ABSTRACT

CZIKA, WENDY ANN. Accounting for Within- and Between-Locus Dependencies in Marker Association Tests. (Advisor: Dr. Bruce Spencer Weir)

The importance of marker association tests has recently been established for locating disease susceptibility genes in the human genome, attaining finer-scaled maps than the linkage variety of tests through the detection of linkage disequilibrium (LD). Many of these association tests were originally defined for biallelic markers under ideal assumptions, with multiallelic extensions often complicated by the covariance among genotype or allele proportions. The well-established allele and genotype case-control tests based on Pearson chi-square test statistics are exceptions since they adapt easily to multiallelic versions, however each of these has its shortcomings. We demonstrate that the multiallelic trend test is an attractive alternative that lacks these limitations. A formula for marker genotype frequencies that incorporates the coefficients quantifying various disequilibria is presented, accommodating any type of disease model. This enables the simulation of samples for estimating the significance level and calculating sample sizes necessary for achieving a certain level of power.

There is a similar complexity in extending the family-based tests of association to markers with more than two alleles. Fortunately, the nonparametric sibling disequilibrium test (SDT) statistic has a natural extension to a quadratic form for multiallelic markers. In the original presentation of the statistic however, information from one of the marker alleles is needlessly discarded. This is necessary for the parametric form of

the statistic due to a linear dependency among the statistics for the alleles, but the non-parametric representation eliminates this dependency. We show how a statistic making use of all the allelic information can be formed.

Obstacles also arise when multiple loci affect disease susceptibility. In the presence of gene-gene interaction, single-marker tests may be unable to detect an association between individual markers and disease status. We implement and evaluate tree-based methods for the mapping of multiple susceptibility genes. Adjustments to correlated p-values from markers in LD with each other are also examined. This study of epistatic gene models reveals the importance of three-locus disequilibria of which we discuss various statistical tests.

ACCOUNTING FOR WITHIN- AND BETWEEN-LOCUS DEPENDENCIES IN MARKER ASSOCIATION TESTS

by

WENDY ANN CZIKA

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

DEPARTMENT OF STATISTICS

Raleigh, NC

2003

APPROVED BY:

Dennis D. Boos	**David A. Dickey**
Dahlia M. Nielsen	**Bruce S. Weir** **Chair of Advisory Committee**
Russell D. Wolfinger	

To Greg

BIOGRAPHY

Personal

- Born April 29, 1974 in Bryn Mawr, Pennsylvania
- Married Gregory J. Czika on October 9, 1999

Education

- B.S. in Mathematics with Computer Science minor, Villanova University, May 1996
- M.Stat, North Carolina State University, May 1998
- Ph.D. in Statistics, North Carolina State University, 2003

  Advisor: Dr. Bruce Weir

  Dissertation: *Accounting for Within- and Between-Locus Dependencies in Marker Association Tests*

Professional Experience

- SAS Programming Intern, Glaxo Wellcome, Research Triangle Park, NC

  August 1996 – August 1997
- Software Tester, SAS Institute, Cary, NC

  May 1996 – December 1999
- Software Developer, SAS Institute, Cary, NC

  December 1999 – present

ACKNOWLEDGEMENTS

I cannot thank my advisor Bruce Weir enough for his endless knowledge, patience, wisdom, and tact. His guidance has been priceless, and it truly has been a pleasure and honor to work with him.

I greatly appreciate all the support and encouragement I received from my committee member and manager, Russell Wolfinger. He has been a wonderful role model for me.

For getting me started in the field of statistical genetics, I owe much thanks to Dahlia Nielsen for giving me the background I needed and answering countless questions and to John Brocklebank for involving me in my first statistical genetics project.

I am very gracious to my colleague at SAS Jack Berry for his insights and collaboration on the mSDT.

Thank you Buffy Hudson-Curtis, Jennifer Gauvin, and George Capuano for helping me get through Measure Theory and the prelim!

And finally, I want to thank all my family and friends who kept believing in me, even after all this time, and especially my husband for tolerating and supporting me throughout this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the landmark paper of Risch and Merikangas (1996) advocating the use of association studies over linkage analysis for locating complex disease genes on a fine-scale genetic map, much discussion about the two classes of association tests, case-control and family-based, has surfaced. These analyses test for a statistical association between the disease and marker genotypes. Since the location of the disease locus is unknown and thus the genotypes at that locus are unobserved, disease phenotype (affected/unaffected with disease) is used as a surrogate for the genotype. Any association found between the disease phenotype and marker genotypes is hoped to be due to linkage disequilibrium between the two loci, that is an association due to linkage. These two association test categories differ in the individuals comprising the samples, with case-control samples containing random samples of individuals from a population and family-based studies including nuclear families with at least one child affected with the disease of interest.

Family-based tests have a clear-cut advantage over case-control tests "because they offer complete robustness to potential population heterogeneity" (Risch and Teng 1998); however, when parental genotypes are unavailable, case-control tests can be substantially more powerful than statistics using unaffected siblings as controls (Risch and Teng 1998; Teng and Risch 1999) while being more easily and conveniently collected. We examine how various types of associations or dependencies among alleles at the same locus or different loci, described below, can affect tests of both sorts.

Many of these association tests were originally defined for biallelic markers under ideal assumptions. The form of these tests is often quite straightforward; a single allele can be used in the analysis, with each yielding the same statistic due to the two-sided nature of the tests and the complete negative correlation between the two allele proportions. However, the natural negative covariance among allele or genotype counts complicates the generalization of such tests to multiallelic versions for markers such as microsatellites; tests cannot be performed on each allele or genotype independently. Additionally, the pair of alleles an individual possesses at a locus may be dependent due to the existence of Hardy-Weinberg disequilibrium in the population; this can be caused by forces such as nonrandom mating or migration. Both of these types of within-locus dependencies can have an effect on the test statistic.

Linkage disequilibrium (LD) is a dependency between alleles at linked loci that facilitates association-based tests. The strength of this between-locus association affects the power of these tests. In addition, markers in LD with each other may have p-values that

are correlated; accounting for these dependencies can serve to remove spurious significant associations. Another sort of dependency to consider is the interaction between genes on disease susceptibility. For diseases that are affected by multiple loci, the effect is not always an additive one; epistasis, or a gene-gene interaction, can occur between genes. Tests applied to single markers at a time may not capture the relationship between multiple loci and a disease, so methods for detecting the multi-gene association with disease status are advantageous in this situation.

In Chapter 2, the multiallelic trend test, extended from the original Armitage linear trend test (1955) by Slager and Schaid (2001b), is examined. This test statistic uses the multinomial distribution of genotypes and the covariances induced by this distribution. First we perform an algebraic comparison of this test to the allele case-control test, which is based on the Pearson chi-square statistic for a contingency table. We also look at the behavior of this test statistic when independent transmission of alleles (HWE) at a marker does not hold, under both the null and alternative hypotheses. Sample sizes for attaining a predefined level of power are calculated taking varying amounts of LD, both gametic and nongametic, into account in the model. In order to perform these analyses, an extension to the equations for marker genotype frequencies presented by Nielsen and Weir (1999) is given, now including single-locus and LD coefficients and applicable for any type of genetic model for the disease.

Association mapping utilizing decision trees, or recursive partitioning, is presented in Chapters 3 and 4. In Chapter 3, the results from association mapping with trees

are compared with those from standard association tests. The approach for combining correlated p-values from sliding windows of markers (Zaykin et al. 2002) in order to smooth out the random noise in the p-values is also examined. Chapter 4 demonstrates a novel approach to analyzing marker data along with other covariates, consisting of tree-based and other data-mining methods combined with conventional statistical and family-based marker tests, again in an attempt to map multiple disease susceptibility genes.

Another alternative to single-marker tests are haplotype-based case-control tests. These test statistics involve both the two-locus allelic associations between individual markers and the disease as well as three-locus haplotypic associations that capture the disequilibrium between two markers and the disease (Nielsen and Weir 2001). Different methods for testing whether the three-locus disequilibria are significant are described in Chapter 5 along with a comparison between an exact method that is introduced and an asymptotic approach.

Chapter 6 introduces a modification of the test statistic for the sibling disequilibrium test (SDT) (Horvath and Laird 1998). The SDT is a nonparametric test that evolved from the popular family-based tests, the transmission/disequilibrium test (TDT) (Spielman, McGinnis, and Ewens 1993) and sib-TDT (S-TDT) (Spielman and Ewens 1998). These tests of linkage and association between a marker and disease locus use the parental genotypes not transmitted (in the TDT) or genotypes of unaffected siblings (in the S-TDT) as controls for affected children. Since they are joint tests of both linkage and associa-

tion, associations between unlinked loci due to population stratification are not detected, an advantage over case-control tests. However, for these tests to remain valid tests of association, only one affected child per family in the TDT and one discordant sibling pair per family in the S-TDT can be used; the SDT can take advantage of the genotypic information from all children in a nuclear family while still testing for association in the presence of linkage. These tests are ideal in their biallelic form, though extensions for multiallelic markers exist (see Spielman and Ewens (1996); Kaplan, Martin, and Weir (1997); Monks, Kaplan, and Weir (1998); and Curtis, Miller, and Sham (1999)). The multiallelic version of the SDT by Horvath and Laird (1998) is based on the multivariate component sign test (Bickel 1965; Randles 1989) that results in a quadratic form with an asymptotic $\chi^2_{(m-1)}$ distribution for a marker with m alleles. A disadvantage of this test is the omission of an arbitrary allele due to a linear dependency among certain statistics for the marker alleles; however, the function of these statistics that is used in the quadratic form removes this dependency, rendering m possible test statistics depending on which allele is dropped. This can be resolved using the test statistic proposed in Chapter 5.

Chapter 2

Properties of the Multiallelic Trend Test

Czika, W. and B. S. Weir (2003). Submitted to *Biometrics*.

2.1 Abstract

Disease genes can be mapped on the basis of associations between genetic markers and disease status, with the case-control design having the advantage of not requiring individuals from different generations. When the marker loci have multiple alleles, there has been debate on whether the power of tests for association increases or decreases. We show here that the multiple-allele version of Armitage's trend test can have increased power over the two-allele version. The trend test has the advantage of remaining valid even when the sampled population is not in Hardy-Weinberg equilibrium. A departure from Hardy-Weinberg proportions means that association tests depend on gametic and nongametic linkage disequilibrium between marker and disease loci, and we illustrate the magnitude of these effects with simulated data.

2.2 Introduction

Chi-square tests based on simple contingency tables are useful tools for the association mapping of disease genes. These tables consist of rows representing those affected with the disease (cases) and those not affected (controls), and columns containing either allele or genotype counts at the marker(s) of interest. The allele and genotype contingency table tests can accommodate markers with any number of alleles. With the allele statistic testing for additive effects of alleles and the genotype statistic testing for both additive and allelic interaction (dominance) effects (Nielsen and Weir 1999), the genotype statistic

can have a χ^2 distribution with up to $m(m-1)/2$ more degrees of freedom (df) than the allele statistic for a marker with m alleles. This often gives the allele test an advantage in the power to detect an association between a marker and disease even in the presence of nonadditive allelic effects. The possible sparseness of the genotype contingency table can also lead to the chi-square test not being valid for the genotype test statistic. However, a drawback of the allele test is its reliance on the assumption of allelic independence (Hardy-Weinberg equilibrium, HWE) in the sample collected. Departures from HWE, depending on whether homozygotes or heterozygotes are in excess, can either decrease or increase, respectively, the size of the test; regardless, the test is no longer valid at the nominal level. With the recent attention given to single nucleotide polymorphisms (SNPs), the Armitage linear trend test (1955) has emerged as a desirable method of testing for associations between biallelic markers and disease status, utilized for example by Devlin and Roeder (1999). Like the standard allele test, only additive allelic effects are tested but, like the genotype test, with the benefit of robustness to departures from HWE. Though its extension to multiallelic markers is not as straightforward and natural as for the contingency-table-based tests, Slager and Schaid (2001b) present a multiallelic version of the linear trend test that has the same power advantage of the allele test over the genotype test due to fewer df. We show here that the multiallelic trend test remains valid when Hardy-Weinberg disequilibrium (HWD) exists in the sample, unlike the allele test, and when HWE holds perfectly, the tests are asymptotically equivalent. Departures from HWE are likely to be accompanied by nongametic disequilibrium between alleles

at disease and marker loci, and we show here how this and linkage disequilibrium affect the trend test. We phrase most of the results in terms of the sample sizes required for obtaining specified power levels.

2.3 Comparison of the Allele and Trend Test Statistics

For the biallelic case, Sasieni (1997) represents the linear trend test statistic in terms of the quantities used in Table 2.1. Marker genotype counts are written as r_i and s_i for cases and controls when individuals have i copies of marker allele M_1. The test statistic X_T^2 is

$$X_T^2 = \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{R(N-R)[N(n_1 + 4n_2) - (n_1 + 2n_2)^2]} \tag{2.1}$$

which has an asymptotic $\chi^2_{(1)}$ distribution like the allele test statistic X_A^2. This latter statistic is the contingency-table statistic for the 2×2 table with row entries $2r_0 + r_1$, $r_1 + 2r_2$ for cases and $2s_0 + s_1$, $s1 + 2s_2$ for controls:

$$X_A^2 = \frac{2N[2N(2r_2 + r_1) - 2R(2n_2 + n_1)]^2}{2R(2N-2R)[2N(n_1 + 2n_2) - (n_1 + 2n_2)^2]} \tag{2.2}$$

Sasieni demonstrated that the ratio of X_A^2 to X_T^2 is $1+(4n_0n_2-n_1^2)/[(n_1+2n_2)(n_1+2n_2)]$, which depends only on the counts in the combined sample. Clearly, the test statistics are equal only when HWE holds in the sample ($n_1^2 = 4n_0n_2$). Otherwise, the allele statistic is larger than the valid trend test statistic if there is an excess of homozygotes and smaller

when there is an excess of heterozygotes, meaning that the test will be conservative for heterozygote excess. The allele test is invalid when there is HWD in either cases or controls because in that case the allele numbers are not binomially distributed (Weir 1996). HWD in either cases or controls is expected when the marker is associated with the disease (Nielsen, Ehm, and Weir 1999), but, as Sasieni noted, HWE in both can result in HWD in the combined sample whenever allele frequencies are different in the two groups.

To extend Sasieni's findings to multiallelic markers, we start by looking at different representations of the biallelic trend test statistic. Slager and Schaid (2001b) present the trend test statistic in the form $u^2/\mathrm{Var}(u)$ where $u = \mathbf{x}'[(1-\phi)\mathbf{r} - \phi\mathbf{s}]$, with $\phi = R/N$, $\mathbf{x}' = (0,1,2)$, $\mathbf{r}' = (r_0, r_1, r_2)$, and $\mathbf{s}' = (s_0, s_1, s_2)$ using the terms in Table 2.1. A small amount of algebra shows the equality of the trend test statistic in this form to the one given in Equation 2.1. Using this representation, there is a natural extension to a multiallelic statistic by substituting a matrix $\mathbf{X}$ for the vector $\mathbf{x}$ to calculate a vector $\mathbf{U}$ instead of the scalar u as $\mathbf{U} = \mathbf{X}'[(1-\phi)\mathbf{r} - \phi\mathbf{s}]$. For a marker with m alleles, the elements X_{ij} of the $[m(m+1)/2 \times (m-1)]$ matrix $\mathbf{X}$ are the number of times the jth allele occurs in the ith genotype (either 0, 1, or 2). Note that the mth column has been omitted because of the linear dependence among all m allele frequencies. The $\mathbf{r}$ and $\mathbf{s}$ vectors are similarly extended to include one row for each of the $m(m+1)/2$ possible genotypes containing the genotype counts for the cases and controls, respectively. The variance of $\mathbf{U}$ is calculated using the multinomial distribution of the genotype counts $\mathbf{r}$ and $\mathbf{s}$ (Slager and Schaid

2001b), which have covariance matrices $(\frac{R}{N}\text{diag}(\mathbf{n}) - \frac{R}{N^2}\mathbf{nn}')$ and $(\frac{N-R}{N}\text{diag}(\mathbf{n}) - \frac{N-R}{N^2}\mathbf{nn}')$ respectively under the null hypothesis and the assumption of unrelated individuals in the sample. The vector $\mathbf{n}$ of length $m(m+1)/2$ contains the overall sample counts for each genotype. This allows $\text{Var}(\mathbf{U})$ to be expressed as $\frac{R(N-R)}{N^3}\mathbf{X}'(N\text{diag}(\mathbf{n}) - \mathbf{nn}')\mathbf{X}$. The quadratic form $\mathbf{U}'[\text{Var}(\mathbf{U})]^{-1}\mathbf{U}$ then has an asymptotic χ^2 distribution with $m-1$ df.

The allele test statistic in Equation 2.2 is not conducive to a multiallelic extension. We need a method of calculating the allele test statistic that can be translated from a scalar to matrix format as with the trend test statistic and, as with the biallelic form, will have the same numerator but a different variance from the trend test statistic. The biallelic test statistic can be derived from the difference of two multinomial proportions: $(\tilde{p}_R - \tilde{p}_S)^2/\text{Var}(\tilde{p}_R - \tilde{p}_S)$ since under the null hypothesis of no marker-disease association we are assuming the alleles in cases and controls come from two independent binomial samples, each with probability that can be estimated with $\bar{p}$, the sample frequency of the corresponding allele. In terms of the notation in Table 2.1, $\tilde{p}_R = (2r_2 + r_1)/(2R)$, $\tilde{p}_S = (2s_2 + s_1)/(2S)$, $\bar{p} = (2n_2 + n_1)/(2N)$, and $\text{Var}(\tilde{p}_R - \tilde{p}_S) = \bar{p}(1-\bar{p})/(2R) + \bar{p}(1-\bar{p})/(2S)$.

In matrix notation the chi-square statistic can be expressed as $(\tilde{\mathbf{p}}_\mathbf{R} - \tilde{\mathbf{p}}_\mathbf{S})'[\text{Var}(\tilde{\mathbf{p}}_\mathbf{R} - \tilde{\mathbf{p}}_\mathbf{S})]^{-1}(\tilde{\mathbf{p}}_\mathbf{R} - \tilde{\mathbf{p}}_\mathbf{S})$, a quadratic form with $\mathbf{U}=\tilde{\mathbf{p}}_\mathbf{R} - \tilde{\mathbf{p}}_\mathbf{S}$. The $\tilde{\mathbf{p}}$ vectors can be expressed in terms of the notation of Slager and Schaid (2001b) for the multiallelic trend statistic

using the fact that $\tilde{\mathbf{p}}_{\mathbf{R}} = \frac{1}{2R}\mathbf{X}'\mathbf{r}$, $\tilde{\mathbf{p}}_{\mathbf{S}} = \frac{1}{2S}\mathbf{X}'\mathbf{s} = \frac{1}{2N-2R}\mathbf{X}'(\mathbf{n}-\mathbf{r})$, and

$$\begin{aligned}
\mathrm{Var}(\tilde{\mathbf{p}}_{\mathbf{R}} - \tilde{\mathbf{p}}_{\mathbf{S}}) &= \left(\frac{1}{2R} + \frac{1}{2S}\right)\begin{bmatrix} \tilde{p}_1(1-\tilde{p}_1) & -\tilde{p}_1\tilde{p}_2 & \cdots & -\tilde{p}_1\tilde{p}_{m-1} \\ -\tilde{p}_2\tilde{p}_1 & \tilde{p}_2(1-\tilde{p}_2) & \cdots & -\tilde{p}_2\tilde{p}_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\tilde{p}_{m-1}\tilde{p}_1 & -\tilde{p}_{m-1}\tilde{p}_2 & \cdots & \tilde{p}_{m-1}(1-\tilde{p}_{m-1}) \end{bmatrix} \\
&= \frac{1}{2R(2N)(2N-2R)}(2N\mathrm{diag}(\mathbf{X}'\mathbf{n}) - \mathbf{X}'\mathbf{n}\mathbf{n}'\mathbf{X})
\end{aligned}$$

Then the allele test statistic for multiallelic markers can be put into a similar form as the multiallelic trend test statistic:

$$X_T^2 = \frac{N^3}{R(N-R)}(\mathbf{r}-\phi\mathbf{n})'\mathbf{X}[N(\mathbf{X}'\mathrm{diag}(\mathbf{n})\mathbf{X}) - \mathbf{X}'\mathbf{n}\mathbf{n}'\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{r}-\phi\mathbf{n})$$

$$X_A^2 = \frac{2N^3}{R(N-R)}(\mathbf{r}-\phi\mathbf{n})'\mathbf{X}[2N\mathrm{diag}(\mathbf{X}'\mathbf{n}) - \mathbf{X}'\mathbf{n}\mathbf{n}'\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{r}-\phi\mathbf{n})$$

As desired, we now have the multiallelic allele and trend statistics as quadratic forms with the same $\mathbf{U}$, but potentially different variances. The requirement for the equality of these two variances and thus the two statistics is

$$2N(\mathbf{X}'\mathrm{diag}(\mathbf{n})\mathbf{X}) - \mathbf{X}'\mathbf{n}\mathbf{n}'\mathbf{X} = 2N\mathrm{diag}(\mathbf{X}'\mathbf{n})$$

This condition can be expressed in more familiar terms, the sample allele and genotype frequencies, $\tilde{p}_i$ and $\tilde{P}_{ij}$ respectively for alleles i and j, as

$$\begin{bmatrix} \tilde{p}_1^2 & \tilde{p}_1\tilde{p}_2 & \cdots & \tilde{p}_1\tilde{p}_{m-1} \\ \tilde{p}_2\tilde{p}_1 & \tilde{p}_2^2 & \cdots & \tilde{p}_2\tilde{p}_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{p}_{m-1}\tilde{p}_1 & \tilde{p}_{m-1}\tilde{p}_2 & \cdots & \tilde{p}_{m-1}^2 \end{bmatrix} = \begin{bmatrix} \tilde{P}_{11} & \frac{1}{2}\tilde{P}_{12} & \cdots & \frac{1}{2}\tilde{P}_{1(m-1)} \\ \frac{1}{2}\tilde{P}_{21} & \tilde{P}_{22} & \cdots & \frac{1}{2}\tilde{P}_{2(m-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}\tilde{P}_{(m-1)1} & \frac{1}{2}\tilde{P}_{(m-1)2} & \cdots & \tilde{P}_{(m-1)(m-1)} \end{bmatrix}$$

which implies that all $m(m-1)/2$ independent sample HWD coefficients in the combined data must be zero in order for these two statistics to be equal. As in the biallelic case, without this assumption of HWE, the multinomial distribution of the alleles in each of the case and control samples, which is assumed for the allele test statistic, does not hold.

2.4 Examination of Significance Levels and Sample Sizes

We now present the results of simulating data for the examination of the size of the trend and allele tests under certain conditions, as well as power/sample-size calculations for the trend test under a variety of disequilibrium values, both within-locus and between the marker and disease loci. We assume a biallelic disease locus with alleles A_1 and A_2, the latter being the high-risk allele, with population frequencies p_1 and p_2, respectively. The disease model is defined using terms from Slager and Schaid (2001a): the population prevalence of the disease K; γ_1 and γ_2, the relative risks of genotypes A_1/A_2 and A_2/A_2, respectively, to A_1/A_1, which has penetrance f_{11}.

We consider markers with alleles M_i with population frequencies q_i, $i = 1, ..., m$. As in Chapman and Wijsman (1998), we assume equifrequent alleles in all markers ($q_i = 1/m$ for all i) to maximize heterozygosity. Thus, for any number of alleles at a marker, there are only two distinct linkage disequilibrium (LD) coefficients. Without loss of generality, we assume the initial disease mutation arose on a haplotype with marker allele M_1, and

we assume maximum LD between the disease and marker loci. The LD coefficients D_{ri} between alleles A_r and M_i have values

$$D_{11} = -D_A \quad , \quad D_{21} = D_A$$

$$D_{1i} = -D_B \quad , \quad D_{2i} = D_B, \;\; i \neq 1$$

where $D_A = (m-1)p_2/m, D_B = -p_2/m$. The usual bounds on linkage disequilibria (Weir 1996) imply that $p_2 \leq 1/m$ for this formulation, so that the disease is supposed to be rare.

We deviate from the Chapman and Wijsman (1998) method of calculating the marker allele and genotype frequencies in cases and controls; instead we use a modification of formulas given by Nielsen and Weir (1999) and Nielsen, Ehm, and Weir (1999). We add to their calculation of the marker genotype frequencies in affected and unaffected individuals the following coefficients (Weir 1979): HWD at the marker locus, denoted d_{ij} for genotype M_i/M_j; HWD at the disease locus, denoted d_{rs} for genotype A_r/A_s; the digenic gametic disequilibrium (LD), previously defined as D_{ri}; and the digenic nongametic disequilibrium, denoted $D_{r/i}$ for allele A_r at the disease locus and M_i at the marker locus. This last term refers to alleles received by an individual from different parents. Trigenic and quadrigenic disequilibria are not included in the model. Using the simplification $\sum_{r,s}(p_r p_s + d_{rs})f_{rs} = K$, marker genotype frequencies in cases and controls can then be computed as

$$\Pr(M_i M_j|\text{Aff.}) \;\; = \;\; \Big\{ \sum_{r,s} [(p_r q_i + D_{ri})(p_s q_j + D_{sj}) + p_r p_s d_{ij} + q_i q_j d_{rs} + d_{ij} d_{rs}$$

$$
\begin{aligned}
& +p_r q_j D_{s/i} + p_s q_i D_{r/j} + D_{r/j} D_{s/i}] f_{rs} \Big\} / K \\
= & \begin{cases} q_i^2 + (2q_i \delta_i^c + \delta_{ii}^c)/K + d_{ij}, & i = j \\ 2q_i q_j + 2[q_i \delta_j^c + q_j \delta_i^c + \delta_{ij}^c]/K + d_{ij}, & i \neq j \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
\Pr(M_i M_j | \text{Unaff.}) = & \Big\{ \sum_{r,s} [(p_r q_i + D_{ri})(p_s q_j + D_{sj}) + p_r p_s d_{ij} + q_i q_j d_{rs} + d_{ij} d_{rs} \\
& + p_r q_j D_{s/i} + p_s q_i D_{r/j} + D_{r/j} D_{s/i}](1 - f_{rs}) \Big\} / (1 - K) \\
= & \begin{cases} q_i^2 - (2q_i \delta_i^c + \delta_{ii}^c)/(1 - K) + d_{ij}, & i = j \\ 2q_i q_j - 2[q_i \delta_j^c + q_j \delta_i^c + \delta_{ij}^c]/(1 - K) + d_{ij}, & i \neq j \end{cases}
\end{aligned}
$$

where δ_i^c and δ_{ij}^c are the composite versions of δ_i and δ_{ij} defined by Nielsen and Weir (1999) that now include the digenic nongametic coefficients and can be expressed as $\delta_i^c = \sum_{r,s} p_r f_{rs}(D_{sj} + D_{s/j})$ and $\delta_{ij}^c = \sum_{r,s} f_{rs}(D_{ri} D_{sj} + D_{r/i} D_{s/j})$. Note that the sum $(D_{sj} + D_{s/j})$ has previously (Weir 1979) been written as Δ_{sj}, the composite linkage disequilibrium coefficient. When HWE holds in the whole population, $d_{ij} = d_{rs} = 0$, it is reasonable to assume that the digenic nongametic disequilibria are also zero. In that case, $\delta_i^c = D_{2i} f_{11}[-p_1 + (p_1 - p_2)\gamma_1 + p_2 \gamma_2]$ and $\delta_{ij}^c = D_{2i} D_{2j} f_{11}(1 - 2\gamma_1 + \gamma_2)$. This formulation accommodates any disease model, and allows us to extend the work of Chapman and Wijsman (1998).

2.4.1 Effect of Hardy-Weinberg Disequilibrium on Significance Level

To demonstrate the effect of HWD on the allele test and the immunity of the trend test to the value of these coefficients, we simulate a multiallelic marker that incorporates departures from HWE. We use a marker with three alleles to create 10,000 samples of 5,000 affected and 5,000 unaffected individuals each under the null hypothesis of no marker-disease association. This null hypothesis can equivalently be expressed for these tests as no LD between the marker and disease locus or no effect of the disease locus on the penetrance of the disease. Thus, only single-locus disequilibria need to be considered in the model. Assuming the same HWD coefficient for all heterozygous genotypes, that is, a single within-population inbreeding coefficient f (Weir 1996), we examine the effect of four values of f (its maximum, average of the maximum and minimum, zero, and minimum) on the significance level of the allele and trend tests at the nominal level $\alpha = 0.05$. The marker genotype probabilities for both cases and controls under this null hypothesis can then be expressed as $\Pr(M_i/M_j) = 2(1-f)/9$ for all $i \neq j$ and $\Pr(M_i/M_i) = (1+2f)/9$ since $q_i = 1/3$ for $i = 1, 2, 3$. SAS/IML® and SAS/Genetics™ software are used for these calculations.

As shown in Table 2.2, the size of the trend test remains close to 0.05 regardless of the value of f; however, the allele test is anticonservative when f is positive and conservative when f is negative, confirming that the finding of Sasieni (1997) for the biallelic case holds in the multiallelic case as well ($f > 0$ corresponds to an excess of homozygotes and

$f < 0$ an excess of heterozygotes).

2.4.2 Power and Sample Sizes Approximations Assuming HWE

Slager and Schaid (2001a) calculate power for the biallelic case using the normal distribution; here we deal with the χ^2 distribution of the multiallelic statistic. Under the alternative hypothesis that case and control genotypes are from two independent multinomial distributions with unequal probabilities $\mathbf{P_R}$ and $\mathbf{P_S}$ respectively, the mean of $\mathbf{U}$ is $\mu_\mathbf{1} = N\phi(1-\phi)\mathbf{X}'(\mathbf{P_R} - \mathbf{P_S})$ and the variance of $\mathbf{U}$ is $\mathbf{\Sigma_1} = N\phi(1-\phi)\mathbf{X}'[(1-\phi)(\text{diag}(\mathbf{P_R}) - \mathbf{P_R}\mathbf{P_R}') + \phi(\text{diag}(\mathbf{P_S}) - \mathbf{P_S}\mathbf{P_S}')]\mathbf{X}$. Then X_T^2 has a $\chi^2_{(m-1)}(\lambda)$ distribution with noncentrality parameter $\lambda = \frac{1}{2}\mu_\mathbf{1}'{\mathbf{\Sigma_1}}^{-1}\mu_\mathbf{1}$. Using these quantities, the power of the trend test is calculated as

$$\text{Prob}(X_T^2 > \chi^2_{(m-1),1-\alpha}(0))$$

under the noncentral χ^2 distribution.

Since an analytic formula for sample size is not tractable as in the biallelic trend test (Slager and Schaid 2001a), we present here numerical results from examining the effects of various disease models on the sample size. It is assumed an equal number of cases and controls are sampled with a total sample size N, chosen to achieve 80% power. First, we assume HWE in which case the multiallelic trend test is equivalent to the allele test and we can extend the work of Chapman and Wijsman (1998) using our equations for calculating marker genotype frequencies to show power of models other than fully penetrant recessive and dominant diseases. Complete LD between the disease

and marker loci is assumed so that the coefficients D_{ri} are defined as above. Tables 2.3-2.6 show the sample sizes required for achieving 80% power using the multiallelic trend test. The numbers were obtained using SAS/IML software. Multiplicative, additive, dominant, and recessive disease models were all examined, defined by K, p_2, γ_1, and γ_2; the penetrance f_{11} for genotype A_1/A_1 can then be extracted using these values (Slager and Schaid 2001a). Note that values shown in bold correspond to the values obtained by Slager and Schaid (2001a) although they are consistently off by a negligible 5 (numbers for disease allele frequency other than 0.50 cannot be compared since the same marker and disease locus is assumed in their work).

We also look at the effect of LD decay over time. Using the model defined by Chapman and Wijsman (1998), our LD coefficients can now be defined as

$$\begin{aligned} D_A &= (m-1)p_2(1-\theta)^t/m \\ D_B &= -p_2(1-\theta)^t/m \end{aligned}$$

For the sake of space, we show in Tables 2.7-2.10 only the results when $\theta = 0.005$ (1 cM) and $t = 40$ generations, corresponding to a loss of about 18% from complete disequilibrium.

2.4.3 Power and Sample Sizes Accounting for Hardy-Weinberg and Nongametic Disequilibrium

As mentioned, examining the power of the multiallelic trend test in the absence of departures from HWE is essentially the same as the power for the allele test. However,

introducing HWD requires a consideration of the digenic nongametic disequilibria as well, complicating the model vastly. Thus we limit our study of the power of the multiallelic trend test with the inclusion of HWD to markers with three alleles, the simplest multiallelic case. We continue to ignore higher-order (trigenic and quadrigenic) disequilibria and keep LD at its maximum, with no decay. We consider for the three-allele case the following as in our significance level calculations: the maximum, average of maximum and minimum, and minimum values for HWD (again assuming the same coefficients for all heterozygous genotypes and the same for all homozygous genotypes) at both the marker and disease locus. Various values for the digenic nongametic coefficients are considered as well. At their maxima, the nongametic coefficients equal the LD coefficients, that is $D_{r/i} = D_{ri}$ and thus the values of the between-locus disequilibrium are twice that of the usual gametic disequilibrium. We also used $D_{r/i} = D_{ri}/2$, half the maximum value. The effects of these disequilibria are shown in Table 2.11. We examine only a single disease model under these specifications: an additive model with $p_2 = 0.1$, $K = 0.01$, $\gamma_1 = 2$, and $\gamma_2 = 3$. From Table 2.11 we can see that the sample sizes required to have 80% power in the multiallelic trend test vary drastically according to the values of the Hardy-Weinberg and digenic nongametic disequilibrium coefficients.

2.5 Discussion

We have shown algebraically that the requirement for the equality of the multiallelic allele and trend test statistics is consistent with that shown by Sasieni (1997) for the biallelic

statistics: the combined sample must have genotype frequencies satisfying the Hardy-Weinberg law. Simulations also confirm the extension of Sasieni's findings on the effect of HWD on the significance level of the allele test to the multiallelic case. Though this effect was expected, the magnitude of the effect was not; even nonsignificant departures from HWE can invalidate the allele test at the nominal significance level. The multiallelic trend test, with the same df as the allele test and same test of hypothesis, is thus a clear choice over the allele test for detecting a significant association between a marker and disease locus.

Additionally, we have extended the work of Chapman and Wijsman (1998) to demonstrate the power advantage of using multiallelic markers, with alleles as equifrequent as possible to maximize heterozygosity, over biallelic SNPs. These findings come as somewhat of a surprise as well, especially for disease models such as an additive model where the effect is relatively small; it is often thought that a multiallelic test will not perform well due to the effect being swamped by the alleles in negative LD with the disease allele. Sample sizes for the multiallelic trend test were examined looking at a variety of disease models assuming different conditions for the within-locus and between-locus disequilibrium coefficients. The dramatic effects that HWD and between-locus disequilibrium can have on power and sample sizes were demonstrated, substantiating the importance of estimating these coefficients prior to testing for association with the disease.

2.6 Acknowledgements

This work was supported in part by NIH Grant GM 45344.

2.7 Tables

Table 2.1: Contingency Table of Genotypes for Case-Control Sample

	Number of M_1 alleles			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Table 2.2: Estimated Significance Level of Tests

f	Trend Test	Allele Test
1	0.0529	0.2267
0.25	0.0490	0.0927
0	0.0474	0.0472
-0.5	0.0511	0.0030

Table 2.3: Sample Sizes for Multiplicative Models with HWE and Complete LD

				Number of Alleles at Marker Locus								
γ_1	γ_2	K	p_2	2	3	4	5	6	7	8	9	10
2	4	0.01	0.01	156939	96791	73371	60530	52294	46505	42185	38821	36119
			0.10	1854	1185	929	792	706	647	604	571	546
			0.50	**131**								
		0.1	0.01	129701	79952	60576	49950	43132	38339	34761	31974	29734
			0.10	1532	974	760	645	573	523	487	459	437
			0.50	**108**								
3	9	0.01	0.01	40010	24790	18878	15645	13576	12126	11047	10210	9539
			0.10	547	359	288	251	229	213	203	195	189
			0.50	**54**								
		0.1	0.01	33065	20467	15570	12891	11176	9973	9077	8382	7824
			0.10	451	293	234	202	183	169	160	153	148
			0.50	**44**								

Table 2.4: Sample Sizes for Additive Models with HWE and Complete LD

				Number of Alleles at Marker Locus								
γ_1	γ_2	K	p_2	2	3	4	5	6	7	8	9	10
2	3	0.01	0.01	160054	98703	74814	61715	53312	47406	42999	39567	36809
			0.10	2201	1397	1090	924	819	747	694	654	623
			0.50	**231**								
		0.10	0.01	132276	81532	61769	50929	43974	39084	35433	32590	30304
			0.10	1819	1150	893	754	667	606	561	527	501
			0.50	**191**								
3	5	0.01	0.01	41587	25758	19608	16244	14091	12582	11459	10587	9888
			0.10	739	477	378	324	291	269	253	241	231
			0.50	**124**								
		0.10	0.01	34369	21267	16174	13387	11602	10350	9418	8694	8113
			0.10	611	391	308	263	235	216	202	191	183
			0.50	**102**								

Table 2.5: Sample Sizes for Dominant Models with HWE and Complete LD

				Number of Alleles at Marker Locus								
γ_1	γ_2	K	p_2	2	3	4	5	6	7	8	9	10
2	2	0.01	0.01	163264	100673	76300	62935	54361	48335	43837	40335	37520
			0.10	2664	1681	1302	1097	968	878	812	762	722
			0.50	**686**								
		0.10	0.01	134930	83161	62997	51938	44841	39851	36126	33225	30892
			0.10	2202	1385	1070	899	790	715	660	617	583
			0.50	**571**								
3	3	0.01	0.01	42411	26264	19990	16557	14361	12821	11674	10784	10071
			0.10	883	565	443	378	337	309	288	273	261
			0.50	**332**								
		0.10	0.01	35051	21686	16490	13646	11825	10547	9596	8856	8264
			0.10	730	464	362	307	273	249	231	218	208
			0.50	**277**								

Table 2.6: Sample Sizes for Recessive Models with HWE and Complete LD

				Number of Alleles at Marker Locus								
γ_1	γ_2	K	p_2	2	3	4	5	6	7	8	9	10
1	2	0.01	0.01	$1.54\text{x}10^9$	$9.45\text{x}10^8$	$7.13\text{x}10^8$	$5.85\text{x}10^8$	$5.03\text{x}10^8$	$4.45\text{x}10^8$	$4.02\text{x}10^8$	$3.68\text{x}10^8$	$3.41\text{x}10^8$
			0.10	157700	97725	74429	61688	53538	47827	43577	40279	37637
			0.50	**408**								
		0.10	0.01	$1.27\text{x}10^9$	$7.81\text{x}10^8$	$5.89\text{x}10^8$	$4.84\text{x}10^8$	$4.16\text{x}10^8$	$3.68\text{x}10^8$	$3.32\text{x}10^8$	$3.04\text{x}10^8$	$2.82\text{x}10^8$
			0.10	130266	80645	61361	50809	44055	39320	35794	33055	30860
			0.50	**334**								
1	3	0.01	0.01	$3.85\text{x}10^8$	$2.36\text{x}10^8$	$1.78\text{x}10^8$	$1.46\text{x}10^8$	$1.26\text{x}10^8$	$1.11\text{x}10^8$	$1.01\text{x}10^8$	$9.22\text{x}10^7$	$8.54\text{x}10^7$
			0.10	40391	25258	19407	16224	14198	12787	11743	10938	10298
			0.50	**146**								
		0.10	0.01	$3.18\text{x}10^8$	$1.95\text{x}10^8$	$1.47\text{x}10^8$	$1.21\text{x}10^8$	$1.04\text{x}10^8$	$9.21\text{x}10^7$	$8.32\text{x}10^7$	$7.62\text{x}10^7$	$7.06\text{x}10^7$
			0.10	33348	20814	15963	13321	11637	10463	9594	8922	8387
			0.50	**119**								

Table 2.7: Sample Sizes for Multiplicative Models with HWE and Decayed LD

				Number of Alleles at Marker Locus								
γ_1	γ_2	K	p_2	2	3	4	5	6	7	8	9	10
2	4	0.01	0.01	234363	144416	109379	90159	77824	69151	62675	57631	53574
			0.1	2773	1760	1372	1162	1030	939	873	822	782
			0.5	200								
		0.1	0.01	193688	119303	90321	74420	64213	57034	51672	47494	44134
			0.1	2291	1448	1125	950	839	763	706	664	630
			0.5	164								
3	9	0.01	0.01	59751	36960	28099	23248	20141	17962	16339	15077	14065
			0.1	820	533	424	367	331	307	290	277	267
			0.5	85								
		0.1	0.01	49381	30520	23184	19166	16592	14785	13439	12392	11552
			0.1	677	437	345	297	266	246	231	220	211
			0.5	69								

Table 2.8: Sample Sizes for Additive Models with HWE and Decayed LD

				Number of Alleles at Marker Locus								
γ_1	γ_2	K	p_2	2	3	4	5	6	7	8	9	10
2	3	0.01	0.01	239020	147275	111536	91931	79349	70501	63894	58747	54609
			0.1	3294	2081	1614	1362	1203	1092	1012	950	901
			0.5	353								
		0.1	0.01	197537	121666	92104	75885	65472	58149	52680	48417	44989
			0.1	2722	1714	1325	1115	982	889	821	769	727
			0.5	291								
3	5	0.01	0.01	62110	38409	29192	24146	20915	18647	16957	15644	14591
			0.1	1111	713	561	479	429	394	369	349	335
			0.5	192								
		0.1	0.01	51331	31718	24088	19909	17231	15351	13950	12861	11986
			0.1	918	585	458	390	347	317	296	279	267
			0.5	159								

Table 2.9: Sample Sizes for Dominant Models with HWE and Decayed LD

				Number of Alleles at Marker Locus								
γ_1	γ_2	K	p_2	2	3	4	5	6	7	8	9	10
2	2	0.01	0.01	243817	150221	113759	93757	80919	71891	65150	59898	55674
			0.1	3991	2509	1937	1627	1431	1294	1194	1117	1055
			0.5	1058								
		0.1	0.01	201503	124100	93942	77393	66770	59298	53717	49368	45870
			0.1	3299	2068	1593	1334	1170	1056	972	907	856
			0.5	878								
3	3	0.01	0.01	63343	39166	29764	24616	21319	19005	17281	15941	14865
			0.1	1329	846	661	562	499	456	425	401	383
			0.5	521								
		0.1	0.01	52350	32344	24561	20297	17565	15647	14217	13105	12213
			0.1	1099	696	542	458	406	369	343	322	306
			0.5	434								

Table 2.10: Sample Sizes for Recessive Models with HWE and Decayed LD

				Number of Alleles at Marker Locus								
γ_1	γ_2	K	p_2	2	3	4	5	6	7	8	9	10
1	2	0.01	0.01	$2.30\text{x}10^9$	$1.410\text{x}10^8$	$1.064\text{x}10^9$	$8.74\text{x}10^8$	$7.51\text{x}10^8$	$6.65\text{x}10^8$	$6.01\text{x}10^8$	$5.50\text{x}10^8$	$5.09\text{x}10^8$
			0.1	235125	145351	110436	91317	79069	70473	64068	59088	55092
			0.5	598								
		0.1	0.01	$1.90\text{x}10^9$	$1.17\text{x}10^9$	$8.79\text{x}10^8$	$7.22\text{x}10^8$	$6.21\text{x}10^8$	$5.50\text{x}10^8$	$4.96\text{x}10^8$	$4.55\text{x}10^8$	$4.21\text{x}10^8$
			0.1	194253	119996	91106	75279	65136	58015	52706	48576	45261
			0.5	491								
1	3	0.01	0.01	$5.75\text{x}10^8$	$3.53\text{x}10^8$	$2.66\text{x}10^8$	$2.19\text{x}10^8$	$1.88\text{x}10^8$	$1.66\text{x}10^8$	$1.50\text{x}10^8$	$1.38\text{x}10^8$	$1.27\text{x}10^8$
			0.1	60132	37427	28628	23827	20764	18623	17035	15806	14824
			0.5	215								
		0.1	0.01	$4.75\text{x}10^8$	$2.92\text{x}10^8$	$2.20\text{x}10^8$	$1.810\text{x}10^8$	$1.55\text{x}10^8$	$1.37\text{x}10^8$	$1.24\text{x}10^8$	$1.14\text{x}10^8$	$1.05\text{x}10^8$
			0.1	49663	30867	23576	19596	17054	15276	13956	12933	12115
			0.5	175								

Table 2.11: Sample Sizes for an Additive Model with HWD and Nongametic Disequilibrium

	Hardy-Weinberg Disequilibrium			
Nongametic Diseq.	Min	0*	Avg	Max
0	718	1397	1737	2757
Half-max	321		774	1228
Max	180		425	690

*Non-zero $D_{r/i}$ are not considered when HWD coefficients are 0

Chapter 3

Performing Association Mapping with Decision Trees

3.1 Introduction

Case-control marker association tests such as the allele test, the genotype test, and the linear trend test, all discussed in Chapter 2, examine one marker at a time; they are single-marker tests that compare marker allele or one-locus genotype frequencies between cases and controls. The drawback to these approaches is that they may be insufficient for uncovering more than one gene that affects a complex disease. For this reason, we present here an examination of decision tree analysis, a recursive partitioning method, which implements case-control tests in such a manner as to capture possible epistatic effects of several genes on disease susceptibility.

Analyzing genetic marker data using decision trees is a logical approach for several other reasons. First, the goal of the analysis; in locating a disease gene through fine-scale mapping with the aid of association techniques, the goal may be dual. In addition to uncovering the markers whose associations with a disease help guide researchers to the location of the disease gene, these markers may also be used to identify individuals who have a greater chance of one day being affected with the disease of interest. These two purposes, understanding the predictive structure of the problem and accurately classifying individuals, makes classification analysis such as decision trees a natural solution (Breiman et al. 1984). In addition, the mere size of data involving genetic markers may render standard statistical techniques, such as regression, ineffective. Breiman et al. (1984) describe a data set possibly too complex for a simple statistical approach as having high dimensionality, a mixture of data types, and nonstandard data structure. This again certainly matches the description of genetic marker data.

The process of creating a decision tree works as follows: using all of the data available, determine which of the independent variables creates the best "split"; that is, the best way of subsetting the data based on the independent variable into categories that are the most homogeneous with respect to the dependent variable, or target. Statistical methods such as goodness-of-fit tests are often used to determine which split is optimal. On each of these new "nodes" that contain a subset of the data, continue this process until some stopping criterion is reached. In this notation, the marker genotypes or alleles are the independent variables, and disease status is the dependent variable or target.

One criterion for choosing which algorithm to implement is the number of branches into which a node in the tree can be split. We opt to consider binary splits only. Since the dependent variable in marker data is binary (disease status: affected or unaffected), this is a logical approach. In addition,

> a multiway split may always be accomplished with a sequence of binary splits on the same input. An algorithm that proceeds in binary steps has the opportunity to split with more than one input and thus will consider more multistep partitions than an algorithm can consider in a single-step multiway split. Too often the data do not clearly determine the number of branches appropriate for a multiway split. The extra branches reduce the data available deeper in the tree, degrading the statistics and splits in deeper nodes. (Neville 1999)

Since we are dealing with binary decision trees and biallelic markers known as single nucleotide polymorphisms (SNPs), a test for classifying individuals with three different genotypes into two categories is needed. Individuals with at least one copy of a particular marker allele can be placed in one category, and those with no copies of the allele placed in the other. We do not know a priori which allele is the one of interest, so the tree-building process performs a preliminary test to determine which two genotype categories are most similar in their proportion of cases and controls, and these two categories are combined together to form one category. It is expected that the heterozygous genotype will always be one of the categories that is combined with one of the homozygous genotype categories. Performing a chi-square test on the contingency table created in this manner is equivalent to Sasieni's "serological" test (1997). This statistic, X_S^2, has an asymptotic χ_1^2 distribution. In terms of the notation used by Nielsen and Weir (1999), X_S^2 is proportional to a linear combination of the terms $[\delta_1/\phi(1-\phi)]^2$ and $[2\delta_2 + \delta_{22}/\phi(1-\phi)]^2$, assuming

that one of the homozygous genotypes M_1/M_1 or M_2/M_2 and the heterozygous genotype M_1/M_2 have been combined into one category, where the terms are defined as

$$\begin{aligned} \delta_i &= \sum_r \alpha_r D_{ri} \\ \delta_{ii} &= \sum_{r,s} D_{ri} D_{si} G_{rs} \end{aligned}$$

with α_r as the additive effect of rth allele at the trait locus, D_{ri} as the coefficient of linkage disequilibrium between marker allele M_i and allele A_r at the trait locus, ϕ the prevalence of the disease in the population, and G_{rs} is the trait of an individual with genotype A_r/A_s. That is, it tests for significance of the additive effect of allele M_i and dominance effect of the M_i/M_i genotype. The genotype chi-square statistic X^2_G tests simultaneously for any additive effects and any dominance effects of the alleles and genotypes, respectively, and has 2 df. The relationship between X^2_S and X^2_G can be shown to be $X^2_S \leq X^2_G$ (Agresti 1990), with the difference between the two statistics dependent on the similarity between the two genotypes that are combined into one category; the more similar they are, the closer the two statistics will be, and thus the more significant X^2_S will be since it has fewer df meaning a possible increase in power. This test will be locally more powerful than the allele and trend tests, discussed in Chapter 2, if the genetic model is nonadditive (Sasieni 1997).

The use of recursive partitioning or trees for gene mapping has surfaced in several applications recently. Data from the Genetic Analysis Workshop 12 were analyzed using tree-based methods by Czika et al. (2001), presented in the following chapter, and Zhang et al. (2001). Zhang and Bonney (2000) also use classification trees for the

association study of a single simulated data set. Here we present results mentioned by Weir et al. (1999) from analyzing SNP data in the *APOE* region with Alzheimer disease and compare these with previous single-marker association analyses. We also use replicates from simulated data in order to perform an analysis of the power of this approach under a multilocus genetic disease model and determine whether the correct size holds using recursive partitioning. We additionally use real SNP data to examine an epistatic disease model using many different pairs of markers, with varying allele frequencies and extent of linkage disequilibrium.

While evaluating decision trees for analyzing genetic marker data, we also examine a method of "smoothing" p-values by taking correlations from neighboring SNPs into account. If p-values are plotted along the map of markers being analyzed, it has been shown that by taking the width of peaks into account in addition to height, the power of linkage studies can be improved (Terwilliger et al. 1997; Goldin et al. 1999; Siegmund 2001). Zaykin et al. (2002) applies a similar methodology to association mapping that we implement here in standard association tests as well as the tree analysis.

3.2 *APOE* Data and Results

The apolipoprotein E (*APOE*) gene on chromosome 19q has been identified as a susceptibility gene for late-onset Alzheimer disease (see Martin et al. (2000) for a description of the work establishing this). Martin et al. (2000) also present an application of using dense SNP maps in the association mapping of complex genes, analyzing ten SNPs in

the *APOE* region. We demonstrate the use of decision trees on the case-control data from this study to see if the SNPs most tightly linked, and significantly associated via standard association tests, with the susceptibility gene are included in the tree.

The default tree-building method of the Enterprise Miner^{TM} (EM) software is applied to the *APOE* data, with the resulting tree displayed in Figure 3.1. The tree first splits the entire data set into two nodes using genotypes from the APOC1S marker, with individuals homozygous for one of the alleles included in the left node and all other individuals (including those with missing genotypes) in the right node. A significant split was then detected in the right node, with the subset of individuals now split on the PRR2 marker. Using the tree to classify individuals, the classification rate improves from 50.7% to 67.1%. Table 3.1 shows that these results confirm the single-marker association tests, with the two markers closest to *APOE* included in the tree as desired.

3.3 Simulations

3.3.1 Data

The data that were simulated to examine the properties of decision trees for analyzing SNP data in order to fine-scale map multiple disease-susceptibility genes consist of ten populations. There are 2,000 total markers, with each of two sets of 1,000 markers spanning approximately one Morgan. The two sets are unlinked so are essentially on two separate chromosomes. Among these SNPs, there are three genes that affect the

penetrance of the disease through dominance and additive effects; the genotypes at the actual disease gene loci are not included in the data.

Three populations were initially simulated, two large populations with 10,000 individuals each, and a smaller population of 1,000 individuals. The large populations admixed at a rate of five percent into the smaller one for 20 generations. Once the admixture was stopped, the smaller population continued random mating for 30 generations, with the population size limited to 10,000.

Initial allele frequencies were generated from a uniform distribution separately for each population. For the three disease genes, the allele frequencies were generated with a uniform distribution for the two alleles. The effect of these genes on the disease penetrance was in the form of additive allele effects, dominance deviations at each of the three loci, and gene-by-gene interactions of the additive allele effects for the three pairs of loci. These random deviations could not exceed ten percent of the maximum allelic effect. Then a random value between one and five percent of the maximum allelic effect was added (or subtracted) to the three-locus genotype penetrance to create three-loci interactions. Only populations with a disease prevalence between 5 and 30% were used. Table 3.2 shows the single-locus penetrances for one of the ten populations, based on the theoretical, not observed, three-locus penetrances.

Recombination was performed by drawing random parents and creating two recombinant gametes from the parental chromosomes. The number of positions was generated from a Poisson distribution with the chromosome length in Morgans as the mean. The

sites for recombination were then distributed uniformly across the chromosome map.

Ten of these populations were simulated. For each of these populations, we took ten samples of size 2,000 with 1,000 affected individuals and 1,000 unaffected individuals for the disease we are simulating. All individuals were genotyped at the 2,000 loci, but the three disease genes were excluded from the analysis.

3.3.2 Methods

There are many different algorithms that can be used to create a decision tree. The method we opt to use for analyzing the simulated data is a simple recursive partitioning algorithm similar to an unsophisticated CHAID-like algorithm (Kass 1980) where the best split is determined using a chi-square test. No pruning of the tree is performed.

In order to account for the multiple testing of dependent hypotheses, several methods were considered. All methods use an approach described by Zaykin et al. (2002) where each marker's p-value is replaced by some function of the p-values of the nearest neighboring markers. Fisher's method (1932) can be used, where assuming independence of tests, the p-value for the ith marker is calculated using the p-value from the $\chi^2_{2(2w+1)}$ distribution for the statistic $t = -2\sum_{j=-w}^{w} \ln(p_{i+j})$. When the assumption of test independence is violated, as it is with these data since we are combining tests from markers that we know are correlated, Fisher's method does not control the type I error rate; that is, it is anti-conservative. An alternative is to apply Simes' method (1986) over a similar window of size $2w + 1$, again centered at the SNP whose p-value is being replaced. For

Simes' method, we use the order statistics of the p-values from the chi-square statistics across the particular window. Then the new p-value p_i^* for the ith SNP is calculated as

$$p_i^* = \min_{1 \leq j \leq n} \left[\frac{n P_{(j)}^i}{j} \right]$$

where $n = 2w+1$ and $P_{(j)}^i$ is the jth order statistic of the p-values in the ith window. This method is in fact conservative when tests are positively correlated (Sarkar and Chang 1997), as we would expect them to be in this situation. Since this is the only method that controls type I error, we used this method alone.

The first step in the analysis is to examine single-marker statistics with Simes' method compared with no correction. Three types of association tests are performed: the allele case-control test, the genotype case-control test, and the linear trend test. For all tests, Sidak's correction is used for the number of tests that are being performed (the adjusted significance level is then 0.0000256). Note that even for tests that implement Simes' method, this correction is still needed since there are still 1,997 p-values being tested for significance. We then compare these with the new statistics created using the methods described above.

Next, decision trees are formed in a similar fashion: with no sort of multiplicity correction, and then using the Simes method. When this method is utilized, it is applied to each node in the tree. The branching of the trees is stopped when either there are five levels in the tree, or no chi-square statistics are above the significance level, whichever occurs first. As with the single-marker tests, a Sidak correction is applied to account for the number of markers being tested for each split in the tree.

In order to evaluate the tests that are performed, we must first define power. Again, the actual polymorphisms that affect the penetrance of the disease are not included in the data; we are hoping to find the SNPs close to these genes. In order to define "close," we have to study the linkage disequilibrium between pairs of markers in these samples to see the number of markers, on average, over which significant linkage disequilibrium extends. Since gametic phase is unknown in this study, haplotype frequencies must be estimated using the EM algorithm (since only the two-locus disequilibrium is examined and the markers are biallelic, this reduces to a solving a cubic equation). From these estimated haplotype frequencies, the linkage disequilibrium coefficient D can be estimated and used in a chi-square statistic to test if D is significantly different from zero (Weir 1979).

3.3.3 Results

The linkage disequilibrium (LD) test was performed on all pairs of markers in the 100 samples. To see how far LD extends on average, p-values from the LD tests for all pairs of markers the same number of markers apart were averaged together. This results in pairs of markers up to 21 markers apart having, on average, significant linkage disequilibrium. We use this to define a power region for evaluating our results; any marker less than 22 markers from a disease gene contributes towards power, while markers outside these three regions contribute towards type I error.

The power and type I error for the single-marker analyses are displayed in Table 3.3. Window sizes indicate the number of markers on either side of a marker that are included

in Simes' method to adjust the p-value at the middle of the sliding window. Since LD between pairs of markers extends across 21 markers on average, we would expect to see differences between the window sizes 21 or less and those more than 21. While there are no clear-cut distinctions between the two groups, it does appear that power begins decreasing at about the window size of 21, and type I error exceeds the 0.05 level near this window size as well. Thus, only window sizes of 20 or less are considered for the rest of the analyses.

The decision trees that were created in analyzing the markers had 2.82 splits on average, or 6.64 nodes in the tree. There is, on average, one less split in trees using Simes' method than those not. Of the 941 trees that had at least one split, almost 46% of the trees made the first split on a marker in the region of gene 1, and 29% were made on a marker in the region of gene 2. Table 3.4 shows how often markers in the region of the disease genes occurred in the trees, using a window size of 10 for Simes' method. For example, 67 of the 100 trees using the window size of 10 did not split on any markers outside any of the gene regions, while 1 tree had seven splits on markers outside all of the gene regions.

Next we compared how often the specific genes were found by the different analyses. For the single-marker tests, each of the three genes is considered "found" if at least one marker in the power region, less than 22 markers from the gene, has a significant test statistic at the Sidak-corrected level 0.0000256. For the decision trees, a gene is "found" if such a marker is included somewhere in the tree. The results are displayed in Table 3.5.

3.4 Analysis of Chromosome 20 Data

With our simulated data, there is only a single specific disease model being examined, so the somewhat discouraging results for the performance of the decision trees can not be generalized. Therefore, we also evaluate the ability of decision trees to locate markers acting epistatically on disease susceptibility using real data containing SNPs from a single chromosome. The data consist of 96 individuals genotyped at 4,427 markers from chromosome 20. A total of 5,827 pairs of markers were used in the following disease model: a single "main" effect marker is fully recessive for the disease, while a second "weak" effect marker, when recessive, causes the main effect marker to act dominantly for the disease (see Table 3.6 for a representation of this model) (Majewski, Li, and Ott 2001). Main effect markers were taken from 59 of the first 1,000 SNPs and 100 different weak effect markers taken from the second 1,000. For the tree analysis, the same method as used on the simulated data was applied though no windowing of the p-values was performed. The trees were limited to a depth of three, that is three total splits, due to the small sample size.

Of the 5,827 trees run on these samples, there were 4,348 that contained a split on a SNP near the main disease gene. Note that "near" is now defined as a marker within 10,000 base pairs (approximately 0.01 cM) of the main or weak disease gene. Of these splits, 1,936 were the first split in the tree, and the other 2,412 splits were made in the second level of the tree, after the first split. A SNP near the weak effect SNP was included in the tree 732 times. Only 92 of these were the first split, and the other 640 were made

after the first split. Table 3.7 shows summary statistics of the allele frequencies for all markers alleles used as main or weak effects, and for those SNPs with a neighboring SNP included in the tree. The wide range of allele frequencies for markers from this chromosome is evident. This table also shows that while the allele frequencies for those SNPs included in the tree are on average slightly higher, the allele frequencies still span almost the entire range of possible values.

There were also 604 trees that included splits on SNPs close to both disease genes. For the disease models for which this occurred, we compared the results from applying single-marker association tests to SNPs in windows around the disease genes. Only 56 of the 604 pairs of disease genes were not found to be significant according to the genotype contingency table test, and of these, 26 pairs had at least one SNP within 10,000 bases of each of the disease genes that was significant ($\alpha = 0.05$). That leaves 30 pairs for which the decision tree successfully located both disease genes when the single-marker association test did not. However, there were 3,021 pairs of disease genes that the genotype test found to be significant in total, without including significance of nearby markers.

3.5 Discussion

We performed three different types of analyses to evaluate the performance of decision trees for the association mapping of a disease gene: first, a confirmation of single-marker tests that found SNPs close to the *APOE* gene that affects susceptibility to Alzheimer's;

then, an examination of the tree method's size and power analyzing simulated data and applying smoothing techniques to p-values; and finally, we created a simple two-locus epistatic disease model using different combinations of SNPs from real data from chromosome 20. While it is important to know that the trees perform as well as single-marker tests in single disease-gene models, the multi-gene, epistatic disease models are those where we hope that the trees will detect interactions between genes that the single-marker tests miss. This was not the result with the simulated data. In this complicated three-gene model, the single-marker tests were detecting two or more of the genes more often than the trees. Also of interest in the analysis of the simulated data was the effect of using Simes' method on sliding windows of p-values to adjust for correlations due to LD. This proved to be a powerful way of taking "peak width" into account while maintaining the nominal level.

We studied a more simplistic epistatic model with two genes only, using real allele frequencies and LD. Looking at over 5,800 pairs of disease genes, we found that the single-marker test based on the genotype contingency table Pearson chi-square was finding both disease genes over half of the time, while the trees only detected both genes approximately one in ten times. The small sample size may have negatively impacted the performance of the trees, but the fact that the single-marker tests are more robust to such limitations further weakens the endorsement for the use of decision trees in association mapping.

3.6 Tables and Figures

Table 3.1: Analysis Results for *APOE* Data

Marker	Distance from APOE (kb)	Association test p-value	Decision Tree
APOC1	10	0.00	First split
PRR2	20-40	0.01	Second split
APOC4-1	60	0.55	–
2050	80	0.78	–
2590	166	0.28	–
42709	250	0.24	–
2526	257	0.24	–
2151	443	0.03	–
582	800	0.23	–
104	841	0.26	–

Table 3.2: Penetrance of Disease for One-Locus Genotypes

Locus	Genotype	Penetrance
Gene 1	0/0	0.3100
	0/1	0.1817
	0/2	0.1742
	1/1	0.0370
	1/2	0.0065
Gene 2	0/0	0.1876
	0/1	0.2974
	0/2	0.4112
	1/1	0.1948
	1/2	0.2973
	2/2	0.3872
Gene 3	0/0	0.2460
	0/2	0.2498
	2/2	0.2553

Table 3.3: Power and Type I Error for Single-Marker Tests Using Simes' Method

Window	Power			Type I Error		
size	Allele	Genotype	Trend	Allele	Genotype	Trend
0	0.1204	0.1161	0.1357	0.0117	0.0109	0.0135
2	0.2933	0.2815	0.3248	0.0250	0.0241	0.0296
4	0.3579	0.3420	0.3905	0.0297	0.0284	0.0348
6	0.3875	0.3716	0.4266	0.0322	0.0311	0.0381
8	0.4083	0.3890	0.4474	0.0349	0.0334	0.0410
10	0.4206	0.3995	0.4568	0.0366	0.0354	0.0436
12	0.4315	0.4079	0.4640	0.0387	0.0375	0.0456
14	0.4357	0.4171	0.4717	0.0408	0.0393	0.0479
16	0.4417	0.4252	0.4758	0.0431	0.0407	0.0501
18	0.4483	0.4315	0.4836	0.0450	0.0428	0.0526
20	0.4522	0.4335	0.4886	0.0464	0.0449	0.0548
22	0.4562	0.4306	0.4868	0.0492	0.0472	0.0568
24	0.4561	0.4305	0.4872	0.0517	0.0492	0.0591
26	0.4569	0.4329	0.4887	0.0538	0.0516	0.0619
28	0.4532	0.4325	0.4904	0.0557	0.0539	0.0647
30	0.4499	0.4300	0.4899	0.0578	0.0559	0.0670
32	0.4498	0.4287	0.4882	0.0602	0.0582	0.0689
34	0.4454	0.4293	0.4858	0.0618	0.0603	0.0712
36	0.4447	0.4296	0.4867	0.0644	0.0624	0.0738
38	0.4453	0.4266	0.4841	0.0668	0.0647	0.0759
40	0.4462	0.4238	0.4816	0.0693	0.0668	0.0782
42	0.4306	0.4210	0.4829	0.0717	0.0689	0.0804

Table 3.4: Number of Trees Containing Markers from Gene Regions

	Times occurred in tree							
Gene Region	0	1	2	3	4	5	6	7
None	67	17	6	6	2	1	0	1
Gene 1	44	14	12	15	8	7	0	0
Gene 2	64	17	10	9	0	0	0	0
Gene 3	94	2	1	1	1	1	0	0

Table 3.5: Number of Samples Finding Specified Genes

		Window size										
Genes found	Test	0	2	4	6	8	10	12	14	16	18	20
1 only	Allele	24	20	20	20	20	22	22	23	23	23	22
	Genotype	20	21	20	22	21	23	23	24	23	23	23
	Trend	23	20	20	20	20	22	22	23	23	23	22
	Tree	17	34	36	40	38	36	36	40	40	36	39
2 only	Allele	10	8	8	10	10	10	10	10	11	12	12
	Genotype	10	10	11	11	11	10	10	9	11	12	13
	Trend	10	8	8	10	10	10	10	10	11	12	12
	Tree	20	17	16	18	19	19	22	23	20	26	23
3 only	Allele	0	0	0	1	1	1	1	1	1	1	1
	Genotype	0	1	1	1	1	1	1	1	1	1	1
	Trend	0	0	0	1	1	1	1	1	1	1	1
	Tree	2	2	2	3	3	3	4	3	4	4	4
Any one gene	Allele	34	28	28	31	31	33	33	34	35	36	35
	Genotype	30	32	32	34	33	34	34	34	35	36	37
	Trend	33	28	28	31	31	33	33	34	35	36	35
	Tree	39	53	54	61	60	58	64	66	64	66	66
1 and 2	Allele	33	35	35	32	34	34	35	34	33	32	32
	Genotype	34	32	34	31	32	30	31	31	29	28	28
	Trend	33	35	35	32	34	34	35	34	33	32	32
	Tree	21	21	20	15	17	17	14	8	11	7	9
1 and 3	Allele	9	10	10	9	9	8	8	8	8	8	8
	Genotype	10	9	8	8	8	8	8	8	8	8	8
	Trend	9	10	10	9	9	8	8	8	8	8	8
	Tree	11	6	5	4	1	3	1	2	1	2	1
2 and 3	Allele	3	2	2	1	2	1	2	2	2	2	2
	Genotype	2	1	1	1	2	2	2	2	2	2	1

Table 3.5: *continued*

		Window size										
Genes found	Test	0	2	4	6	8	10	12	14	16	18	20
	Trend	2	2	2	1	2	2	2	2	2	2	2
	Tree	1	0	0	0	0	0	0	0	0	0	0
Any two genes	Allele	45	47	47	42	45	43	45	44	43	42	42
	Genotype	46	42	43	40	42	40	41	41	39	38	37
	Trend	44	47	47	42	45	44	45	44	43	42	42
	Tree	33	27	25	19	18	20	15	10	12	9	10
All three genes	Allele	16	14	13	13	10	9	8	8	8	8	8
	Genotype	15	13	10	10	8	8	7	7	7	7	
	Trend	17	14	13	13	10	9	8	8	8	8	8
	Tree	6	1	1	1	1	0	0	2	0	0	0

Table 3.6: Disease Penetrance Given Two-Locus Genotypes

	Locus 2		
Locus 1	B/B	B/b	b/b
A/A	0	0	0
A/a	0	0	1
a/a	1	1	1

Table 3.7: Allele Frequencies for Chromosome 20 SNPs

	Allele Frequencies		
SNPs	Min	Mean	Max
All used as main disease gene	0.0053763	0.4889745	0.9947917
Main genes used in tree	0.0053763	0.5440823	0.9947917
All used for weak disease gene	0.0052083	0.4723178	0.9895833
Weak genes used in tree	0.0052083	0.6181344	0.9791667

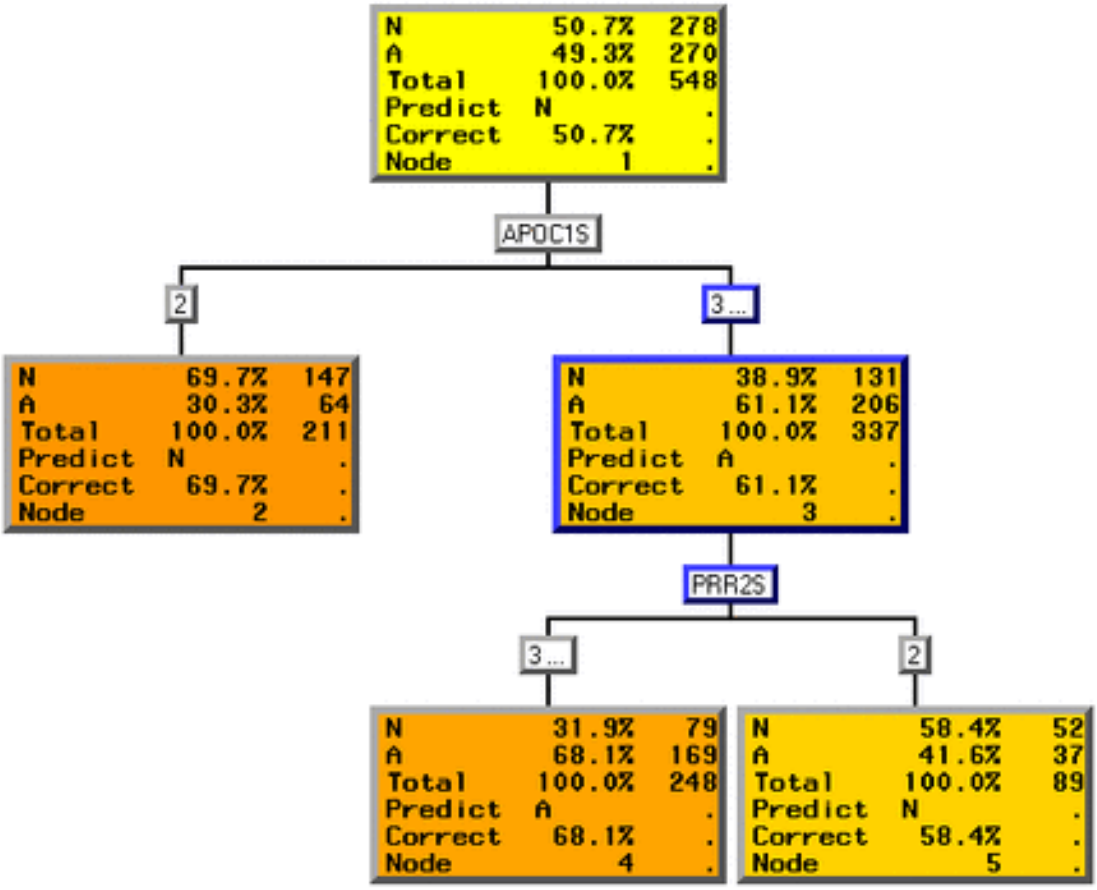

Figure 3.1: Tree Created for *APOE* Data

Chapter 4

Applying Data Mining Techniques to the Mapping of Complex Disease Genes

Czika, W. A., B. S. Weir, S. R. Edwards, R. W. Thompson, D. M. Nielsen, J. C. Brocklebank, C. Zinkus, E. R. Martin, and K. E. Hobler (2001). *Genetic Epidemiology 21*, S435-S440.

4.1 Abstract

The simulated sequence data for the Genetic Analysis Workshop 12 were analyzed using data mining techniques provided by SAS Enterprise Miner^{TM} Release 4.0 in addition to traditional statistical tests for linkage and association of genetic markers with disease status. We examined two ways of combining these approaches to make use of the covariate data along with the genotypic data. The result of incorporating data mining techniques with more classical methods is an improvement in the analysis, both by correctly classifying the affection status of more individuals and by locating more SNPs related to the disease, relative to analyses that use classical methods alone.

4.2 Introduction

With the identification of hundreds of thousands of single nucleotide polymorphisms (SNPs) well underway, many issues have materialized that may reduce the effectiveness of standard techniques that use these markers to locate disease genes. Multiple hypothesis testing, dependencies between tests performed, and managing the sheer volume of the data are among the obstacles that can be encountered. Methods that analyze single markers at a time are particularly inadequate for the mapping of complex disease genes, because they fail to capture interactions between genes. These problems make data mining (Westphal and Blaxton 1998), the process of uncovering patterns in large amounts of data, a natural approach for the analysis of genetic data. Decision trees, a popular

data mining method which have a number of applications, are especially useful for finding interactions between variables. In this paper, we demonstrate the use of trees in modeling the covariate data, reducing the number of SNPs, and categorizing the quantitative risk factors into binary variables. Our goal then is to combine these methods with logistic regression and SNP linkage and association tests to produce better results than analyses using the traditional methods alone.

4.3 Methods

Three replicates of the sequence data for the candidate genes from the general population were analyzed. Replicate 1 was used for model training, replicate 42 was selected as the validation data and used for model assessment, and replicate 3 was used as the test data set to obtain a final, unbiased estimate of the generalization error of each model. For modeling and association methods, the data was subset to include only the 165 founders since only unrelated individuals should be used. The answers were available to us during analysis. Before incorporating the genetic factors into the analysis, preliminary analyses were performed to determine the effect of the five quantitative traits, the two environmental factors, age of examination, household membership, and gender on the affection status of the individuals. Because model interpretability was deemed more important than prediction accuracy, the Decision Tree and the Regression tools of Enterprise Miner were used in favor of the Neural Network tool and model ensembles to predict the affection status. For the decision trees, the chi-square test was used as the splitting criterion

($p < 0.20$) to recursively partition the data. The logistic regression models selected the subset of predictors from all possible steps in a stepwise regression based on the step that minimized the misclassification rate in the validation data. Entry and stay p-values of 0.15 were used. Models were first created using only the quantitative risk traits Q1 through Q5. We then added the following variables into the analysis: environmental factors, age of examination, gender, household membership, and the interaction of these variables with Q1 through Q5. Because the trait means are noticeably higher in females than in males, we attempted to improve the classification rate by fitting separate logistic regression models for males and females.

The next step in the analysis was to incorporate the sequence data from the seven candidate genes for the general population into the analysis of the phenotypic data. SNPs were generated from the candidate genes based on variants identified in sequencing ten individuals from the three chosen replicates. The primary goal of this analysis was to find SNPs associated and linked with affection status. The Reconstruction-Combined TDT (Knapp 1999) was used to test for linkage of the SNPs with affection status. We used a continuity correction of 0.5 for this test. For the RC-TDT, we used all of the pedigree data from replicate 1. However, because the data represent an extended pedigree with related nuclear families, the test's validity for association does not hold. Thus, separate association tests needed to be performed on unrelated individuals; we therefore used the subset of data that was created for the covariate analysis. We conducted tests for the composite linkage disequilibrium between SNPs (Weir 1996), tests of Hardy-

Weinberg equilibrium for each SNP, and chi-square tests of association of SNPs with disease phenotype using SNP alleles. In order to combine the genotypic information with the covariate model, we first wanted to reduce the number of SNPs that were found to be linked and associated with affection status. We implemented a tree-based method to decrease the number of SNPs used in subsequent modeling. This method takes multiple testing and associations between neighboring SNPs into account through a windowing algorithm. Thus, splits in the tree were formed based on a new statistic for each SNP, which is a result of combining the chi-square value from the SNP itself and the neighboring SNPs' chi-square values. The SNPs that formed splits in the tree were then added to the covariate regression or tree models in an attempt to improve classification rates of individuals. Our other approach to combining data mining with linkage and association methods was to examine the relationship between the candidate genes and the risk factors. Our analysis tools are set up to handle a binary phenotype only; thus, we first binned these quantitative factors into two categories based on their optimal relationship with the disease status. This was performed using a binary decision tree in the Transform Variables node of Enterprise Miner. Once this was done, we performed the same linkage and association analyses on these new binary phenotypes that we had performed on affection status to find additional genetic information.

4.4 Results

Figure 4.1 displays the tree diagram produced by Enterprise Miner using only the five quantitative traits as model inputs. Each node in the tree diagram contains three columns that contain the following information: the target values (A=Affected and U=Unaffected); the target percentages and counts for the training data (replicate 1); and the target percentages and counts for the validation data (replicate 42). The root or top node contains all of the data. The first split was made on Q1. The left-most terminal leaf in the tree diagram indicates that individuals with values of Q1 less than 19.075 have a greater tendency to be unaffected than individuals with higher Q1 values. The individuals in the other nonterminal node were further split based on values of Q5 less than 37.62, or greater than or equal to 37.62. No other significant splits were found based on choosing the subtree that minimizes the misclassification rate in the validation data.

The stepwise logistic regression model provides a better overall fit than the decision tree (Table 4.1). For the model using the quantitative traits alone, Q2 was found to be a significant input in addition to Q1 and Q5. Table 4.3 reports the parameter estimates and related statistics for this model. The odds of being affected increase by factors of 1.428, 1.226, and 1.392 for each unit increase in the values of Q1, Q2, and Q5 respectively.

After including the other covariates in the analysis with their interactions with Q1 through Q5, we found Q1, Q5, and the interaction of Q5 by age of examination to be significant. The parameter estimates from this step are shown in Table 4.2. Thirteen

of 45 affected individuals are misclassified, and seven of the 120 unaffected individuals are misclassified in the training data. It is important to note that the interaction of age of examination by Q4 and the interaction of environmental factor 2 by Q4 are both significant when using the training data exclusively to build the model. When including the other phenotypic factors in the analysis, the decision tree node still segmented the data based only on Q1 and Q5 as shown in Figure 4.1.

In the logistic regression models that were fit separately for males and females, only half of the 12 affected males in the training data are correctly classified. On the other hand, only one of the 68 unaffected males is incorrectly classified. Significant model effects for males include Q1, Q5, age of examination by environmental factor 2, and age of examination by Q5. For the females, 29 of the 33 affected individuals are correctly classified, and 49 of 52 unaffected individuals are correctly classified in the training data. Significant model effects include Q1, Q5, age of examination by Q1, age of examination by Q4, and age of examination by Q5.

All tests of linkage and association are reported on replicate 1 though similar results were found on replicates 3 and 42. The association map for candidate gene 1 in Figure 4.2 gives us a picture of the composite linkage disequilibrium between SNPs, with the color on the off-diagonal corresponding to the significance of the p-values for testing the null hypothesis of no linkage disequilibrium between SNPs. The p-values for testing SNPs for Hardy-Weinberg equilibrium are represented by the coloron the diagonal, and the shape on the diagonal indicates the presence or absence of a SNP's association with disease

status at a significance level of 0.05. For example, the yellow plus symbol in the bottom left-hand corner of the figure indicates that SNP2 has a significant association with affection status, and we would not reject the hypothesis that SNP2 is in Hardy-Weinberg equilibrium. We can determine that there is evidence of linkage disequilibrium between SNP2 and SNP3 by the red square to its right. Table 4.4 displays the results of these linkage tests using the RC-TDT and chi-square association tests on SNP alleles. Nearly half of the SNPs in candidate gene 1 show both significant linkage and association with affection status (42 out of the 47 have p-values below 0.01 for both tests). Our tree-based method, which is applied to reduce the number of SNPs found to be linked and associated with the disease, split the data with respect to affection status only at the SNP located at base pair 76 in candidate gene 1. When we added this sole SNP to the quantitative risk factors found in the previous regression model, the percentage of correctly classified individuals improves from 87.9 to 95.8 for the training data, and from 85.5 to 92.7 for the validation data.

In our examination of linkage and association of the binary variables formed from the quantitative traits, the most noticeable relationship occurs between candidate gene 6 and Q1. Ten out of a total of 30 SNPs in candidate gene 6 have p-values below 0.01 for both the association test and linkage test with the binary Q1 that was created. Candidate gene 6 also has several SNPs, primarily in the 6000-8000 bp range, linked and associated with the binned Q2 variable. With several SNPs in candidate gene 2 displaying significant linkage and association p-values with the binned Q5, there is also evidence of a relationship in

this gene-trait pair. Thus, we found candidate genes that have a significant effect on all three quantitative risk factors significantly related to affection status.

4.5 Discussion

Genetic data are now as vast as the business data for which data mining was originally developed, making data mining a natural approach for analyzing the GAW data. Through preliminary logistic regression and decision tree runs using Enterprise Miner software, we determined that the quantitative risk factors Q1, Q2, and Q5 are most important in differentiating the affected and unaffected patients; in the actual model, these are the three risk factors with the strongest effect on liability. The other phenotypic variables were not found to be significant predictors in the absence of the candidate genes. Using the RC-TDT and the chi-square test for association, 47 SNPs in candidate gene 1 were identified as linked and associated with disease status. This is in fact the only candidate gene that has a direct effect on liability according to the true model. We used data mining algorithms to bring the phenotypic and genotypic analyses together to glean more information from the data. A tree-based method was implemented to reduce the number of markers that were considered for the regression model; it selected only the marker at base pair 76. This marker has a significant p-value for testing the hypothesis of no linkage and is less than 500 base pairs from the site of the true gene that directly affects liability. By incorporating this SNP with the risk factors, we were able to improve the classification rate in our regression model by approximately seven percent in both the training and

validation data sets. In addition to characterizing the affection status, we also examined the relationship between the genotypic information and risk factors. Since we had already found several risk factors that were significantly related to disease status, we examined which genes were linked and associated with these phenotypes. A decision tree was applied to each of the continuous traits to form binary variables, making this analysis possible. This produced a new set of SNPs from the candidate genes that have an indirect effect on affection status that we otherwise would not have found. The relationships that we found between the quantitative traits and the candidate genes (candidate gene 6 with Q1 and Q2, and candidate gene 2 with Q5) are again consistent with the true model. Our analysis demonstrates that the integration of common data mining techniques with traditional genetic statistical methods is a valuable tool for detecting and describing complex genotypic-phenotypic relationships.

4.6 Acknowledgements

This work was supported in part by NIH grant GM45344.

4.7 Tables and Figures

Table 4.1: Fit Statistics for the Decision Tree and Logistic Regression Models Using Q1-Q5 as Model Inputs

Statistic	Train Rep 1	Validation Rep 42	Test Rep 3
Tree			
Misclassification Rate	0.152	0.176	0.188
Average Squared Error	0.112	0.140	0.158
Logistic Regression			
Misclassification Rate	0.121	0.145	0.157
Average Squared Error	0.093	0.098	0.120

Table 4.2: Parameter Estimates from Stepwise Logistic Regression Using Q1-Q5 as Model Inputs

Parameter	Estimate*	Standard Error	Wald Chi-square	P > Chi-square
Intercept	-23.936	3.955	36.63	0.0001
Q1	0.357	0.106	11.23	0.0008
Q2	0.203	0.093	4.80	0.0285
Q5	0.330	0.070	22.50	0.0010

*Parameter estimates are from replicate 1

Table 4.3: Parameter Estimates for the Stepwise Logistic Regression Using All of the Phenotypic Factors as Model Inputs

Parameter	Estimate*	Standard Error	Wald Chi-Square	P > Chi-Square
Intercept	-27.476	4.7163	33.94	0.0001
Q1	0.587	0.1157	25.78	0.0001
Q5	0.556	0.1142	23.70	0.0001
Q5xAge	-0.002	0.0008	9.51	0.0020

*Parameter estimates are from replicate 1

Table 4.4: SNPs from the Candidate Genes Showing Linkage and Association with Affection Status

Candidate gene	# of SNPs	# of SNPs showing linkage w/ disease*	# of SNPs showing association w/ disease*	# of SNPs showing both*
1	114	52	49	47
2	71	0	2	0
3	59	1	0	0
4	111	2	28	0
5	37	1	1	0
6	30	8	13	5
7	153	2	5	0

*$\alpha = .05$

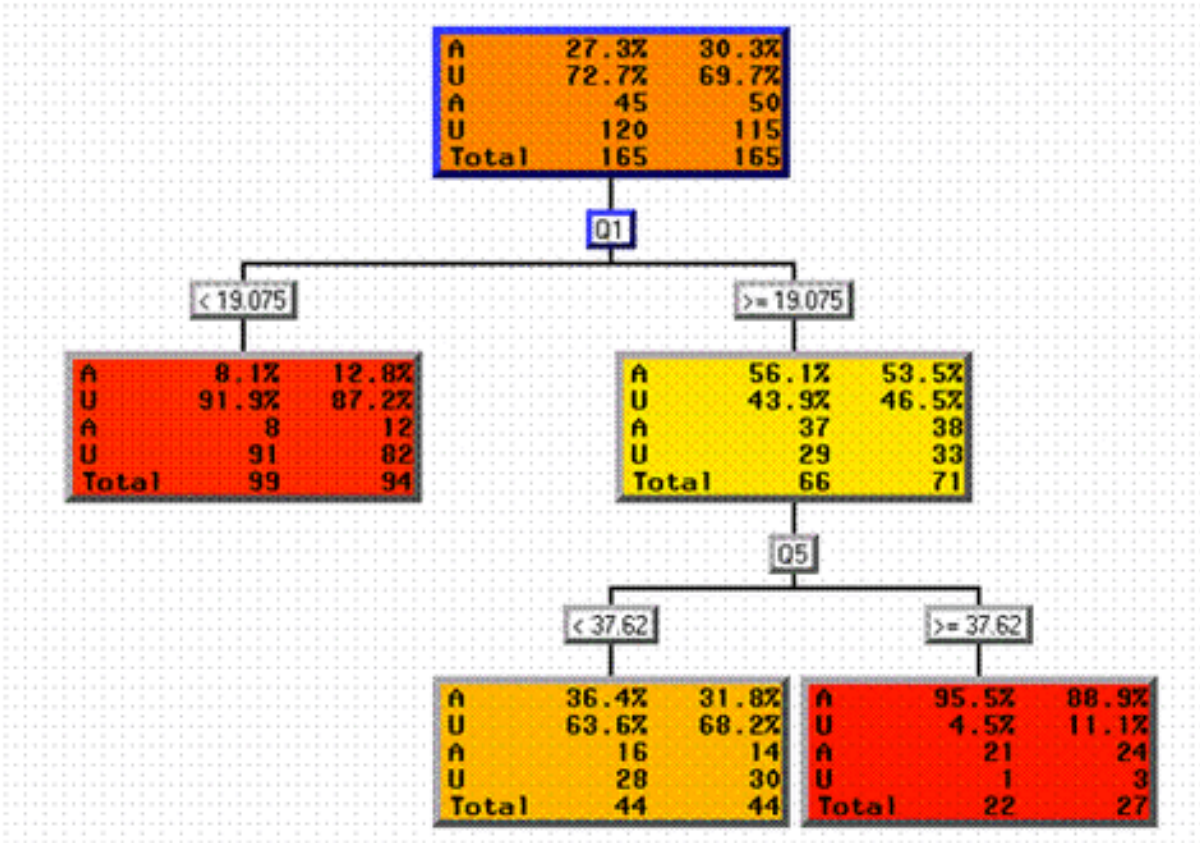

Figure 4.1: Tree Diagram Using Only Q1-Q5 as Predictors of Affection Status

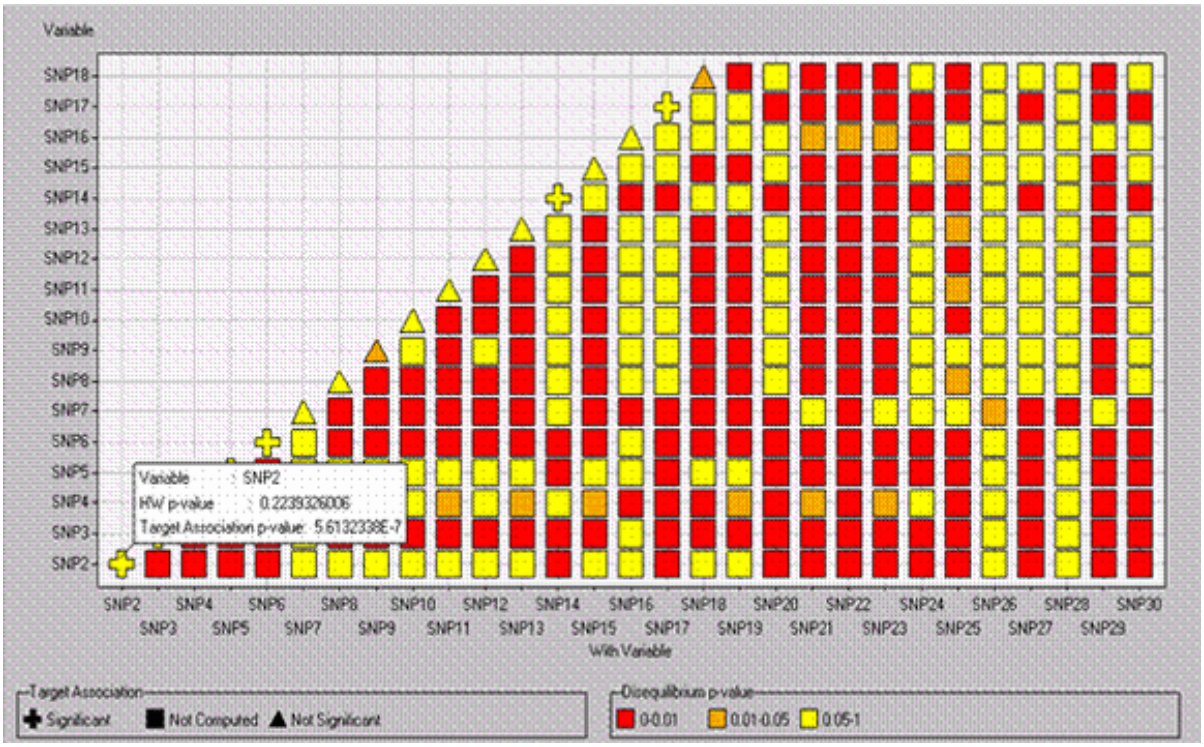

Figure 4.2: Association Map of Candidate Gene 1

Chapter 5

Testing for Three-Locus Disequilibrium

5.1 Introduction

In Chapter 2, we discussed single-marker case-control tests that can be used for the association mapping of a disease gene. We then mentioned the possible inadequacy of these tests for detecting multiple genes that act epistatically on disease susceptibility and suggested decision trees as an alternative in Chapter 3. A third category of association tests to consider is haplotype-based case-control tests. Nielsen et al. (2002) describe the following situations in which haplotype tests can be more powerful than single-marker tests: two markers within a gene may jointly affect disease susceptibility; or markers may have a haplotype structure such that no two-locus linkage disequilibrium (LD) exists

between two individual markers and the disease gene, but their three-locus disequilibrium is significant. The construction of two haplotype-based case-control tests is given by Nielsen et al. (2002), and then simulations are used to show that for many patterns of allele frequencies and LD, the haplotype tests outperform the single-marker tests that use allele or genotype counts.

We showed in Chapter 2 how the strength of LD between the marker and disease loci can positively affect the power of single-marker tests, and thus it is important to have a sense of how far LD extends across the portion of the genome that is being considered. Similarly, both the two- and three-locus disequilibria impact the power of the haplotype-based tests as we show below. In both situations, these disequilibria cannot be directly estimated since only disease phenotypes, not genotypes, are observed, but nonetheless, the identification of disequilibria patterns, facilitated by testing for nonzero disequilibria, among the markers in the genotyped regions can serve to characterize the behavior of these coefficients in the portion of the chromosome(s) being studied. Statistics testing for significance of two-locus disequilibria have been well-defined (Weir 1979; Weir 1996), but several tests used for detecting nonzero three-locus disequilibria have unresolved issues that we discuss here, while proposing new methods for testing the hypothesis of no three-locus disequilibria.

We presented a formulation for the association of genotypes at a marker locus with the presence or absence of a disease in Chapter 2. Genotypic associations depend on linear and quadratic functions of gametic and nongametic linkage disequilibria between

marker and disease alleles, whereas allelic associations depend only on linear functions of these disequilibria. For marker allele M_i, we can obtain the following frequencies among cases and controls by summing over genotypes:

$$\begin{aligned} \Pr(M_i|\text{Aff.}) &= q_i + \delta_i^c/K \\ \Pr(M_i|\text{Unaff.}) &= q_i - \delta_i^c/(1-K) \end{aligned}$$

where K is the disease prevalence in the population, q_i is the frequency of M_i in the whole population, and the composite association measure δ_i^c depends on the susceptibilities f_{rs} of all disease genotypes A_r/A_s:

$$\delta_i^c = \sum_{r,s} p_r f_{rs} \Delta_{is}$$

with Δ_{is} representing the composite linkage disequilibrium between M_i and A_s. When Hardy-Weinberg proportions hold, the usual linkage disequilibrium coefficient D_{is} can be used in place of Δ_{is} so that $\delta_i^c = \delta_i$.

Now consider a second marker **N** with alleles N_k. Nielsen and Weir (2001) showed that, under a random-mating model, two-locus marker haplotype frequencies within affected and unaffected individuals are

$$\begin{aligned} \Pr(M_i N_k|\text{Aff.}) &= P_{ik} + [q_k \delta_i^{(M)} + q_i \delta_k^{(N)} + \delta_{ik}^{(MN)}]/K \\ \Pr(M_i N_k|\text{Unaff.}) &= P_{ik} - [q_k \delta_i^{(M)} + q_i \delta_k^{(N)} + \delta_{ik}^{(MN)}]/(1-K) \end{aligned}$$

where the association measures δ for each marker are distinguished with a superscript:

$$\delta_i^{(M)} = \sum_{r,s} p_r f_{rs} D_{si}$$

$$\delta_k^{(N)} = \sum_{r,s} p_r f_{rs} D_{sk}$$

The new marker haplotype association measure $\delta_{ik}^{(MN)}$ is

$$\delta_{ik}^{(MN)} = \sum_{r,s} p_s D_{rik} f_{rs}$$

Allelic associations involve two-locus disequilibria, and two-locus haplotype associations involve both two- and three-locus gametic disequilibria. These disequilibria are

$$\begin{aligned} D_{ri} &= P_{ri} - p_r q_i \\ D_{rk} &= P_{rk} - p_r q_k \\ D_{rik} &= P_{rik} - q_i D_{rk} - q_k D_{ir} - p_r D_{ik} - p_r q_i q_k \end{aligned}$$

The three-locus disequilibrium measure was introduced by Bennett (1954), and this chapter considers methods for drawing inferences about the measure. Evidently, the difference between marker haplotype frequencies in cases and controls depends on this quantity:

$$\begin{aligned} \Pr(M_i N_k|\text{Aff.}) - \Pr(M_i N_k|\text{Unaff.}) &= q_k[\Pr(M_i|\text{Aff.}) - \Pr(M_i|\text{Unaff.})] + q_i[\Pr(N_k|\text{Aff.}) \\ &\quad - \Pr(N_k|\text{Unaff.})] + \delta_{ik}^{(MN)}/[K(1-K)] \end{aligned}$$

and any advantage of the two-locus analysis over the two single-locus analyses depends on $\delta_{ik}^{(MN)}$, and hence on D_{rik}. Since we can use observed estimates of allele and haplotype frequencies in the cases and controls in this formula, we can solve for an estimate of $\delta_{ik}^{(MN)}$, but our interest is in predicting the behavior of haplotype case-control tests from

values of D_{rik}. Note that, as with the two-locus disequilibria, the D_{rik} can be shown to decay over t generations from the initial disequilibrium D^0_{irk} as

$$D^t_{irk} \;=\; D^0_{irk}[(1-1/N)(1-2/N)(1-\theta_{\mathrm{MA}})(1-\theta_{\mathrm{AN}})]^t$$

(Hill 1976) assuming **A** is located between **M** and **N**; θ_{MA} and θ_{AN} are the recombination fractions of **A** with **M** and **N**, respectively; and a population of size $N/2$ comprising N chromosomes.

5.2 Example

To give an example of a comparison between single-marker and haplotype-based case-control tests, we now create a data set based on the estimated haplotype frequencies for the *APOE* data used by Fallin et al. (2001). Their data contain 210 individuals affected with Alzheimer disease and 159 individuals unaffected with the disease. Single-marker and haplotype-based tests using haplotype frequencies estimated from the EM algorithm were performed on eight SNPs in the *APOE* gene region in their analysis, including a SNP containing an actual disease allele. We focus on markers M1 with alleles C and T and M2 with alleles A and G. According to the results given by Fallin et al. (2001) as well as the results from analyzing the data set we created to mimic theirs, neither of these markers shows a significant association with Alzheimer disease based on allele case-control tests at the 0.05 level. However, using the two-locus haplotype frequencies that we inferred from the estimated four-locus haplotype frequencies given by Fallin et al. (2001), Table 5.2

shows the significant chi-square statistic from the haplotype case-control test. We can estimate the association parameters $\delta_C^{(M1)}$, $\delta_A^{(M2)}$ and $\delta_{CA}^{(M1,M2)}$ by plugging in the observed allele frequencies in the total sample and in cases to give

$$\begin{aligned} \tilde{\delta}_C^{(M1)} &= 0.00661 \\ \tilde{\delta}_A^{(M2)} &= -0.00595 \\ \tilde{\delta}_{CA}^{(M1,M2)} &= -0.01513 \end{aligned}$$

The magnitude of the haplotype association measure gives an indication why the haplotype-based test statistic was more significant than the single-marker ones. This measure is a function of the three-locus disequilibria between markers M1 and M2 and the disease locus.

5.3 Existing Asymptotic Tests

The involvement of three-locus disequilibria in the two-marker-haplotype association tests suggests that it may be of interest to examine patterns of three-locus disequilibrium among marker loci. Various approaches have been suggested for testing that this quantity is zero (Hill 1975; Hill 1976; Smouse 1974; Weir 1996), and these are now reviewed. The approaches of Hill and Weir, however, ignore the fact that the parameter may be bounded away from zero so that testing for a zero value may not be appropriate.

Suppose we have three markers **A**, **B**, and **C** with two alleles each: A, a; B, b; and

C, c, respectively. For the additive model of disequilibrium

$$P_{ABC} = p_A p_B p_C + p_A D_{BC} + p_B D_{AC} + p_C D_{AB} + D_{ABC}$$

Hill (1976) showed that, in the two allele case, the goodness-of-fit chi-square test statistic for the hypothesis of complete independence: $H_0 : P_{ABC} = p_A p_B p_C$ when the sample has n chromosomes can be calculated using sample allele and haplotype frequencies as

$$\begin{aligned} X^2_H &= n\left(\frac{(\tilde{P}_{ABC} - \tilde{p}_A\tilde{p}_B\tilde{p}_C)^2}{\tilde{p}_A\tilde{p}_B\tilde{p}_C} + \frac{(\tilde{P}_{ABc} - \tilde{p}_A\tilde{p}_B\tilde{p}_c)^2}{\tilde{p}_A\tilde{p}_B\tilde{p}_c} + \frac{(\tilde{P}_{AbC} - \tilde{p}_A\tilde{p}_b\tilde{p}_C)^2}{\tilde{p}_A\tilde{p}_b\tilde{p}_C}\right. \\ &\quad + \frac{(\tilde{P}_{Abc} - \tilde{p}_A\tilde{p}_b\tilde{p}_c)^2}{\tilde{p}_A\tilde{p}_b\tilde{p}_c} + \frac{(\tilde{P}_{aBC} - \tilde{p}_a\tilde{p}_B\tilde{p}_C)^2}{\tilde{p}_a\tilde{p}_B\tilde{p}_C} + \frac{(\tilde{P}_{aBc} - \tilde{p}_a\tilde{p}_B\tilde{p}_c)^2}{\tilde{p}_a\tilde{p}_B\tilde{p}_c} \\ &\quad \left. + \frac{(\tilde{P}_{abC} - \tilde{p}_a\tilde{p}_b\tilde{p}_C)^2}{\tilde{p}_a\tilde{p}_b\tilde{p}_C} + \frac{(\tilde{P}_{abc} - \tilde{p}_a\tilde{p}_b\tilde{p}_c)^2}{\tilde{p}_a\tilde{p}_b\tilde{p}_c}\right) \\ &= n\left(r^2_{AB} + r^2_{AC} + r^2_{BC} + r^2_{ABC}\right) \end{aligned}$$

where

$$r^2_{AB} = \frac{\tilde{D}^2_{AB}}{\tilde{p}_A\tilde{p}_a\tilde{p}_B\tilde{p}_b} \tag{5.1}$$

$$r^2_{AC} = \frac{\tilde{D}^2_{AC}}{\tilde{p}_A\tilde{p}_a\tilde{p}_C\tilde{p}_c} \tag{5.2}$$

$$r^2_{BC} = \frac{\tilde{D}^2_{AB}}{\tilde{p}_B\tilde{p}_b\tilde{p}_C\tilde{p}_c} \tag{5.3}$$

$$r^2_{ABC} = \frac{\tilde{D}^2_{ABC}}{\tilde{p}_A\tilde{p}_a\tilde{p}_B\tilde{p}_b\tilde{p}_C\tilde{p}_c} \tag{5.4}$$

The 4 df statistic X^2_H has therefore been partitioned into four terms, the first three of which are the 1 df statistics for testing two-locus disequilibria. This suggests that the fourth term is a 1 df chi-square statistic for testing $H_0 : D_{ABC} = 0$. Hill (1976) points

to the controversy surrounding this form of Lancaster (1951) partitioning. The three two-locus disequilibria are not independent, and this is missed by this partitioning.

Hill (1976) and Smouse (1974) instead consider a multiplicative model, where the null hypothesis is $Z_{ABC} = 1$ with

$$Z_{ABC} = \frac{P_{ABC}P_{Abc}P_{aBc}P_{abC}}{P_{ABc}P_{AbC}P_{aBC}P_{abc}}$$

corresponding to a common measure of association in a $2\times2\times2$ contingency table (Goodman 1969; Fienberg 1970). This can be tested using a likelihood approach. There are seven df and seven parameters in the full model: $p_A, p_B, p_C, D_{AB}, D_{AC}, D_{BC}, D_{ABC}$. Note that sample allele and haplotype counts are represented by an n with the corresponding subscript(s). The full unconstrained maximum likelihood is therefore easy to calculate:

$$L_1 \propto \tilde{P}_{ABC}^{n_{ABC}} \tilde{P}_{ABc}^{n_{ABc}} \tilde{P}_{AbC}^{n_{AbC}} \tilde{P}_{aBC}^{n_{aBC}} \tilde{P}_{Abc}^{n_{Abc}} \tilde{P}_{aBc}^{n_{aBc}} \tilde{P}_{abC}^{n_{abC}} \tilde{P}_{abc}^{n_{abc}}$$

with the eight three-locus gametic probabilities given in Table 5.1, and a likelihood ratio test can be constructed by comparing this to the likelihood under the null hypothesis:

$$L_0 \propto \tilde{p}_A^{n_A} \tilde{p}_B^{n_B} \tilde{p}_C^{n_C} (1-\tilde{p}_A)^{n-n_A} (1-\tilde{p}_B)^{n-n_B} (1-\tilde{p}_C)^{n-n_C}$$

with only three parameters giving $X_L^2 = -2(\ln L_0 - \ln L_1) \sim \chi_4^2$ assuming Hardy-Weinberg equilibrium. Analogous to the X_H^2 statistic described above, this statistic is testing for any two- or three-locus associations and can be partitioned into four 1 df chi-square statistics. These four statistics, however, do not test for the three individual two-locus associations and the three-locus disequilibrium as with the X_H^2 statistic like we would

hope; rather, they involve marginal and conditional associations so the genetic interpretation of these tests is not as straightforward as with the goodness-of-fit approach. Additionally, two of these likelihood ratio chi-square statistics involve estimates of haplotype frequencies allowing for pairwise, but no three-locus, associations for which there is no formula. Iterative procedures given for example by Fienberg (1970) must be used.

An explicit test statistic for the three-locus disequilibrium was given by Weir (1996):

$$X^2_W = \frac{\tilde{D}^2_{ABC}}{\widehat{\mathrm{Var}}(\tilde{D}_{ABC})}$$

where, under the assumption that $D_{ABC} = 0$

$$\begin{aligned}\widehat{\mathrm{Var}}(\tilde{D}_{ABC}) = \frac{1}{n}\Big\{&\tilde{p}_A(1-\tilde{p}_A)\tilde{p}_B(1-\tilde{p}_B)\tilde{p}_C(1-\tilde{p}_C) + 6\tilde{D}_{AB}\tilde{D}_{AC}\tilde{D}_{BC} \\ &+ \tilde{p}_A(1-\tilde{p}_A)[(1-2\tilde{p}_B)(1-2\tilde{p}_C)\tilde{D}_{BC} - \tilde{D}^2_{BC}] \\ &+ \tilde{p}_B(1-\tilde{p}_B)[(1-2\tilde{p}_A)(1-2\tilde{p}_C)\tilde{D}_{AC} - \tilde{D}^2_{AC}] \\ &+ \tilde{p}_C(1-\tilde{p}_C)[(1-2\tilde{p}_A)(1-2\tilde{p}_B)\tilde{D}_{AB} - \tilde{D}^2_{AB}]\Big\}\end{aligned}$$

The problem with this equation is that the observed gametic frequencies may not allow D_{ABC} to be zero, and this can lead to negative values for the variance expression when D_{ABC} is set to zero. D_{ABC} can be shown to be constrained by $L \leq D_{ABC} \leq U$, where the bounds are defined as

$$\begin{aligned}L &= \max(-S, P_{AB} + P_{AC} - p_A - S, P_{AB} + P_{BC} - p_B - S, P_{AC} + P_{BC} - p_C - S) \\ U &= \min(p_A - S, p_B - S, p_C - S, 1 - p_A - p_B - p_C + P_{AB} + P_{AC} + P_{BC} - S)\end{aligned}$$

with $S = p_A p_B p_C + p_A D_{BC} + p_B D_{AC} + p_C D_{AB}$. These bounds can be shown to be equal to those given by Thomson and Baur (1984). Both L and U can be positive or negative.

We mentioned previously that of particular interest is the situation where the three two-locus disequilibria D_{AB}, D_{AC}, and D_{BC} are zero but there is significant three-locus disequilibria, in which case the haplotype tests are more powerful than the single-marker association tests. It is important to note that when there are no two-locus associations, L is negative and U is positive, assuming both alleles are observed at each of the three marker loci in the sample, and the estimate of the variance of D_{ABC} when all two- and three-locus disequilibria are zero is simply the product of the six allele frequencies at the three loci divided by the sample size, yielding X_W^2 equivalent to nr_{ABC}^2.

5.4 An Exact Method

Calculating the exact probabilities of the possible values for D_{ABC} holding the sample size n, allele frequencies, and two-locus haplotype frequencies constant can provide a means for testing hypotheses about D_{ABC}. That is, D_{ABC} is estimated for all sets of valid three-locus haplotype counts given the single- and two-locus counts. Valid haplotype counts can be produced by using integers ranging from $n(L+S)$ to $n(U+S)$ for n_{ABC}, then the other seven haplotype counts are given by subtraction for each of the n_{ABC}. To calculate probabilities of the haplotype counts under the null distribution that allows for two-locus disequilibrium but no three-locus disequilibrium, in general, an iterative procedure must be used to arrive at the estimates of the three-locus haplotype frequencies (Hill 1976). However, when the two-locus disequilibria are zero, mentioned above as a situation of interest, we can use the product of the observed allele frequencies for the three-locus

haplotype frequencies. The probability of observing each of these sets of eight three-locus haplotype counts under the null hypothesis is then given by the multinomial distribution as

$$\Pr(n_{ABC}, n_{ABc}, \ldots, n_{abc}) = \frac{\frac{n!}{n_{ABC}!\cdots n_{abc}!}(\tilde{p}_A\tilde{p}_B\tilde{p}_C)^{n_{ABC}}\cdots(\tilde{p}_a\tilde{p}_b\tilde{p}_c)^{n_{abc}}}{\sum_{n_{ABC}=n(L+S)}^{n(U+S)}\frac{n!}{n_{ABC}!\cdots n_{abc}!}(\tilde{p}_A\tilde{p}_B\tilde{p}_C)^{n_{ABC}}\cdots(\tilde{p}_a\tilde{p}_b\tilde{p}_c)^{n_{abc}}}$$

since not all values $0, \ldots, n$ are necessarily valid for each of the three-locus haplotype counts. A p-value for testing $\mathrm{H}_0 : D_{ABC} = 0$ can be calculated by summing the probabilities of all D_{ABC} possible for the fixed allele and two-locus haplotype counts that are greater or equal in absolute value to the absolute value of the observed $\tilde{D}_{ABC}$.

We now describe an analysis performed to compare this exact method with the asymptotic test statistic X_W^2. Using all possible combinations of allele frequencies from the set $\{0.1, 0.2, \ldots, 0.8, 0.9\}$ for the three marker loci, we generated the distribution of D_{ABC} given these allele frequencies and setting the two-locus disequilibria, D_{AB}, D_{AC}, and D_{BC}, to 0. There were 165 such distinct combinations of allele frequencies. A sample size of $n = 100$ haplotypes was used. Of these combinations, 25 had only two distinct sets of haplotype counts possible, and thus only two possible values for D_{ABC} given the allele and two-locus haplotype counts. The maximum number of distinct values, 26, for D_{ABC} occurred when $p_A = p_B = p_C = 0.5$. Using the probability formula given above for each of the three-locus disequilibria, we show the distribution of D_{ABC} for several sets of allele frequencies in Figures 5.1-5.4. Comparing the p-values from these two approaches, p-values from the asymptotic χ_1^2 distribution are significant at level 0.05 for 584 out of 1074 values of D_{ABC} from the 165 runs, while 488 p-values using the exact distribution

are significant.

5.5 Discussion

With marker case-control tests expanding from analyzing single loci at a time to using haplotype information, it is clear that a similar extension from the analysis of two-locus disequilibrium to three-locus disequilibrium is necessary. These disequilibria give the corresponding tests the power to detect an association. Though tests for three-locus associations were developed 30 years ago, there remain disadvantages to both the goodness-of-fit approach, which does not test for a true three-way, contingency-table type of association, and the likelihood approach, which can be computationally intensive and does not have as natural of a genetic interpretation. The explicit method of testing $D_{ABC} = 0$ of Weir (1996) using Fisher's formula for the variance of this quantity is limited by the bounds of D_{ABC} possibly resulting in a negative variance. These bounds also greatly restrict the possible values for D_{ABC} so that even for a sample size of 100, the normal distribution may not hold well. Using the exact test based on multinomial probabilities is an alternative approach that does not rely on asymptotic theory, though other than for the case when there are no two-locus associations, iterative procedures are again required to estimate three-locus haplotype frequencies accounting for the pairwise, but not the three-locus, disequilibria. Future work will include the development of alternative testing procedures.

5.6 Tables and Figures

Table 5.1: Three-Locus Gamete Probabilities

	1	D_{BC}	D_{AC}	D_{AB}	D_{ABC}
P_{ABC}	$p_A p_B p_C$	p_A	p_B	p_C	1
P_{ABc}	$p_A p_B (1-p_C)$	$-p_A$	$-p_B$	$(1-p_C)$	-1
P_{AbC}	$p_A (1-p_B) p_C$	$-p_A$	$(1-p_B)$	$-p_C$	-1
P_{Abc}	$p_A (1-p_B)(1-p_C)$	p_A	$-(1-p_B)$	$-(1-p_C)$	1
P_{aBC}	$(1-p_A) p_B p_C$	$(1-p_A)$	$-p_B$	$-p_C$	-1
P_{aBc}	$(1-p_A) p_B (1-p_C)$	$-(1-p_A)$	p_B	$-(1-p_C)$	1
P_{abC}	$(1-p_A)(1-p_B) p_C$	$-(1-p_A)$	$-(1-p_B)$	p_C	1
P_{abc}	$(1-p_A)(1-p_B)(1-p_C)$	$(1-p_A)$	$(1-p_B)$	$(1-p_C)$	-1

Table 5.2: Contingency Table for M1 and M2 Haplotypes

	CA	CG	TA	TG	Total
Cases	51	169	199	1	420
Controls	57	101	140	20	318
Total	108	270	339	21	738
Chi-Square=31.421, DF=3, P-value$<$ 0.0001					

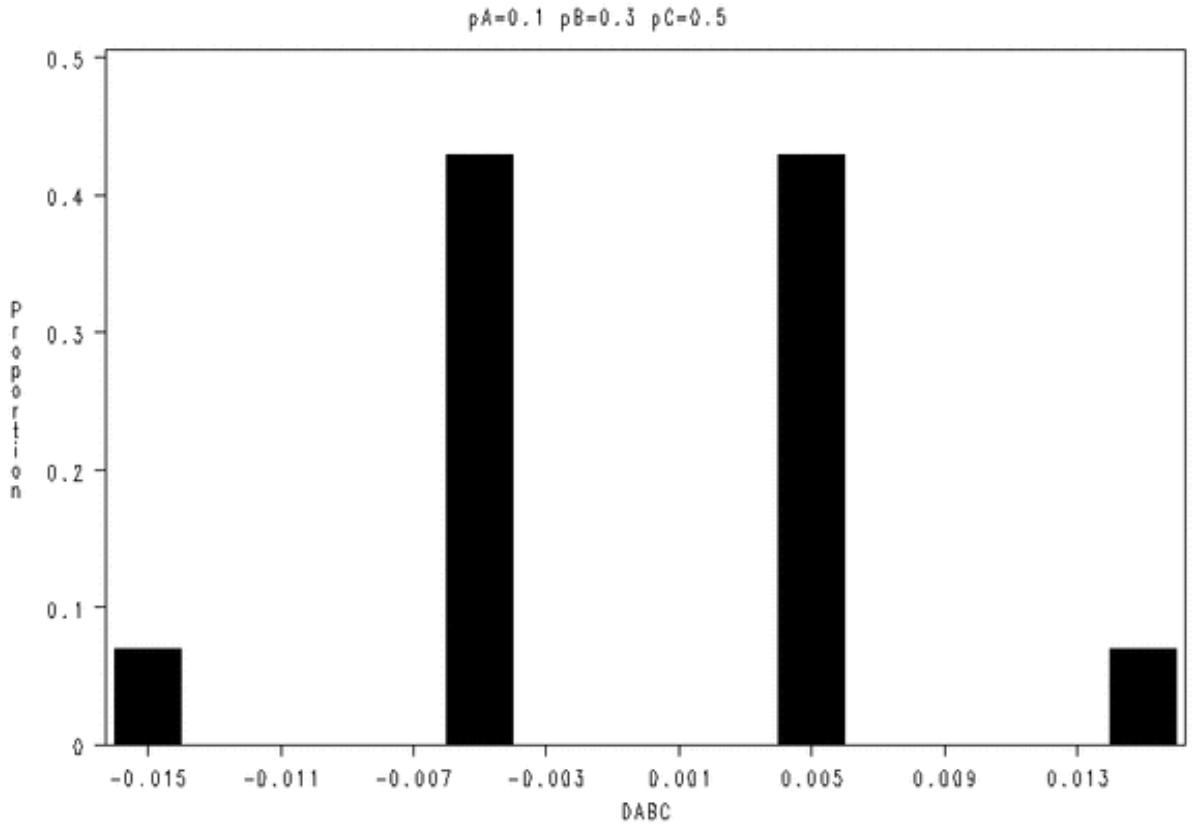

Figure 5.1: Distribution of D_{ABC} when $D_{AB} = D_{AC} = D_{BC} = 0$ and $p_A = 0.1$, $p_B = 0.3$, $p_C = 0.5$

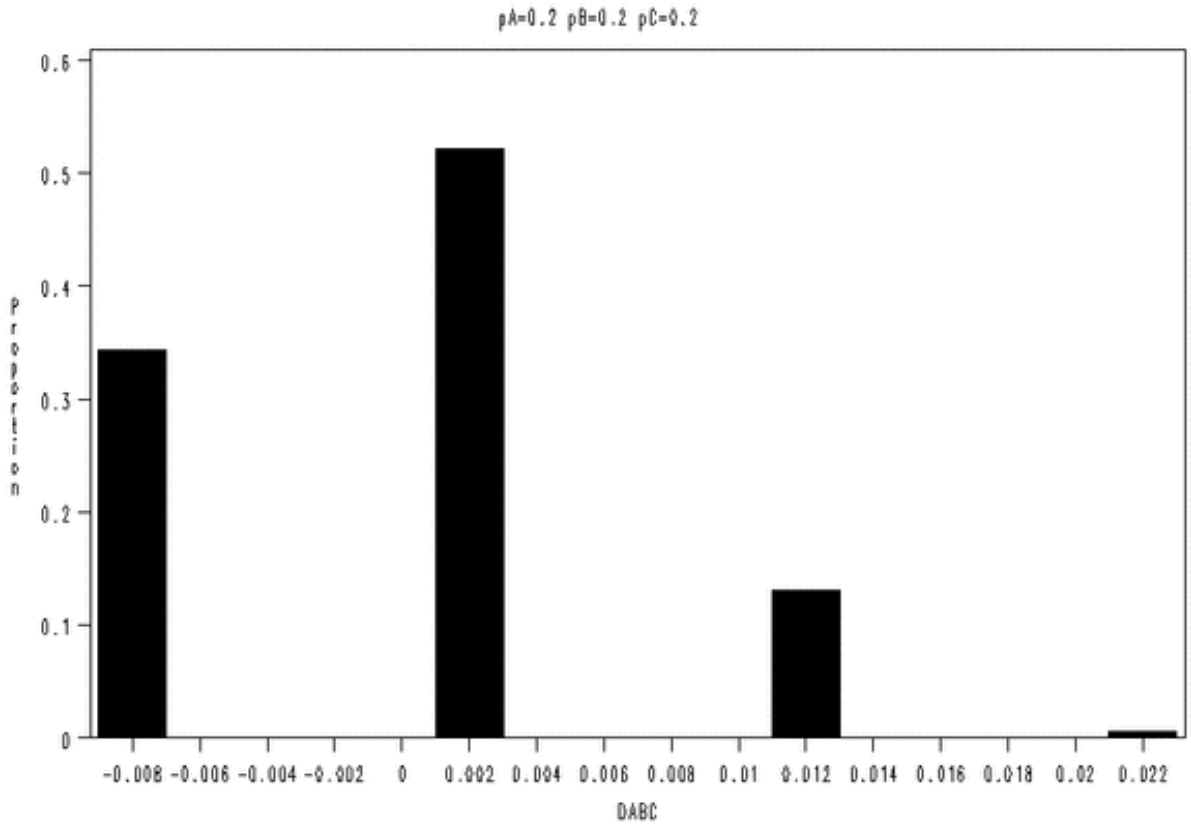

Figure 5.2: Distribution of D_{ABC} when $D_{AB} = D_{AC} = D_{BC} = 0$ and $p_A = 0.2$, $p_B = 0.2$, $p_C = 0.2$

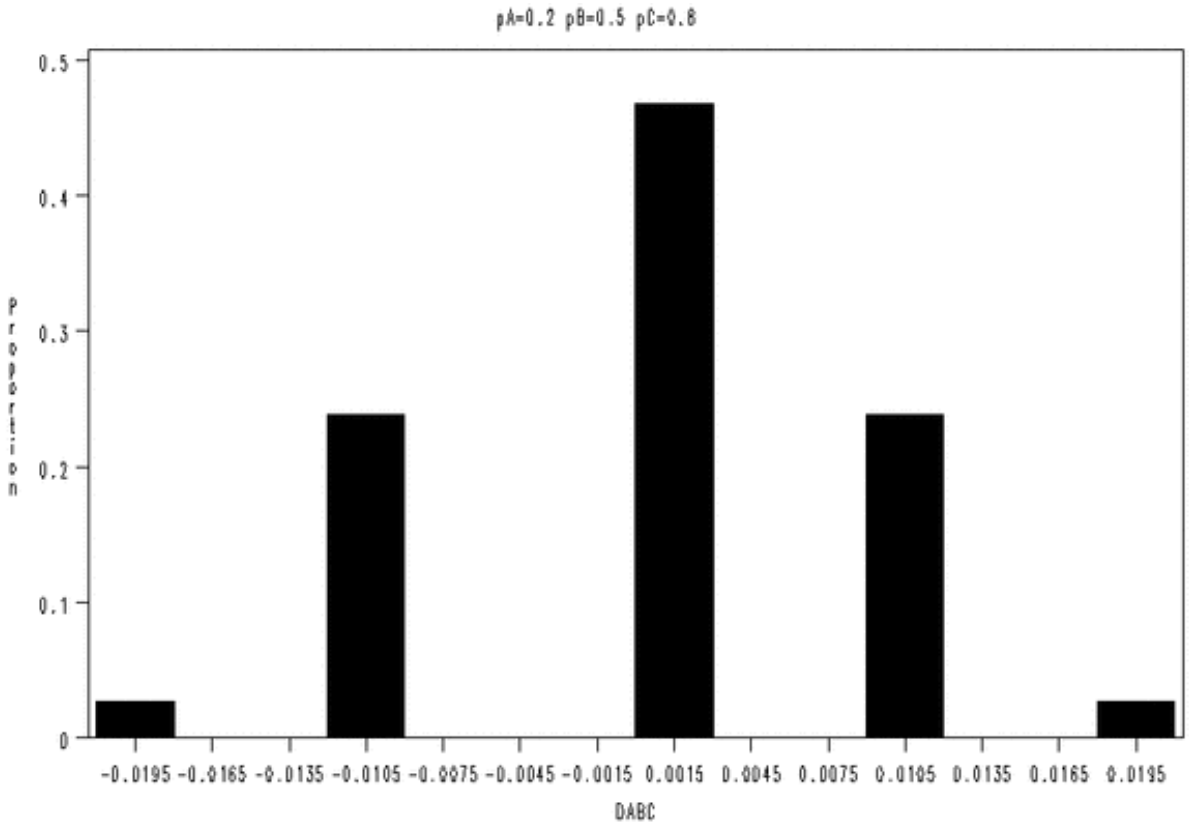

Figure 5.3: Distribution of D_{ABC} when $D_{AB} = D_{AC} = D_{BC} = 0$ and $p_A = 0.2$, $p_B = 0.5$, $p_C = 0.8$

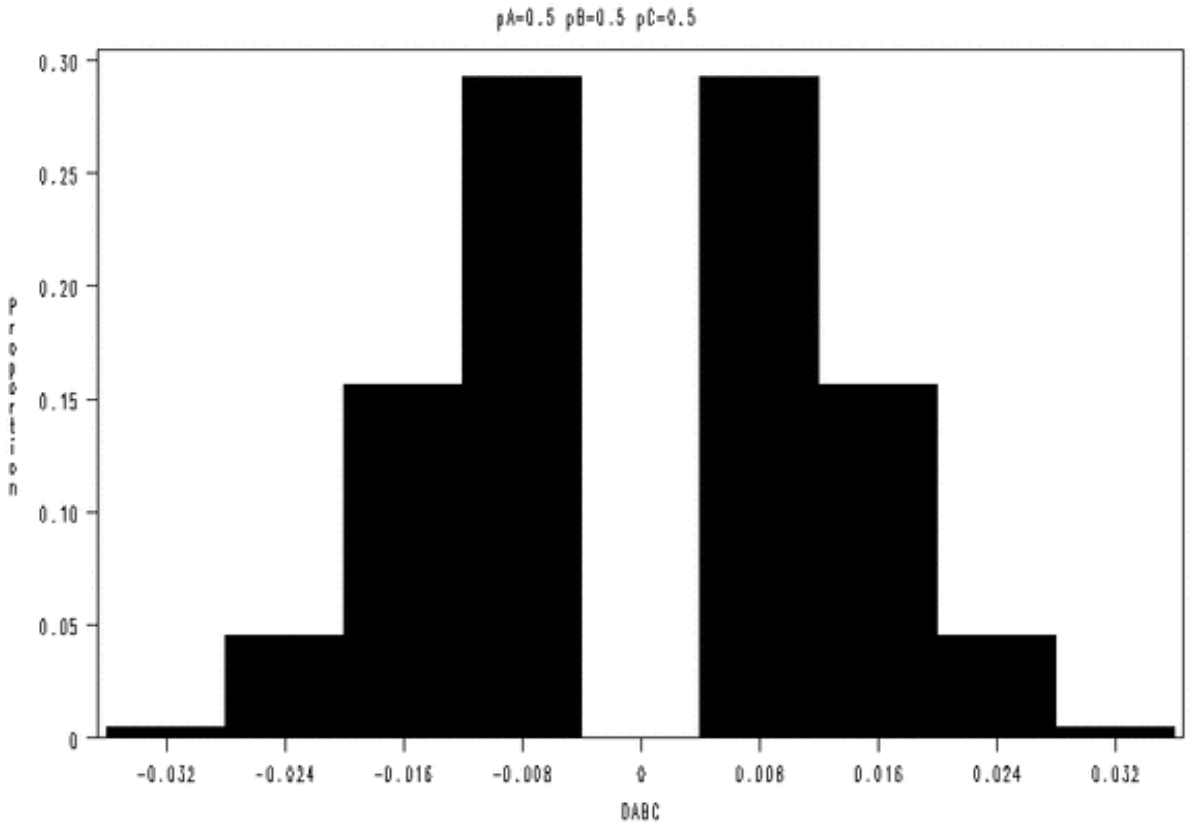

Figure 5.4: Distribution of D_{ABC} when $D_{AB} = D_{AC} = D_{BC} = 0$ and $p_A = 0.5$, $p_B = 0.5$, $p_C = 0.5$

Chapter 6

Using All Alleles in the Multiallelic Versions of the SDT and Combined SDT/TDT

Czika, W. and J. J. Berry (2002). *American Journal of Human Genetics 71*, 1235-1236.

6.1 Introduction

Horvath and Laird's sibling disequilibrium test (SDT) provides a nonparametric approach to testing genetic markers for both linkage and association with a disease (1998). The advantage over its parametric alternatives is its validity as a test of association when using sibships containing more than one affected sibling and/or more than one unaffected sibling. Horvath and Laird introduced an SDT for multiallelic markers and a biallelic combined SDT/TDT when some parental genotypic information is available. Curtis, Miller, and Sham (1999) later developed a multiallelic combined SDT/TDT. The multiallelic versions of these tests are designed for situations in which there is no *a priori* knowledge of which allele at a marker may have an effect on disease status; otherwise, a biallelic test can be performed on the allele of interest versus all other alleles collapsed into one. A problem with the multiallelic extensions is the statistic varies depending on which allele is omitted from the analysis. We present an alternative multiallelic SDT (mSDT) that takes into account all the allelic information and is consistent with the biallelic approach. This method can also be applied to the combined SDT/TDT.

6.2 Justification for the mSDT

In calculating the multiallelic versions of both the SDT and combined SDT/TDT, the statistics $d^j, j = 1, ..., m$ for a marker with m alleles are used. In the SDT, $d^j = \sum_i d_i^j$ where the d_i^j represent the difference between the average number of times allele j occurs

in an affected sibling and the average number of times it occurs in an unaffected sibling within sibship i (Horvath and Laird 1998); for the combined SDT/TDT, d^j is the difference between the number of times allele j is transmitted and the number times it is not transmitted from a heterozygous parent to an affected child (Sham 1997). As discussed in Stuart (1955), a quadratic form of the d^j can be used to create a statistic with an asymptotic chi-square distribution. It is noted that since $\sum_{j=1}^{m} d^j = 0$, the degrees of freedom (df) for the distribution are $(m-1)$. Furthermore, since using all m columns of the variance-covariance matrix creates a singularity and thus the matrix is uninvertible, the natural solution is to eliminate one of the d^j and the corresponding row and column in the variance-covariance matrix to make it full-rank. The invariance of the chi-square statistic according to which variate (d^j) is omitted from the statistic is demonstrated by Stuart (1955).

To create a nonparametric test, $S_i^j = \text{sgn}(d_i^j)$ is used in place of d_i^j, where $\text{sgn}(d) = -1, 0, 1$ for $d <, =, > 0$ respectively. Though the sum of the quantities $d_i^j, j = 1, ..., m$, is 0 for each sibship $i = 1, ..., N$ and $S_i^1 = -S_i^2$ in the biallelic case, for more than two alleles, the sum over j of the S_i^j is not similarly linearly constrained within a sibship. In fact, the S_i^j can sum over j to either -1, 0, or 1. Despite this fact, multiallelic extensions to the SDT and combined SDT/TDT are formed by arbitrarily dropping one of the $S^j = \sum_{i=1}^{N} S_i^j$ from the analysis. The resulting $\chi^2_{(m-1)}$ test statistic is no longer invariant to which allele's information has been omitted since there is no linear dependency among the S^j; information is being unnecessarily discarded. Furthermore, the variance-covariance

matrix $\mathbf{W}$ for $\mathbf{S} = (S^1, ..., S^m)$ is nonsingular (exceptions are discussed below) before omitting any of the m alleles. Thus, when all m alleles are used, a valid test statistic can still be created as $\mathbf{S'W^{-1}S}$, which has an asymptotic $\chi^2_{(m)}$ distribution (Hettmansperger 1984; Randles 1989).

There are, as mentioned, situations in which $\mathbf{W}$ will not be full-rank. Among these are

1. the biallelic case: in this case, the S^j are constrained since there is a perfect negative correlation between S_i^1 and S_i^2 for all i ($\sum_{j=1,2} S_i^j = 0$ for all i)

2. If there exists at least one allele j such that $S_i^j = 0$ for all N sibships. Thus, this allele will have a row and column of 0s in $\mathbf{W}$ creating a singularity.

3. $\sum_{j=1}^m S_i^j = C$, the same constant, for all N sibships

For these situations, we recommend the use of the Moore-Penrose generalized inverse (g-inverse) of the variance-covariance matrix $\mathbf{W}$, $\mathbf{W}^-$. This is a unique generalized inverse of $\mathbf{W}$ that satisfies the following conditions (Rao and Mitra 1971; Searle 1971): $\mathbf{WW}^-$ and $\mathbf{W}^-\mathbf{W}$ are symmetric; $\mathbf{W}^-\mathbf{WW}^- = \mathbf{W}^-$; and $\mathbf{WW}^-\mathbf{W} = \mathbf{W}$. It is worth noting that the last two scenarios listed for a singular variance-covariance matrix are possible with the original SDT statistic even after having omitted one allele from the analysis, in which case the statistic cannot be calculated since $\mathbf{W}$ is uninvertible.

When using $\mathbf{W}^-$ in place of $\mathbf{W^{-1}}$ in the quadratic form, the test statistic $\mathbf{S'W}^-\mathbf{S}$ still has an asymptotic chi-square distribution, now with df equal to the rank of $\mathbf{W}$ (Rao and

Mitra 1971). Note that for the biallelic case, using Horvath and Laird's notation (1998), the mSDT gives $\mathbf{S} = (b - c, c - b)$ and the $\mathbf{W}$ matrix will be of the form:

$$\begin{bmatrix} b+c & -(b+c) \\ -(b+c) & b+c \end{bmatrix}$$

The g-inverse is then calculated as

$$\begin{bmatrix} 1/(4b+4c) & -1/(4b+4c) \\ -1/(4b+4c) & 1/(4b+4c) \end{bmatrix}$$

which yields a chi-square statistic of $(b-c)^2/(b+c)$ with 1 df, the same as the usual biallelic statistic.

6.3 Summary and Example

To summarize our approach, we suggest modifying Horvath and Laird's SDT statistic (1998) and the combined SDT/TDT of Curtis, Miller, and Sham (1999) in the following manner to calculate the statistic for the mSDT:

1. use all m alleles in the $\mathbf{S}$ vector and $\mathbf{W}$ matrix
2. to create the chi-square statistic, use $\mathbf{W}^-$ in place of $\mathbf{W^{-1}}$ (note that these are identical when $\mathbf{W}$ is full-rank)
3. use rank($\mathbf{W}$) as the df for the chi-square distribution

We give an example here using simulated data from GAW9 (Hodge 1995). As in Spielman and Ewens (1998) and Knapp (1999), we focus on multiallelic markers D1G31

and D5G23, each which contain an actual disease allele, M8 and M7 respectively. Table 6.1 shows the results of analyzing the data using the original Horvath and Laird SDT method where each allele is dropped in turn. Also shown are the results from analyzing the data using our mSDT approach. Note that each marker has eight alleles so p-values from the SDT are based on a χ^2_7 distribution, while the mSDT p-values are from a χ^2_8 distribution since the variance-covariance matrices for both markers are full-rank. This example is not intended as any sort of power comparison, but merely to illustrate that there is not necessarily a loss of power by introducing an additional degree of freedom. The other thing to note from this table is the variation of the SDT p-values depending on which allele is dropped. While for marker D5G23, all test statistics are highly significant, we can see quite a discrepancy between the SDT statistic for marker D1G31 when dropping allele M8 and any of the other seven SDT statistics. The mSDT approach will always give a unique chi-square statistic, regardless of whether $\mathbf{W}$ is full-rank or not. This method will be available in a future release of SAS/Genetics$^{\text{TM}}$.

6.4 Tables

Table 6.1: SDT and mSDT Statistics for Two Markers Linked and Associated with Disease

	D1G31		D5G23	
Allele Dropped	Chi-Square	*P*-Value	Chi-Square	*P*-Value*
M1	23.115255	0.001628	52.441075	0.000048
M2	23.543802	0.001370	52.365979	0.000049
M3	23.239746	0.001548	52.382481	0.000049
M4	23.621073	0.001328	51.086058	0.000088
M5	23.661028	0.001307	52.546616	0.000046
M6	23.648748	0.001313	53.238694	0.000033
M7	23.417311	0.001441	45.631132	0.001031
M8	14.806102	0.038567	51.811979	0.000064
mSDT	23.667390	0.002605	53.455015	0.000088

*p-values multiplied by 10^4

Bibliography

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.

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

Bennett, J. H. (1954). On the theory of random mating. *Annals of Eugenics 18*, 311–317.

Bickel, P. J. (1965). On some asymptotic competitors to Hotellings t^2. *Annals of Mathematical Statistics 36*, 160–173.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth, Inc.

Chapman, N. H. and E. M. Wijsman (1998). Genome screens using linkage disequilibrium tests: Optimal marker characteristics and feasibility. *American Journal of Human Genetics 63*, 1872–1885.

Curtis, D., M. B. Miller, and P. C. Sham (1999). Combining the sibling disequilibrium test and transmission/disequilibrium test for multiallelic markers. *American*

Journal of Human Genetics 64, 1785–1786.

Czika, W. A., B. S. Weir, S. R. Edwards, R. W. Thompson, D. M. Nielsen, J. C. Brocklebank, C. Zinkus, E. R. Martin, and K. E. Hobler (2001). Applying data mining techniques to the mapping of complex disease genes. *Genetic Epidemiology 21*, S435–S440.

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

Fallin, D., A. Cohen, L. Essioux, I. Chumakov, M. Blumenfeld, D. Cohen, and N. J. Schork (2001). Genetic analysis of case/control data using estimated haplotype frequencies: Application to APOE locus variation and Alzheimer's disease. *Genome Research 11*, 143–151.

Fienberg, S. E. (1970). The analysis of multidimensional contingency tables. *Ecology 51*, 419–433.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Goldin, L. R., G. A. Chase, and A. F. Wilson (1999). Regional inference with averaged P values increases the power to detect linkage. *Genetic Epidemiology 17*, 157–164.

Goodman, L. A. (1969). On partitioning χ^2 and detecting partial association in three-way contingency tables. *Journal of the Royal Statistical Society: Series B 31*, 486–498.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*. New York: John Wiley & Sons, Inc.

Hill, W. G. (1975). Tests for association of gene frequencies at several loci in random mating diploid populations. *Biometrics 31*, 881–888.

Hill, W. G. (1976). *Population Genetics and Ecology*, Chapter Non-Random Association of Neutral Linked Genes in Finite Populations, pp. 339–376. New York: Academic Press.

Hodge, S. E. (1995). An oligogenic disease displaying weak marker associations: A summary of contributions to problem 1 of GAW9. *Genetic Epidemiology 12*, 545–554.

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

Kaplan, N. L., E. R. Martin, and B. S. Weir (1997). Power studies for the transmission/disequilibrium tests with multiple alleles. *American Journal of Human Genetics 60*, 691–702.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*, 119–127.

Knapp, M. (1999). The transmission/disequilibrium test and parental-genotype reconstruction: The reconstruction-combined transmission/disequilibrium test. *Ameri-*

can Journal of Human Genetics 64, 861–870.

Lancaster, H. O. (1951). Complex contingency tables treated by the partition of χ^2. *Journal of the Royal Statistical Society: Series B 13*, 242–249.

Majewski, J., H. Li, and J. Ott (2001). The Ising model in physics and statistical genetics. *American Journal of Human Genetics 69*, 853–862.

Martin, E. R., J. R. Gilbert, E. H. Lai, J. Riley, A. R. Rogola, B. D. Slotterbeck, C. A. Sipe, J. M. Grubber, L. L. Warren, P. M. Conneally, A. M. Saunders, D. E. Schmechel, I. Purvis, M. A. Pericak-Vance, A. D. Roses, and J. M. Vance (2000). Analysis of association at single nucleotide polymorphisms in the *APOE* region. *Genomics 63*, 7–12.

Monks, S. A., N. L. Kaplan, and B. S. Weir (1998). A comparative study of sibship tests of linkage and/or association. *American Journal of Human Genetics 63*, 1507–1516.

Neville, P. (1999). Decision trees for predictive modeling. Internal SAS Paper. Available by request through Technical Support at SAS Institute, Cary, NC.

Nielsen, D. M., M. G. Ehm, and B. S. Weir (1999). Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *American Journal of Human Genetics 63*, 1531–1540.

Nielsen, D. M., M. G. Ehm, D. Zaykin, and B. S. Weir (2002). Higher order LD and haplotype-based tests of association. Unpublished manuscript.

Nielsen, D. M. and B. S. Weir (1999). A classical setting for associations between

markers and loci affecting quantitative traits. *Genetical Research 74*, 271–277.

Nielsen, D. M. and B. S. Weir (2001). Association studies for general disease models. *Theoretical Population Biology 60*, 253–263.

Randles, R. H. (1989). A distribution-free multivariate sign test based on interdirections. *Journal of the American Statistical Association 84*, 1045–1050.

Rao, C. R. and S. K. Mitra (1971). *Generalized Inverse of Matrices and Its Applications*. New York: John Wiley & Sons, Inc.

Risch, N. and K. Merikangas (1996). The future of genetic studies of complex human diseases. *Science 273*, 1516–1517.

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

Sarkar, S. K. and C.-K. Chang (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association 92*, 1601–1608.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics 53*, 1253–1261.

Searle, S. R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.

Sham, P. (1997). Transmission/disequilibrium tests for multiallelic loci. *American Journal of Human Genetics 61*, 774–778.

Siegmund, D. (2001). Is peak height sufficient? *Genetic Epidemiology 20*, 403–408.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

Slager, S. L. and D. J. Schaid (2001a). Case-control studies of genetic markers: Power and sample size approximations for Armitage's test for trend. *Human Heredity 52*, 149–153.

Slager, S. L. and D. J. Schaid (2001b). Evaluation of candidate genes in case-control studies: A statistical method to account for related subjects. *American Journal of Human Genetics 68*, 1457–1462.

Smouse, P. E. (1974). Likelihood analysis of recombinational disequilibrium in multiple-locus gametic frequencies. *Genetics 76*, 557–565.

Spielman, R. S. and W. J. Ewens (1996). The TDT and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics 59*, 983–989.

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

Spielman, R. S., R. E. McGinnis, and W. J. Ewens (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics 52*, 506–516.

Stuart, A. (1955). A test of homogeneity of the marginal distributions in a two-way classification. *Biometrika 42*, 412–416.

Teng, J. and N. Risch (1999). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Research 9*, 235–254.

Terwilliger, J. D., W. D. Shannon, G. M. Lathrop, J. P. Nolan, L. R. Goldin, G. A. Chase, and D. E. Weeks (1997). True and false positive peaks in genomewide scans: Applications of length-biased sampling to linkage mapping. *American Journal of Human Genetics 61*, 430–438.

Thomson, G. and M. Baur (1984). Third order linkage disequilibrium. *Tissue Antigens 24*, 250–255.

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

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates, Inc.

Weir, B. S., J. C. Brocklebank, P. M. Conneally, M. G. Ehm, J. R. Gilbert, J. H. Goodnight, W. A. Hassler, E. R. Martin, D. M. Nielsen, M. A. Pericak-Vance, A. R. Rogala, A. D. Roses, A. M. Saunders, D. E. Schmechel, B. D. Slotterbeck, J. M. Vance, and D. Zaykin (1999). A data-mining approach to fine-scale gene mapping. In *The 49^{th} Annual Meeting of the American Society of Human Genetics*.

Westphal, C. and T. Blaxton (1998). *Data Mining Solutions: Methods and Tools for*

Solving Real-World Problems. New York: John Wiley & Sons, Inc.

Zaykin, D. V., L. A. Zhivotovsky, P. H. Westfall, and B. S. Weir (2002). Truncated product method for combining p-values. *Genetic Epidemiology 22*, 170–185.

Zhang, H. and G. Bonney (2000). Use of classification trees for association studies. *Genetic Epidemiology 19*, 323–332.

Zhang, H., C.-P. Tsai, C.-Y. Yu, and G. Bonney (2001). Tree-based linkage and association analyses of asthma. *Genetic Epidemiology 21*, S317–S322.