance lends itself easily to this modification. We have implemented X-APL in a freely available software package. The X-APL software package is written in C++ and available for several computer platforms. It can be publicly accessed at http://www.chg.duke.edu/research/software.html.

3.6 Acknowledgements

We gratefully acknowledge generous Funding from NIH Grants NS051355 and NS39764, We are also grateful for participation of PD patients and their families. We also thank two anonymous reviewers for helpful comments on the manuscript.

3.7 Tables

Table 3.1 Genetic models used in simulations

Model of inheritance	Disease marker allele frequency	GRR[a]	Disease prevalence
Recessive:			
RecA	0.25	5.00	0.0063
RecB	0.25	4.00	0.0059
RecC	0.25	3.00	0.0056
RecD	0.25	2.50	0.0054
RecE	0.25	2.00	0.0053
RecF	0.25	1.50	0.0052
Multiplicative:			
MultA	0.15	7.50	0.0064
MultB	0.15	6.25	0.0060
MultC	0.15	4.00	0.0053
MultD	0.15	3.06	0.0049
MultE	0.15	2.25	0.0046
MultF	0.15	1.56	0.0043
NullModel	0.25	1.00	0.0023

[a]GRR for female = f_{DD}/f_{dd}. GRR for male = f_D/f_d. f is the penetrance function for disease allele D.

Table 3.2 Scenarios simulated for different family structures and genetic effects.

Scenario	Sample	Male/Female GRR[b]
Scenarios for type I error		
1[a]	300 AAU (1000 AAU)	Equal
2	250 AAU + 250 AAU*	Equal
3[a]	300 AAU (1000 AAU)	Unequal
4[a]	300 AAU (1000 AAU)	Sex-limited (GRR=1 in one sex)
Scenarios for power		
5	250 AAU	Equal
6	250 AAU	Sex-limited (GRR=1 and lower disease prevalence in one sex)

Note- AAU families have 2 affected and one unaffected sibling with parental genotypes missing. AAU* families are the same as AAU families but have genotype data available for one parent.

[a]1000 AAU families were used for haplotype analyses.

[b]GRR and prevalences are determined based on the disease models in Table 1.

Table 3.3 Type I error of XRC-TDT and XS-TDT for 5000 simulated data sets.

		Type I error[a]			
Number of	Inheritance	XS-TDT		XRC-TDT	
Families	Model	$\alpha = 0.05$	$\alpha = 0.005$	$\alpha = 0.05$	$\alpha = 0.005$
300 AAU					
	RecA	0.060	0.006	0.070	0.011
	MultA	0.062	0.006	0.088	0.016
600 AAU					
	RecA	0.060	0.006	0.077	0.011
	MultA	0.060	0.007	0.078	0.016

[a]Proportion of data sets with p-value $\leq \alpha$.

Table 3.4 Type I error of XMCPDT for 5000 simulated data sets.

Family structure	Type I error[a]	
	$\alpha = 0.05$	$\alpha = 0.005$
50 AU[b] + 300 AU[d]	0.077	0.011
50 AAU[b] + 300 AAU[d]	0.077	0.011
100 AAU[b] + 250 AAU[d]	0.058	0.006
250 AAU[c] + 250 AAU[d]	0.059	0.008

[a]Proportion of data sets with p-value $\leq \alpha$.
[b]Both parental genotypes are available.
[c]One of the parents is missing
[d]Both parents are missing

Table 3.5 Type I error of the X-APL tests for 5000 simulated data sets.

	X-APL test (Type I error[a])			X-APL global haplotype test (Type I error[a])		
Scenarios and Model of inheritance	ALL[c]	MALE[d]	FEMALE[e]	ALL[c]	MALE[d]	FEMALE[e]
Scenario 1:						
RecA	0.047	0.052	0.051	0.056	0.052	0.051
RecB	0.050	0.054	0.055	0.057	0.055	0.057
MultA	0.049	0.049	0.051	0.049	0.052	0.056
MultB	0.047	0.053	0.047	0.053	0.051	0.054
Scenario 2:						
RecA	0.046	0.054	0.044	0.045	0.051	0.055
MultA	0.046	0.052	0.047	0.048	0.053	0.051
Scenario 3:						
RecA in Male, RecC in Female	0.053	0.045	0.051	0.048	0.050	0.055
MultA in Male, MultC in Female	0.048	0.045	0.049	0.048	0.045	0.053
Scenario 4[b]:						
RecA	-	-	0.053	-	-	0.048
MultA	-	-	0.053	-	-	0.052

[a]Proportion of data sets with p-value ≤ 0.05.
[b]Disease loci have effect on males but not on females. Thus type I errors occur only in females.
[c]X-APL test using all data
[d]X-APL male test
[e]X-APL female test

Table 3.6 Power estimates for X-APL test using all data and separate tests for males and females.

Scenarios		X-APL test[a]		
	Sex ratio[b] (Male:Female)	ALL	MALE	FEMALE
		$\alpha = 0.050$	$\alpha = 0.025$	$\alpha = 0.025$
Scenario 5: (RecD in both sexes)	1:1	0.944	0.688	0.140
Scenario 6:				
RecD in Male, NullModel in Female	7:3	0.990	0.997	0.023[c]
NullModel in Male, RecD in Female	3:7	0.214	0.023[c]	0.218

[a]Proportion of data sets with p-value $\leq \alpha$.
[b]Sex ratio for males to females in the affected siblings is calculated by the prevalence of males : the prevalence of females in the population.
[c]Estimates type I error for the test since disease loci have no effect in the sex.

Table 3.7 Type I error of X-APL tests with HWE deviations.

		X-APL test (type I error)[a]			X-APL Global Haplotype test (type I error)[a]
HWE GOF[b]	Model	ALL	MALE	FEMALE	GLOBAL
0.036	RecA	0.052	0.046	0.043	0.053
	RecB	0.048	0.053	0.047	0.050
	MultA	0.047	0.043	0.046	0.052
	MultB	0.052	0.047	0.057	0.051
2.479	RecA	0.046	0.048	0.052	0.073
	RecB	0.049	0.047	0.049	0.068
	MultA	0.048	0.046	0.052	0.075
	MultB	0.054	0.045	0.050	0.070
5.055	RecA	0.053	0.046	0.050	0.122
	RecB	0.043	0.043	0.048	0.115
	MultA	0.043	0.043	0.052	0.096
	MultB	0.052	0.044	0.048	0.100

300 AAU families were simulated with HWE deviations for RecA, RecB, MultA and MultB models without parental genotypes.

[a]Proportion of data sets with p-value ≤ 0.05.

[b]HWE goodness-of-fit statistic. Larger value means larger deviation from HWE.

Table 3.8 XS-TDT, XPDT, and X-APL results for MAOB gene analysis

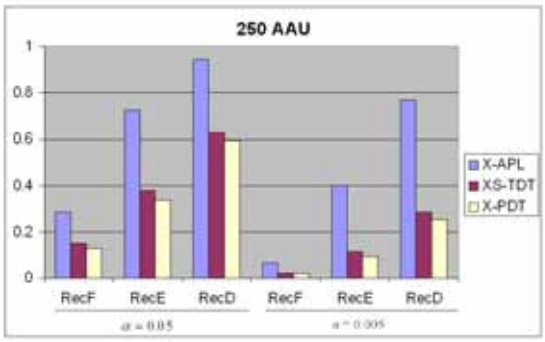
			p-value	
Marker	*N*[a]	XS-TDT[b]	XPDT	X-APL
RS3027452				
OVERALL	774	0.634	0.676	0.063
MALE	530	0.595	0.673	0.519
FEMALE	329	0.175	0.349	0.036

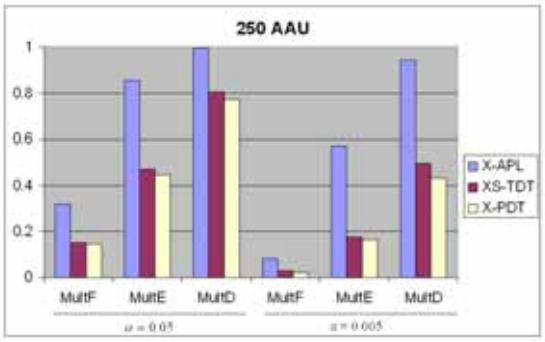
[a]Number of families.
[b]Asymptotic p-values for XS-TDT were used.

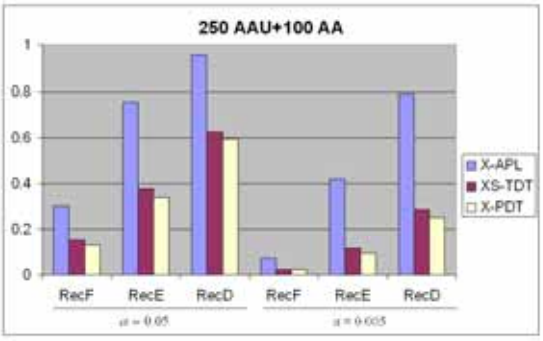
3.8 Figures

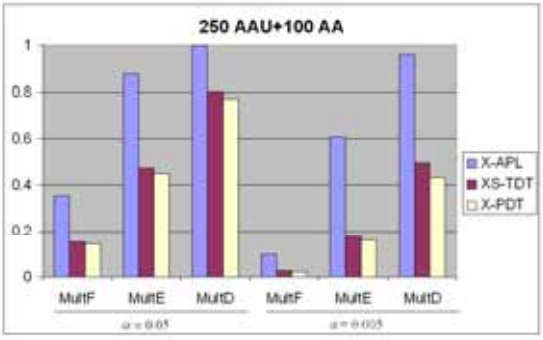
Figure 3.1 Power comparison for single-marker analysis.

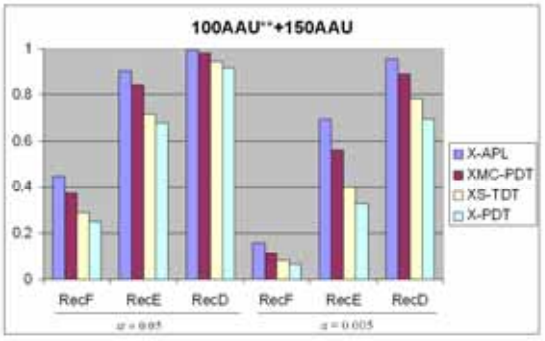
Power estimates for the single-marker X-APL, XS-TDT, XPDT and XMCPDT tests based on 2000 simulated data sets. All families were simulated without parental data except AAU** families which both parents are present and AAU* families which have one parental genotype missing. Power is calculated for significance level (α) of 0.05 and 0.005 for recessive and multiplicative models.

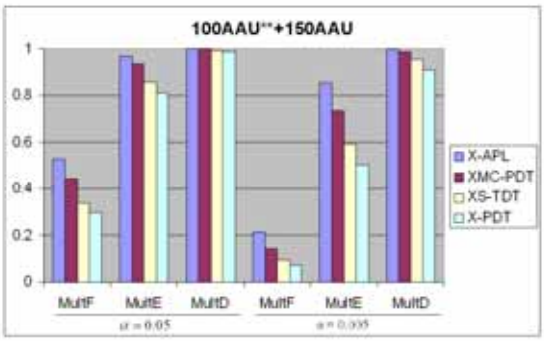

250 AAU
1
0.8
0.6
0.4
0.2
0
X-APL
XS-TDT
X-PDT
RecF
RecE
RecD
RecF
RecE
RecD
α = 0.05
α = 0.005

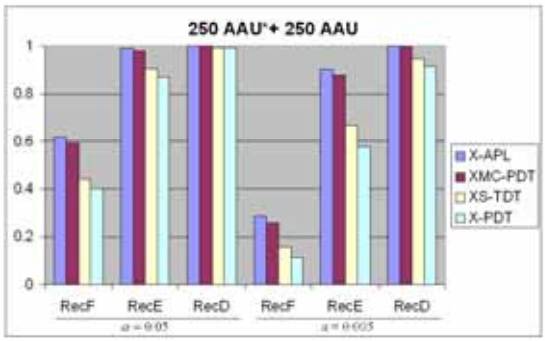

250 AAU
1
0.8
0.6
0.4
0.2
0
X-APL
XS-TDT
X-PDT
MultF
MultE
MultD
MultF
MultE
MultD
α = 0.05
α = 0.005

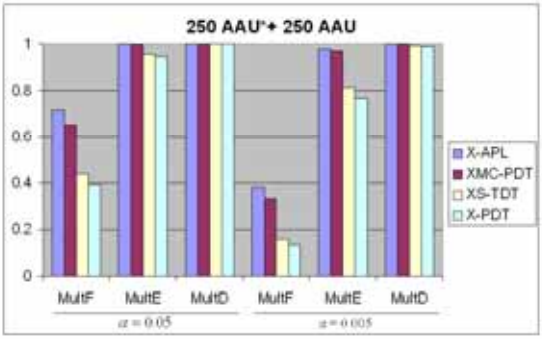

250 AAU+100 AA
1
0.8
0.6
0.4
0.2
0
X-APL
XS-TDT
X-PDT
RecF
RecE
RecD
RecF
RecE
RecD
α = 0.05
α = 0.005


250 AAU+100 AA
1
0.8
0.6
0.4
0.2
0
X-APL
XS-TDT
X-PDT
MultF
MultE
MultD
MultF
MultE
MultD
α = 0.05
α = 0.005


100AAU**+150AAU
1
0.8
0.6
0.4
0.2
0
X-APL
XMC-PDT
XS-TDT
X-PDT
RecF
RecE
RecD
RecF
RecE
RecD
α = 0.05
α = 0.005


100AAU**+150AAU
1
0.8
0.6
0.4
0.2
0
X-APL
XMC-PDT
XS-TDT
X-PDT
MultF
MultE
MultD
MultF
MultE
MultD
α = 0.05
α = 0.005


250 AAU*+ 250 AAU
1
0.8
0.6
0.4
0.2
0
X-APL
XMC-PDT
XS-TDT
X-PDT
RecF
RecE
RecD
RecF
RecE
RecD
α = 0.05
α = 0.005


250 AAU*+ 250 AAU
1
0.8
0.6
0.4
0.2
0
X-APL
XMC-PDT
XS-TDT
X-PDT
MultF
MultE
MultD
MultF
MultE
MultD
α = 0.05
α = 0.005

Chapter 4

Interpretation of simultaneous linkage and family-based association tests in genome screens

Ren-Hua Chung, Elizabeth R. Hauser, Eden R. Martin
In Press, Genetic Epidemiology

4.1 Abstract

Linkage and association analyses have played important roles in identifying susceptibility genes for complex diseases. Linkage tests and family-based tests of association are often applied in the same data to help fine-map disease loci or validate results. This paradigm increases efficiency by making maximal use of family data sets. However, it is not intuitively clear under what conditions association and linkage tests performed in the same data set may be correlated. Understanding this relationship is important for interpreting the combined results of both tests. We used computer simulations and theoretical statements to estimate the correlation between linkage statistics (affected sib pair maximum LOD scores) and family-based association statistics (PDT and APL) under various hypotheses. Different types of pedigrees were studied: nuclear families with affected sib pairs, extended pedigrees and incomplete pedigrees. Both simulation and theoretical results showed that when there is no linkage or no association, the linkage and association tests are not correlated. When there is linkage and association in the data, the two tests have a positive correlation. We concluded that when linkage and association tests are applied in the same data the type I error rate of neither test will be affected and that power can be increased by applying tests conditionally.

4.2 Introduction

Linkage analyses are used to find chromosomal regions that do not recombine with a proposed disease locus. Linkage test statistics test whether traits co-segregate with genetic markers within families more often than expected under the hypothesis of no linkage. For

example, the allele-sharing method proposed in Kruglyak et al. [1996] examines the excess allele sharing between affected sib pairs. The identified regions of linkage are then more likely to contain genes related to the disease. However, linkage analysis can lack power for common diseases caused by multiple genes and environmental factors [Cardon and Bell, 2001]. Association analysis can be used as a complementary method to linkage analysis and may be more powerful for common complex diseases [Risch and Merikangas, 1996]. Linkage tests often identify very large chromosomal regions, often as large as 40cM. Association analyses examine the linkage disequilibrium (LD) between the disease and marker alleles on a much finer scale. In large random-mating populations, LD generally spans only small distances and the markers used for association analysis are often very tightly spaced. These properties mean that association analysis provides a higher resolution for locating disease genes than linkage analysis. To take advantage of their complementary properties, a common strategy for identifying complex disease genes is to conduct linkage analyses first and then follow significant results with family-based tests for association at a denser panel of markers in an attempt to further localize the disease gene [Cardon and Bell, 2001]. Using this strategy, many studies have found significant association results from regions that showed high linkage peaks [Martin et al., 2000a; van der Walt et al., 2004; Oliveira et al., 2005; Connelly et al., 2006].

Recently advanced technology and reduced genotyping costs have made whole-genome association analyses of hundreds of thousands of single nucleotide polymorphism (SNP)

markers possible. With the completion of PHASE I of the HAPMAP project [International HapMap Consortium, 2003; Altshuler et al., 2005], about 6 million new SNPs were genotyped to promote the discovery of high-quality SNPs and to define LD structures in the human genome as a framework for the whole-genome association analyses. Whole-genome association analyses can be performed without information from linkage analyses. However, a large sample size is required to compensate for the power lost from multiple comparison corrections required for the huge number of hypothesis tests. This multiple-testing issue is a challenging problem for whole-genome association analysis [Carlson et al., 2004]. Approaches to judiciously limit the number of markers in a whole genome association analysis are needed. One approach is to use a panel of tag SNPs [Hirschhorn et al., 2005]. Narrowing the set of markers through linkage analysis is another approach.

We may also use a combination of linkage and association tests as complementary lines of evidence pointing to a susceptibility locus [Schwab et al., 2002; Mansur et al., 2004; McQueen et al., 2005]. The transmission/disequilibrium test (TDT) [Spielmen et al., 1993], which is a classic family-based test for linkage or association, has been widely used. The TDT applied to triad data (families with parents and one affected sibling) tests the null hypothesis of no linkage or no association. Rejection of the hypothesis means that both linkage and association are present in the data. Several modifications of TDT have generalized this test [Martin et al., 1997; Abecasis et al., 2000; Martin et al., 2000b]. When significant results are found using the TDT, linkage tests other than TDT may be performed

in the same data to verify if the TDT results are true positives. The convergence of positive results from both tests can provide more confidence that markers identified by the TDT are true positives and help eliminate false positives. Several studies have used this strategy to verify their association results [Deng et al., 2002; Schork et al., 2002].

A novel approach for whole-genome association analyses also uses linkage test results to weight the p-values of association tests and this approach shows more power than association tests alone if the linkage tests are informative [Roeder et al., 2006]. Further, we must keep in mind that due to the limitation of association analyses for finding rare variants associated with the diseases, linkage analyses will still remain essential [Wang et al., 2005]. Even in the age of whole-genome association studies, linkage analyses will still play an important role in identifying disease genes. The combination of association tests and linkage tests provides a powerful strategy for identifying genes for complex diseases. Understanding the relationship between these tests is essential for interpreting the results.

With these strategies, the same data are often used to perform both tests of linkage and family-based association tests. If there is correlation between linkage and family-based association test statistics, the results obtained from the tests may be difficult to interpret together, particularly when one test is performed based on the results from the other test. For example, if there is correlation between linkage and family-based association test statistics when there is linkage and no association, then association tests performed based on

significant results from linkage tests may tend to be liberal. To help interpret the results from linkage and association tests conducted on the same data, it is desirable to know when the tests are correlated. It has been shown analytically that the TDT, as a test of linkage, and the mean haplotype allele-sharing test (MHS) [Blackwelder and Elston, 1985] of linkage are independent under the null hypothesis of no linkage [Spielman et al., 1993]. However, it is unclear what the relationship is between family-based tests that focus on testing for association and allele-sharing linkage tests, particularly when parents are missing.

To examine these questions, we used computer simulations to estimate the correlation between association and linkage tests in general pedigrees. Two types of general pedigrees were investigated: extended and incomplete pedigrees. In extended pedigrees, the software packages MERLIN [Abecasis et al., 2002] for linkage and PDT for association were used. For incomplete families, the software packages MERLIN for linkage and APL [Chung et al., 2006] for association were used. We also derived expressions for the correlation between the pedigree disequilibrium test (PDT), a test for association in pedigree data [Martin et al., 2000b], and the allele-sharing test for linkage [Kruglyak et al., 1996] in affected sib pair families.

4.3 Methods

We used computer simulations to examine the correlation between linkage and association tests. Two different types of general pedigrees were studied including the extended pedigrees and incomplete pedigrees.

4.3.1 Computer simulations

The software SIMLA [Bass et al., 2004] was used to generate simulated data sets. Estimates of correlation between linkage and association test statistics are each based on 5000 replicate data sets. We considered sample sizes of 200 or 400 families per data set. For extended pedigree analyses, families had the structure shown in Figure 1 under the constraint that each family has at least an affected sib pair. Additional family members were assigned an affection status randomly, dependent upon the disease model. Disease model parameters were based on estimates for *APOE-4* and Alzheimer disease [Martin et al., 2000a]. The disease allele frequency was set to 0.15 and penetrances were f_0=0.006, f_1=0.03 and f_2=0.1, where f_i is the probability of being affected given that a person carries i copies of the disease allele. For incomplete pedigrees, we generated nuclear families with two affected sibs and one unaffected sib and all parents are assumed to have missing genotypes.

We considered four cases for linkage and association between markers and disease loci: *Case 1*: neither association nor linkage; *Case 2*: linkage and no association; *Case 3*: association and no linkage; *Case 4*: both linkage and association. Each of the four cases is biologically

realistic and we describe possible reasons that cause the four cases: *Case 1*: the marker and the disease loci are on different chromosomes in a large, random-mating population such that they are neither linked nor associated; *Case 2*: The marker and the disease loci are closely located on the same chromosome so they are linked. However, the population has undergone sufficient random mating such that there is no association between the marker and the disease loci at the population level; *Case 3*: Association between markers and disease loci is maintained by other forces such as population admixture, recent mutation or inbreeding rather than the linkage [Spielman et al., 1993]. *Case 4*: The marker and the disease loci are closely located on a chromosome so they are linked and the linkage maintains an initial association between the alleles at the markers and disease loci.

Case 3 in SIMLA was simulated under the assumption that there is association in the first generation and the association decays quickly during each generation because the loci are unlinked. By the third generation, there is actually no association present. Thus, the allele frequencies are different between parents and children. When parents are missing in this case, MERLIN does not correctly estimate their allele frequencies. A frequency file which contains correct parental allele frequencies was provided for MERLIN with option "-f" to guide MERLIN for calculating the LOD scores properly in our simulations.

Association parameters and allele frequencies were chosen based on observed values for association with *APOE-4* and SNPs in APOE region [Martin et al., 2000a]. For markers

linked with disease, we assumed complete linkage between the markers and disease locus. Table I shows the values of the recombination fraction between the disease and marker loci, the level of association (measured by D' and r^2) between marker and disease alleles, and the marker allele frequencies for the four cases examined. The disease allele frequency and the disease penetrance model were fixed for the four cases we examined.

4.3.2 Extended pedigrees

We first consider extended pedigrees with three generations and with affected and unaffected siblings. We used the software MERLIN for linkage tests and the software PDT for association tests. MERLIN implements the nonparametric LOD score proposed in Kong and Cox [1997], which is a modification of the allele-sharing statistic. MERLIN calculates LOD scores for extended pedigrees efficiently using the sparse gene flow trees [Abecasis et al., 2002]. Hence, it is suitable for our simulation purpose. The option "--npl" in MERLIN was used to calculate the nonparametric LOD scores. The software PDT computes the T_{PDT} statistic and is appropriate for extended pedigrees.

4.3.3 Incomplete pedigrees

For linkage analysis, when the parental data are missing, the inheritance vectors in the allele-sharing method need to be estimated. The perfect-data-approximation method was used to estimate the inheritance vectors [Kruglyak et al., 1996]. However, the allele-sharing statistic was shown to be conservative for incomplete pedigrees [Kong and Cox, 1997]. Kong and

Cox's LOD score modified the allele-sharing statistic and was shown to have a reasonable type I error rate. The software MERLIN implements the Kong and Cox's LOD score (with the option "--npl") and was used for the linkage tests in our simulations.

The association in the presence of linkage (APL) method can infer missing parental genotypes in the linkage regions and test for association in nuclear families [Martin et al., 2003]. The APL and PDT statistics are analogous for nuclear families with affected sib pairs and with parental genotypes. When parental genotypes are missing, APL infers all possible parental mating types to compute the APL statistic. APL accounts for the variance associated with inferring missing parental genotypes in the denominator of the APL statistic using a robust bootstrap procedure [Chung et al., 2006].

Due to the ability of APL to infer missing parental genotypes, APL was used instead of PDT for the association tests for incomplete pedigrees. Calculating the correlation between the allele-sharing statistic and the APL statistic is not straightforward when parents are missing. Hence, we relied on simulations to estimate the correlation between these two statistics. Typically APL accounts for linkage by taking the identity-by-decent (IBD) parameters into consideration when inferring missing parental mating-types. It is of particular interest to investigate if this IBD-based inference induces correlation between the APL statistic and the MERLIN linkage statistic when linkage is present.

4.4 Results

Considering a simple example of families with an affected sib pair and genotyped parents, we first present theoretical arguments to demonstrate the correlation between the allele-sharing statistic for linkage and the PDT statistic for association. We then show the computer simulation results for more general pedigrees including the extended pedigrees and incomplete pedigrees.

4.4.1 Affected Sib Pairs with Parents

We consider N nuclear families with two affected siblings and both parents genotyped at a marker locus with two alleles M_1 and M_2. For each heterozygous parent in the fth family, define the random variables

I_f=1 if M_1 is transmitted to both offspring

= 0 otherwise

and

J_f=1 if M_2 is transmitted to both offspring

= 0 otherwise

For the linkage test, we consider the nonparametric scoring function S_{pairs} proposed in Kruglyak et al. [1996]. Let S_f be a function that measures the number of pairs of alleles IBD in the fth family. Then S_f is simply the sum of I_f and J_f. The S_{pairs} statistic can be written as the following form:

$$S_{pairs} = \frac{\sum_{f=1}^{N}\left[S_f - \mu_f\right]}{\sqrt{\sum_{f=1}^{N}\hat{\sigma}_f^2}} = \frac{\sum_{f=1}^{N}\left[(I_f + J_f) - \mu_f\right]}{\sqrt{\sum_{f=1}^{N}\hat{\sigma}_f^2}} \quad (1)$$

where μ_f is the expected value of S_f under the null hypothesis H_0 of no linkage and σ_f^2 is the variance of S_f. Under H_0, the statistic is asymptotically normal with mean 0 and variance 1.

As a family-based test for association, we consider the pedigree disequilibrium test (PDT). Unlike the original TDT, the PDT is a valid test for association even when there are multiple affected individuals in families. The PDT can use data from related families from extended pedigrees. The PDT statistic is the sum of two quantities comparing counts of a specific allele, say M_1 in families. One is the difference between the number of times the allele is transmitted and the number of times that the allele is not transmitted from parents to the affected children. The other is the difference in the numbers of times the allele is transmitted in affected and unaffected siblings. For affected-sib-pair families, only the difference of transmissions from parents to the children is used. The difference can be calculated by (I_f - J_f). The PDT statistic can be written as:

$$T_{PDT} = \frac{\sum_{f=1}^{N}(I_f - J_f)}{\sqrt{\sum_{f=1}^{N}(I_f + J_f)}} \quad (2)$$

The statistic tests the compound null hypothesis H_0: No linkage or No association. If there is either no linkage or no association in the data, the test statistic is asymptotically normal with mean 0 and variance 1.

We begin by examining the correlation between S_{pairs} and T_{PDT} under various null hypotheses. Since both statistics are asymptotically normal, showing the covariance is equal to 0 implies independence. If there is either *no linkage or no association*, both the expected value of S_{pairs} and T_{PDT} is 0. Hence, the product of the expected values of the two statistics is 0. Then the covariance can be written as:

$$Cov(S_{pairs}, T_{PDT}) = E(S_{pairs} T_{PDT})$$

$$= E\left(\left(\frac{\sum_{f=1}^{N}\left[(I_f + J_f) - \mu_f\right]}{\sqrt{\sum_{f=1}^{N}\hat{\sigma}_i^2}}\right)\left(\frac{\sum_{f=1}^{N}(I_f - J_f)}{\sqrt{\sum_{f=1}^{N}(I_f + J_f)}}\right)\right)$$

(3)

When all N families have the same family structure, the statistics are independently identically distributed (iid) across families, which gives:

$$Cov(S_{pairs}, T_{PDT}) = N \times E\left(\left(\frac{\left[(I_f + J_f) - \mu_f\right]}{\sqrt{\sum_{f=1}^{N}\hat{\sigma}_f^2}}\right)\left(\frac{(I_f - J_f)}{\sqrt{\sum_{f=1}^{N}(I_f + J_f)}}\right)\right)$$

(4)

where *f* can be any family in the *N* families. For large samples, the denominators converge to a constant. Then it suffices to examine:

$$E[((I_f + J_f) - \mu_f)(I_f - J_f)] = E[(I_f + J_f)(I_f - J_f)] - \mu_f E[(I_f - J_f)]$$

(5)

In order to calculate the expected values in equation (5), we need to consider the joint distribution of the random variables I_f and J_f. We consider two situations. If there is no association (but possibly linkage), the probability of transmitting M_1 and transmitting M_2 to both siblings should be equal. Then for all *f*, the distribution of (*If , Jf*) is

$$(I_f, J_f) = \begin{cases} (1,0) & r \\ (0,0) & 1-2r \\ (0,1) & r \end{cases}$$

(6)

where *r* is a probability depending on the amount of linkage, and it is not difficult to show that expression in equation (5) is 0. If there is no linkage (but possibly association), the alleles are independently transmitted. By assuming there is no transmission distortion, the distribution from above holds with r=1/4 and the expression in equation (5) is 0. It follows that, for large samples, the covariance between the statistics converges to 0 if there is either no linkage or no association. Therefore, we expect the allele-sharing and PDT statistics to be uncorrelated (asymptotically) under the null hypothesis (*Case 1*, *2* and *3*) for families with only affected sib pairs.

4.4.2 Extended pedigrees

Table II shows the estimated correlation coefficients between MERLIN (linkage test) and PDT (association test) for extended pedigrees. These results show that if there is either no linkage or no association (*Case 1*, *2*, and *3*), there is little correlation between the PDT and MERLIN tests. Furthermore, the correlation gets closer to 0 when sample size increases. If there is both linkage and association (*Case 4*), then the PDT and the MERLIN tests are correlated and the correlation between the linkage test and the association test statistics increases with increasing levels of LD (*D*' and r^2) between the disease and marker alleles for fixed disease allele frequency.

We then studied two conditional testing procedures that are often used for real data analyses and examined the type I error rate of these conditional tests: (1) Association analyses are performed based on significant linkage results under *Case 2*; (2) Linkage analyses are performed based on significant association results under *Case 3*. We estimated the conditional type I rates for PDT when MERLIN is significant, and conversely for MERLIN when PDT is significant. Since the tests are uncorrelated, we expect the type I error rate for either PDT or MERLIN will not be affected even though one test is performed conditional on the other test. Cutoffs of p-value less than 0.05 for both PDT and MERLIN were used to declare significance. Table III shows as expected that the type I error rates are not significantly different from the expected rates of 0.05 for the PDT and MERLIN. We also estimated the two conditional type I error rates under *Case 1*. The results in Table III

confirmed our arguments again that when there is no linkage or no association, the conditional type I error rates for either test will be the same as their unconditional type I error rates.

Under *Case 4*, we estimated the conditional power for PDT conditioned on significant MERLIN test results and the conditional power for MERLIN conditioned on significant PDT test results. We observed that the conditional power for PDT and MERLIN is modestly greater than their unconditional power (Results not shown). For example, the conditional power for PDT for marker 3 is 0.951 and the unconditional power is 0.936. The conditional power for MERLIN for marker 3 is 0.607 and the unconditional power is 0.597. The power was calculated under the significance level of 0.001.

4.4.3 Incomplete pedigrees

Table IV shows the estimated correlation coefficients between MERLIN (linkage test) and APL (association test) for incomplete pedigrees. Results were similar to those in extended pedigrees. In the cases of either no linkage or no association (*Case 1*, *2*, and *3*), little correlation is observed between the MERLIN and APL tests and the correlation is closer to 0 for the larger sample size. There is significant correlation between the MERLIN and APL tests when both linkage and association are present. The correlation between the linkage test and the association test statistics increases with increasing levels of LD between the disease and marker alleles for fixed allele frequency as well.

The conditional test type I error rates for APL when MERLIN is significant and for MERLIN when APL is significant were also estimated in the same manor as in extended pedigrees. The results are shown in Table V. The type I error rates are not significantly different from the expected rates of 0.05. This also demonstrates that even though APL infers the missing parental genotypes using the IBD parameters, the APL statistic is not correlated with the MERLIN linkage statistic when there is either no linkage or no association.

Under Case 4, we also estimated the conditional power for APL conditioned on significant MERLIN test results and the conditional power for MERLIN conditioned on significant APL test results. The same patterns were observed as seen with PDT and MERLIN (Results not shown). For example, the conditional power for APL for marker 3 is 0.616 and the unconditional power is 0.548. The conditional power for MERLIN for marker 3 is 0.993 and the unconditional power is 0.883. The power was calculated under the significance level of 0.001.

4.5 Discussion

Linkage and family-based association analyses have been applied widely for identifying disease susceptibility genes. Both tests are often used jointly in data analyses. Linkage analyses can identify candidate regions for association analyses. Linkage tests can also be used to validate the results of family-based association tests. This study was inspired by the

fact that, although both tests are often applied in the same data sets, the correlation between the tests has not been investigated. Hence, we use theoretical and computer simulation studies to examine the correlation between the tests under different linkage and association hypotheses for different family structures.

Our theoretical results showed that the allele-sharing statistic for linkage and the PDT statistic for association are not correlated when there is either no linkage or no association for nuclear families with affected sib pairs. We have further verified these results for extended pedigrees and incomplete pedigrees using computer simulations. The computer simulation results showed that the MERLIN test for linkage and the PDT test for association are uncorrelated if there is either no linkage or no association for extended pedigrees. If, however, the marker is both linked to the disease locus and has an associated allele, the linkage and association tests can be correlated and this correlation increases with increasing association. For incomplete pedigrees, it was not intuitively clear that the APL test, which infers missing parental genotypes by taking linkage into account using the IBD parameters, would be independent of the linkage statistic when linkage is present. However, the same pattern of joint and conditional independence was observed between the MERLIN test for linkage and the APL test for association. This observation is particularly important since the APL can analyze families with only ASPs and maximizes the use of family-based association tests in families ascertained under an ASP ascertainment strategy.

A consequence of the low correlation between linkage and association tests when there is either no linkage or no association is that type I error rates are not increased for either test even if we condition on the other test being significant in the same data. Thus, applying linkage and association tests in the same data set does not compromise statistical validity of the tests. This assures us that results will not be misinterpreted when linkage and association analyses are performed in the same data. Furthermore, the chance of both tests being significant for a marker in the same data if there is neither linkage nor association is small. Specifically, it is the product of the significance levels of the two tests since we verified the two tests are independent (asymptotically) when there is neither linkage nor association. That is, for 0.05-level tests, the probability of two false-positives at a marker is 0.0025. This result indicates that we have smaller chance of obtaining false positives by performing the linkage and association tests jointly in the same data and requiring that both be significant.

For whole-genome association studies, multiple testing corrections can greatly compromise power. Van Steen et al. [2005] suggest a two-stage design for a genome-wide family-based association test where a group of SNPs with the highest estimated power are selected in stage one and these SNPs are tested by family-based association test (FBAT) [Laird et al., 2000] in stage 2. This strategy was successfully applied in Herbert et al. [2006] for identifying a common genetic variant associated with obesity. Our results have implications for the multiple-testing issue for whole-genome association studies. Assume families with multiple affected siblings are genotyped in the data. After significant markers are identified by whole-

genome family-based association tests, we suggest that linkage statistics can be used as an adjunct test statistic to better identify false positives and maintain a lower type I error rate overall.

When there is both linkage and association, the linkage and association tests are correlated, and the stronger the levels of association, the more likely the tests are to both be positive. We observed that the conditional power of one test conditioned on the significant test results of the other is greater than the unconditional power under *Case 4*. This raises an interesting question: In the real data analysis, do both significant linkage and association results imply a higher likelihood that a true positive has been identified? Bayesian approaches may provide a framework to address this question. This question needs to be investigated further and the answers will enhance interpretation of results from real data analysis.

4.6 Acknowledgement

The work was supported in part by generous funding from the National Institute of Health, NS51355 and MH59528. We are grateful to Meredyth Bass and several members of the Duke Center for Human Genetics for helpful discussions.

4.7 Tables

Table 4.1 Parameters for simulated data.

	Θ[a]	D'[b]	r^{2}[c]	p[d]
Case 1: No linkage and no association				
Marker 1	0.5	0	0	0.2
Marker 2	0.5	0	0	0.5
Case 2: Linkage and no association				
Marker 1	0	0	0	0.2
Marker 2	0	0	0	0.5
Case 3: Association and no linkage				
Marker 1	0.5	1	1	0.15
Marker 2	0.5	0.718	0.472	0.29
Marker 3	0.5	0.531	0.298	0.36
Case 4: Linkage and association				
Marker 1	0	1	1	0.15
Marker 2	0	0.718	0.472	0.29
Marker 3	0	0.531	0.298	0.36

The disease model was simulated based on estimates for APOE-4 and Alzheimer disease with allele frequency 0.15 for the four cases.

[a]recombination fraction between the marker and disease locus

[b]LD measure $D' = D/\max(D)$, where $D = P_{AB}-P_AP_B$. P_{AB} is the gamete frequency of allele A and allele B in two loci. P_A and P_B are allele frequencies for A and B, respectively.

[c]LD measure $r^2 = (P_{AB}P_{ab} - P_{Ab}P_{aB})\Big/\sqrt{P_AP_aP_BP_b}$, where alleles A and a are the two copies of alleles at the same locus and allele B and b are the two copies of alleles at another locus.

[d]marker allele frequency

Table 4.2 Correlation coefficient between MERLIN and PDT statistics.

	Correlation coefficient	
	N = 200[a]	N = 400
Case 1: No linkage and no association		
Marker 1	0.008	0.007
Marker 2	-0.015	0.006
Case 2: Linkage and no association		
Marker 1	-0.005	0.002
Marker 2	0.011	0.001
Case 3: Association and no linkage		
Marker 1	-0.009	-0.003
Marker 2	-0.008	0.002
Marker 3	-0.007	0.007
Case 4: Linkage and association		
Marker 1	0.449	0.450
Marker 2	0.263	0.260
Marker 3	0.165	0.178

[a]number of families

Table 4.3 Type I error rates for PDT given a significant MERLIN test and MERLIN given a significant PDT.

	P(PDT $p < 0.05$ | MERLIN $p < 0.05$)[a]	
	N = 200	N = 400
Case 1: No linkage and no association[c]		
Marker 1	0.051	0.058
Marker 2	0.048	0.049
Case 2: Linkage and no association		
Marker 1	0.046	0.048
Marker 2	0.047	0.049
	P (MERLIN $p < 0.05$ | PDT $p < 0.05$)[b]	
Case 1: No linkage and no association[c]		
Marker 1	0.049	0.055
Marker 2	0.043	0.044
Case 3: Association and no linkage[c]		
Marker 1	0.047	0.042
Marker 2	0.046	0.046
Marker 3	0.053	0.048

[a]The probability of PDT p-value less than 0.05 given MERLIN p-value less than 0.05
[b]The probability of MERLIN p-value less than 0.05 given PDT p-value less than 0.05
[c]20000 replicate data sets were generated to better approximate the probability

Table 4.4 Correlation coefficient between MERLIN and APL statistics.

	Correlation coefficient	
	N = 200	N = 400
Case 1: No linkage and no association		
Marker 1	0.019	0.008
Marker 2	-0.021	-0.025
Case 2: Linkage and no association		
Marker 1	-0.018	-0.009
Marker 2	0.019	0.007
Case 3: Association and no linkage		
Marker 1	0.018	0.009
Marker 2	0.013	-0.026
Marker 3	0.016	-0.002
Case 4: Linkage and association		
Marker 1	0.498	0.493
Marker 2	0.260	0.277
Marker 3	0.156	0.164

Table 4.5 Type I error for APL given significant MERLIN test and MERLIN given significant APL.

	P(APL p < 0.05 | MERLIN p < 0.05)[a]	
	200 AAU	400 AAU
Case 1: No linkage and no association[c]		
Marker 1	0.055	0.047
Marker 2	0.050	0.056
Case 2: Linkage and no association		
Marker 1	0.055	0.054
Marker 2	0.041	0.047
	P (MERLIN p < 0.05 | APL p < 0.05)[b]	
Case 1: No linkage and no association[c]		
Marker 1	0.051	0.044
Marker 2	0.045	0.058
Case 3: Association and no linkage[c]		
Marker 1	0.049	0.050
Marker 2	0.047	0.043
Marker 3	0.058	0.047

[a]The probability of APL p-value less than 0.05 given MERLIN p-value less than 0.05
[b]The probability of MERLIN p-value less than 0.05 given APL p-value less than 0.05
[c]20000 replicate data sets were generated to better approximate the probability

4.8 Figures

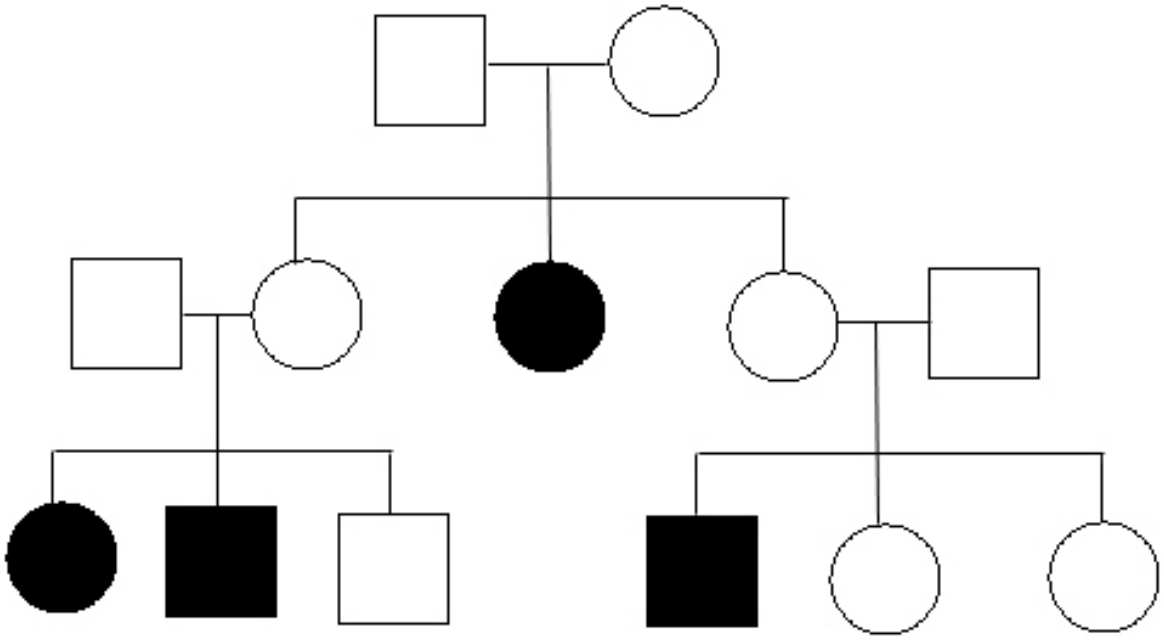


Figure 4.1 Pedigree structure in simulations.

Squares represent males and circles represent females. Filled shapes represent affected individuals. There is at least one affected sib pair in the third generation.

Chapter 5

Conclusions

In summary, we present extensions and examination of the robustness of the APL method, which is a powerful family-based association method. A novel variance estimator based on the bootstrap approach was proposed for the extended APL, which allows for a mixture of different nuclear family structures and extension to the use of three affected siblings. The IBD parameters between each pair in the three affected siblings were considered when inferring missing parental genotypes. The IBD parameters were estimated jointly with allele frequency by the EM algorithm. The original APL test was also extended to multiple-marker haplotype analysis.

Simulations were used to demonstrate that the extended APL has a correct type I error rate for both single-marker and haplotype analyses. For nuclear families with missing parents, our simulation results also show that APL has more power than other family-based association tests such as PDT, FBAT/HBAT and PDTPHASE. We also evaluated the effects of rare alleles or haplotypes on the APL test. We suggested that the variance for the APL test can be used as a guideline to decide whether the APL test is valid or not. Generally, variance for the APL statistic greater than 5 suggests that APL maintains the correct type I error rate. We also used simulations to examine the robustness of APL toward the departure from the HWE assumption. We concluded that a single-marker APL test is robust for samples with allele frequencies that deviate from HWE. For the global haplotype APL test, the departure of the haplotype frequencies from HWE can inflate the type I error rate.

We also present X-APL, an extension from APL to X-chromosome markers. Inherited from the properties of APL, X-APL properly infers missing parental genotypes in the linkage regions by taking the IBD parameters into consideration. X-APL can also use multiple affected siblings for association tests. Since diseases can have different effects on males and females, X-APL also provides sex-specific tests for males and females. X-APL can also perform single-marker and haplotype analyses. X-APL is the only method we are aware of testing haplotypes for association on the X chromosome.

We also used simulations to demonstrate that X-APL has a correct type I error rate for single-marker and haplotype tests. Separate tests for males and females were also shown to have a correct type I error rate. Simulations showed that X-APL consistently has more power than other family-based association methods for a single marker on the X chromosome such as XS-TDT, XPDT, and XRCTDT. We also examined the robustness of X-APL toward the departure of allele and haplotype frequencies from the HWE assumption. Similar patterns were observed in X-APL and APL. X-APL is robust for the departure of allele frequencies from HWE for single-marker test. For the global haplotype test, the departure of haplotype frequencies from HWE can inflate the type I error rate. X-APL was also applied to a real data set that contains families with Parkinson disease. A significant marker was identified by X-APL in the candidate gene MAOB for the female-specific test, but not in the overall test.

We also explored the correlation between the linkage and association statistics. Our simulation and theoretical results showed that when there is either no linkage or no association, the linkage and association statistics are not correlated. When there is both linkage and association, the two statistics are positively correlated. Hence, we concluded that when both tests are applied on the same data, the statistical validity for both tests is not compromised. Moreover, when there is linkage and association, the conditional power can be higher when one test is performed based on the results of the other test.

Some future extensions are possible for the APL and X-APL test. APL and X-APL test for association in nuclear families. However, in extended pedigrees in which several related nuclear families have affected siblings, using the overall extended pedigrees can be more informative. The APL and X-APL can be potentially extended to extended pedigrees using a similar strategy in PDT. The extended pedigrees can be partitioned into several related nuclear families and the statistic is calculated based on the sum of the statistics from the related nuclear families. Since APL infers missing parental genotypes in the nuclear families, the correlation between the related nuclear families should be accounted for in the variance estimate. This strategy needs to be investigated for APL and X-APL through theoretical proof and computer simulations.

Covariates are often used in population case-control studies to increase the power of finding disease genes. For example, for late-onset disease, age can be used as a covariate in the

regression model for the disease [Pankratz et al., 2006]. For family-based designs for association, covariates can also be included for analysis. However, family structures and the correlation between individuals in each family should be considered when creating the models with covariates [Pfeifer et al., 2001; Neuhaus et al., 2006]. Schaid [1996] proposed a score test for triad family structures that generalized the TDT statistic. Liu et al. [2002] generalized the score test proposed in Schaid [1996] that can incorporate covariates in the test. Millstein et al. [2005] proposed a likelihood-ratio test that can handle families with two affected siblings with covariates. APL and X-APL can be extended to include covariates into analysis based on the methods developed above. The IBD parameters estimated in APL and X-APL can be used to model the correlation between affected siblings. Other approaches such as the Ordered-Subset Analysis (OSA) [Hauser et al., 2004] proposed for linkage analysis may also be applied in APL and X-APL. The OSA can find the optimal subset based on covariates that provide maximum information of linkage. The power before and after including covariates into analysis should be compared.

As mentioned in Chapters 2 and 3, the allele frequency and IBD parameters for the haplotype test are estimated under the global null hypothesis that none of the haplotypes are associated with the disease. Hence, the global tests for all haplotypes are provided in APL and X-APL. Separate tests for each haplotype are not appropriate in this design framework. However, separate tests for each haplotype can be more informative than the global test if one single haplotype has a strong effect on the disease. In order to perform the separate test, the

parameters need to be estimated based on the observed model. This strategy requires further investigation.

References

Abecasis GR, Cookson WOC, Cardon LR (2000) Pedigree tests of transmission disequilibrium. Eur J Hum Genet 8:545–551.

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genet 30:97-101.

Agresti A (2002) Categorical data analysis. Wiley, New York.

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P; International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(7063):1299-320.

Antonarakis SE, McKusick VA (2000) OMIM passes the 1,000-disease-gene mark. Nat Genet 25(1):11.

Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nat Rev Genet 7(4):277-82.

Bass M, Martin ER, Hauser ER (2004) Pedigree generation for analysis of genetic linkage and association. Pac Symp Biocomput 93-103.

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2:85-97.

Botstein D, White DL, Skolnick M, Davis RW (1980) Constructing of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331.

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33:228-237.

Cardon LR and Bell JI. (2001) Association study designs for complex diseases. Nature Rev Genet 2:91-99.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429(6990):446-52.

Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177.

Connelly JJ, Wang T, Cox JE, Haynes C, Wang L, Shah SH, Crosslin DR, Hale AB, Nelson S, Crossman DC, Granger, CB, Haines JL, Jones, CJ, Vance JM, Goldschmidt PJ, Kraus WE, Hauser ER, Gregory SG (2006) GATA2 is associated with familial early-onset coronary artery disease. PLoS Genet 2(8): e139.

Cooperative Human Linkage Center (1994) A comprehensive human linkage map with centimorgan density. Science 265:2049-2054.

Cornelisse CJ, Cornelis RS, Devilee P (1996) Genes responsible for familial breast cancer. Pathol Res Pract 192(7):684-93.

Costa P, Checkoway H, Levy D, Smith-Weller T, Franklin GM, Swanson PD, Costa LG (1997) Association of a polymorphism in intron 13 of the monoamine oxidase B gene with Parkinson disease. Am J Med Genet 74(2):154-156.

Curtis D (1996) Genetic dissection of complex traits. Nat Genet 12:356357.

Deng HW, Shen H, Xu FH, Deng HY, Conway T, Zhang HT, Recker RR (2002) Tests of linkage and/or association of genes for vitamin D receptor, osteocalcin, and parathyroid hormone with bone mineral density. J Bone Miner Res. 17(4):678-86.

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

Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60(3):155-66.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P et al. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380(6570):152-4.

Ding J, Lin S, Liu Y (2006) Monte carlo pedigree disequilibrium test for markers on the X chromosome. Am J Hum Genet 79(3):567-573.

Dobyns WB (2006) The pattern of inheritance of X-linked traits is not dominant or recessive, just X-linked. Acta Paediatr Suppl 95(451):11-15.

Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25:115-121.

Efron B, Tibshirani RJ (1993) An introduction to the Bootstrap. Chapman & Hall/CRC.

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision and admixture. Am J Hum Genet 57:455–64.

Ewens WJ, Spielman RS (2005) What is the significance of a significant TDT? Human Heredity 60:206-210.

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy, reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227-233.

Farrer MJ (2006) Genetics of Parkinson disease: paradigm shifts and future prospects. Nat Rev Genet 7(4):306-18.

Feller W (1971) An Introduction to Probability Theory and Its Applications, Vol. 2, 3rd ed. New York: Wiley.

Fentiman IS, Fourquet A, Hortobagyi GN (2006) Male breast cancer. Lancet 367(9510):595-604.

Fisher RA (1935a) The detection of linkage with dominant abnormalities. Ann Eugen 6:187-201.

Fisher RA (1935b) The detection of linkage with recessive abnormalities. Ann Eugen 6:339-351.

Frank TS, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, Gumpper KL, Scholl T, Tavtigian SV, Pruss DR, Critchfield GC (2002) Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals. J Clin Oncol 20(6):1480-90.

Gartler SM (1983) Mammalian X-chromosome inactivation. Ann Rev Genet 17:155-90.

Gillander EM, Pearson JV, Sorant AJM, Trent JM, O'Connell JR, Bailey-Wilson JE (2006) Am J Hum Genet 79:458-468.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 37(5):549-54.

Hauser ER, Mooser V, Crossman DC, Haines JL, Jones CH, Winkelmann BR, Schmidt S, et al. (2003) Design of the genetics of early onset cardiovascular disease (GENECARD) study. Am Heart J 145(4):602-613.

Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, Colditz G, Hinney A, Hebebrand J, Koberwitz K, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn JN, Laird NM, Lenburg ME, Lange C, Christman MF (2006) A common genetic variant is associated with adult and childhood obesity. Science. 312(5771):279-83.

Hirschhorn JN and Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6(2):95-108.

Ho SL, Kapadi AL, Ramsden DB, Williams AC (1995) An allelic association study of monoamine oxidase B in Parkinson's disease. Ann Neurol 37(3):403-405

Horvath SM and Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886-1897.

Horvath S, Laird NM, Knapp M (2000) The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. Am J Hum Genet 66(3):1161-7.

Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM (2004) Family-based tests fro associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol 26:61-69.

International HapMap Consortium (2003) The International HapMap project. Nature 426:689-796.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-45.

Kang SJ, Scott WK, Li YJ, Hauser M, van der Walt JM, Fujiwara K, Vance JM, Martin ER (2006) Family based case-control study of MAOA and MAOB polymorphisms in Parkinson disease. Mov Disord. In press.

Kaplan NL, Martin ER, Weir BS (1997) Power studies for the transmission/disequilibrium tests with multiple alleles. Am J Hum Genet 60(3):691-702.

Knapp M (1999) The transmission/disequilibrium test and parental genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. Am J Hum Genet 64:861-870.

Kong A and Cox NJ (1997) Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61: 1179-1188.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347-1363.

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21-24.

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27:234-236.

Kurth JH, Kurth MC, Poduslo SE, Schwankhaus JD (1993) Association of a monoamine oxidase B allele with Parkinson's disease. Ann Neurol 33(4):368-372

Laird NM, Horvath S, Xu X (2000) Implementing a unified approach to family-based tests of association. Genet Epidemiol 19 Suppl 1:S36-42.

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

Lake SL, Blacker D, Laird NM (2000) Family-based tests of association in the presence of linkage. Am J Hum Genet 67:1515-1525.

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265(5181):2037-48.

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241247.

Liu Y, Tritchler D, Bull SB (2002) A unified framework for transmission-disequilibrium test analysis of discrete and continuous traits. Genet Epidemiol. 22:26-40.

Lyon M (1961) Gene action in the X-chromosome of the mouse (Mus musculus L). Nature 190:372-373.

Mansur AH, Bishop DT, Holgate ST, Markham AF, Morrison JFJ (2004) Linkage/association study of a locus modulating total serum IgE on chromosome 14q13–24 in families with asthma. Thorox 59: 876-882.

McQueen MB, Murphy A, Kraft P, Su J, Lazarus R, Laird NM, Lange C, Steen KV (2005) Comparison of linkage and association strategies for quantitative traits using the COGA dataset. BMC Genet 6:s96.

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61:439–448.

Martin ER, Monks SA, Warren LL, Kaplan NL (2000a) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 67:146-154.

Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM. (2000b) SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am J Hum Genet 67(2):383-94.

Martin ER, Bass MP, Hauser ER, Kaplan NL (2003) Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Hum Genet 73:1016-1026.

McGinnis R, Shifman S, Darvasi A (2002) Power and efficiency of the TDT and case-control design for association scans. Behav Genet 32(2):135-144.

Monks SA, Kaplan NL, Weir BS (1998) A comparative study of sibship tests of linkage and/or association. Am J Hum Genet 63:1507–1516.

Morgan TH. (1910) Sex-linked inheritance in Drosophila. Science 32:120-2.

Morris AP, Curnow RN, Whittaker JC (1997) Randomization tests of disease-marker associations. Ann Hum Genet 61:47-58.

Morris RW, Kaplan NL (2003) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23(3):221-33.

Morton NE and MacLean CJ (1974) Analysis of family resemblance. 3. Complex segregation of quantitative traits. Am J Hum Genet 26(4):489-503.

Morton NE (1998) Significance Levels in Complex Inheritance. Am J Hum Genet 62:690-697.

Neuhaus JM, Scott AJ, Wild CJ (2006) Family-specific approaches to the analysis of case-control family data. Biometrics 62(2):488-494.

Oliveira SA, Li YJ, Noureddine MA, Zuchner S, Qin X, Pericak-Vance MA, Vance JM (2005) Identification of risk and age-at-onset genes on chromosome 1p in Parkinson disease. Am J Hum Genet 77(2):252-64.

Ormiston W (1996) Hereditary breast cancer. Eur J Cancer Care 5(1):13-20.

Owen MJ, Craddock N and O'Donovan MC (2005) Schizophrenia: genes at last? Trends in genetics 21(9):518-25.

Pericak-Vance MA, Bebout JL, Gaskell PC, Yamaoka LH, Hung WY, Alberts MJ, Panguluri RC, Long LO, Chen W, Wang S, Coulibaly A, Ukoli F, Jackson A, Weinrich S, Ahaghotu C, Isaacs W, Kittles RA (2004) COX-2 gene promoter haplotypes and prostate cancer risk. Carcinogenesis 25(6):961-6.

Pankratz N, Nichols WC, Uniacke SK, Halter C, Murrell J, Rudolph A, Shults CW, et al. (2003) Genome-wide linkage analysis and evidence of gene-by-gene interactions in a sample of 362 multiplex Parkinson disease families. Hum Mol Genet 12(20):2599-608.

Pankratz N, Byder L, Halter C, Rudolph A, Shults CW, Conneally PM, Foroud T, Nichols WC (2006) Presence of an APOE4 allele results in significantly earlier onset of Parkinson's disease and a higher risk with dementia. Mov Disord 21(1):45-49.

Penrose LS (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugen 6:133-138.

Pfeifer RM, Gail MH, Pee D (2001) Inference for covariates that accounts for ascertainment and random genetic effects in family studies. Biometrika 88**:**933–948.

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

Rao CR (1971) Generalized inverse of matrices and its applications. Wiley, New York, 1971.

Risch N, Botstein D (1996) A manic depressive history. Nat Genet 12:351353.

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

Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using Linkage Genome Scans to Improve Power of Association in Genome Scans. Am J Hum Genet 78:243-252.

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M (2005) The DNA sequence of the human X chromosome. Nature 434:325-337.

Schmidt M, Hauser ER, Martin ER, Schmidt S (2005) Extension of the SIMLA package for generating pedigrees with complex inheritance patterns: environmental covariates, gene-gene and gene-environment interaction. Stat Appl Genet Mol Biol 4(1).

Schork NJ, Gardner JP, Zhang L, Fallin D, Thiel B, Jakubowski H, Aviv A (2002) Genomic association/linkage of sodium lithium countertransport in CEPH pedigrees. Hypertension. 40(5):619-28.

Schwab SG, Hallmayer J, Freimann J, BeLerer B, Albus M, Borrmann-Hassenbach M, Segman RH, Trixler M, Rietschel M, Maier W, Wildenauer DB (2002) Investigation of linkage and association/linkage disequilibrium of HLA A-, DQA1-, DQB1-, and DRB1-alleles in 69 sib-pair- and 89 trio-families with schizophrenia. Am J Med Genet 114(3):315-20.

Scott WK, Nance MA, Watts RL, Hubble JP, Koller WC, Lyons K, Pahwa R, et al. (2001) Complete genomic screen in Parkinson disease: evidence for multiple genes. JAMA 286(18):2239-2244.

Shahedi K, Lindstrom S, Zheng SL, Wiklund F, Adolfsson J, Sun J, Augustsson-Balter K, Chang BL, Adami HO, Liu W, Gronberg H, Xu J (2006) Genetic variation in the COX-2 gene and the association with prostate cancer risk. Int J Cancer 119(3):668-72.

Shao Y, Wolpert CM, Raiford KL, Menold MM, Donnelly SL, Ravan SA, Bass MP, et al. (2002) Genomic screen and follow-up analysis for autistic disorder. Am J Med Genet 114(1):99-105.

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin region and insulin-dependent diabetes mellitus. AmJ Hum Genet 52:506–516.

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458.

Syvanen AC (2006) Toward genome-wide SNP genotyping. Nat Genet 37 Suppl:S5-10.

Tabor HK, Risch NJ, Myers RM (2002) Opinion: Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet 3(5):391-7.

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933.

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299-1320.

Todd JA (2001) Human genetics. Tackling common disease. Nature 411(6837):537-539.

Vallender EJ, Pearson NM, Lahn BT (2005) The X chromosome: not just her brother's keeper. Nat Genet 37:343-345.

van der Walt JM, Noureddine MA, Kittappa R, Hauser MA, Scott WK, McKay R, Zhang F, Stajich JM, Fujiwara K, Scott BL, Pericak-Vance MA, Vance JM, Martin ER (2004) Fibroblast growth factor 20 polymorphisms and haplotypes strongly influence risk of Parkinson disease. Am J Hum Genet 74(6):1121-7.

Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C (2005) Genomic screening and replication using the same data set in family-based association testing. Nat Genet. 37(7):683-91.

Vincent JB, Melmer G, Bolton PF, Hodgkinson S, Holmes D, Curtis D, Gurling HM (2005) Genetic linkage analysis of the X chromosome in autism, with emphasis on the fragile X region. Psychiatr Genet 15(2):83-90.

Wang WYS, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6(2):109-18.

Wang S, Detera-Wadleigh SD, Coon H, Sun CE, Goldin LR, Duffy DL, Byerley WF, Gershon ES, Diehl SR (1996) Evidence of linkage disequilibrium between schizophrenia and the SCa1 CAG repeat on chromosome 6p23. Am J Hum Genet 59(3):731-6.

Weeks D, Lathrop G (1995) Polygenic disease: methods for mapping complex disease traits. Trends Genet 11:513–519.

Weinberg C (1999) Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 64:1186–1193.

Weiss KM, Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? Nat Genet 26(2):151-7.

Whittemore AS and Tu IP (1998) Simple, robust linkage tests for affected sibs. Am J Hum Genet. 62:1228-1242.

Wu RM, Cheng CW, Chen KH, Lu SL, Shan DE, Ho YF, Chern HD (2001) The COMT L allele modifies the association between MAOB polymorphism and PD in Taiwanese. Neurology 56(3):375-382.