2007

Haplotype-Based Association Studies: Approaches to Current Challenges

Mark A. Levenstien

HAPLOTYPE-BASED ASSOCIATION STUDIES: APPROACHES TO CURRENT CHALLENGES

A Thesis Presented to the Faculty of
The Rockefeller University
in Partial Fulfillment of the Requirements for
the degree of Doctor of Philosophy
by

Mark A. Levenstien
June 2007

HAPLOTYPE-BASED ASSOCIATION STUDIES: APPROACHES TO CURRENT CHALLENGES

Mark Levenstien, Ph.D.
The Rockefeller University 2007

Haplotype-based association studies have greatly aided researchers in their attempts to map genes. However, current designs of haplotype-based association studies lead to several challenges from a statistical perspective. To reduce the number of variants, some researchers have employed hierarchical clustering. This thesis starts by addressing the multiple testing problem that results from applying a hierarchical clustering procedure to haplotypes and then performing a statistical test for association at each of the steps in the resulting hierarchy. Applying our method to a haplotype case-control dataset, we find a global *p*-value. Relative to the minimum *p*-value over all steps in the hierarchy, the global *p*-value is markedly inflated. The second challenge involves the inherent errors present when prediction programs are employed to assign haplotype pairs for each individual in a haplotype-based association study. We examined the effect of these misclassification errors on the false positive rate and power for two association tests—the standard likelihood ratio test (LRT_{std}) and a likelihood ratio test that allows for the misclassification inherent in the haplotype inference procedure (LRT_{ae}). Our simulations indicate that 1) for each statistic permutation methods maintain the correct type I error; 2) specific multilocus genotypes that are misclassified as the incorrect haplotype pair are consistently misclassified throughout each entire dataset; and 3) a significant power gain exists for the LRT_{ae} over the LRT_{std} for a subset of the parameter settings. The LRT_{ae} showed the greatest benefit over the LRT_{std} when the cost of

phenotyping was very high relative to the cost of genotyping. This situation is likely to occur in a replication study as opposed to a whole genome association study. The third challenge addressed by this thesis involves the uncertainty regarding the exact distribution of the likelihood ratio test (LRT) statistic for haplotype-based association tests in which many of the haplotype frequency estimates are zero or very small. By simulating datasets with known haplotype frequencies and comparing the empirical distribution with various theoretical distributions, we characterized the distribution of the LRT statistic as a χ^2 distribution where the degrees of freedom are related to the number of the haplotypes with nonzero frequency estimates. Awareness of the potential pitfalls and the strategies to address them will increase the effectiveness of haplotype-based association as a gene-mapping tool.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my mentor, Dr. Jürg Ott, for guiding me and providing me with support throughout my graduate education. His leadership style granted me sufficient autonomy to explore a variety of approaches to the scientific questions I was investigating and created an environment in our laboratory where I could be productive in my scientific pursuits.

Secondly, I would like to thank the members of my Faculty Advisory Committee. Specifically, I want to recognize the high level of support offered by both Dr. Mary Jeanne Kreek and Dr. Joel Cohen. Both were willing to find time to schedule additional meetings to help me stay on course and offered constructive comments that guided my research. I would also like to thank Dr. Marcella Devoto who graciously agreed to join the committee as the external examiner.

I want to acknowledge the contributions of Dr. Derek Gordon and Dr. Yaning Yang. The research presented in this thesis would not be possible without their guidance. They acted as soundboards for my ideas, and their input was instrumental in developing the work. In addition, I would like to thank Chad Haynes for the computer programming support he provided and Katherine Montague for tremendous attention to detail in proofreading this thesis as well as helping me to navigate the administrative process at Rockefeller.

I would like to thank both current and past members of the Ott Laboratory here at Rockefeller who contributed to this research indirectly by creating such an enjoyable

work atmosphere. I appreciate and value all the friendships that I have developed over the course of my graduate studies.

Finally, I would like to thank my wife, Laurel, who has been a steadfast source of support throughout my career as a graduate student. From easing my transition from engineering to the biological sciences to editing the final versions of this thesis, she has offered emotional as well as scientific support every step of the way.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Full Name
AC	Adenocarcinoma
AD test	Anderson-Darling test
ASP	Affected sib pair test
cdf	Cumulative density function
C.I.	Confidence interval
df	Degrees of freedom
DAF	Disease allele frequency
DLBCL	Diffuse large B-cell lymphoma
DNA	Deoxyribonucleic acid
EM algorithm	Expectation-Maximization algorithm
FDR	False discovery rate
FWER	Family-wise error rate
GC	Genomic control
H_0	Null hypothesis
H_1	Alternative hypothesis
HHR	Haplotype relative risk
HHRR	Haplotype-based haplotype relative risk
HWE	Hardy Weinberg equilibrium
KS test	Kolmogorov-Smirnov test
LCLC	Large cell lung cancer
LD	Linkage disequilibrium
LOD	Logarithm of odds
LRT	Likelihood ratio test
LRT_{ae}	Likelihood ratio test allowing for errors
LRT_{em}	Likelihood ratio test using the EM algorithm
LRT_{std}	Standard likelihood ratio test
MAF	Minor allele frequency
ncp	Noncentrality parameter

NCBI	National Center for Biotechnology Information
pdf	Probability density function
PAWE	Power for association with error
PCR	Polymerase chain reaction
QT clustering	Quality threshold clustering
rs number	Reference single nucleotide polymorphism number
RFLP	Restriction fragment length polymorphism
SA	Structured association
SCC	Squamous cell carcinoma
SCLC	Small cell lung cancer
SNP	Single nucleotide polymorphism
TDT	Transmission disequilibrium test
VNTR	Variable number of tandem repeat

CHAPTER 1: BACKGROUND

1.1 A historical perspective of genetic mapping

Over the past the twenty-five years, human gene mapping has developed into a highly effective tool for localizing mutations which lead to disease. The startling rate of advancement in molecular biology has provided the field with genetic and physical maps of excellent quality while more efficient computational algorithms and more powerful computing systems have permitted researchers to analyze larger datasets containing a greater number of marker loci.

A wide variety of statistical approaches and study designs has been employed in the effort to map human disease genes. Although the statistical methodology had been developed (Morton 1955) and the algorithm refined (Elston and Stewart 1971) as well as incorporated into software (Ott 1974) much earlier, linkage studies experienced new levels of popularity and successfully mapped many disease genes starting in the 1980s (Gusella et al. 1983; Monaco and Kunkel 1988; Kerem et al. 1989). Such studies proved to be successful for mapping Mendelian disorders—disorders whose genetic basis involves a single major gene. These diseases show high penetrance (individuals possessing one or two copies of the mutant allele at the disease locus have a high probability of showing the disease phenotype) and tend to follow classical modes of inheritance. Specifically, linkage analysis has been instrumental in localizing genes responsible for cystic fibrosis (Eiberg et al. 1985; Knowlton et al. 1985; Wainwright et al. 1985; Schmiegelow et al. 1986; Kerem et al. 1989; Riordan et al. 1989), Duchenne muscular dystrophy (Monaco and Kunkel 1988), Huntington disease (Gusella et al.

1983), Charcot-Marie-Tooth disease (Ouvrier 1996), retinitis pigmentosa (Sullivan and Daiger 1996), certain forms of early-onset breast cancer (Hall et al. 1990), and certain forms of Alzheimer disease (Levy-Lahad and Bird 1996; Rademakers et al. 2005), among other Mendelian disorders. Although linkage analysis has been shown to be an effective tool for Mendelian disorders, the linkage results provide wide candidate regions which require additional fine-mapping typically performed using linkage analysis (or association studies) with a denser marker map in the region where linkage was initially detected. As part of a fine-mapping analysis to narrow a candidate region, researchers may reconstruct the haplotypes for the family and identify the largest section of the haplotype shared by the affected study individuals (Seri et al. 1999; Bolino et al. 2000; Lo Nigro et al. 2000; Paluru et al. 2003). In this context, haplotype phasing is determined using the familial relationships and minimizing the number of recombination events.

In addition to linkage analysis, other statistical methods have greatly aided in the mapping of human traits. Association studies (case-control studies) aim to find a genetic variant that appears with the disease state more often than it should by chance alone. When a new mutation arises in a population, the alleles at nearby polymorphic sites on the mutated chromosome will be initially coupled with the mutant allele. As the mutation is inherited by new generations, recombination events will eventually cause this coupling effect to decay. However, a state of coupling or linkage disequilibrium (LD) may remain detectable if the disease and marker loci are in sufficient proximity to one another so that recombination events between the two loci are rare, and, consequently, the decay of the coupling effect is very slow (Ott 1999). That is, a certain genetic variant may be

associated with the disease state such that the frequency of the genetic variant is higher in cases than in controls.

Linkage and association differ fundamentally. For linkage analysis, the specific genetic variants (alleles) serve as a means to examine the linkage properties of the region or estimate the amount of recombination between the marker and the disease. In contrast, for association studies the genetic variants (alleles, genotypes, or haplotypes) themselves are the center of the test and may be directly responsible for the disease phenotype. That is, the reason an association is detected between a genetic variant and a disease is that the variant itself causes the disease state (direct association) or is in high LD with a mutation that causes the disease state (indirect association) (Ott 1999; Cordell and Clayton 2005).

The first major finding from genetic association studies occurred in the late 1960s and early 1970s when a number of researchers detected an association between a number of different diseases and the HLA (human leukocyte antigen) loci. Perhaps the best known of these associations is that between ankylosing spondylitis and HLA-B27 because of the large number of studies able to replicate the finding (Brewerton et al. 1973; Schlosstein et al. 1973; Levitin et al. 1976; Brautbar et al. 1977; Contu et al. 1977). These discoveries generated increased interest in association studies. However, because the LD required to detect association exists over a short distance from the marker locus, such association findings were rare until genetic maps with a higher density of markers were developed. Since the regions over which LD can be detected do not extend as far as linkage peaks, association studies traditionally have been utilized in human genetics in order to fine map after an initial linkage analysis has implicated candidate genes for follow up studies. A major disadvantage of association studies is that they are

susceptible to inflated false positive rates in the presence of population stratification or admixture (Simpson 1951; Li 1955; Gorroochurn et al. 2004; Heiman et al. 2004) since differences in allele (genetic variant) frequencies between cases and controls may only be the result of differences in ethnicity between the case and control populations rather than differences related to the disease state itself.

As a way to protect against this situation, other methods such as the haplotype relative risk (HHR) test (Rubinstein et al. 1981; Falk and Rubinstein 1987; Thomson et al. 1989), the haplotype-based haplotype relative risk (HHRR) test (Terwilliger and Ott 1992), and the Transmission Disequilibrium Test (TDT) (Spielman et al. 1993; Spielman and Ewens 1996; Ewens and Spielman 2005) applied family-based controls rather than population-based controls. In particular, because of its ability to use the genetic information from multiple affected siblings, the TDT and other family-based association methods gained popularity and successfully aided in mapping genes for many diseases including psoriasis (Helms et al. 2003; Helms et al. 2005) and sitosterolemia (Lee et al. 2001; Gordon et al. 2004). Although the TDT solved the problem caused by working with admixed populations (excluding the situation involving extreme admixture (Lazzeroni and Lange 1998)), the test requires additional sample collection and increased costs to maintain the same power as association studies (Morton and Collins 1998). In addition, the TDT has the undesirable property of an increased false positive rate in the presence of genotyping error (Mitchell et al. 2003) or absence of parental genotype data (Curtis and Sham 1995). Simulation studies have shown that genotyping errors and missing parental genotypes interact to increase the false positive rate of the TDT (Barral et al. 2005).

Several factors have contributed to a rise in popularity of the case-control association studies despite the issues related to population stratification. New methods have been developed which can account for population structure in case-control studies and, therefore, avoid the spurious associations between genes and disease that result from admixed populations. These methods utilize additional genetic markers to correct for the stratification using one of two approaches—genomic control (GC) (Devlin and Roeder 1999; Bacanu et al. 2000; Devlin et al. 2000; Devlin et al. 2001) and structured association (SA) (Pritchard et al. 2000a; Pritchard et al. 2000b; Pritchard and Donnelly 2001; Kohler and Bickeboller 2006). The GC approach assumes that the effects of population stratification should be equal across the entire genome. From the test results of many polymorphisms at genomic regions unlikely to harbor a disease gene, the GC approach estimates the amount of "overdispersion" or inflatedness present in the statistic used to detect association. This estimate is then used to correct the test statistic in regions under consideration for association. Simulation studies have shown that under some circumstances GC methods may not completely eliminate the inflation in false positive rate due to population stratification and under other circumstances may significantly reduce power (Shmulewitz et al. 2004). In contrast to GC, the SA approach uses many polymorphisms to classify individuals into subpopulations with high degrees of genetic similarity. (The method proposed by Pritchard *et al.* allows for admixture in the sense that individuals may be classified as possessing the genetic ancestry of several different subpopulations.) With subpopulations established, the next second step in the SA approach performs a test for association that conditions on the inferred subpopulation membership.

Besides the development of these defenses against the dangers of population stratification and the relative ease and cost-efficiency offered by case-control study designs, in the late 1990s new technological resources, including sequence data from the Human Genome Project, facilitated single nucleotide polymorphism (SNP) discovery efforts and high-throughput SNP genotyping (Collins et al. 1998). Theoretical studies show genome-wide association mapping to be a very powerful strategy for localizing genes related to complex traits (Risch and Merikangas 1996; Risch 2000). This potential for genome-wide association using SNP markers has prompted companies, such as Affymetrix Inc. and Illumina, Inc., to develop as well as manufacture gene arrays and platforms with the ability to provide genotypes for thousands of SNPs. In addition, the work of the International HAPMAP Project, an organization dedicated to describing the patterns of human genetic variation by developing a map of the linkage disequilibrium in the human genome, provides a valuable resource for efficient SNP selection for custom chip studies (International HapMap Consortium 2003; International HapMap Consortium 2005). Furthermore, large-scale genome-wide association studies have already proven successful for identifying genes related to age-related macular degeneration and obesity (Klein et al. 2005; Herbert et al. 2006). A final reason researchers are attracted to the case-control design is the plausibility that in the future large databases containing genome-wide information for controls will facilitate highly efficient case-control studies.

As described briefly above, genetic association tests aim to detect an association between a genetic variant and the disease state. Although traditionally the genetic variant under investigation has been an allele (Botstein and Risch 2003), a genotype, a single haplotype, or even a diplotype (haplotype pair) can be the focus of a case-control genetic

association test. The alleles present at multiple genetic markers inherited from the same parent form a haplotype (Ott 1999). Often haplotypes are comprised of alleles on the same chromosome (Brumfield et al. 2003). Each of the forms (allelic, genotypic, haplotypic, and diplotypic) of association testing may have advantages under specific circumstances. For instance, test statistics that utilize single allele frequencies (or single haplotype frequencies) may not be valid when the genotype frequencies (or diplotype frequencies) deviate from Hardy Weinberg Equilibrium (HWE) (Sasieni 1997). However, allelic (or haplotypic) tests of association generally are more powerful than their genotypic (or diplotypic) counterparts because these tests have fewer degrees of freedom (Agresti 1996).

With the advent of the HAPMAP project (International HapMap Consortium 2003; International HapMap Consortium 2005), the popularity of a relatively recently developed form of genetic association analysis, haplotype-based case-control genetic association studies, has grown markedly. It has been suggested that association studies utilizing haplotypes formed from SNPs may be more powerful than single locus association (Martin et al. 2000; Akey et al. 2001; Fallin et al. 2001; Morris and Kaplan 2002; Zaykin et al. 2002; Botstein and Risch 2003; Clark 2004; Clayton et al. 2004; De La Vega et al. 2005; Ellis et al. 2005) . One reason haplotypes may provide a power advantage over single SNPs in association studies is that the combined effects of multiple sequence variants on promoter activity or protein structure (and/or function) may precipitate the disease phenotype (Devlin and Roeder 1999; Drysdale et al. 2000; Joosten et al. 2001). A second reason stems from a mathematical finding. It has been shown that case-control genetic association studies are most powerful when the genetic variant under

consideration possesses a frequency in the population identical to that of the disease mutation (Abel and Muller-Myhsok 1998; Tu and Whittemore 1999; Pfeiffer and Gail 2003; Zondervan and Cardon 2004). Therefore, if the frequency of a single haplotypic variant more closely matches the frequency of the disease mutation than the frequency of any allele at any of the marker loci comprising the haplotype, a haplotype-based association test should be more powerful than an allelic association test (Martin et al. 2000; De La Vega et al. 2005). However, haplotype-based association tests also present some disadvantages. Techniques for directly observing haplotypes are expensive so more often haplotypes are inferred from multilocus genotypes using statistical methods. Also, since haplotypes generally have a large number of genetic variants compared to genotypes or single alleles, haplotype-based association tests either possess more degrees of freedom or face a larger multiple testing problem than tests involving a single locus.

1.2 Background for statistical tests

Regardless of the genetic variant under investigation, several approaches can be taken to test for association. One option relies on a $2 \times s$ contingency table, where s is the total number of genetic variants, to record counts of the genetic information for cases and controls. From the counts in the contingency table, a Pearson χ^2 statistic, which compares the observed counts with those expected under the assumption of independence between case status and the genetic variant, can be computed (Pearson 1900; Agresti 1996). An alternative approach is to calculate the likelihood ratio test (LRT) statistic, which is twice the difference between the log-likelihood of the data under the assumption that an association exists (between case status and genetic variant) and the log-likelihood

of the data under the assumption that no association exists (Fisher 1922b; Fisher 1925; Edwards 1992). When formulating the likelihood in terms of direct observations, such as genotypes, the multinomial distribution is used. However, when formulating the likelihood in terms of a quantity with missing data, such as haplotypes, frequencies may be estimated from the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) or other efficient methods and incorporated in the likelihood expression.

Both the Pearson χ^2 and the likelihood ratio approaches are examples of hypothesis testing involving two mutually exclusive hypotheses—the null hypothesis (H_0) and the alternative hypothesis (H_1). The test evaluates the data available to determine whether sufficient evidence exists to reject the null hypothesis in favor of the alternative hypothesis. For genetic association tests, the null hypothesis is that no association exists between the genetic variant and the disease state whereas the alternative hypothesis is that such an association does exist. The general goal of a statistical test is to maximize power while controlling for type I error. Power is the probability of a test yielding a positive result (i.e. rejecting the null hypothesis) when in fact the null hypothesis is false. In other words, power represents a test's ability to find true positives. On the other hand, type I error (or the false positive rate) represents the probability that a test which rejects the null hypothesis will do so incorrectly. Closely related to type I error, the statistical significance or *p*-value related to a test result represents the cumulative probability of achieving an equivalent or more extreme test result when H_0 is true. Prior to a single statistical test, type I error is set to a commonly accepted threshold (significance level) such as 0.05. A test statistic with an associated *p*-value less than this threshold results in rejecting H_0 in favor of H_1. For example, if one

performs 100 tests on a population for which H_0 is true using the 0.05 threshold for type I error, on average the results from five of the tests would lead to incorrectly rejecting H_0.

The statistical significance, or *p*-value, of a test result can be evaluated in several ways in the context of case-control association studies. However, all approaches aim to determine the distribution of the test statistic under the null hypothesis and then use this distribution to compute the probability of achieving a test result (when the H_0 is true) equivalent to or more extreme than the test result calculated from the data. One approach relies on using a null distribution determined by classical statistics. For instance, according to statistical theory under the null hypothesis of no association both the Pearson χ^2 statistic and the LRT statistic follow a central χ^2 distribution asymptotically for large sample sizes (Agresti 1996). The number of degrees of freedom associated with the central χ^2 distribution equals one less than the number of genetic variants present in the sample for the Pearson χ^2 test and equals the difference between the number of free parameters estimated under H_1 and H_0 for the likelihood ratio test. It has been shown that when Cochran's rule is followed (more than five observations in each cell of the contingency table), this approach is reliable (Cochran 1952). A second approach employs permutation testing to generate the distribution of the test statistic under the null hypothesis and to determine its statistical significance (Fisher 1935; Pitman 1937; Pitman 1938). In permutation testing, many null replicates of the original dataset are created by randomly reassigning case-control labels to the individuals in the study. Then the test statistic is computed for each replicate dataset, and the distribution of these test statistics represents the distribution of the test statistic under the null hypothesis. While this empirical approach provides extremely accurate *p*-values when a very large number of

permutations is used (regardless of the sample size), it is computationally intensive. A third approach is Fisher's exact test (Fisher 1922a) which employs the hypergeometric distribution to express the probability of contingency tables equivalent to and more extreme than the contingency table for the dataset. Like the permutation approach, Fisher's exact test produces accurate *p*-values even with extremely small sample sizes. However, large sample sizes or datasets associated with well-balanced tables can lead to difficulty in executing the test.

Like computing statistical significance, there are multiple ways to determine the power of a test for genetic association. In order to compute the power of a statistical test, one must know the distribution of the test statistic when the alternative hypothesis is true. In addition, the alternative hypothesis must be formulated in terms of parameters such that the power associated with given parameter values can be established. Once the distribution is determined, one can compute the probability that a test statistic computed under H_1 will be equivalent to or exceed the value of the test statistic associated with the type I error threshold. One approach to finding this distribution relies on classical statistics. For instance, according to statistical theory under the alternative hypothesis of association, both the Pearson χ^2 statistic and the LRT statistic follow a noncentral χ^2 distribution asymptotically for large sample sizes (Mitra 1958; Hogg and Craig 1995; Agresti 1996). This distribution is defined by two parameters—the degrees of freedom (*df*) and the noncentrality parameter (*ncp*). While the degrees of freedom can be computed as described above in the discussion regarding statistical significance, the noncentrality parameter can be computed as a function of frequencies belonging to the genetic variants under investigation (in cases and controls separately), the number of

cases, and the number of controls (see http://linkage.rockefeller.edu/derek/pawe2.html) (Mitra 1958; Sham 1998; Gordon et al. 2002). Once parameters defining the noncentral χ^2 distribution are known explicitly, the power for a given significance level can be determined analytically. Another approach for finding the distribution of the test statistic under the alternative hypothesis requires data simulation. In this empirical approach, power is computed by generating thousands of datasets under a model where there is an association and finding the proportion of simulated datasets that produce a test statistic equivalent to or more extreme than the value of the test statistic associated with the type I error threshold. For a very large number of generated datasets, simulation methods provide accurate power estimates; however, the cost of this accuracy is increased computational time.

1.3 Multiple Testing

Recall the example describing type I error in which 100 tests were performed on a population for which the null hypothesis is true using the 0.05 threshold for type I error. On average, the results from five of the tests would lead to incorrectly rejecting the null hypothesis. A somewhat analogous situation arises when it is desirable to perform a family of tests on the same dataset. This analogous situation complicates the interpretation of type I error. When many tests are performed on the same dataset, each additional test provides another opportunity for a spurious positive result. Consequently, the probability of at least one of the tests yielding a false positive result is higher than the type I error threshold employed for each individual test. To protect against this phenomenon, classical comparison procedures strive to control the family-wise error rate

(FWER) or the probability of incorrectly rejecting any null hypothesis in a group of tests under simultaneous consideration.

Several statistical methods have been developed to control the FWER. These methods can be classified as single-step methods and stepwise methods (Westfall and Young 1993). For single-step methods, such as the Bonferroni correction and the Šidák method (Šidák 1967), equivalent multiplicity adjustments are applied to the *p*-values for all tests, regardless of the ordering of the observed *p*-values. Since the Bonferroni and Šidák methods assume that the tests are independent of one another, they can be conservative when this assumption is false. In contrast, stepwise methods, such as step-up and step-down procedures, permit different adjustments for different tests depending on the ordering of the observed *p*-values (Westfall and Young 1993). In recent years, improved computing technology has facilitated the use of resampling methods, such as bootstrapping, Monte Carlo simulations, and permutation resampling. Specifically, Westfall and Young have contributed several resampling methods for multiple testing (Westfall and Young 1993). While these methods are attractive in that they often can effectively capture the correlation structure of the tests and allow for increased power, they may be computationally expensive.

More recently, procedures have been developed which control the false discovery rate (FDR) rather than the FWER (Benjamini and Hochberg 1995). Such procedures aim to ensure that on average the proportion of false positives among all positive results is within an acceptable limit. Only for cases where the number of true null hypotheses equals the total number of hypotheses examined are the FDR and the FWER criteria equivalent. Otherwise, as more null hypotheses are false, the FDR becomes smaller.

Thus, procedures that control the FDR often have greater power than classical multiple comparison procedures aimed at controlling the FWER (Benjamini and Hochberg 1995). These procedures seem promising and may prove to offer an advantage over the current practices.

In the search for susceptibility genes for human disease, multiple testing has posed a formidable obstacle. Over the last two decades, the discovery of new varieties of polymorphic genetic markers has aided the effort to localize disease genes. With Restriction Fragment Polymorphisms (RFLPs), Variable Number of Tandem Repeat (VNTR) markers, microsatellite markers, and SNPs, researchers have millions of markers at their disposal and continue to discover more (Sachidanandam et al. 2001; Venter et al. 2001). As a result, researchers perform tests of linkage and association for large numbers of haplotypes, alleles, or genotypes at regular intervals across entire chromosomes or genomes (Risch and Merikangas 1996). Although this comprehensive approach improves the likelihood of testing in an area of the genome where true linkage or linkage disequilibrium (LD) exists, it requires a multiplicity of testing—one test (or more) at each marker. To control the false positive rate, appropriate genome-wide LOD score thresholds have been created for tests of linkage under both homogeneity and heterogeneity (Morton 1955; Terwilliger and Ott 1994; Lander and Kruglyak 1995; Huang and Vieland 2001). In addition, to adjust for multiple testing for other tests, such as association tests, the Affected Sib Pair Test (ASP), and the Transmission Disequilibrium Test (TDT), researchers apply other forms of correction to *p*-values (Lander and Kruglyak 1995; Miller 1997). Over the past few years, procedures which control the FDR have been applied to genetic mapping (Weller et al. 1998; Devlin et al.

2003; Sabatti et al. 2003) and the analysis of differential gene expression (Storey and Tibshirani 2001; Reiner et al. 2003; Yang et al. 2003). In spite of the difficulties imposed by multiple comparisons, genome-wide testing has successfully localized many Mendelian disorders (Gusella et al. 1983; Kerem et al. 1989; Saunders et al. 1993). However, prominent medical conditions, such as diabetes, heart disease, schizophrenia, and bipolar disorder, appear not to follow Mendelian patterns of inheritance but rather involve interactions with the environment and/or other genes. In these situations, adjusting for multiple testing severely compromises the power of the test since testing for main effects and interactions across the genome results in an unwieldy number of comparisons (Dupuis et al. 1995).

1.4 Hierarchical Clustering

In addition to the multiple testing issues mentioned above, other methods employed to organize genetic marker data also introduce a multiplicity of testing. Hierarchical (agglomerative) clustering is an information theoretical method that sequentially merges samples based on the pair-wise similarity of a given measurement to form common groups until all samples are contained in a single group (Hastie et al. 2001). The method has many applications and is widely used in the analysis of biological data. For example, researchers testing for association between haplotypes and disease have employed hierarchical clustering as a means to reduce a large number of haplotypes to a manageable number of haplotype classes with the aim to increase statistical power (Hoehe et al. 2000). With an increasing number of marker loci, the number of possible haplotypes grows exponentially so that many of these haplotypes tend to have low

frequency. This situation is relatively common when examining haplotypes within candidate genes because of the availability of dense SNP marker maps, in which the spacing between markers often is less than one kilobase (Sachidanandam et al. 2001). In comparisons of haplotype frequencies between case and control individuals, the corresponding contingency tables are therefore often sparse and difficult to interpret. Hierarchical clustering then allows researchers to merge haplotypes into classes that are easier to handle. At each step within the hierarchy, either implicitly or explicitly, researchers tend to interpret results and eventually focus on that set of classes providing the most significant result. Testing at each of the different clustering steps within a hierarchical structure also represents a form of multiple comparisons; therefore, the minimum *p*-value evaluated over many steps is too small to represent the experiment-wise significance level.

Many methods, including hierarchical clustering, can be applied to partition a dataset into subgroups whose elements share common characteristics. Several methods of non-hierarchical clustering or partitional clustering exist. Some common partitional methods include k-means clustering (MacQueen 1967), quality threshold (QT) clustering (Heyer et al. 1999), and fuzzy clustering (Dunn 1973). In addition, hierarchical clustering has two varieties—1) agglomerative (bottom-up) clustering in which the groups are built up at each progressive stage of clustering so that each item starts in its own group and 2) divisive (top-down) clustering in which a single group exists initially and items are removed as clustering progresses. Once a hierarchical clustering procedure constructs a dendrogram or tree diagram representing the grouping structure, the dataset can be divided into any number of groups by selecting the appropriate clustering stage or

level. In order to construct the grouping structure, the procedures must define the distance between groups. One option, single linkage, defines the distance between two groups as the minimum pair-wise distance between any item in the first group and any item in the second group whereas another option, complete linkage, finds distance as the maximum pair-wise distance between any item in the first group and any item in the second group. A third option, average linkage, uses the average of all pair-wise distances between items in the two groups (Johnson 1967). In addition, there are several metrics for determining pair-wise distance between individual items. Common distance metrics include Euclidean distance, squared Euclidean distance, Manhattan distance, and the correlation coefficient.

1.5 Estimation, inference, and haplotype-based association

Methods which apply techniques, such as allele-specific long-range PCR and somatic cell hybrid construction, from molecular biology for explicit determination of phased haplotypes are available (Papadopoulos et al. 1995; Michalatos-Beloin et al. 1996; Clark et al. 1998; Yan et al. 2000; Douglas et al. 2001; Patil et al. 2001; Burgtorf et al. 2003; Ding and Cantor 2003; Horan et al. 2003; Hoppe et al. 2004; Proudnikov et al. 2004; Yu et al. 2004; Hoppe et al. 2006; Proudnikov et al. 2006). However, because current molecular haplotyping methods are expensive and not amenable to automation, in practice phased haplotypes are rarely determined explicitly. Instead, statistical methods for gene mapping estimate haplotype frequencies from multilocus genotype data and often provide haplotype assignments or calls for individuals (Clark 1990; Xie and Ott 1993; Terwilliger and Ott 1994; Excoffier and Slatkin 1995; Hawley and Kidd 1995;

Long et al. 1995; Zhao et al. 2000; Stephens et al. 2001b; Zhao and Sham 2002; Stephens and Donnelly 2003). Since the parental origins of the two alleles comprising any single genotype are not directly observed, constructing phased haplotypes from multilocus genotypes can be complicated. Consider two SNP marker loci where *A* and *a* represent the alleles at the first locus while *B* and *b* represent the alleles at the second locus. One can assign haplotype pairs unequivocally for all possible multilocus genotypes except for the double heterozygote *AaBb*. For instance, the multilocus genotype *AaBB*, must derive from the haplotype pair *AB* and *aB*. In contrast, the multilocus genotype *AaBb* either derives from the haplotype pair *AB* and *ab* or the haplotype pair *Ab* and *aB*. Such ambiguous cases occur for any multilocus genotype possessing two or more loci with a heterozygote. As with fine mapping in linkage studies, knowledge of the parental genotypes can greatly simplify the problem of phasing. However, for case-control association studies, the sampling design involves unrelated individuals, and, consequently, parental genotypes are rarely collected. Therefore, the procedure utilized to estimate haplotype frequencies treats each individual as an independent observation.

Several methods have been developed to estimate haplotype frequencies for non-familial study designs. While the first method developed for haplotype estimation is based on the principle of maximum parsimony (Clark 1990; Wang and Xu 2003), methods that rely on the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) for a likelihood approach (Xie and Ott 1993; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995) or use a Bayesian approach applying a prior based on coalescence theory (Stephens et al. 2001b; Stephens and Donnelly 2003) or a Dirichlet prior (Niu et al. 2002) are more commonly used. Although it is a relatively

straightforward method, Clark's method has several disadvantages among which are its inability to provide unique solutions and its sensitivity to deviations from HWE (Niu et al. 2002; Niu 2004). In contrast, the EM algorithm-based and Bayesian approaches have been shown to be relatively robust to such deviations (in spite of the fact that the EM algorithm-based approaches assume HWE) (Niu et al. 2002; Niu 2004). Each of these approaches has been implemented in statistical software. Specifically, the parsimony-based methods are implemented in HAPINFERX and HAPAR (Wang and Xu 2003); the EM algorithm-based methods are implemented in SNPHAP (see Electronic Resource Information), HAPLO (Hawley and Kidd 1995), and PL-EM (Qin et al. 2002); and the Bayesian approaches are implemented in PHASE (Stephens et al. 2001b) (see Electronic Resource Information) and HAPLOTYPER (Niu et al. 2002).

As with other procedures for statistical estimation, the accuracy of haplotype frequency estimates depends on several factors including "sample size, number of loci studied, allele frequencies, and locus-specific allelic departures from Hardy-Weinberg and linkage equilibrium" (Fallin and Schork 2000). Furthermore, these factors also affect the accuracy of phased haplotype inference or phased haplotype calls (Niu 2004). Several researchers have investigated the accuracy of haplotype inference procedures by applying them to real and simulated data sets (Tishkoff et al. 2000; Clark et al. 2001; Xu et al. 2002; Stephens and Donnelly 2003; Adkins 2004; Kang et al. 2004; Niu 2004; Xu et al. 2004; Heid et al. 2005; Sabbagh and Darlu 2005; Zhang et al. 2005; Marchini et al. 2006; Proudnikov et al. 2006). In addition, Douglas *et al.* found molecular haplotyping provided large efficiency advantages over haplotype inference from multilocus genotypes under the condition of linkage equilibrium between marker loci (Douglas et al. 2001).

Using simulation studies, Schaid extended the work of Douglas *et al.* to conditions with linkage disequibrium between the markers. The studies found that the advantage of molecular haplotyping over haplotype inference decreased with increasing LD (Schaid 2002). Similar studies have investigated the power advantage of molecular haplotyping over haplotype estimation techniques for genetic association studies (O'Hely and Slatkin 2003; Thomas et al. 2004).

As described earlier, multiple statistical methods, such as the Pearson χ^2 test and the likelihood ratio test, are available to perform tests of case-control association. However, since the original observations (multilocus genotypes) lack phase information, the testing situation is a bit more complex. This additional complexity results in issues unique to tests of haplotype-based association as compared with other genetic association tests. In accord with the earlier description, the likelihood ratio test for haplotype-based association involves calculating the likelihood of the data in terms of the estimated haplotype frequencies (Xie and Ott 1993; Fallin et al. 2001). However, some haplotypic variants may be estimated to have a small frequency despite the fact that none of the study participants comprising the sample possess them. The effect of this situation on the distribution of the resulting test statistic under both null and alternative hypotheses remains unclear. One still expects that the test statistic will follow a central χ^2 distribution under H_0 and a noncentral χ^2 distribution under H_1. However, the degrees of freedom associated with either χ^2 distribution are no longer well defined.

Also analogous to the earlier discussion, an alternative method for haplotype-based association relies on the use of a contingency table containing the case-control counts for each inferred haplotype. The counts in the contingency table can be

determined either by inferring phased haplotypes for each individual or by multiplying each haplotype frequency estimate by the total number of haplotypes in the study. With data in the completed contingency table, either a Pearson χ^2 test or a likelihood ratio test can be performed. Many researchers find this second method with the contingency table appealing since it applies the same format as classic genotypic and allelic case-control studies and explicitly accounts for each phased haplotype. As a result, many researchers employ this method in practice (Hoehe et al. 2000; Maksymowych et al. 2003; Xu et al. 2004; Hindorff et al. 2006; Proudnikov et al. 2006). In the event that all phased haplotypes have been called correctly, this method can provide additional power (Cox and Hinkley 1974; Little and Rubin 1987). This situation is analogous to tests of association using allele estimates from individual genotypes as compared with allele frequency estimates from DNA-pooling data (Johnson et al. 2001).

However, misclassifications can lower a study's power and/or affect the false positive rate. The act of calling haplotype pairs from multilocus genotypes in the phase ambiguous situation is similar to the act of dichotomizing continuous measures. Royston *et al.* document a loss in power when dichotomizing continuous predictor variables in a regression analysis (Royston et al. 2006). In the context of a haplotype-based association study utilizing the contingency table design, a misclassification results when the haplotype pair called for an individual is not the true underlying haplotype pair. Non-differential misclassification occurs when the misclassification rates are the same in cases and controls. When non-differential misclassification exists, the test suffers a loss in power but the false positive rate remains unchanged (Mote and Anderson 1965; Gordon et al. 2002). In contrast, differential misclassification inflates the test's false positive rate

and may diminish its power (Clayton et al. 2005). In addition to errors due to the statistical procedure, misclassification of the multilocus genotypes will lead to miscalling haplotype pairs. In the absence of differential genotype misclassification, all haplotype misclassification should be non-differential when haplotype frequency distributions are the same in cases and controls, i.e. under the null hypothesis.

This thesis addresses several of the challenges currently confronting investigators conducting haplotype-based association studies. Chapter 2 examines the multiple testing problem that results from applying a hierarchical clustering procedure to haplotypes and then performing a statistical test for association at each of the steps in the resulting hierarchy. The proposed approach to overcome this challenge is creating an experiment-wise statistic of interest and finding its significance. Chapter 3 explores the consequences of the errors present when haplotype prediction programs are employed to assign haplotype pairs for each individual in commonly used tests of haplotype-based association. While there have been several studies aimed at evaluating the accuracy of haplotype inference and haplotype frequency estimation procedures (Fallin and Schork 2000; Tishkoff et al. 2000; Clark et al. 2001; Xu et al. 2002; Stephens and Donnelly 2003; Niu 2004; Xu et al. 2004; Sabbagh and Darlu 2005; Marchini et al. 2006), no systematic study has documented the effects of haplotype misclassification on the false positive rate and power. In this chapter, we compare the performance of a test statistic that utilizes a double-sampling procedure to account for haplotype misclassification with the standard likelihood ratio test statistic. Chapter 4 investigates the uncertainty regarding the exact distribution of the likelihood ratio statistic under the null hypothesis of no association for haplotype-based association tests in which many of the haplotype

frequency estimates are zero or very small. In this chapter, we characterize the distribution of the LRT statistic by simulating null datasets with known haplotype frequencies and comparing the empirical distribution with various theoretical distributions. Finally, chapter 5 draws some conclusions from these studies and discusses future directions for research related to haplotype-based association.

CHAPTER 2: HIERARCHICAL CLUSTERING AND GLOBAL SIGNIFICANCE

2.1 Introduction

New techniques in the biological sciences, like high throughput genotyping, microarray chip assays, and an explosion of online databases, have created a wealth of information regarding biological systems. The burgeoning discipline of bioinformatics illustrates the need for data organization and the development of statistically sound methods for analysis. An increasingly common issue for a variety of applications in biology is the artificial inflation of statistical significance associated with multiple testing.

With the increasing amount of data generated in molecular genetics laboratories, it is often difficult to make sense of results because of the vast number of different outcomes or variables studied. Examples include haplotypes comprised of large numbers of loci and expression levels for large numbers of genes. It is then natural to group observations into smaller numbers of classes that allow for an easier overview and interpretation of the data. This grouping is often carried out in multiple steps with the aid of hierarchical cluster analysis, each step leading to a smaller number of classes by combining similar observations or classes.

For example, researchers testing for association between haplotypes and disease have employed hierarchical clustering to reduce a large number of haplotypes to a manageable number of haplotype classes with the aim to increase statistical power (Hoehe et al. 2000). With an increasing number of marker loci, the number of possible

haplotypes grows exponentially so that many of these haplotypes tend to have low frequency. In comparisons of haplotype frequencies between case and control individuals, the corresponding contingency tables are often sparse and difficult to interpret. Several strategies, such as pooling the rarest categories to form a single haplotype class (Sham and Curtis 1995; Schaid et al. 2002; Zhao et al. 2003) and using haplotype diversity criteria for SNP selection (Johnson et al. 2001; Jannot et al. 2004) (http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/htsnp.pdf), have been suggested to reduce the number of classes. Unlike these alternatives, hierarchical clustering allows researchers to merge haplotypes, based on sequence similarities, into classes that are easier to handle. Initially, each haplotype is considered to be its own class. With each step in the clustering process, haplotype classes are merged based on the pair-wise similarity of the allele sequences comprising the haplotypes contained within each class until all samples are contained in a single haplotype class. At each step in the resulting hierarchy, either implicitly or explicitly, researchers tend to interpret results and eventually focus on the set of classes providing the "best" (most significant) result. While this approach makes sense, the overall statistical significance of the experiment must include the clustering process, which modifies the grouping structure of the data.

Another example of hierarchical clustering is its application in microarray analyses (Eisen et al. 1998; Alon et al. 1999; Gasch et al. 2000). Often clustering of arrays based on microarray expression data is utilized to distinguish tumor subclasses, which have clinical implications (Golub et al. 1999; Chung et al. 2002). In many of these studies involving microarray expression data from tumor specimens, researchers are

interested in examining survival information for the subjects who contributed the samples and comparing the survival curves between groups formed by the hierarchical clustering procedure (Alizadeh et al. 2000; Bhattacharjee et al. 2001; Garber et al. 2001; Sorlie et al. 2001; Guo et al. 2006; Perreard et al. 2006). After performing cluster analysis on the expression data, researchers tend to concentrate their attention on the step in the resulting hierarchy with the most striking difference in survival between patient groups and evaluate this result without taking into account the grouping structure at the other steps.

Here we propose an analysis method that properly takes the process of clustering into account. We achieve this by defining the strongest result or, equivalently, the smallest *p*-value, occurring in the course of clustering as the statistic of interest and computing its associated (experiment-wise) empirical significance level. The methods developed in this chapter will be applied to three previously published datasets in which hierarchical clustering has been employed. One of these datasets involves a haplotype-based association analysis while the other two datasets refer to survival analyses of groups of individuals determined by microarray expression measurements.

The problem of testing group differences sequentially is in the framework of multiple testing. Historically, both genetic association studies and microarray studies have been plagued with multiple testing problems. In the case of association studies, multiple testing occurs because researchers perform tests of association for large numbers of haplotypes, alleles, or genotypes across entire chromosomes or genomes (Risch and Merikangas 1996). In the case of microarray data analysis, researchers sequentially test thousands of genes for differential expression. Testing at each of the different clustering steps within a hierarchical structure also represents a form of multiple comparisons;

therefore, the experiment-wise type I error is inflated. Various correction methods such as Bonferroni, step-up, and step-down have been employed to adjust for the multiplicity of testing (Reiner et al. 2003). These procedures appear to work well only when the tests in the sequence are independent or weakly correlated. Since the tests within the hierarchy possess a nested structure, these procedures are inappropriate for our situation. As mentioned above, here we propose an alternative solution by defining a single test statistic, for which we evaluate the experiment-wise statistical significance.

2.2 Methods

Local *p*-values. Consider multiple steps in hierarchical clustering. For each of n steps of the hierarchy, we calculate our statistic of interest depending on the application. In the case of haplotype-based association tests, we compute the Pearson χ^2 (Agresti 1996) for a $2 \times s$ contingency table (case/control individuals versus s haplotypes or haplotype classes) while, in the case of survival analyses, we compute the log-rank statistic (Kalbfleisch and Prentice 1980). We represent these statistics as a vector, $\vec{X} = (X_1, X_2, ..., X_n)$, where X_i represents the statistic obtained at the i^{th} step in the clustering process. To make statistics from different applications comparable, we compute the empirical significance level, p_i, associated with X_i and call this a local *p*-value.

We approximate these local empirical significance levels via permutation analysis. These permutation methods involve randomly permuting labels for each individual as follows. For haplotype-based association tests, we permute the case/control labels (Zhao et al. 2000; Zhao and Sham 2002) while for survival analyses, we permute

failure times and censorship statuses jointly. For each permutation of the dataset, we cluster the permuted samples as illustrated by the dendrogram and calculate a null statistic based on the permuted samples at each step in order to generate the null distribution for the statistic. We can represent the collection of null statistics calculated from each of m permutations of the data at each of n steps within the hierarchy as the matrix,

$$\mathbf{X}_{\mathbf{null}} = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{m1} & \cdots & X_{mn} \end{bmatrix},$$

where the entry appearing in the i^{th} row and the j^{th} column, X_{ij}, is the statistic of interest computed from the i^{th} permutation of the data at the j^{th} step in the hierarchy. At each step of the hierarchy, by comparing the statistic we computed from the data with the null statistics we computed from the m permutations, we calculate a local p-value, p_j, as the proportion of permutation samples with a null statistic at least as large as the observed statistic. That is, the local p-value, p_j, is the proportion of null statistics in the j^{th} column of $\mathbf{X}_{\mathbf{null}}$ that are greater than or equal to the statistic, X_i, calculated from the data at the j^{th} step in the hierarchy. We represent the local p-values as the vector, $\vec{p} = (p_1, p_2, ..., p_n)$.

Permutation (randomization) samples allow one to conveniently approximate the sampling distribution of test statistics under the null hypothesis (the "null distribution"). Ideally, permutation tests are based on the total of all permutations but in practice we usually can only collect a random sample from these permutations. The number m of permutation samples should be large enough to adequately represent the sample space of permutations. For the haplotype data (example 1), at each step we compared approximated p-values obtained with different values of m to exact p-values calculated

using the statistical software package *StatXact 5* (see Electronic Resource Information). For the first few steps in the hierarchy, values of m on the order of 10,000 were sufficient to provide p-values very close to the correct ones. However, at later steps, agreement was only obtained with $m = 100{,}000$, presumably because at early steps the total number of permutations is much smaller than at later steps. Table 2.1 displays the local p-values for example 1 computed both with our method using 100,000 permutation samples and with Pearson's exact test as implemented in *StatXact 5*. The calculations for the two survival analyses (examples 2 and 3) were also performed with $m = 100{,}000$.

Table 2.1 Comparison of local *p*-values computed using our method with *p*-values computed using exact tests

Step	Local *p*-Value	95% C.I.	Exact Test *p*-Value
0	0.5275	(0.5244, 0.5306)	0.5270
1	0.4736	(0.4705, 0.4767)	0.4739
2	0.4710	(0.4679, 0.4741)	0.4718
3	0.3533	(0.3503, 0.3563)	0.3532
4	0.3930	(0.3900, 0.3960)	0.3928
5	0.2844	(0.2816, 0.2872)	0.2825
6	0.2726	(0.2700, 0.2754)	0.2706
7	0.2229	(0.2203, 0.2255)	0.2205
8	0.1502	(0.1480, 0.1524)	0.1501
9	0.1282	(0.1261, 0.1303)	0.1289
10	0.1166	(0.1146, 0.1186)	0.1165
11	0.0929	(0.0911, 0.0947)	0.0929
12	0.0668	(0.0653, 0.0684)	0.0674
13	0.0425	(0.0413, 0.0438)	0.0433
14	0.0292	(0.0282, 0.0303)	0.0298
15	0.1659	(0.1636, 0.1682)	0.1659
16	0.2362	(0.2336, 0.2388)	0.2379
17	0.1486	(0.1464, 0.1508)	0.1500
18	0.1089	(0.1070, 0.1108)	0.1099
19	0.0477	(0.0464, 0.0490)	0.0482
20	0.0424	(0.0412, 0.0437)	0.0423

Legend for Table 2.1: This table displays the local *p*-values computed for example 1 using both our method with 100,000 permutation samples and Pearson's exact test as implemented in *StatXact 5*. In addition, the table provides the 95% confidence interval for the *p*-value estimates computed by our method. The zeroth step refers to the data before any clustering is performed. For the test at the zeroth step, the Monte Carlo method (500,000 tables sampled) in *StatXact 5* was employed to find the *p*-value since the problem was too large for the exact test.

Global *p*-value. In order to gain an empirical significance assessment for the entire experiment, we define a single statistic, that is, the smallest of the local *p*-values, $\min_i(p_i)$ (Hoh et al. 2001). To assess the empirical significance level (global *p*-

value), p_{min}, associated with this statistic we generate the null distribution of $\min_i(p_i)$ from the matrix of null statistics, $\mathbf{X_{null}}$. In this matrix, we consider each row (replicate dataset) in turn as observed data and evaluate these data based on the remaining $m - 1$ null data as described above for m null data. That is, for each of these "null observed" permutation samples a minimum p-value is obtained at whatever step it occurs. This leads to a set of m null values for $\min_i(p_i)$. The proportion of these values at least as small as the observed $\min_i(p_i)$ represents the global significance level, p_{min}, associated with our single experiment-wise statistic. Since this approach requires that the p-values be ordered, starting with the most significant, it could be considered a step-down p-value adjustment procedure similar to the procedure developed by Westfall and Young (Westfall and Young 1993). If $p_{min} \leq 0.05$ then we say that the experiment (at least one of the steps in the clustering process) is significant at the 5% level.

It is also of interest to compare the global p-value with the significance level, p_0, of the association or log-rank statistic before clustering since clustering is only beneficial when $p_{min} < p_0$. It may well happen that the smallest p-value, $\min_i(p_i)$, at one of the steps in the course of clustering is smaller than p_0 but the clustering process is such that this smallest p-value has a high probability of occurring by chance. In that case, one will find that $p_{min} > p_0$. For example, observing a minimum p-value smaller than 0.05 and interpreting it as significant is fallacious when this small p-value is easily obtained with probability $p_{min} > 0.05$.

Statistics of interest. As mentioned above, in the case of association studies between haplotypes and disease we employ the Pearson χ^2 to test each step of the

hierarchy for association (Agresti 1996). However, in the case of survival analyses, our statistic of interest is the log-rank statistic (Kalbfleisch and Prentice 1980). It provides an overall comparison of the Kaplan-Meier survival curves for two or more groups of subjects. For r groups, the log-rank statistic asymptotically follows a central χ^2 distribution with $r - 1$ degrees of freedom under the null hypothesis of equality of survival curves.

Validation of the algorithm. In order to validate this method, we analyzed several datasets with a strategy nearly identical to the one described above. The only difference was that this second strategy relies on the theoretical χ^2 distribution to determine the local p-values. Since the use of the theoretical χ^2 distribution for finding statistical significance is valid only for non-sparse datasets, we analyzed several non-sparse datasets with both procedures and compared the results.

In addition, we validated our method using an analytical approach. Suppose we have n steps in the hierarchy formed by clustering, and a test is performed at each step. Then under the null hypothesis, local p-values at all steps of clustering are standard uniform random variables. We can express the global p-value as

$$\Pr\left\{\min_i (p_i)^{null} \leq \min_i (p_i)^{obs}\right\} \tag{2.1}$$

or the probability that value for the minimum of the local p-values from data under the null hypothesis, $\min_i (p_i)^{null}$, is less than or equal to the value of the minimum of the local p-values from the observed data, $\min_i (p_i)^{obs}$. Applying a basic axiom of probability involving complementary events (Ross 2002), we can alter expression (2.1) to become the expression,

$$1 - \Pr\left\{p_1 > \min_i (p_i)^{obs}, p_2 > \min_i (p_i)^{obs}, \ldots, p_n > \min_i (p_i)^{obs}\right\}. \quad (2.2)$$

Under the assumption of independence, this expression simplifies to the Bonferroni correction. However, since the tests are correlated due to the nested structure of the hierarchy, we must pursue an alternate approach. We would like to use the correlation structure between steps to determine this joint probability. Although the multivariate uniform distribution from which the null local *p*-values are derived does not have a one to one correspondence between the correlation structure and the probability density function (pdf), the multivariate normal distribution does have this property. We apply the inverse normal cumulative density function (cdf) to transform the null local *p*-values from standard uniform random variables to a multivariate normal distribution with variance-covariance matrix, **V**, and mean vector, $\vec{u}$. We uniquely define the multivariate normal distribution by setting $\vec{u}$ to be a vector composed entirely of 0 values and using the transformed local *p*-values to estimate **V**. Thus, after the transformation of variables, we can rewrite expression (2.2) as

$$1 - \Pr\left\{Y_1 > \Phi^{-1}\left(\min_i (p_i)^{obs}\right), Y_2 > \Phi^{-1}\left(\min_i (p_i)^{obs}\right), \ldots, Y_n > \Phi^{-1}\left(\min_i (p_i)^{obs}\right)\right\}, \quad (2.3)$$

where each Y_i is the transformed null local *p*-value at the i^{th} step and Φ^{-1} is the inverse standard normal cdf. Because of symmetry, the expression becomes

$$1 - \Pr\left\{-Y_1 < -\Phi^{-1}\left(\min_i (p_i)^{obs}\right), -Y_2 < -\Phi^{-1}\left(\min_i (p_i)^{obs}\right), \ldots, -Y_n < -\Phi^{-1}\left(\min_i (p_i)^{obs}\right)\right\}. \quad (2.4)$$

Since -Y_i also follows a multivariate normal distribution with **V** and $\vec{u}$, the quantity can be expressed as a function of the cdf of this multivariate normal distribution as in expression (2.5).

$$1 - \text{cdf}\left[-\Phi^{-1}\left(\min_i (p_i)^{obs}\right), -\Phi^{-1}\left(\min_i (p_i)^{obs}\right), \ldots, -\Phi^{-1}\left(\min_i (p_i)^{obs}\right)\right] \quad (2.5)$$

With this analytical approach, we examined two datasets—one consisting of two steps as a result of clustering (calibration) and another consisting of nine steps as a result of clustering (analysis). Table 2.2 and Table 2.3 display the contingency tables at each step of clustering for the two-step and nine-step datasets, respectively. For both datasets, we used Mathematica v.4.2 to compute the cdf in expression (2.5). In our estimate of **V**, the diagonal elements were rounded to the value 1. For the two-step dataset, we found an explicit value for the global *p*-value using the analytical method. For the nine-step dataset, we were unable to analytically determine an explicit value for the global p-value due to limitations of software. Instead, we established an upper and a lower bound by applying the analytical approach twice—once using the minimum pair-wise covariance estimate for all off-diagonal elements of **V** and a second time using the maximum pair-wise estimate for all off-diagonal elements of **V**. Thus, the first calculation, assuming the minimum correlation structure, provides a lower bound while the second calculation, assuming the maximum correlation structure, provides an upper bound. We compared the results applying the analytical approach with those from our original algorithm for determining p_{min}. For the validation, we used 10,000 permutation datasets.

Table 2.2 Contingency tables for two-step dataset used for method validation

Step	Group	Number of Cases	Number of Controls
0	Group 1	32	18
	Group 2	24	16
	Group 3	20	30
1	Group 1	56	34
	Group 2	20	30

Legend for Table 2.2: This collection of contingency tables displays the case-control counts for the haplotype classes present at each of the two steps of clustering for a dataset used to validate the method employed to find the global *p*-value. The zeroth step refers to the data before any clustering is performed.

Table 2.3 Contingency tables for nine-step dataset used for method validation

Step	Group	Number of Cases	Number of Controls	Step	Group	Number of Cases	Number of Controls
0	Group 1	10	10	3	Group 1	10	10
	Group 2	7	8		Group 2	7	8
	Group 3	11	14		Group 3	11	14
	Group 4	6	9		Group 4	6	9
	Group 5	12	13		Group 5	18	22
	Group 6	6	9		Group 6	25	20
	Group 7	17	13		Group 7	23	17
	Group 8	8	7	4	Group 1	10	10
	Group 9	11	9		Group 2	7	8
	Group 10	12	8		Group 3	17	23
1	Group 1	10	10		Group 4	18	22
	Group 2	7	8		Group 5	25	20
	Group 3	11	14		Group 6	23	17
	Group 4	6	9	5	Group 1	17	18
	Group 5	12	13		Group 2	17	23
	Group 6	6	9		Group 3	18	22
	Group 7	17	13		Group 4	25	20
	Group 8	8	7		Group 5	23	17
	Group 9	23	17	6	Group 1	17	18
2	Group 1	10	10		Group 2	17	23
	Group 2	7	8		Group 3	18	22
	Group 3	11	14		Group 4	48	37
	Group 4	6	9	7	Group 1	17	18
	Group 5	12	13		Group 2	35	45
	Group 6	6	9		Group 3	48	37
	Group 7	25	20	8	Group 1	52	63
	Group 8	23	17		Group 2	48	37

Legend for Table 2.3: This collection of contingency tables displays case-control counts for the haplotype classes present at each of the nine steps of clustering for a dataset used to validate the method employed to find the global *p*-value. The zeroth step refers to the data before any clustering is performed.

2.3 Results

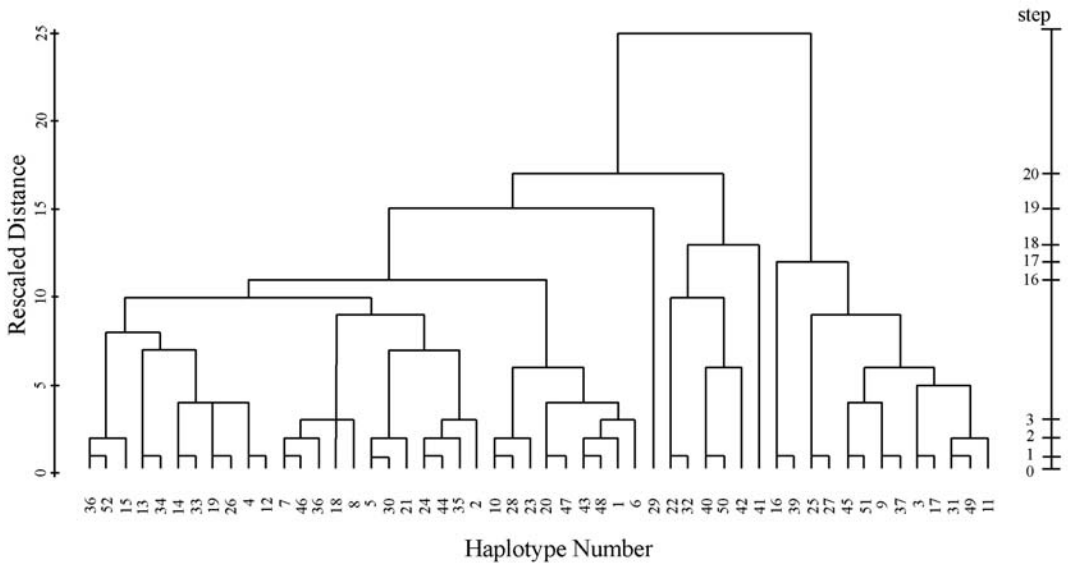
To demonstrate our approach on real data, we reanalyze the following three previously published datasets.

Example 1 (haplotype data). The first dataset consists of 52 statistically predicted haplotypes in 172 African-American study participants (137 case and 35 control individuals) (Hoehe et al. 2000). The aim of that case-control study was to test for association between haplotypes at 25 single-nucleotide polymorphism (SNP) loci in the human μ opioid receptor gene (OPRM1) and substance dependence. The large number of haplotypes was difficult to interpret and appeared to create a situation with insufficient power to detect association. Thus, hierarchical clustering was applied to the 52 haplotypes. These were sequentially grouped according to the procedure CLUSTER (method = BAVERAGE, measure = SEUCLID) from the SPSS software package for Windows (Hoehe et al. 2000). For each step of the resulting dendrogram shown by Figure 2.1, the hierarchical clustering procedure designates which haplotypes are clustered to form haplotype classes. At each step of the hierarchy an association test was performed between haplotype classes and disease status. As the clustering progressed, the number of classes became smaller and smaller.

Using the same clustering methods and resulting hierarchical structure, we apply our algorithm for assessing local and global *p*-values in this dataset. Our *p*-values differ somewhat from the ones previously published (Hoehe et al. 2000) but the patterns of the local *p*-values across the clustering steps shown in Figure 2.2 and in the publication by Hoehe *et al*. (Hoehe et al. 2000), respectively, are highly comparable. Based on $m = 100{,}000$ permutation samples (see section 2.2), we calculate local *p*-values for hierarchical clustering steps zero through 20, where zero represents the step with un-clustered haplotypes and 20 represents the step where only two haplotype groups remain. We find the smallest *p*-value, $\min_i(p_i) = 0.0292$, at step 14 (Figure 2.2). Thus, one is

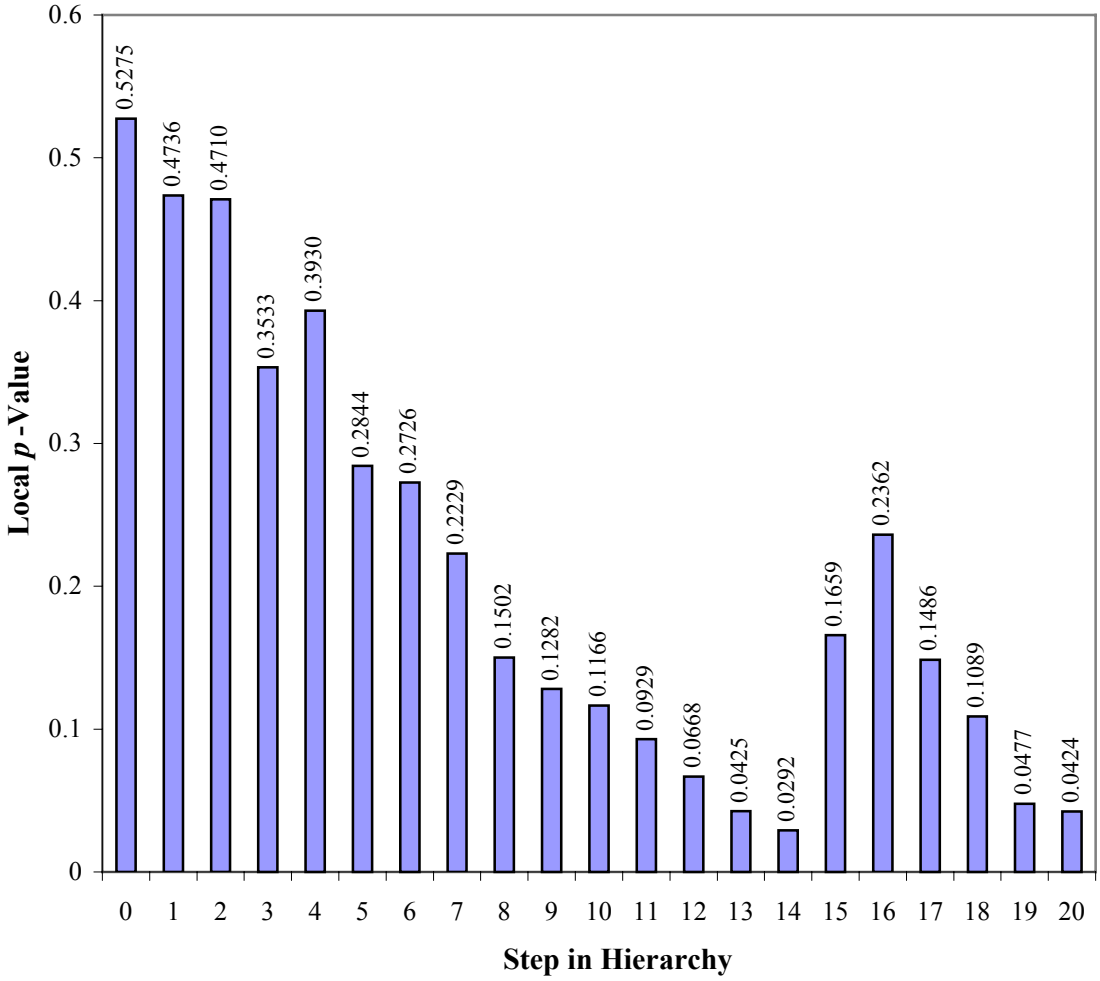
tempted to declare this result borderline significant at the 5% level. However, the (global) significance level associated with this smallest *p*-value turns out to be $p_{min} = 0.1328$, that is, there is more than a 13% random chance (unrelated to association between haplotypes and disease) to find at any step in the hierarchy a minimum *p*-value at least as small as the value of 0.0292 found for the observed data. This result leaves the experiment statistically non-significant. Since clustering produced an experiment-wise significance level of p_{min} less than the initial pre-clustering significance level of $p_0 = 0.5275$, the clustering process did provide a benefit for this dataset (even though the results from clustering were not statistically significant).

Figure 2.1 Dendrogram created by clustering data from Hoehe *et al.* (Hoehe et al. 2000)



Legend for Figure 2.1: This schematized dendrogram reflects the process of clustering case-control observations based on the similarity of haplotype data as measured by the squared Euclidean distance. Distances between haplotype classes are approximated (not to scale) by the vertical axis. Along the bottom of the dendrogram are the identification numbers for the inferred haplotypes as described by Hoehe *et al.* (Hoehe et al. 2000).

Figure 2.2 Results from haplotype-based association tests applied to all steps of the hierarchical structure formed by clustering data from Hoehe *et al*. (Hoehe et al. 2000)



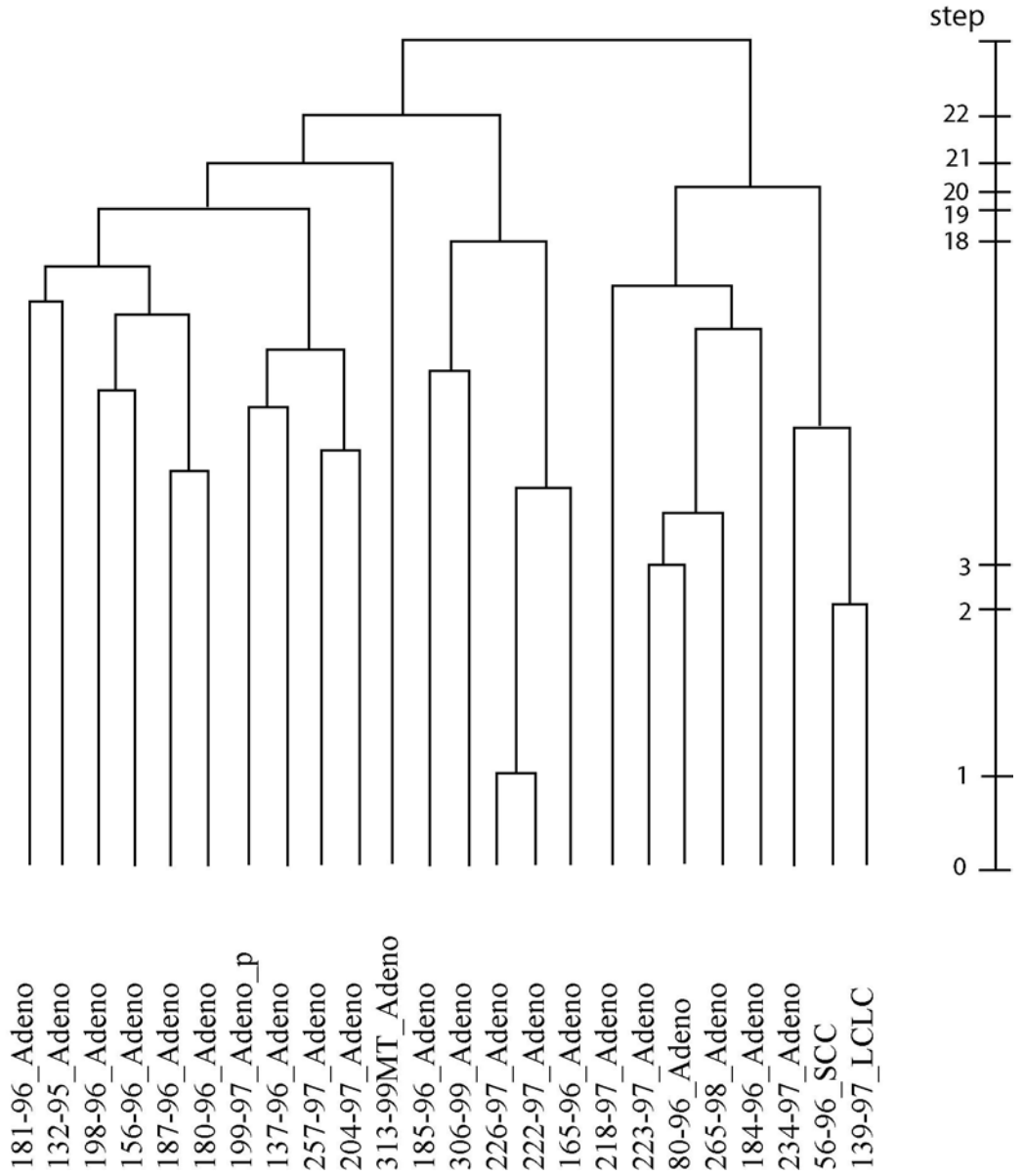
Legend for Figure 2.2: This bar graph presents the local *p*-values we computed at all steps with hierarchical structure.

Example 2 (lung cancer data). This dataset contains expression levels for 835 unique genes represented by 918 cDNA clones in tissues harvested from lung cancer patients and normal individuals (Garber et al. 2001). Specifically, expression levels are measured in 41 adenocarcinomas (ACs), 16 squamous cell carcinomas (SCCs), five large cell lung cancers (LCLCs), five small cell lung cancers (SCLCs), five normal lung

samples, and one normal fetal lung sample. Based on the Complete Linkage method and Pearson's correlation coefficient as a measure of similarity in the CLUSTER software, hierarchical cluster analysis was performed to group the samples according to the degree of similarity present in the gene expression data. In the resulting dendrogram, the AC samples appeared in three distinct clusters. The aim of the study was to examine whether the groups of AC samples created by the hierarchical clustering procedure correlated with clinical outcomes of the AC patients, that is, whether the Kaplan-Meier survival curves differed for these groups (Garber et al. 2001).

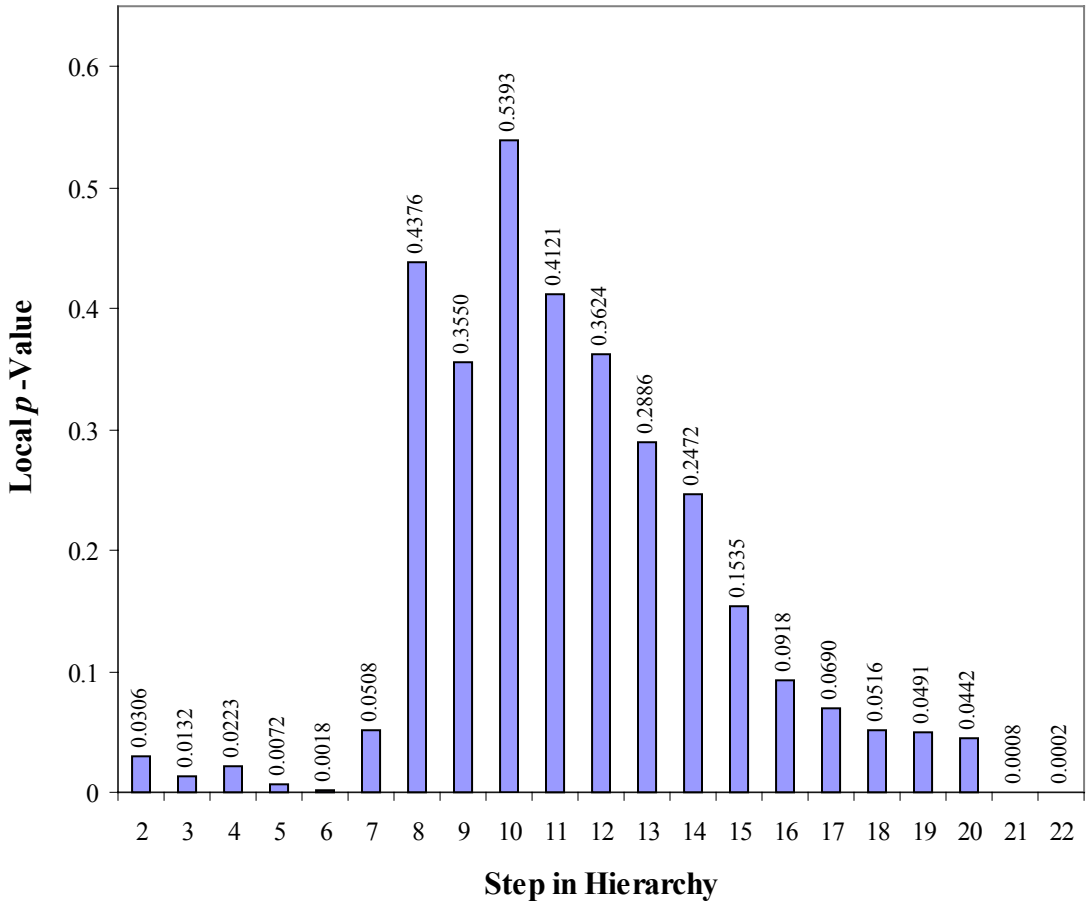
Again, using the same clustering methodology as in the publication (Garber et al. 2001), we apply this technique to their AC data and work with the resulting hierarchical structure for assessing the local and global *p*-values. The dendrogram in Figure 2.3 details the hierarchical clustering of the data (for the 24 AC samples from patients with reported survival information) for steps zero through 22. For each step in the hierarchy we calculate a log-rank statistic and the corresponding local *p*-value (m = 100,000 permutation samples). Figure 2.4 graphically presents these local *p*-values. We exclude the first two clustering steps (0 and 1) from the figure and further assessments because insufficient variability in the log-rank statistic at these steps does not permit meaningful calculation of local *p*-values. (At the zeroth step of clustering, each patient from the survival analysis is in his/her own group.) At step 22, we observe the minimum local *p*-value of 0.0002, and we calculate the global *p*-value for this dataset to be 0.0027. Thus, the experiment shows a statistically significant result, and clustering was effective. It reduced the initial *p*-value of 0.0306 at step 2 to the global significance level of $p_{min} = 0.0027$.

Figure 2.3 Dendrogram created by clustering data from Garber *et al.* (Garber et al. 2001)



Legend for Figure 2.3: This schematized dendrogram reflects the process of clustering microarray samples according to the similarity of their gene expression profiles as measured by the Pearson correlation coefficient. Distances between array sample clusters are approximated (not to scale) by the vertical axis. Along the bottom of the dendrogram are the microarray tissue samples from individuals for which survival data was available (Garber et al. 2001).

Figure 2.4 Results from log-rank tests applied to steps of the hierarchical structure formed by clustering data from Garber *et al.* (Garber et al. 2001)



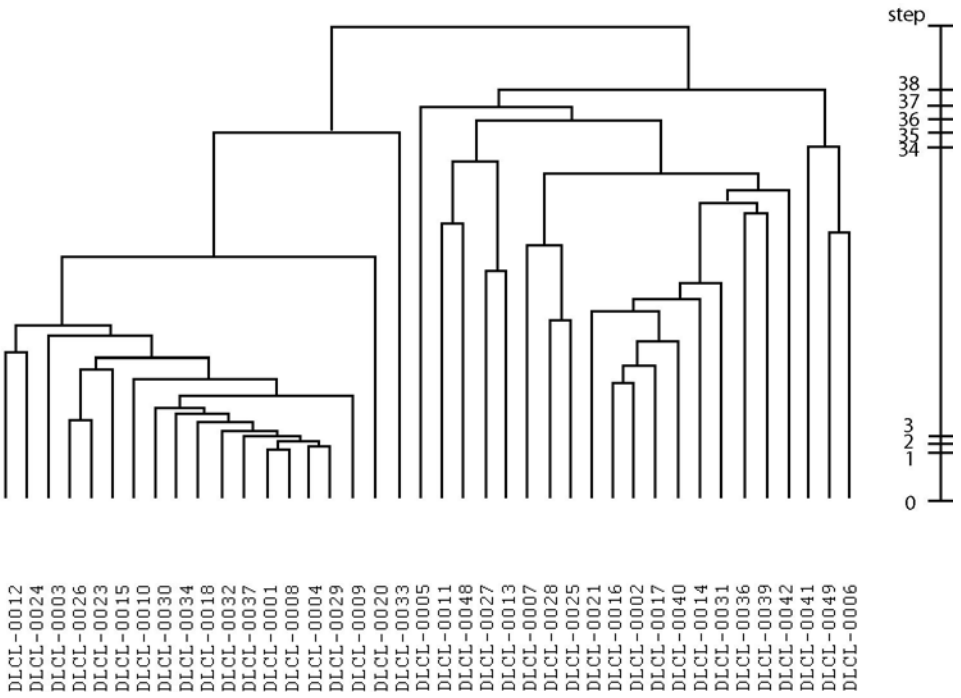
Legend for Figure 2.4: This bar graph displays the local *p*-values we computed at each step within the structure created by hierarchical clustering.

Example 3 (lymphoma data). The third dataset contains expression levels of cDNA clones from genes expressed in germinal center B-cells for 47 samples of diffuse large B-cell lymphoma (DLBCL) (Alizadeh et al. 2000). Hierarchical clustering was performed with the CLUSTER program and the Pearson correlation coefficient as its similarity measure to group the samples by similarity of gene expression levels for all genes expressed in germinal center B-cells. The resulting dendrogram shows two main

branches, one containing samples with expression patterns similar to those of germinal center B-cells and one containing samples with expression patterns similar to those of activated B-cells. To examine the clinical relevance of this subdivision of DLBCL, a Kaplan-Meier survival analysis for the two groups of patients was performed based on the dendrogram's penultimate clustering step (Alizadeh et al. 2000).

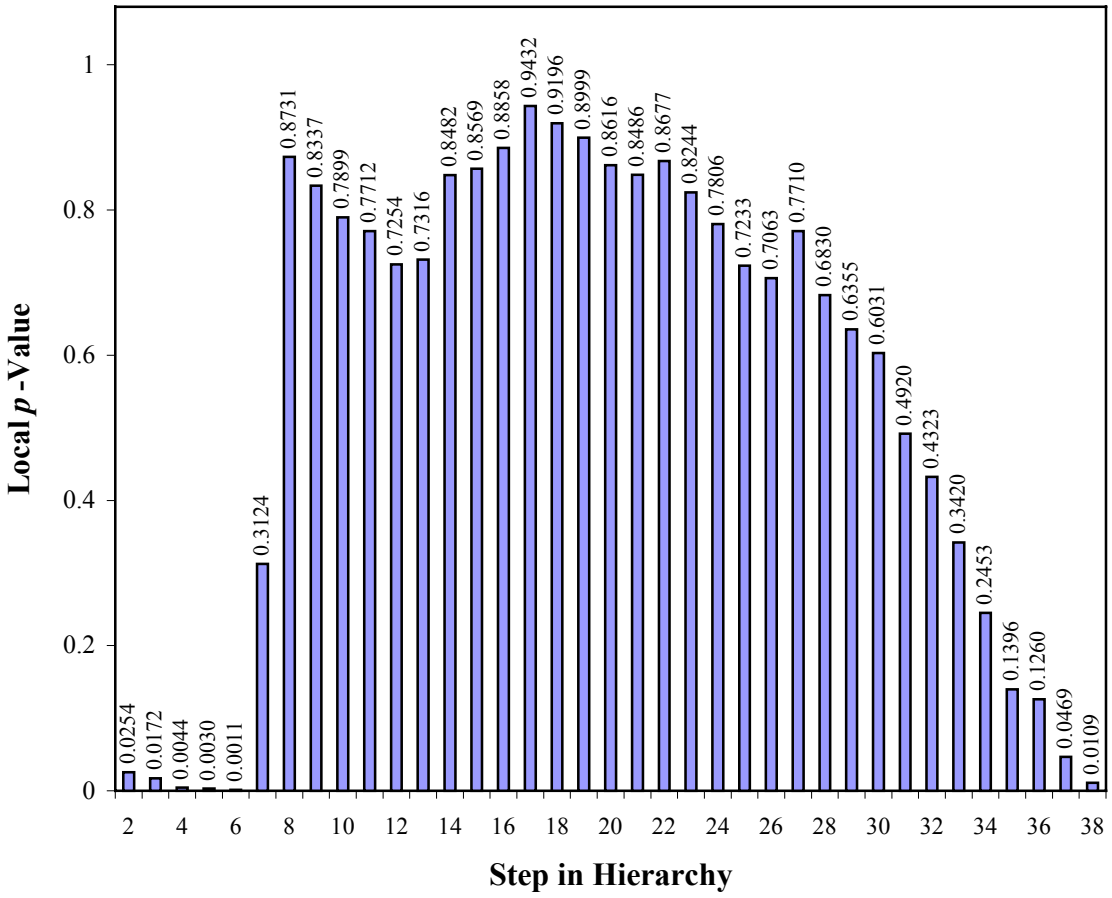
As with the other datasets, we cluster the data with the same method as published (Alizadeh et al. 2000) and use the resulting hierarchical structure for calculations of log-rank statistics and associated local p-values (m = 100,000 permutation samples) at different steps in the hierarchy. The dendrogram in Figure 2.5 provides the order of clustering (for the 40 DLBCL samples from patients with reported survival information) for steps zero through 39 while Figure 2.6 graphically presents local p-values at the different clustering steps. As in Example 2, we observe a very small variance in the log-rank statistic at the first two clustering steps and, therefore, exclude these steps from further analysis. At step 6, we observe the minimum local p-value of 0.0011, with an associated global p-value of $p_{min} = 0.0167$. This result is statistically significant at the 5% level, and clustering has contributed to an increase in significance because un-clustered or only minimally clustered data show much lower significance (higher p-value).

Figure 2.5 Dendrogram created by clustering data from Alizadeh *et al.* (Alizadeh et al. 2000)



Legend for Figure 2.5: This schematized dendrogram reflects the process of clustering microarray samples according to the similarity of their gene expression profiles as measured by the Pearson correlation coefficient. Distances between array sample clusters are approximated (not to scale) by the vertical axis. Along the bottom of the dendrogram are the microarray tissue samples from individuals for which survival data was available (Alizadeh et al. 2000).

Figure 2.6 Results from log-rank tests applied to steps of the hierarchical structure formed by clustering data from Alizadeh *et al.* (Alizadeh et al. 2000)



Legend for Figure 2.6: This bar graph displays the local *p*-values we computed at each step within the structure created by hierarchical clustering.

Validation of the algorithm. We found agreement between our original algorithm for determining p_{min} and the global *p*-value determined with the analytical approach (see section 2.2). Applying the analytical approach for the two-step dataset (see Table 2.2) resulted in a global *p*-value of 0.026 while our original algorithm computed p_{min} to be 0.023. The 95% confidence interval for this estimate is [0.020, 0.026]. For the nine-step dataset (see Table 2.3), the analytical approach produced lower and upper

bounds for p_{min} of 0.248 and 0.657, respectively. Our original algorithm computed p_{min} to be 0.371. A set of naïve bounds can also be created. By assuming perfect pair-wise correlation between tests, one finds the lower bound to be the minimum *p*-value, which for the nine-step dataset was 0.149. In contrast, by assuming the tests at each step to be independent of one another, one finds the upper bound to be the Bonferroni corrected *p*-value, which for the nine-step dataset was 1.0. Thus, the bounds established by using the multivariate normal distribution were a substantial improvement over the naïve bounds.

2.4 Discussion

In hierarchical clustering, evaluating the minimum local *p*-value in isolation, outside of the context of the larger hierarchical structure used to create the data, can drastically affect the interpretation of test results. For example, even though the haplotype data show an apparently significant result with a minimum *p*-value of 0.0292, our analysis demonstrates that clustering the same data, but without association between haplotypes and disease, has a high chance of obtaining such a "significant" result. In fact, that chance is $p_{min} = 0.1328$, which represents the actual significance level of the experiment. On the other hand, as examples 2 and 3 show, clustering can improve the significance of a result and provide a result that is statistically significant.

How can we explain that in some cases clustering is beneficial while in other cases it is not? Presumably, some datasets possess an underlying heterogeneity; that is, such datasets are composed of samples from multiple distinct populations. If the information used for clustering (haplotypes for example 1 and gene expression patterns for examples 2 and 3) is related to the information used to perform the statistical test (in

our examples, proportions of cases to controls and survival times), hierarchical clustering will detect the heterogeneity. Otherwise, the clustering process is random and any heterogeneity detected is artificial. Our approach allows one to distinguish between these two situations. If the clustering process is random because the information used for clustering and test statistic are unrelated (or because the dataset is homogeneous), a large p_{min} will result indicating that any small local p-values probably occurred only by chance. In contrast, if the clustering process is directed by a measurement strongly related to the test statistic, a small p_{min} will result indicating that any heterogeneity found within the hierarchy is most likely real.

Often when hierarchical clustering is applied to a dataset, it is of interest to determine the true number of classes present. This situation commonly arises in the analysis of microarray data. For instance, as in examples 2 and 3, in the study of human cancers, researchers often utilize microarray expression data to cluster samples. From the hierarchical structure created by clustering, it may be of interest to distinguish the optimum number of tumor subclasses that are most clinically relevant. Several statistics-based methods have been utilized to estimate the true number of groups from such microarray expression datasets (Horimoto and Toh 2001; Dudoit and Fridlyand 2002). However, such methods rely solely on the expression data itself. Alternatively, it may prove practical in such microarray expression studies to consider additional information available, such as survival data, for each sample to distinguish clinically relevant subclasses. Employing our procedure of calculating the local p-values for a test statistic at multiple steps within the hierarchy and then selecting the step where the minimum of these p-values occurs as the basis for determining the true number of classes which exist

for a given dataset may provide an advantage over existing methods. Of course, if such a method for determining the true number of classes is applied, the global *p*-value will provide an assessment of its significance. However, applying our procedure to some datasets, such as the data in Example 3, results in determining a large number of true classes. In fact, the number of classes determined may be so large that the use of these expression-based tumor subclasses in clinical diagnosis may not provide a benefit. Therefore, in order to increase the practicality of our method, it may prove necessary to eliminate some of the lower steps in the hierarchy from eligibility for selecting the minimum local *p*-value and the calculation of its significance.

Besides determining subclasses for biological samples, hierarchical clustering is often employed in the context of microarray expression studies in order to identify groups of genes that are regulated in a similar manner. In these cases, clustering is performed on the genes rather than on the samples. Our method relies on two sets of data – one for clustering and a second for the statistical test. Since the samples possess both expression data across genes and survival data, our method is applicable to hierarchies created by clustering on samples. However, genes only possess expression data across samples, and, consequently, our method is inappropriate for analyzing the significance of hierarchies created by clustering on genes.

Our approach may be viewed as a contribution to the problem of multiple testing. We address this problem by defining a single experiment-wise statistic whose associated empirical significance level represents the overall significance of the experiment. For the cases we have examined, the experiment refers to performing a test at each step in a hierarchy created by clustering. However, the meaning of experiment can be expanded to

reflect other practices adopted by researchers. For example, researchers may apply several clustering algorithms involving various combinations of clustering methods and distance measures before finalizing their choice of clustering algorithm. Since this practice introduces an additional test at each step within each of the trial hierarchies, it compounds the effect of multiple testing. Additionally, in some situations researchers may be interested in testing for heterogeneity among groups with multiple measurements. For instance, when searching for clinically relevant subclasses of cancer, researchers may examine groups for differences in survival times as well as differences in physical characteristics of the tumor cells. Both sets of information may be clinically relevant; however, to correct for the additional testing, the meaning of the experiment in calculating p_{min} must be expanded to reflect the entire process employed by the researcher. Of course, it is possible that the process of hierarchical clustering forms medically relevant groups that do not display heterogeneity for any of the measurements collected. In this case, our strategy will not find these groups as the true grouping structure for the samples.

Several other methods addressing multiple comparison problems have been proposed and are in current use. In particular, as an alternative to the classical significance level, p, the false discovery rate (FDR) has become rather popular (Reiner et al. 2003). However, it is important to keep in mind that p and FDR are not really comparable. The classical significance level, p, is the conditional probability of a significant test result given the null hypothesis is true (the expected proportion of false positive results among all "false" results, i.e., results obtained under the null hypothesis) while FDR is the conditional probability of the null hypothesis being true given a

significant test result (the expected proportion of false positive results among all "positive" results, i.e., significant test results). Future research will have to determine which of these various approaches to eliminate the effects of multiple testing is most effective.

CHAPTER 3: ARE MOLECULAR HAPLOTYPES WORTH IT? A COST EFFECTIVE METHOD FOR TREATING MISCLASSIFICATION IN HAPLOTYPE-BASED ASSOCIATION

3.1 Introduction

While clustering haplotype data may create a situation that requires correction for the multiplicity of testing, other haplotype-based association studies in which no clustering is employed face complications as well. One such complication is the issue of haplotype misclassification.

Although recent advances in molecular biology have produced techniques to unequivocally ascertain phased haplotypes (Michalatos-Beloin et al. 1996; Clark et al. 1998; Yan et al. 2000; Douglas et al. 2001; Patil et al. 2001; Burgtorf et al. 2003; Ding and Cantor 2003; Horan et al. 2003; Hoppe et al. 2004; Proudnikov et al. 2004; Hoppe et al. 2006), such molecular haplotyping techniques are seldom employed due to their expense and incongruity to automation. A more pragmatic alternative is to estimate haplotype frequencies or infer haplotype pairs by applying statistical methods to multilocus genotypes (Clark 1990; Xie and Ott 1993; Terwilliger and Ott 1994; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995; Zhao et al. 2000; Stephens et al. 2001b; Zhao and Sham 2002; Stephens and Donnelly 2003). Although knowledge of parental genotypes can simplify the problem, for haplotype-based association studies the sample design generally calls for collecting DNA on unrelated individuals, and, in this case, the statistical methods for haplotype estimation must consider each individual as an independent observation.

For these non-familial study designs, several methods are available to estimate haplotype frequencies and/or infer haplotype pairs. The main methods follow one of two approaches—1) relying on the EM algorithm (Dempster et al. 1977) for a likelihood approach (Xie and Ott 1993; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995) and 2) using a Bayesian approach to apply a prior based on coalescence theory (Stephens et al. 2001b; Stephens and Donnelly 2003) or a Dirichlet prior (Niu et al. 2002). The EM algorithm-based methods are implemented in SNPHAP (see Electronic Resource Information), HAPLO (Hawley and Kidd 1995), and PL-EM (Qin et al. 2002) while the Bayesian approaches are implemented in PHASE (Stephens et al. 2001b) (see Electronic Resource Information) and HAPLOTYPER (Niu et al. 2002). Several investigators have examined the accuracy of these approaches for both haplotype frequency estimation and haplotype inference (Fallin and Schork 2000; Tishkoff et al. 2000; Clark et al. 2001; Xu et al. 2002; Stephens and Donnelly 2003; Adkins 2004; Kang et al. 2004; Niu 2004; Xu et al. 2004; Heid et al. 2005; Sabbagh and Darlu 2005; Zhang et al. 2005; Marchini et al. 2006; Proudnikov et al. 2006).

Many statistical methods are available to perform tests of haplotype-based case-control association. One method calculates the likelihood of the data in terms of the estimated haplotype frequencies. An alternative method relies on the use of a contingency table containing the case-control counts for each inferred haplotype. Since it applies the same format as the classic genotypic and allele case-control studies and accounts for each phased haplotype explicitly, many researchers prefer the latter approach. One can determine the counts in the contingency table by inferring phased haplotypes for each individual (or by multiplying each haplotype frequency estimate by

the total number of haplotypes in the study). Once the contingency table contains the haplotype (or diplotype) counts, a Pearson χ^2 test or a likelihood ratio test can be performed. However, the counts entered in the contingency table may misrepresent the true situation since inferred haplotypes (and haplotype estimates) are prone to errors. These haplotype misclassification errors may affect the behavior of the statistical test performed.

Thus, the purpose of this work is to address the effects of haplotype misclassification on the false positive rate and power of commonly used tests of haplotype-based association. Specifically, this research aims to 1) classify the nature of the misclassification present in calling phased haplotypes; 2) determine the appropriateness of using the asymptotic χ^2 distribution and permutation methods to evaluate the significance of the test statistics we employ; and 3) compare the power of our test statistic which accounts for haplotype misclassification with the power of the standard likelihood ratio test statistic when the costs are fixed.

3.2 Methods

Test statistics. In order to detect an association between a haplotype pair and disease status, we employed two statistical tests on $2 \times k$ contingency tables where k is the number of haplotype pair categories found by inference. These tests include the standard likelihood ratio test (LRT_{std}) and a likelihood ratio test that employs a double-sampling approach to allow for the misclassification inherent in the haplotype inference procedure (LRT_{ae}). The LRT_{std} is a likelihood ratio statistic that treats the called haplotype pairs as observations, and as a result the likelihood is the multinomial

distribution where the called haplotype pairs are the categories (Agresti 1996). The LRT_{ae} statistic is a likelihood ratio statistic that employs a double-sampling procedure to account for the misclassification present in a haplotype inference. On all the individuals in the study, there is a fallible measure (Tenenbein 1970; Tenenbein 1972), the haplotype pairs inferred from the multilocus genotypes, and on a subset of these individuals, there is a second measure which is considered to be infallible (Tenenbein 1970; Tenenbein 1972), molecular haplotypes. By comparing the fallible data with infallible data, the LRT_{ae} procedure estimates the misclassification rates present in the fallible data and incorporates this information into the likelihood calculation (Gordon et al. 2004).

Computation of the LRT_{std} and LRT_{ae} statistics. For completeness, details regarding the LRT_{std} and LRT_{ae} statistics including notation and computation as described by Gordon *et al.* are provided in this section. We present the mathematical basis for computation of the LRT_{std} and LRT_{ae} statistics. This work largely follows from the original publication on the LRT_{ae} statistic (Gordon et al. 2004). The primary difference is that, in this work, we assume only misclassification in haplotype pairs (called "genotypes" in the original publication) and assume no misclassification of phenotype. Because we do not collect a second phenotype measurement, we assume that all phenotype classifications are correct. We begin with some notation.

Notation. For all terms, the index i' is either 0 (case) or 1 (control) and the integer indices j and j' range from 1 through k inclusive, where k is the number of haplotype pairs.

We use prime superscripts to distinguish true categories from observed categories. For example, j' refers to the true haplotype pair classification for an individual. Also, we use the superscript t to denote "true" (as compared with observed) when referring to either an event or a parameter. For example, the notation $X^t_{j'}$ represents the event that an individual's true haplotype pair classification is j', whereas the notation X_j represents the event that an individual's observed haplotype pair classification is j (see below). Similarly, the notation $p^t_{i'j'}$ represents the true probability of the haplotype pair j' for individuals with (true) phenotype classification i', whereas the notation $p_{i'j}$ represents the observed probability of the haplotype pair j for individuals with (true) phenotype classification i'. With this notation, we may distinguish between the events X^t_0 and X_0 and the probabilities p^t_{01} and p_{01}.

$n^{(1)}_{i'j'j}$ = Number of individuals with (true) phenotype category i', true haplotype pair category j', and observed haplotype pair category j. (These individuals are double-sampled on haplotype pair classification.)

$n^{(1)}_{i'j'+} = \sum_j n^{(1)}_{i'j'j}$.

$n^{(2)}_{i'j}$ = Number of individuals with (true) phenotype category i' and observed haplotype pair category j.

$$n^{(2)}_{i'+} = \sum_j n^{(2)}_{i'j}$$

$$n^{(2)}_{+j} = \sum_{i'} n^{(2)}_{i'j}$$

$n = \sum_{i'}\sum_{j'}\sum_j n^{(1)}_{i'j'j} + \sum_{i'}\sum_j n^{(2)}_{i'j}$; Note that n is the total sample size.

$Y_{i'}^t$ =Event that an individual has phenotype $i', (i' = 0,1)$.

X_j =Event that an individual has observed haplotype pair $j, 1 \le j \le k$.

$X_{j'}^t$ =Event that an individual has true haplotype pair $j', 1 \le j' \le k$.

$X_{i'j'}^t$ =Event that an individual has phenotype $i', (i' = 0,1)$ and true haplotype pair $j', 1 \le j' \le k$.

$q_{i'}^t = \Pr(Y_{i'}^t)$ = True sampling frequency of phenotype i'.

$p_{i'j} = \Pr(X_j \mid Y_{i'}^t)$ = Observed population frequency of haplotype pair j for individuals with true phenotype i'.

$p_{i'j'}^t = \Pr(X_{j'}^t \mid Y_{i'}^t)$ = True population frequency of haplotype pair j' for individuals with phenotype i'.

$p_{*j'}^t = \Pr(X_{j'}^t)$ = True population frequency of haplotype pair j' under the null hypothesis that $p_{0j'}^t = p_{1j'}^t = p_{*j'}^t$.

$p_{*j} = \Pr(X_j)$ = Observed population frequency of haplotype pair j under the null hypothesis that $p_{0j'}^t = p_{1j'}^t = p_{*j'}^t$.

Note: For each $i', \sum_{j'} p_{i'j'}^t = \sum_{j} p_{i'j} = 1$. Also, $q_0^t + q_1^t = 1$.

$\theta_{j'j} = \Pr(X_j \mid X_{j'})$

Note: When $j' \neq j$, these parameters are referred to as misclassification parameters (Tenenbein 1972; Gordon et al. 2002). We make use of the double-sample data structure to determine estimates of haplotype pair misclassification values $\theta_{j'j}$. The misclassification parameter estimates $\hat{\theta}_{j'j}$ are $\hat{\theta}_{j'j} = m_{j'j} / m_{j'+}$ (see below).

$m_{j'j}$ $(1 \le j, j' \le k)$ = The number of individuals that have been classified by the fallible method as haplotype pair j and by the infallible method as haplotype pair j'.

$m_{j'+} = \sum_j m_{j'j}$.

$\ln(L_{1,std})$ = Log-likelihood of data when not correcting for misclassification, where haplotype pair frequencies $p_{i'j}$ are allowed to differ among different phenotype classes (i.e., p_{0j} is not necessarily equal to p_{1j} for every j) (also see equation 3.1b below).

$\ln(L_{0,std})$ = Log-likelihood of data when not correcting for misclassification, where haplotype pair frequencies $p_{i'j}$ are constrained to be equal among different phenotype classes (i.e., $p_{0j} = p_{1j} = p_{*j}$ for every j) (also see equation 3.1b).

$\ln(L_{1,ae})$ = Log-likelihood of data as represented in equation (3.4), where haplotype pair frequencies $p^t_{i'j'}$ are allowed to differ among different phenotype classes. (i.e., $p^t_{0j'}$ is not necessarily equal to $p^t_{1j'}$ for every j')

$\ln(L_{0,ae})$ = Log-likelihood of data as represented in equation (3.4) below, where haplotype pair frequencies $p^t_{i'j'}$ are constrained to be equal among different phenotype classes. (i.e., $p^t_{0j'} = p^t_{1j'} = p^t_{*j'}$ for every j')

Log-likelihood of observed data and likelihood ratio test statistics

We compute the log-likelihood of the observed data under the null and alternative hypotheses, allowing for error. The null hypothesis we test is $H_0 : p^t_{0j'} = p^t_{1j'} = p^t_{*j'}$ for all

haplotype pairs j'. The alternative hypothesis is $H_1 : p^t_{0j'} \neq p^t_{1j'}$ for at least one j'. Under either hypothesis, we have, by definition, the log-likelihood of the data given by:

$$\ln(L_{ae}) = \sum_{i'}\sum_{j}\sum_{j'} n^{(1)}_{i'j'j} \ln[\Pr(Y^t_{i'}, X_j, X^t_{j'})] + \sum_{i'}\sum_{j} n^{(2)}_{i'j} \ln[\Pr(Y^t_{i'}, X_j)], \tag{3.1a}$$

where the notation $\Pr(A, B, C, ...)$ is the probability of observing event *A* and event *B* and event *C* and so forth and $n^{(1)}_{i'j'j}, n^{(2)}_{i'j}$ represent the counts for different categories of double-sample information (see above for definitions of all notation). For example, $n^{(2)}_{i'j'j}$ is the number of individuals who have been double-sampled for haplotype pair classification and who have true phenotype classification i', true haplotype pair classification j', and observed haplotype pair classification j. In equation (3.1a), the subscript i' runs over all phenotype classifications ($0 \le i' \le 1$) and the subscripts j, j' run over all haplotype pair classifications ($1 \le j, j' \le k$).

When a double-sample has not been collected or when we assume that there is no error in the data, equation (3.1a) reduces to:

$$\begin{aligned} \ln(L_{std}) &= \sum_{i'}\sum_{j} n^{(2)}_{i'j} \ln[\Pr(Y^t_{i'}, X_j)] \\ &= \sum_{i'}\sum_{j} n^{(2)}_{i'j} \ln(p_{i'j} q^t_{i'}) \\ &= \sum_{i'}\sum_{j} n^{(2)}_{i'j} [\ln(p_{i'j}) + \ln(q^t_{i'})]. \end{aligned} \tag{3.1b}$$

A key assumption in our work is that the observed haplotype pair is only dependent on the underlying true haplotype pair and not on phenotype so that $\Pr(X_j \mid X^t_{j'}, Y^t_{i'}) = \Pr(X_j \mid X^t_{j'})$. It follows that:

$$\begin{aligned}\Pr(Y_{i'}^t, X_j, X_{j'}^t) &= \Pr(X_j \mid X_{j'}^t, Y_{i'}^t)\Pr(X_{j'}^t, Y_{i'}^t) \\ &= \Pr(X_j \mid X_{j'}^t)\Pr(X_{j'}^t \mid Y_{i'}^t)\Pr(Y_{i'}^t) \\ &= \theta_{j'j} p_{i'j'}^t q_{i'}^t.\end{aligned} \tag{3.2}$$

Using equation (3.2) and the fact that

$$\Pr(Y_{i'}^t, X_j) = \sum_{j'} \Pr(Y_{i'}^t, X_j, X_{j'}^t), \tag{3.3}$$

we may rewrite the log-likelihood (3.1a) as:

$$\ln(L_{ae}) = \sum_{i'}\sum_{j'}\sum_{j} n_{i'j'j}^{(1)} \ln[\theta_{j'j} p_{i'j'}^t q_{i'}^t] + \sum_{i'}\sum_{j} n_{i'j}^{(2)} \ln[\sum_{j'} \theta_{j'j} p_{i'j'}^t q_{i'}^t]. \tag{3.4}$$

From equation (3.4) we can determine the log-likelihood of the data under H_1 using the EM algorithm estimates of $p_{i'j'}^t$ and $q_{i'}^t$ (see (Gordon et al. 2004)). Similarly, we can determine the log-likelihood of the data under H_0 using the EM algorithm estimates of $p_{*j'}^t$ and $q_{i'}^t$. The estimates of $q_{i'}^t$ may differ under the null and alternative hypotheses.

It follows from equation (3.4) that the log-likelihoods $\ln(L_{0,ae})$ and $\ln(L_{1,ae})$ (equation (3.1a)) are completely determined by misclassification parameters $\theta_{j'j}$, the true parameters $p_{i'j'}^t, p_{*j'}^t, q_{i'}^t$, and sample counts $(n_{i'j'j}^{(1)}, n_{i'j}^{(2)})$. In the previous sentence, $\ln(L_{0,ae})$ refers to the situation under the null hypothesis, where the terms $p_{i'j'}^t$ in equation (3.4) are replaced by $p_{*j'}^t$. In contrast, $\ln(L_{1,ae})$ refers to situation under the alternative hypothesis, where the terms $p_{i'j'}^t$ in equation (3.4) remain. Our test of H_1 versus H_0 is a likelihood ratio test (Kendall et al. 1994), which we call the *l*ikelihood *r*atio *t*est *a*llowing for *e*rror, or LRT_{ae}. It is given by

$$\text{LRT}_{\text{ae}} = 2[\ln(L_{1,ae}) - \ln(L_{0,ae})], \tag{3.5a}$$

where $\ln(L_{1,ae})$ and $\ln(L_{0,ae})$ are determined using equation (3.4) with the EM algorithm estimates of the various parameters. Asymptotically, LRT_{ae} is distributed as χ^2_{k-1}, where the degrees of freedom (df) are $k-1$ for a set of k observed haplotype pairs (Gordon et al. 2004). For small samples or in situations where the asymptotic distribution may not hold, we can compute *p*-values via permutation (Gordon et al. 2004; Proudnikov et al. 2006).

The standard likelihood ratio test, denoted LRT_{std}, that does not make any correction, has its log-likelihoods computed solely from the observed data. That is,

$$\text{LRT}_{\text{std}} = 2[\ln(L_{1,std}) - \ln(L_{0,std})], \tag{3.5b}$$

where the log-likelihoods under the null and alternative hypotheses are computed using the estimates $\hat{p}_{i'j} = n^{(2)}_{i'j} / n^{(2)}_{i'+}$, $\hat{p}_{*j} = (n^{(2)}_{0j} + n^{(2)}_{1j})/n$, $\hat{q}_{i'} = n^{(2)}_{i'+}/n$, ($n^{(2)}_{i'+} = \sum_j n^{(2)}_{i'j}$) that are then substituted into equation (3.1b) (Rice and Holmans 2003). When there is no correction for misclassification, there is no need to compute $\hat{q}_{i'}$ under both the null and alternative hypothesis, as the terms with $\hat{q}_{i'}$ will cancel in the expression for the difference of the log-likelihoods (equation (3.5b)).

Permuted and asymptotic *p*-values. We applied two methods for evaluating the *p*-value or statistical significance of each statistic. The first method relies on using the central χ^2 distribution to find the *p*-value since, according to statistical theory under the null hypothesis of no association, twice the natural logarithm of the likelihood ratio follows the central χ^2 distribution asymptotically for large sample sizes (Agresti 1996). In addition, it has been shown that when Cochran's rule is followed (more than five observations in each cell of the contingency table), the presence of non-differential

misclassification does not affect the distribution of the likelihood ratio test statistic under the null hypothesis of no association (Mote and Anderson 1965; Gordon et al. 2004). The second method employs permutation testing to generate the distribution of the test statistic under the null hypothesis and to determine its statistical significance. In this thesis, *p*-values found with the former and latter approaches are referred to as asymptotic *p*-values and permutation *p*-values, respectively.

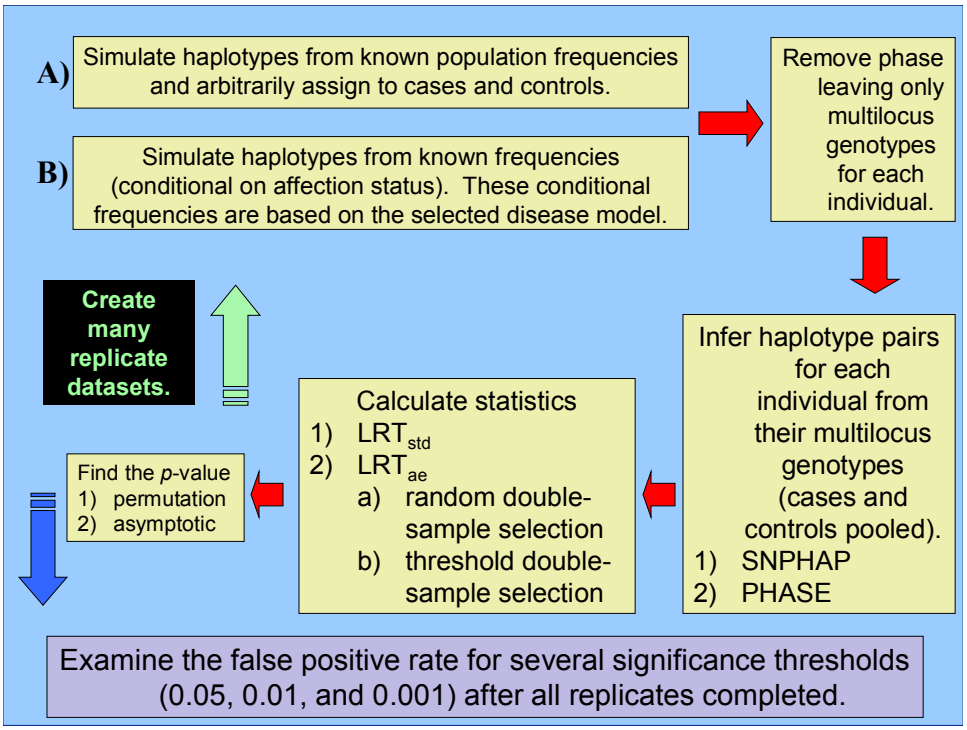
Description of data generation and analysis. To investigate the behavior of these test statistics for a variety of situations, we applied these statistical tests to many simulated datasets. Figure 3.1 illustrates the procedure we used to simulate the data and to evaluate the false positive rate and power at fixed significance levels for each statistic. For the analysis of each replicate dataset simulated, the multilocus genotype data from cases and controls were pooled to infer haplotype pairs for each individual. Individuals were assigned the haplotype pair with the highest posterior probability. The posterior probability of a given haplotype pair is defined as the probability of that haplotype pair being the true haplotype pair conditioned on the observed multilocus genotypes. For example, consider two SNP marker loci where *A* and *a* represent the alleles at the first locus while *B* and *b* represent the alleles at the second locus. The posterior probability of the haplotype pair, *AB* and *ab*, can be expressed as

$$\Pr\{AB, ab = true \mid Aa, Bb = observed\} = \frac{\Pr\{AB\}\Pr\{ab\}}{\Pr\{AB\}\Pr\{ab\} + \Pr\{Ab\}\Pr\{aB\}}$$

by applying Bayes' Theorem and simplifying. The EM algorithm can be used to estimate the probability of each haplotype in the posterior probability expression from multilocus genotypes.

The inferred haplotypes are sufficient for the computation of LRT_{std}; however, LRT_{ae} requires additional information in the form of molecular haplotypes for a subset of the individuals in the study. We employed two alternative procedures for selecting individuals for the double-sample (individuals with molecular haplotypes in addition to genotypes). In one selection scheme, individuals were selected randomly. In the other selection scheme, individuals possessing the most ambiguity in their statistically inferred haplotype pairs were prioritized in selecting the double-sample. Specifically, we double-sampled those individuals with the smallest posterior probabilities associated with their inferred haplotype pair up to a posterior probability threshold, δ, of 0.85 or until the number of individuals specified by the maximum double-sample proportion was reached. Therefore, under this second selection scheme the number of individuals double-sampled varied between replicate datasets. In this thesis, the former and latter procedures for determining the double-sample are referred to as random and threshold double-sample selection, respectively.

Figure 3.1 Schematic flow chart illustrating the procedure for data simulation and analysis



Legend for Figure 3.1: This schematic flow chart illustrates the procedure employed for computing (A) type I error and (B) power by way of data simulation.

Two SNP scenario

Evaluation of false positive rate for permutation and asymptotic p-values. For the simplest non-trivial case, the scenario where the haplotype under evaluation includes two SNPs, we applied a fractional factorial design (Box et al. 1978) to perform a comprehensive study of type I error. For the type I error, haplotype pairs were inferred using both SNPHAP v 1.3.1 (see Electronic Resource Information) and PHASE v 2.1.1 (Stephens et al. 2001b) (see also Electronic Resource Information). Table 3.1 contains the fractional factorial design settings for the study of type I error for the scenario involving two SNP markers. We consider a $1/2(2^g)$ fractional factorial design, where

$g = 6$. Because of redundancy, we were able to reduce the number of experimental runs from 32 to 18. For instance, under the null hypothesis of no association, a run with 1000 cases and 250 controls is equivalent to a run with 250 cases and 1000 controls (with all other factors having equal settings to those for the first run). During each run, 10,000 replicate datasets were simulated. We performed the 18 runs with both of the two alternative procedures for selecting the double-sample – random and threshold double-sample selection. For the threshold double-sample selection method, δ was 0.85, and the maximum double-sample proportion was set to the value of α in the fractional factorial design.

Table 3.1 Fractional factorial design parameter settings for the study of type I error assuming the haplotype under investigation contains two SNP markers

Description of parameter	Low	High
Number of cases	250	1000
Number of controls	250	1000
Minor allele frequency at locus 1	0.1	0.5
Minor allele frequency at locus 2	0.1	0.5
LD between locus 1 and 2 (measured by *D'*)	0	0.9
Proportion of individuals double-sampled (α)	0.25	0.75

Legend for Table 3.1: This table presents the settings for all parameters considered in the type I error simulations assuming the haplotype under investigation contains two SNP markers. We consider a $1/2\,(2^g)$ fractional factorial design, where $g = 6$. The number of experimental runs was reduced from 32 to 18 due to redundancy. *D'* is the standardized linkage disequilibrium measure. The simulations included 10,000 replicates, and haplotype pairs were inferred using both SNPHAP v 1.3.1 and PHASE v 2.1.1. LRT_{ae} was computed with the random and threshold double-sample selection methods for all 18 runs in the fractional factorial design. For the threshold double-sample selection method, δ was 0.85, and the maximum double-sample proportion was set to the value of α in the fractional factorial design.

To evaluate each test statistic's ability to maintain the correct type I, we examined the distribution of the *p*-values computed for data simulated under the null hypothesis of no association. We performed two goodness-of-fit tests, the Kolmogorov-Smirnov (KS) and the Anderson-Darling (AD) tests (DeGroot 1991), to determine whether the *p*-values deviate significantly from the standard uniform distribution and examined the false positive rate for significant thresholds of 0.05, 0.01, and 0.001 .

Evaluation of power for fixed cost. We also evaluated the behavior of these statistics under the hypothesis that a disease allele at an unobserved locus exists in linkage disequilibrium (LD) with the haplotype under study. Table 3.2 contains the factorial design settings for the power study in the scenario involving two SNP markers. The factorial design includes three factors, disease model, genotype relative risk (Schaid and Sommer 1993) for the homozygote genotype (R_2), and the disease allele frequency (*DAF*). Each factor contains two levels. For the disease model factor, the two levels are a dominant disease model and a multiplicative disease model. The dominant disease model requires that $R_2 = R_1$ while the multiplicative disease model requires that $R_2 = R_1^2$, where R_1 and R_2 are the genotype relative risks for the heterozygote and homozygote genotypes, respectively. Specifically, the genotype relative risks are defined as the following. If the penetrances, f_i, are defined by

$f_i = \Pr(\text{affected} \mid i \text{ copies of disease allele})$, where $i = 0,\ 1,\ 2$, the genotype relative risks, R_1 and R_2, are defined by

$R_1 = f_1 / f_0$ and $R_2 = f_2 / f_0$, respectively (Schaid and Sommer 1993).

Table 3.2 Factorial design parameter settings for the study of power assuming the haplotype under investigation contains two SNP markers

Description of parameter	Low	High
Disease model	dominant	multiplicative
Genotype relative risk of homozygote (R_2)	2	3.5
Disease allele frequency (*DAF*)	0.07	0.27

Legend for Table 3.2: This table presents the settings for all parameters considered in the power simulations assuming the haplotype under investigation contains two SNP markers. We consider a 2^g factorial design, where $g = 3$. The dominant disease model requires that $R_2 = R_1$ while the multiplicative disease model requires $R_2 = R_1^2$, where R_1 and R_2 are the genotype relative risks for the heterozygote and homozygote genotypes, respectively. For the random double-sample selection method, the proportion of individuals double-sampled (α) was 0.75 while a haplotype pair posterior probability threshold (δ) of 0.85 and a maximum double-sample proportion of 0.75 were used for the threshold double-sample selection method. The cost ratio of molecular haplotyping to genotyping (r) was 5. For each combination of settings, 1000 replicate datasets comprised of 500 cases and 500 controls were simulated. The disease prevalence was 0.025; the LD between the disease allele and the linked haplotype was 0.9 (measured by *D'*); and the population haplotype frequencies were 0.05, 0.15, 0.25, and 0.5. The haplotype with a frequency of 0.05 was linked to the disease allele when *DAF* = 0.07, and the haplotype with a frequency 0.25 was linked to the disease allele when *DAF* = 0.27. Haplotype pairs were inferred using both SNPHAP v 1.3.1 and PHASE v 2.1.1.

As with the study of type I error, we inferred the haplotypes for the power simulations with both SNPHAP v 1.3.1 and PHASE v 2.1.1. The proportion of individuals double-sampled, α, for the LRT_{ae} method (random double-sample selection) was set at 0.75. For the threshold double-sample selection, δ was set at 0.85, and the maximum double-sample proportion was 0.75. In the power simulations, the conditional (on case status) haplotype frequencies were found from the specified disease model

parameters by a method described by Sham and subsequently by De La Vega *et al.* (Sham 1998; De La Vega et al. 2005) (also see the Power for Association with Error (PAWE) website at http://linkage.rockefeller.edu/derek/pawe1.html). However, we selected a specific haplotype to be in LD with the disease allele. For completeness, details regarding the conditional haplotype frequencies including notation and computation as described by De La Vega *et al*. (2005) are provided at the end of this section. During each run, 1000 replicate datasets comprised of 500 cases and 500 controls were simulated. For these simulations, the disease prevalence was 0.025; the LD between the disease allele and the linked haplotype was 0.9 (measured by *D'* (Lewontin 1964)); and the population haplotype frequencies were 0.05, 0.15, 0.25, and 0.55. Selection of the specific haplotype in LD with the disease allele depended on the disease allele frequency (*DAF*). The haplotype occurring with a frequency most similar to that of the disease allele was selected. Thus, haplotypes with frequencies of 0.05 and 0.25 were selected as the variant in LD with the disease when the *DAF* was set at 0.07 and 0.27, respectively. As with the evaluation of the false positive rate, we performed all 8 runs from the factorial design using both random and threshold double-sample selection.

To compare the power of the two test statistics, we evaluated the power of the statistics under fixed cost conditions. Since the LRT_{ae} requires the additional cost associated with obtaining molecular haplotypes on a subset of the samples, we reduced the number of samples when the LRT_{ae} statistic was applied so that the same total cost would be incurred as for the runs with the LRT_{std}. The reduced sample size for the LRT_{ae} sample was computed using equation (3.6),

$$N^{DS} = \frac{N\left(\frac{C_p}{C_g} + 1\right)}{1 + \frac{C_p}{C_g} + r\alpha}, \tag{3.6}$$

where N^{DS} is the sample size for the LRT_{ae}; N is the sample size for the LRT_{std}; C_p is the cost of phenotyping; C_g is the cost of genotyping; r is the cost ratio of molecular haplotyping to genotyping (C_{mh}/C_g); and α is the proportion of individuals in the LRT_{ae} sample which have molecular haplotypes determined (double-sampling proportion). We consider the phenotyping costs, C_p, to include costs associated with ascertainment and diagnosis. We illustrate fixed cost sample sizes for the following example. With settings of $C_p/C_g = 25$, $r = 5$, $\alpha = 0.75$, and $N = 1000$ for the LRT_{std} method, then the corresponding total sample size for the LRT_{ae} method, N^{DS}, is 874. The reader should note that the reduced sample size results from the additional cost incurred by double-sampling 75% of the total sample for the LRT_{ae} method. If $C_p/C_g = 1000$, note this term will dominate the expression in equation (3.6) and the fixed cost sample size, N^{DS}, will not differ greatly from the sample size for the LRT_{std}, N. All power simulations were performed under fixed cost conditions. Since the double-sample proportion, α, varies from replicate to replicate when the threshold double-sample selection method is employed, we first performed several test runs to determine the mean double-sample proportion, $\overline{\alpha}$. Using $\overline{\alpha}$, we computed N^{DS}*, the total sample size for the LRT_{ae} determined from the expectation of α. For a specific disease model, we performed a comprehensive study of the power difference between the LRT_{ae} and LRT_{std} for the situation of a haplotype comprised of two SNPs.

Computation for conditional haplotype frequencies. For completeness, here we illustrate how the conditional (on case status) haplotype frequencies were computed as described by De La Vega et al. (2005).

List of notation

Marker Loci:

h_i = population haplotype frequency of the i^{th} haplotype (out of w possible haplotypes)

Disease Locus:

p_d = allele frequency of disease-causing allele at the disease locus

p_+ = allele frequency of the wild-type allele at the disease locus

Disease-Marker Haplotypes:

$h_{+,j}$ = frequency of disease-marker haplotype containing the wild-type allele (+) at the disease locus and the marker haplotype j. This is the probability that the wild-type allele is on the same chromosome as a given haplotype j.

$h_{d,j}$ = frequency of disease-marker haplotype containing the disease allele (d) at the disease locus and the marker haplotype j. This is the probability that the disease allele is on the same chromosome as a given haplotype j.

Disequilibrium Parameters:

D' = standardized LD parameter (Lewontin 1964), ($0 \le D' \le 1$)

$D_{max} = min\left[p_d(1-h_z), p_+ h_z\right]$, where z is the haplotype selected to be LD with the disease allele

Penetrances:

f_0 = Pr{affected|++ at disease locus}

f_1 = Pr{affected|+d at disease locus}

f_2 = Pr{affected|*dd* at disease locus}

Conditional Probabilities:

I_{i,j_1,j_2} = Pr{individual possesses marker haplotypes j_1 and j_2|affection status i}, $1 \le j_1, j_2 \le w$, i = 0 (affected) or 1 (unaffected)

I_{i,j_3} = Pr{individual possesses marker haplotype j_3|affection status i}, $1 \le j_3 \le w$, $i = 0$ (affected) or 1 (unaffected)

Prevalence:

ϕ = disease prevalence = $(1-p_d)^2 f_0 + 2p_d(1-p_d)f_1 + {p_d}^2 f_2$

LD Pattern and Disease-Marker Haplotype Frequencies:

Because the number of linkage disequilibrium parameters increases as the number of haplotypes increases (Lewontin 1964), we simplify the analysis by constructing an LD pattern that, for estimated haplotype frequencies $h_1,\ldots,h_w$, is a function of a single parameter *D'*. Note that *D'* can vary between 0 and 1, where 0 represents linkage equilibrium and 1 represents complete linkage disequilibrium. We now describe the LD pattern for a "selected" haplotype, *z*, where $1 \le z \le w$. By "selected", we mean that haplotype *z* is in positive LD with the disease allele (occurs in phase with the disease allele more often than under linkage equilibrium conditions). The LD pattern is given by:

$$\mathrm{LD}(j) = \begin{cases} D' \times D_{max}, j = z \\ \dfrac{h_j}{1-h_z} \times D' \times D_{max}, 1 \le j \le w, j \ne z \end{cases} \tag{3.7}$$

Using the LD pattern described above, for the "selected" haplotype z, we write the following two equations for the disease-marker haplotype frequencies:

$$\begin{aligned} h_{+,j} &= p_+ h_j - \mathrm{LD}(j), \\ h_{d,j} &= p_d h_j + \mathrm{LD}(j) \end{aligned} \tag{3.8}$$

Since $h_{+,j}$ and $h_{d,j}$ are the probabilities that the j^{th} marker haplotype resides on the same chromosome as the wild-type and disease allele, respectively, $h_{d,j} \geq h_{+,j}$ for $j = z$ (the haplotype "selected" to be LD with the disease allele).

Applying the definition of conditional probability and the law of total probability, we can write the conditional haplotype pair frequencies as

$$I_{i,j_1,j_2} = \begin{cases} \left[f_0 h_{+,j_1} h_{+,j_2} + f_1 \left(h_{+,j_1} h_{d,j_2} + h_{d,j_1} h_{+,j_2} \right) + f_2 h_{d,j_1} h_{d,j_2} \right] / \phi, i = 0 \\ \left[(1-f_0) h_{+,j_1} h_{+,j_2} + (1-f_1) \left(h_{+,j_1} h_{d,j_2} + h_{d,j_1} h_{+,j_2} \right) + (1-f_2) h_{d,j_1} h_{d,j_2} \right] / (1-\phi), i = 1 \end{cases} \quad (3.9)$$

and the conditional haplotype frequencies as

$$I_{i,j_1} = I_{i,j_1,j_1} + 0.5 \sum_{\substack{j_2=1 \\ j_1 \neq j_2}}^{w} I_{i,j_1,j_2} \, . \quad (3.10)$$

We used these conditional (on case status) haplotype frequencies as the generating frequencies in our simulations under the alternative hypothesis (power runs). Using the above equations, we were able to compute these conditional haplotype frequencies from the disease prevalence (ϕ), the disease allele frequency (p_d or DAF), the disease model, the genotype relative risk for the homozygote (R_2), the population haplotype frequencies (h_j), the "selected" haplotype (z), and the LD between the "selected" haplotype, and the disease allele (measured by D'). For example, suppose we set $\phi = 0.025$, $p_d = 0.07$, $R_2 = 3.5$, $h_1 = 0.05$, $h_2 = 0.25$, $h_3 = 0.15$, $h_4 = 0.55$, z = haplotype 1, and D' (between d and the haplotype 1) = 0.9 and use a dominant disease model. Since the dominant disease model requires that $R_2 = R_1$, the genotype relative risk for the heterozygote (R_1) must also be 3.5. Using the definitions for the genotype relative risks ($R_1 = f_1 / f_0$ and $R_2 = f_2 / f_0$) and the definition of the disease prevalence, we find that $f_0 = 0.019$, $f_1 = 0.065$, and $f_2 = 0.065$. Then by using the definition of D_{max} as well as equations (3.7)

and (3.8), we find each disease-marker haplotype frequency ($h_{+,j}$ *and* $h_{d,j}$). Now with the penetrances (f_0, f_1, and f_2) and the disease-marker haplotype frequencies ($h_{+,j}$ *and* $h_{d,j}$), we can use equation 3.9 to find the conditional haplotype pair frequencies (I_{i,j_1,j_2}). Finally, applying equation (3.10) we find the conditional haplotype frequencies, $I_{01} = 0.123,\ I_{11} = 0.048,\ I_{02} = 0.231,\ I_{12} = 0.250,\ I_{03} = 0.139,\ I_{13} = 0.150,\ I_{04} = 0.508,$ and $I_{14} = 0.551$. Note that the frequency of the "selected" haplotype (haplotype 1) is elevated in cases (relative to the population frequency for this haplotype) while the frequency of all other haplotypes is lowered in cases (relative to the populations frequencies for these haplotypes).

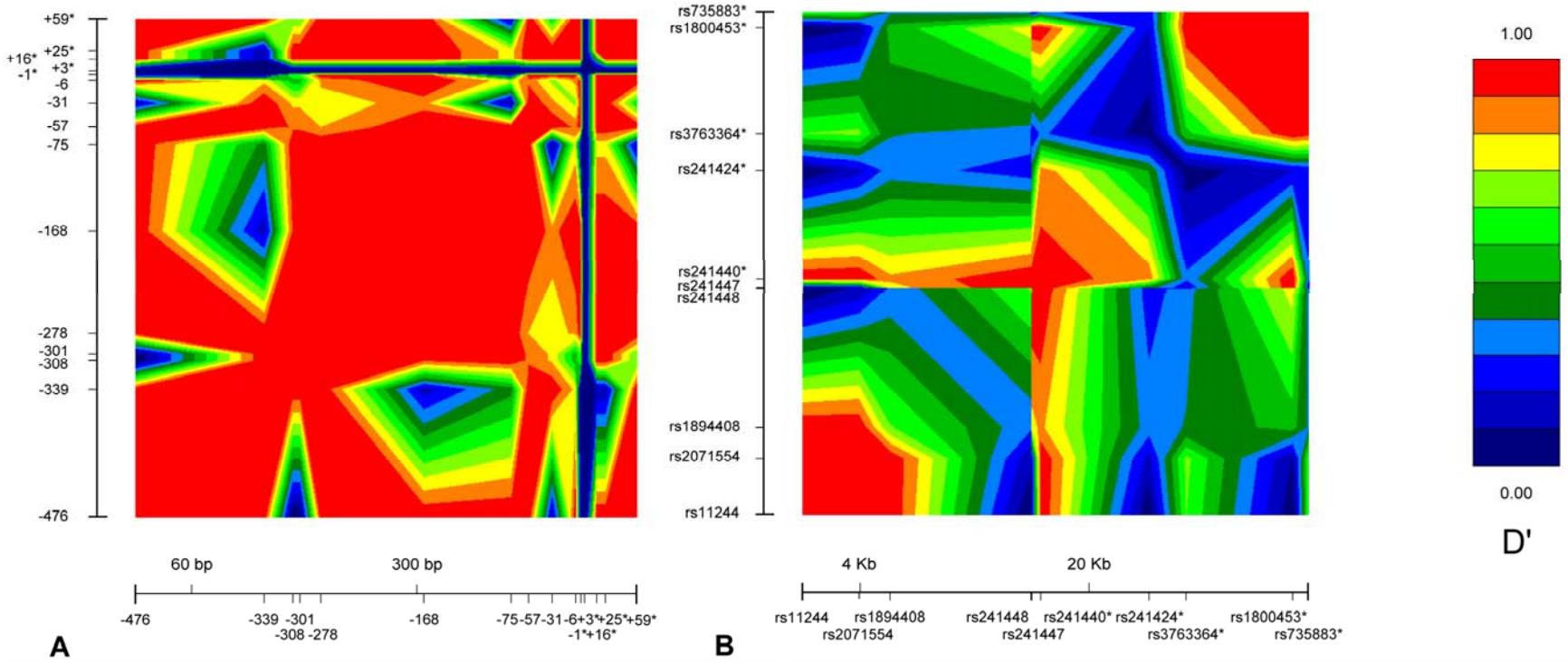
Multi-SNP scenario

Evaluation of false positive rate and power for fixed costs. Through additional simulations, we investigated the behavior of these statistics when applied to haplotypes comprised of larger numbers of SNPs. Because these simulations required additional computational time, we only utilized SNPHAP v 1.3.1 (see Electronic Resource Information) for inferring haplotypes. Our simulations were based on haplotype frequencies from two datasets – 1) a dataset of molecular haplotypes with very high levels of pair-wise LD between markers (Horan et al. 2003) and 2) a dataset of multilocus genotypes from the TAP2 gene within the major histocompatibility complex, a region with low pair-wise LD between markers (International HapMap Consortium 2003; International HapMap Consortium 2005) (see also Electronic Resource Information), hereafter referred to as the Horan and HapMap TAP2 datasets, respectively. Figure 3.2 displays the inter-marker LD for each of these two datasets using GOLD plots (Abecasis

and Cookson 2000). For the Horan dataset, we determined the generating population haplotype frequencies for our simulations directly using the counting method (Ott 1999). For the HapMap TAP2 dataset, we found the generating population haplotype frequencies for our simulations indirectly using SNPHAP v 1.3.1 (see Electronic Resource Information). In the latter case, haplotype frequencies were estimated from the parents of each trio in the Yoruba population group from the International HapMap Project. For the type I error simulation studies, 1000 replicate datasets containing 250 cases and 250 controls were simulated. For the type I error runs based on the Horan data and the HapMap TAP2 data, we simulated haplotypes comprised of 15 SNPs and 10 SNPs, respectively, while for the power runs, we simulated haplotypes comprised of 5 SNPs (Horan et al. 2003; International HapMap Consortium 2003; International HapMap Consortium 2005). Figure 3.2 specifies the SNPs we utilized from each dataset in the type I error and power runs. For the Horan dataset, we provide the SNP markers' positions (relative to the transcription start site of the GH1 gene) while for the HapMap TAP2 dataset, we provide the name of the SNP marker. As a result, we simulated haplotypes using 17 haplotype variants with frequencies greater than 0.01 for both the Horan and HapMap TAP2 type I error simulations. In addition, we simulated haplotypes using 5 and 10 haplotype variants with frequencies greater than $1/(2t)$, where t is the total number of individuals ($t = 153$ for the Horan dataset and $t = 60$ for the HapMap TAP2 dataset), for the Horan and HapMap TAP2 power simulations, respectively. For each scenario, we normalized the frequencies so that they summed to unity. As with the power studies for the two SNP scenario, the selection of the specific haplotype in LD with the disease allele depended on the *DAF*. The rationale for the selection procedure is

provided in section 3.3 addressing multi-SNP power. For multi-marker type I error and power studies, we employed both the random and threshold double-sample selection methods in computing the LRT_{ae} statistic. When the random double-sample selection method was utilized, the double-sample proportion, α, was 0.75. When the threshold double-sample method was utilized, the setting of $\delta = 0.85$ was used, and the maximum proportion of individuals included in the double-sample was 0.75.

Figure 3.2 GOLD plots for the Horan and HAPMAP TAP2 datasets



Legend for Figure 3.2: These GOLD plots (Abecasis and Cookson 2000) show the pair-wise intermarker LD in terms of *D'* for (A) 15 SNP markers within the proximal promoter region of human pituitary expressed growth hormone (GH1) and (B) 10 SNP markers with the TAP2 gene. In (A), the SNP markers are listed as their position relative to the transcription start site of the GH1 gene whereas in (B), the SNP markers are listed by their National Center for Biotechnology Information (NCBI) reference SNP (rs) numbers. Physical distances are provided. All SNP markers displayed were included in the type I error study while SNP markers accompanied by an asterisk (*) were included in the power study.

Identifying the nature of haplotype pair misclassification. For all the simulations performed, we recorded the details of the misclassifications that occurred. Specifically, for every replicate we computed the misclassification rates,

$\theta_{j'j}$ = Pr{observed haplotype pair classification is j| true haplotype classification is j'},

where $j' \neq j$ (Gordon et al. 2004). Previous research studying genotype misclassification rates in tests of genotypic association provides the motivation for ascertaining these values (Kang et al. 2004). This notation is also used in the description of the LRT_{std} and LRT_{ae} statistics.

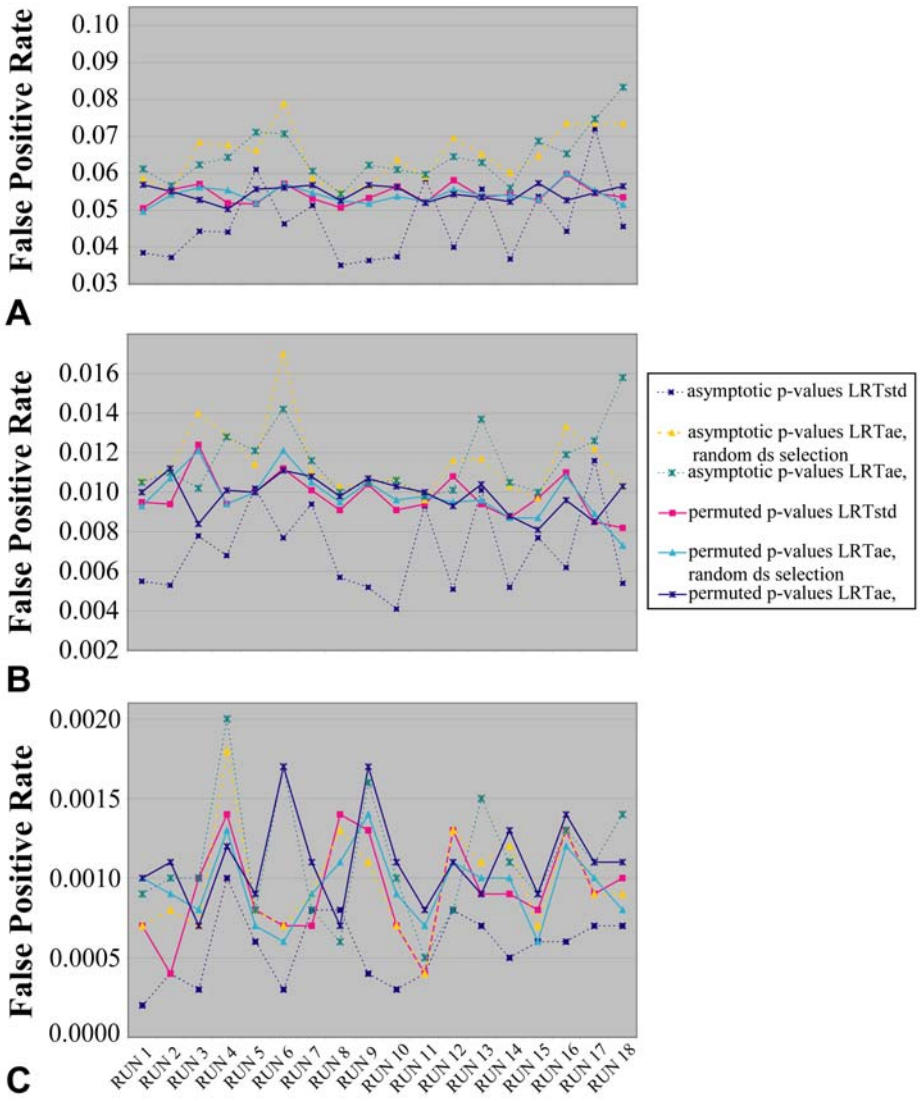
3.3 Results

Two SNP scenario. Our type I error and power results from the simulations utilizing SNPHAP v 1.3.1 and PHASE v 2.1.1 for the haplotype inference were almost identical. Although we present graphs and tables that display the results where SNPHAP v 1.3.1 provided the haplotype inference, the reader should note that similar results were found using PHASE v 2.1.1 for the haplotype inference.

Evaluation of false positive rates for permutation and asymptotic p-values. The type I error simulations demonstrated that the approach for determining statistical significance is critical for maintaining the correct false positive rate. While KS and AD test results indicated that the distribution of permutation *p*-values was consistent with the standard uniform distribution, they indicated that the distribution of asymptotic *p*-values did not resemble the standard uniform distribution. These results were reinforced by the false positive rates we found. For all the simulation runs displayed in Table 3.1,

Figure 3.3 shows the false positive rate for various significance thresholds for LRT_{std} and LRT_{ae} (using the random and threshold double-sample selection methods) association tests in which statistical significance was indicated by permutation and asymptotic *p*-values. The graph in Figure 3.3A shows that asymptotic *p*-values for LRT_{ae} are anti-conservative while those for LRT_{std} fluctuate between conservative and anti-conservative values when a significance threshold of 0.05 is applied. In contrast, the permutation *p*-values for both statistics consistently maintain the nominal significance level of 0.05. We found that the asymptotic and permuted *p*-values demonstrated similar behavior for significance thresholds of 0.01 and 0.001. However, for the 0.001 significance threshold, the *p*-values appear more scattered due to the scale at this extreme significance threshold. Haplotype pairs were inferred using SNPHAP v 1.3.1 for the simulation results displayed in the graph. These results are not surprising since several simulation parameter settings have expected cell counts of less than five counts, violating Cochran's rule (Cochran 1952).

Figure 3.3 Line graph illustrating estimates of the false positive rate at various significance levels for LRT_{std} and LRT_{ae}



Legend for Figure 3.3: The line graphs show estimates of the false positive rate at the (A) 0.05 significance level, (B) 0.01 significance level, and (C) 0.001 significance level for LRT_{std} and LRT_{ae} with *p*-values determined by both permutation and the asymptotic central χ^2 distribution. The 18 runs correspond to the combinations of parameter settings described in Table 3.1. For all 18 runs, LRT_{ae} was computed with the random and threshold double-sample selection methods. When the threshold double-sample method was utilized to compute LRT_{ae}, the setting of $\delta = 0.85$ was used, and the maximum proportion of individuals included in the double-sample was the value for α specified by the fractional factorial design. Haplotype pairs were inferred using SNPHAP v 1.3.1 for the simulation results displayed in the graph.

Evaluation of power for fixed cost. Based on the results for the false positive rates, we conclude that power can only be evaluated using the permutation *p*-values. We compare the power of LRT_{ae} (using the random and threshold double-sample selection methods) to LRT_{std}. Table 3.3 presents summary statistics for the power difference (LRT_{ae} power – LRT_{std} power) at various significance levels for the two cost ratios $C_p / C_g = 25$ and $C_p / C_g = 1000$ using the 8 parameter settings from the factorial design (Table 3.2). Note that in all runs, we set the cost ratio of molecular haplotyping to genotyping, *r*, to be 5, and the proportion of individuals to be double-sampled, α, to be 0.75 (for the random double-sample selection method). The values reported correspond to the simulations utilizing SNPHAP v 1.3.1.

Table 3.3 Summary statistics for power difference ($LRT_{ae} - LRT_{std}$) at various significance levels

		Significance Level = 0.05		Significance Level = 0.01		Significance Level = 0.001	
DS Selection Method	Summary Statistic	$C_p / C_g = 25$	$C_p / C_g = 1000$	$C_p / C_g = 25$	$C_p / C_g = 1000$	$C_p / C_g = 25$	$C_p / C_g = 1000$
random	minimum	-0.061	-0.004	-0.062	0.001	-0.056	0.000
	median	0.004	0.014	0.005	0.019	-0.007	0.021
	maximum	0.036	0.105	0.033	0.089	0.025	0.135
threshold	minimum	-0.010	-0.004	0.001	0.003	-0.001	0.000
	median	0.043	0.045	0.048	0.048	0.064	0.068
	maximum	0.126	0.162	0.117	0.123	0.151	0.152

Legend for Table 3.3: This table presents summary statistics for the power difference between the LRT_{ae} and LRT_{std} methods (p-values evaluated using permutation) at the 0.05, 0.01, and 0.001 significance levels. Results are shown for LRT_{ae} computed using both the random and threshold double-sample selection methods. The methods are compared for fixed costs where the power for LRT_{ae} is computed under two conditions, 1) the cost ratio of phenotyping to genotyping (C_p/C_g) is 25 and 2) the cost ratio of phenotyping to genotyping (C_p/C_g) is 1000. The sample size for LRT_{std}, N, is 1000 (500 cases, 500 controls). For the LRT_{ae} statistic, settings of $\alpha = 0.75$ (random double-sample selection method) and $r = 5$ were used. When the threshold double-sample selection method was utilized to compute LRT_{ae}, the setting of $\delta = 0.85$ was used, and the maximum proportion of individuals included in the double-sample was 0.75. Haplotype pairs were inferred using SNPHAP v 1.3.1.

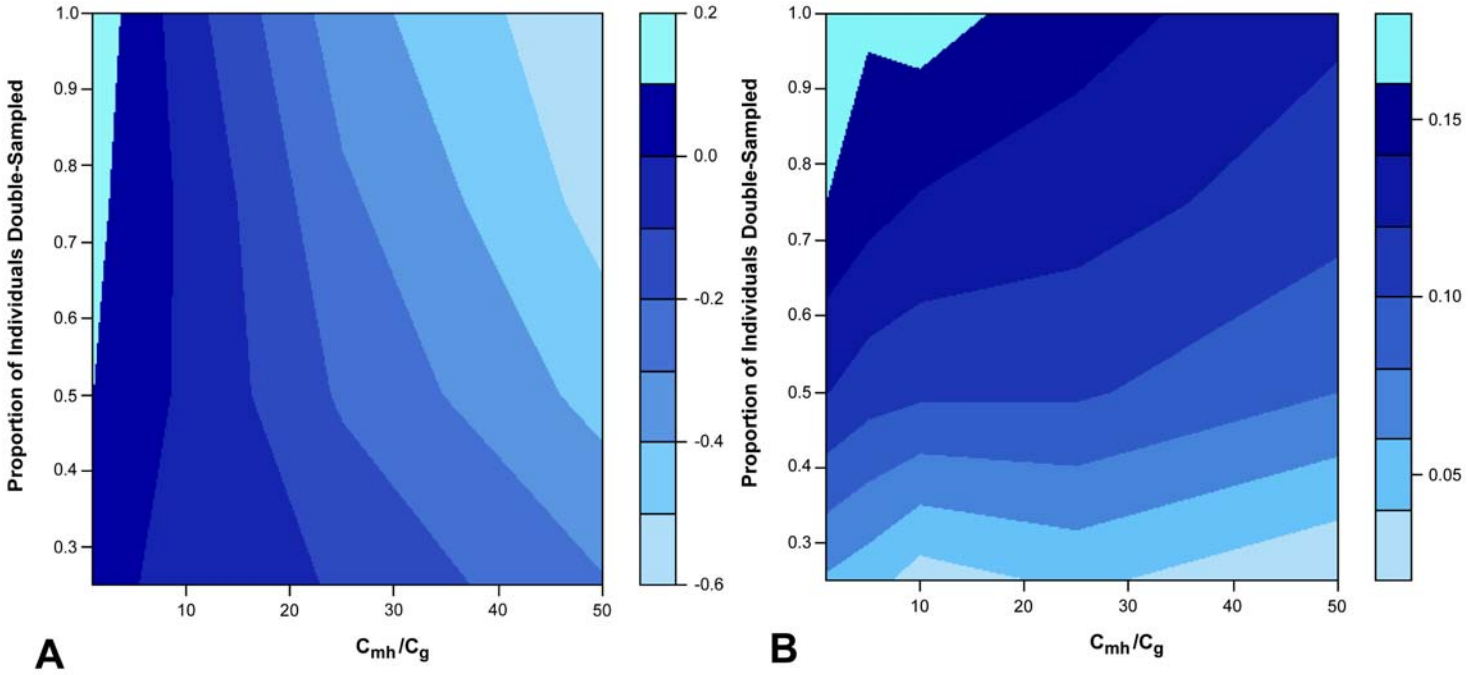
For the random double-sample selection method, the minimum power difference occurred when $C_p / C_g = 25$ for a dominant disease model with $R_2 = 2$ and $DAF = 0.27$ at a significance level of 0.01. For these settings, the LRT_{ae} power was 0.544 and LRT_{std} power was 0.606. The maximum power difference occurred when $C_p / C_g = 1000$ for a dominant disease model with $R_2 = 3.5$ and $DAF = 0.07$ at a significance level of 0.001. For these settings, the LRT_{ae} power was 0.910 and LRT_{std} power was 0.775.

For the threshold double-sample selection method, the minimum power difference occurred when $C_p / C_g = 25$ for a dominant disease model with $R_2 = 2$ and $DAF = 0.27$ at a significance level of 0.05. For these settings, the LRT_{ae} power was 0.821 and LRT_{std} power was 0.831. The maximum power difference occurred when $C_p / C_g = 1000$ for a dominant disease model with $R_2 = 2$ and $DAF = 0.07$ at a significance level of 0.05. For these settings, the LRT_{ae} power was 0.573 and LRT_{std} power was 0.411.

Power difference as a function of double-sample proportion and cost ratio. In the spirit of response surface analysis for factorial design (Box et al. 1978), we performed a more thorough analysis of the parameter settings that provided the maximum power difference when LRT_{ae} was computed with the random double-sample selection method. These parameter settings are a dominant disease model with $R_2 = 3.5$ and $DAF = 0.07$. These settings provided the additional benefit of power results greater than 75% for both the LRT_{ae} and LRT_{std} methods at the 0.05, 0.01, and 0.001 significance levels for both cost ratios of $C_p / C_g = 25$ and $C_p / C_g = 1000$. The analysis involved computation of the LRT_{ae} with the random double-sample selection method. Figure 3.4 displays the two-dimensional contour plots of the power difference between the LRT_{ae} and the LRT_{std} as a

function of *r*, the cost ratio of molecular haplotyping to genotyping, and α, the proportion of individuals double-sampled. These power differences are computed for the fixed parameter settings of $C_p / C_g = 25$ (Figure 3.4A) and $C_p / C_g = 1000$ (Figure 3.4B) at significance level $= 0.001$ for the disease model described immediately above. The values of *r* considered in the contour plots are 1, 5, 10, 25, and 50 while the values of α considered are 0.25, 0.50, 0.75, and 1.0. One should note that $\alpha = 1.0$ indicates all individuals in the study are double-sampled regardless of phase ambiguity. Simulations were performed with 1000 replicates and 10,000 permutations for each combination of parameters, and SNPHAP v 1.3.1 was utilized for the haplotype inference. The sample size for the LRT_{std}, *N*, was 1000 (equal numbers of cases and controls).

Figure 3.4 Contour plots of the power difference between LRT_{ae} and LRT_{std} at a significance level of 0.001 (two SNP scenario)



Legend for Figure 3.4: The contour plots display the power difference between the LRT_{ae} and LRT_{std} at various settings for the cost ratio of molecular haplotyping to genotyping (r) and the proportion of individuals double-sampled (α). Power is compared at the 0.001 significance level. The cost ratio of phenotyping to genotyping for (A) is 25 while the cost ratio of phenotyping to genotyping for (B) is 1000. The two SNP scenario is examined for the parameter settings that provided the maximum power difference for factorial design (Table 3.2) using the random double-sample selection method. Generating haplotype frequencies for cases and controls were based on a dominant disease model with $\phi = 0.025$, $R_2 = 3.5$, and $DAF = 0.07$, as well as, population haplotype frequencies of 0.05, 0.25, 0.15, and 0.55. The haplotype with a frequency of 0.05 was placed in LD ($D' = 0.9$) with the disease allele. LRT_{ae} was only computed with the random double-sample selection method. Haplotype pairs were inferred using SNPHAP v 1.3.1.

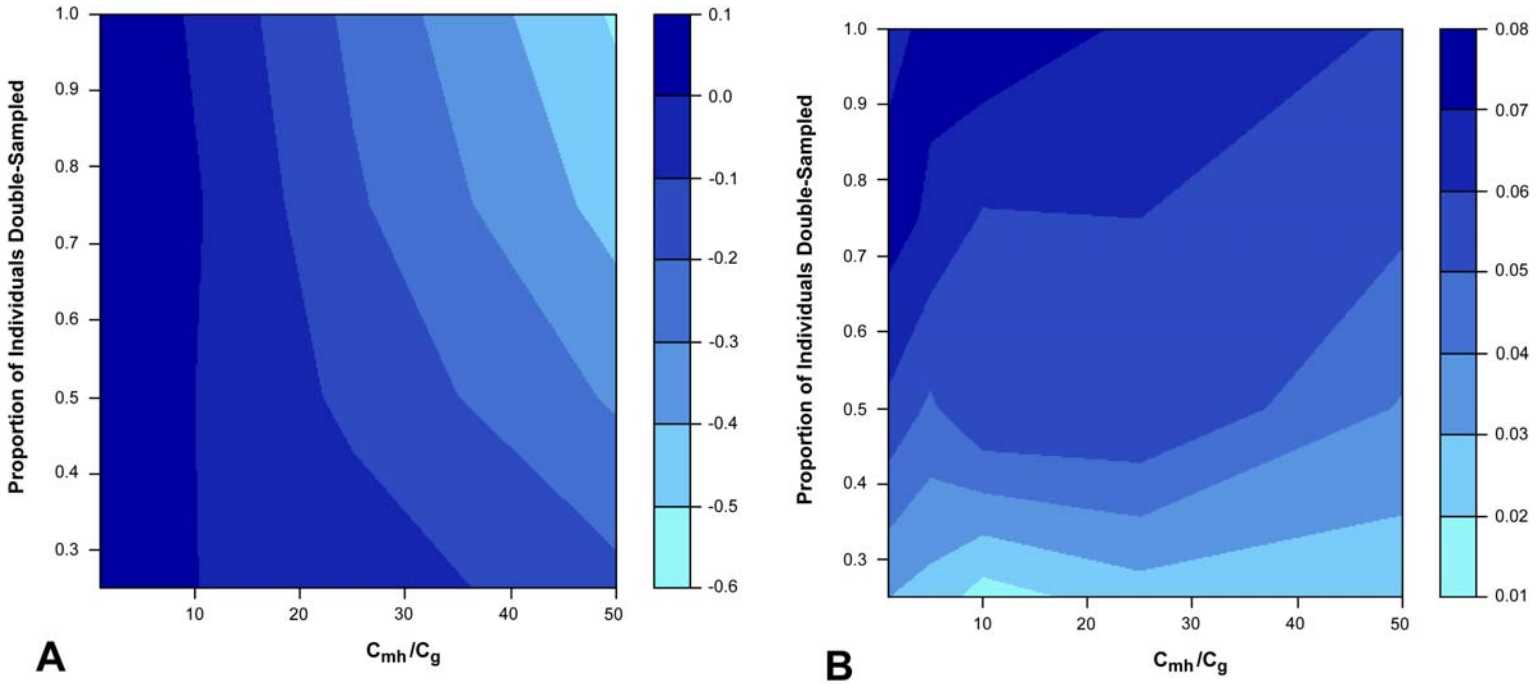
Figure 3.4A shows that the LRT_{ae} provides a power advantage over the LRT_{std} when r is less than 10 and α is greater than 0.5. The maximum power gain is 0.16 and occurs when r and α are 1.0. Conversely, when r is greater than 10, LRT_{ae} is less powerful than LRT_{std} for these parameter settings. The maximum power loss is 0.58 and occurs when r is 50 and α is 1.0. Note that for these values the total sample available for the LRT_{ae} method, N^{DS} (equation (1)), is 342 while the total sample available for the LRT_{std} method, N, is 1000.

Figure 3.4B illustrates that LRT_{ae} is always at least as powerful as LRT_{std} when $C_p / C_g = 1000$. We observe a minimum power gain of 0.02 when r is 50 and α is 0.25 and a maximum power gain of 0.17 when r and α are 1.0. Furthermore, Figure 3.4B indicates that for any cost ratio, r, increasing the double-sampling proportion, α, always increases the power gain with the maximum power gain occurring when $\alpha = 1.0$.

Figures 3.5 and 3.6 display similar contour plots of the power difference between the LRT_{ae} and the LRT_{std} (as a function of r, the cost ratio of molecular haplotyping to genotyping, and α, the proportion of individuals double-sampled) using the same parameters as above at significance levels of 0.01 and 0.05, respectively. Again, these power differences are computed for fixed parameter settings of $C_p / C_g = 25$ (Figures 3.5A and 3.6A) and $C_p / C_g = 1000$ (Figures 3.5B and 3.6B). Figures 3.5 and 3.6 show that the results using significance thresholds of 0.01 and 0.05 are similar to those using a significance threshold of 0.001 (Figure 3.4). When r is less than 10 (same as for Figure 3.4A), Figures 3.5A and 3.6A show that LRT_{ae} provides a power advantage over LRT_{std}. As the significance level increases, the power advantage decreases. Thus,

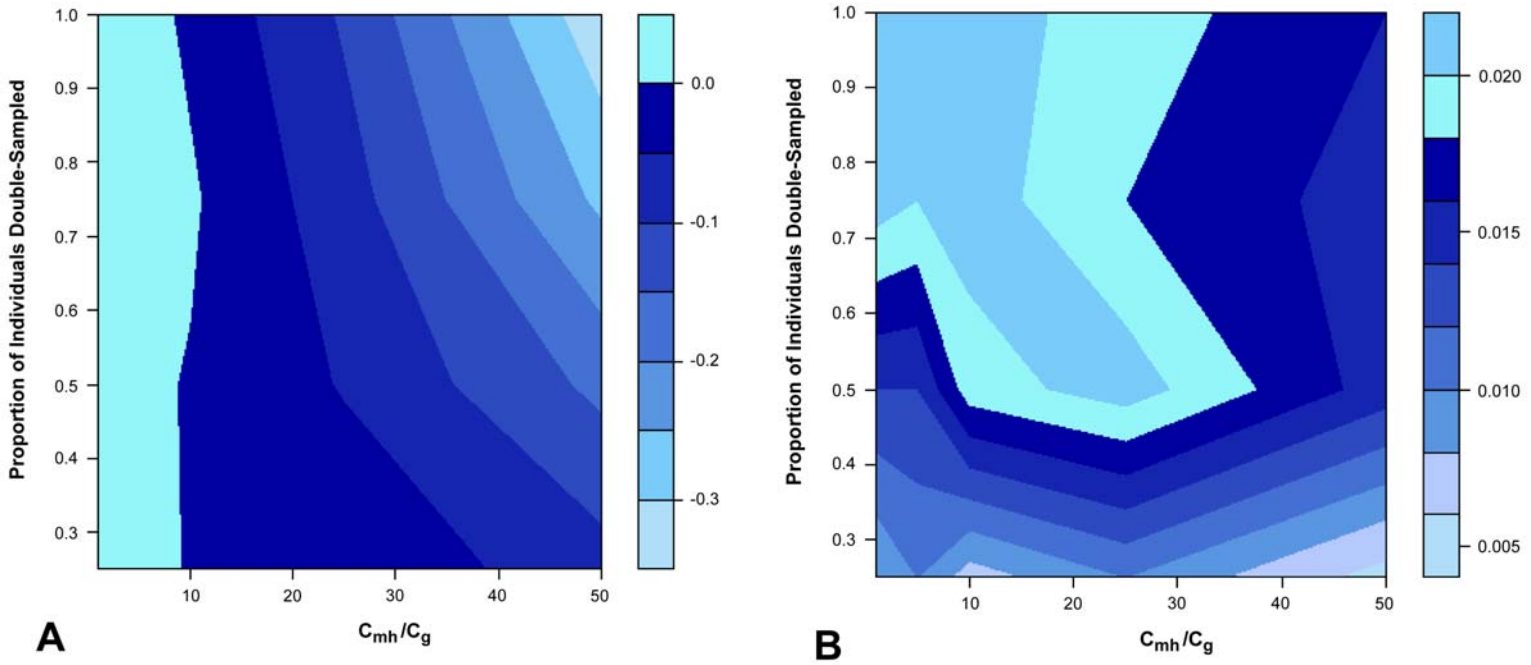
the power advantage is greatest for the 0.001 significance level, less for the 0.01 significance level, and least for the 0.05 significance level. Figures 3.5B and 3.6B illustrate that when $C_p / C_g = 1000$, LRT_{ae} is always more powerful than LRT_{std} at significance levels of 0.01 and 0.05, respectively. Again, the power advantage decreases as the significance level increases.

Figure 3.5 Contour plots of the power difference between LRT_{ae} and LRT_{std} at a significance level of 0.01 (two SNP scenario)



Legend for Figure 3.5: The contour plots display the power difference between LRT_{ae} and LRT_{std} at various settings for the cost ratio of molecular haplotyping to genotyping (r) and the proportion of individuals double-sampled (α). Power is compared at the 0.01 significance level. The cost ratio of phenotyping to genotyping for (A) is 25 while the cost ratio of phenotyping to genotyping for (B) is 1000. The two SNP scenario is examined for the parameter settings that provided the maximum power difference for factorial design (Table 3.2) using the random double-sample selection method. Generating haplotype frequencies for cases and controls were based on a dominant disease model with $\phi = 0.025$, $R_2 = 3.5$, and $DAF = 0.07$, as well as, population haplotype frequencies of 0.05, 0.25, 0.15, and 0.55. The haplotype with a frequency of 0.05 was placed in LD ($D' = 0.9$) with the disease allele. LRT_{ae} was only computed with the random double-sample selection method. Haplotype pairs were inferred using SNPHAP v 1.3.1

Figure 3.6 Contour plots of the power difference between LRT_{ae} and LRT_{std} at a significance level of 0.05 (two SNP scenario)



Legend for Figure 3.6: The contour plots display the power difference between LRT_{ae} and LRT_{std} at various settings for the cost ratio of molecular haplotyping to genotyping (r) and the proportion of individuals double-sampled (α). Power is compared at the 0.05 significance level. The cost ratio of phenotyping to genotyping for (A) is 25 while the cost ratio of phenotyping to genotyping for (B) is 1000. The two SNP scenario is examined for the parameter settings that provided the maximum power difference for factorial design (Table 3.2) using the random double-sample selection method. Generating haplotype frequencies for cases and controls were based on a dominant disease model with $\phi = 0.025$, $R_2 = 3.5$, and $DAF = 0.07$, as well as, population haplotype frequencies of 0.05, 0.25, 0.15, and 0.55. The haplotype with a frequency of 0.05 was placed in LD ($D' = 0.9$) with the disease allele. LRT_{ae} was only computed with the random double-sample selection method. Haplotype pairs were inferred using SNPHAP v 1.3.1.

Multi-SNP scenario

Evaluation of false positive rates for permutation and asymptotic p-values. Table 3.4 displays our estimates of the false positive rate using various significance thresholds (0.05, 0.01, and 0.001) and the results of the KS test for the Horan and HapMap TAP2 dataset-based simulations. Again, only the permuted *p*-values resemble the standard uniform distribution. In addition, the permuted *p*-values maintain the nominal significance level while the asymptotic *p*-values are anti-conservative.

Table 3.4 False positive rate estimates for simulations with generating population haplotype frequencies based on the Horan and HAPMAP TAP2 datasets

Sig. Level	*p*-Value Type	Statistic	DS Selection Method	Horan Dataset			HAPMAP TAP2 Dataset		
				False Positive Rate	95% C.I.	KS *p*-Value	False Positive Rate	95% C.I.	KS *p*-Value
0.05	asymptotic	LRT_{std}	N/A	0.396	(0.366, 0.427)	< 0.001	0.424	(0.393, 0.455)	< 0.001
		LRT_{ae}	random	0.500	(0.469, 0.532)	< 0.001	0.659	(0.629, 0.688)	< 0.001
			threshold	0.490	(0.459, 0.522)	< 0.001	0.632	(0.601, 0.662)	< 0.001
	permuted	LRT_{std}	N/A	0.062	(0.048, 0.079)	0.931	0.041	(0.030, 0.055)	0.770
		LRT_{ae}	random	0.053	(0.040, 0.069)	0.718	0.047	(0.035, 0.062)	0.665
			threshold	0.051	(0.038, 0.067)	0.143	0.048	(0.036, 0.063)	0.267
0.01	asymptotic	LRT_{std}	N/A	0.122	(0.102, 0.144)	< 0.001	0.154	(0.132, 0.178)	< 0.001
		LRT_{ae}	random	0.181	(0.158, 0.206)	< 0.001	0.336	(0.307, 0.366)	< 0.001
			threshold	0.168	(0.145, 0.193)	< 0.001	0.314	(0.285, 0.344)	< 0.001
	permuted	LRT_{std}	N/A	0.014	(0.008, 0.023)	0.931	0.010	(0.005, 0.018)	0.770
		LRT_{ae}	random	0.010	(0.005, 0.018)	0.718	0.008	(0.004, 0.016)	0.665
			threshold	0.004	(0.001, 0.010)	0.143	0.012	(0.006, 0.021)	0.267
0.001	asymptotic	LRT_{std}	N/A	0.020	(0.012, 0.031)	< 0.001	0.013	(0.007, 0.022)	< 0.001
		LRT_{ae}	random	0.033	(0.023, 0.046)	< 0.001	0.062	(0.048, 0.079)	< 0.001
			threshold	0.026	(0.017, 0.038)	< 0.001	0.070	(0.055, 0.088)	< 0.001
	permuted	LRT_{std}	N/A	0.003	(0.001, 0.009)	0.931	0.004	(0.001, 0.010)	0.770
		LRT_{ae}	random	0.003	(0.001, 0.009)	0.718	0.002	(0.000, 0.007)	0.665
			threshold	0.001	(0.000, 0.006)	0.143	0.000	(0.000, 0.003)	0.267

Legend for Table 3.4: This table presents estimates of the false positive rate and the corresponding 95% confidence intervals for the LRT_{std} and LRT_{ae} statistics (asymptotic and permuted *p*-values) for various significance levels. The generating population haplotype frequencies for the simulations were based on the Horan and HAPMAP TAP2 datasets (as described extensively in the Methods section). Simulations for 1000 replicate datasets containing 250 cases and 250 controls were performed. LRT_{ae} was computed with the random and threshold double-sample selection methods. When the random double-sample selection method was utilized, a setting of $\alpha = 0.75$ was used. When the threshold double-sample method was utilized to compute LRT_{ae}, the setting of $\delta = 0.85$ was used, and the maximum proportion of individuals included in the double-sample was 0.75. The table also displays *p*-values for the Kolmogorov-Smirnov Test (KS Test) which tests the null hypothesis that the *p*-values computed for each statistic are drawn from a standard uniform distribution. Haplotype pairs were inferred using SNPHAP v 1.3.1.

Evaluation of power for fixed cost. In our power study for haplotypes comprised of five SNPs, we again utilized the disease model parameter settings that provided the maximum power difference (LRT_{ae} power – LRT_{std} power) for the two SNP factorial design (Table 3.2) with LRT_{ae} computed using random double-sample selection. These parameter settings are a dominant disease model with $R_2 = 3.5$ and $DAF = 0.07$. We based the population haplotype frequencies on the Horan and HapMap TAP2 datasets as described in the Methods section. For each dataset, we selected the haplotype with a frequency closest to 0.05 as the haplotype in LD with the disease allele. By this choice of haplotype, we approximated the frequency of the linked haplotype for the two SNP scenario (see section 3.2) when $DAF = 0.07$. As with two SNP power study, the LD between the disease allele and the linked haplotype was 0.9 (measured by *D'*) (Lewontin 1964). The cost ratio of molecular haplotyping to genotyping (r) was 5. When the random double-sample selection method was utilized to compute LRT_{ae}, the double-sample proportion (α) was 0.75. When the threshold double-sample method was utilized to compute LRT_{ae}, the setting of $\delta = 0.85$ was used, and the maximum proportion of individuals included in the double-sample was 0.75.

For the Horan dataset, the power estimates for LRT_{std} and LRT_{ae} were almost identical at the 0.05, 0.01, and 0.001 significance levels for cost ratios (C_p / C_g) of both 1000 and 25 (results not shown). The high pair-wise intermarker LD present in the Horan dataset causes the haplotype inference to occur with almost complete fidelity. In the absence of misclassification, the LRT_{ae} statistic reduces to LRT_{std}. Therefore, the high degree of similarity in power for these statistics is not surprising.

For the HAPMAP TAP2 dataset, Table 3.5 displays the power estimates and the corresponding 95% confidence intervals for the LRT_{std} and LRT_{ae} methods at the 0.05, 0.01, and 0.001 significance levels assuming fixed costs. When $C_p / C_g = 1000$, LRT_{ae} provides a substantial power benefit over LRT_{std} with the power difference ranging from 6% and 7% at a significance level of 0.05 to 14% and 21% at a significance level of 0.001 for random double-sample selection and threshold double-sample selection, respectively. When $C_p / C_g = 25$, the advantage of LRT_{ae} over LRT_{std} is still substantial for threshold double-sample selection but more modest for random double-sample selection. For the three significance levels under investigation, the power difference ranged from 7% to 22% and 1% to 3.5% for threshold and random double-sample selection, respectively.

Table 3.5 Power estimates for simulations with generating population haplotype frequencies based on the HAPMAP TAP2 datasets

Significance Level	Statistic	DS Selection Method	C_p/C_g	Power	95% C.I.
0.05	LRT_{std}	N/A	N/A	0.858	(0.835, 0.879)
	LRT_{ae}	random	1000	0.919	(0.900, 0.935)
			25	0.868	(0.845, 0.888)
		threshold	1000	0.924	(0.906, 0.940)
			25	0.935	(0.918, 0.950)
0.01	LRT_{std}	N/A	N/A	0.666	(0.636, 0.695)
	LRT_{ae}	random	1000	0.801	(0.775, 0.825)
			25	0.701	(0.672, 0.729)
		threshold	1000	0.804	(0.778, 0.828)
			25	0.817	(0.792, 0.841)
0.001	LRT_{std}	N/A	N/A	0.405	(0.374, 0.436)
	LRT_{ae}	random	1000	0.546	(0.515, 0.577)
			25	0.421	(0.390, 0.452)
		threshold	1000	0.613	(0.582, 0.644)
			25	0.626	(0.595, 0.656)

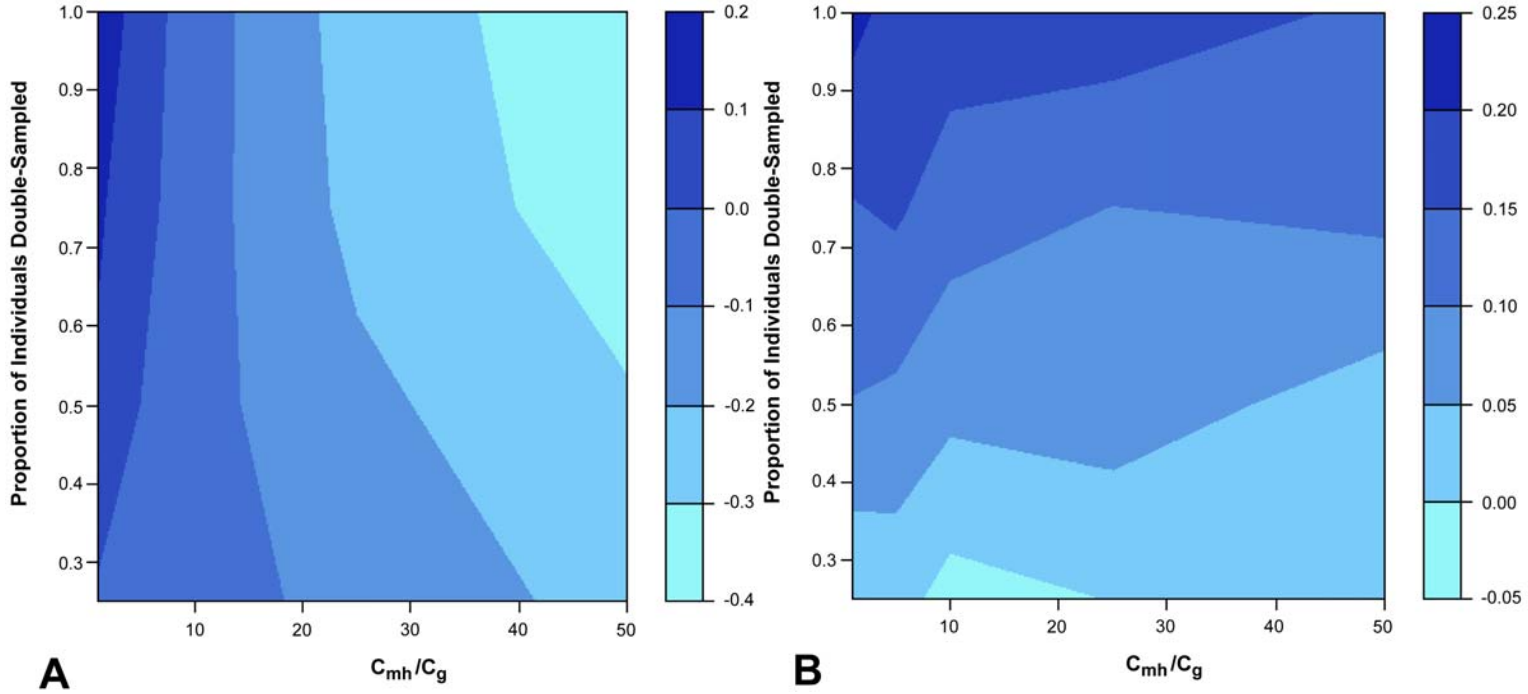
Legend for Table 3.5: This table presents power estimates and the corresponding 95% confidence intervals for LRT_{std} and LRT_{ae} statistics (permuted *p*-values) at various significance levels. The simulations were performed under fixed costs such that the number of samples when LRT_{ae} is applied is reduced according to equation (3.6). The generating population haplotype frequencies for the simulations were based on the HAPMAP TAP2 dataset (as described extensively in the Methods section). The disease model is dominant with $R_2 = 3.5$, disease prevalence = 0.025, $DAF = 0.07$, and D' between the disease allele and the associated haplotype = 0.9. Settings of $\alpha = 0.75$ (random double-sample selection method) and $r = 5$ were used. When the threshold double-sample method was utilized to compute LRT_{ae}, the setting of $\delta = 0.85$ was used, and the maximum proportion of individuals included in the double-sample was 0.75. Haplotype pairs were inferred using SNPHAP v 1.3.1.

We found that the median power gain of LRT_{ae} over LRT_{std} for the threshold double-sample selection method was consistently greater than that for the random double-sample selection method for the runs associated with the factorial design settings displayed in Table 3.2 and the HAPMAP TAP2 power simulations (see Tables 3.3 and 3.5). Furthermore, the power gain for the threshold double-sample selection method occurred for either setting of C_p / C_g. For the threshold double-sample selection method, $\overline{\alpha}$ was small (less than 21%) in our simulations so that our computed N^{DS}* corresponded to 963 individuals.

Power difference as a function of double-sample proportion and cost ratio. As we did for the two SNP scenario, we performed a more thorough analysis to explore the effect of varying the cost ratio of molecular haplotyping to genotyping (r) and the double-sample proportion (α) on the power difference between LRT_{ae} and LRT_{std} for the multi-SNP scenario. Again, we used the parameter settings of a dominant disease model with $R_2 = 3.5$ and $DAF = 0.07$. As before, ϕ was set to 0.025, and the haplotype with a frequency closest to 0.05 was placed in LD ($D' = 0.9$) with the disease allele. The population haplotype frequencies were those from the HAPMAP TAP2 dataset (haplotype comprising 5 SNPs). The analysis involved computation of LRT_{ae} with the random double-sample selection method. Figures 3.7, 3.8, and 3.9 display the two-dimensional contour plots of the power difference between LRT_{ae} and LRT_{std} as a function of r and α at significance levels of 0.001, 0.01, and 0.05, respectively. These power differences are computed for the fixed parameter settings of $C_p / C_g = 25$ (Figures 3.7A, 3.8A, and 3.9A) and $C_p / C_g = 1000$ (Figures 3.7B, 3.8B, and 3.9B). The values of r considered in the contour plots are 1, 5, 10, 25, and 50 while the values of α

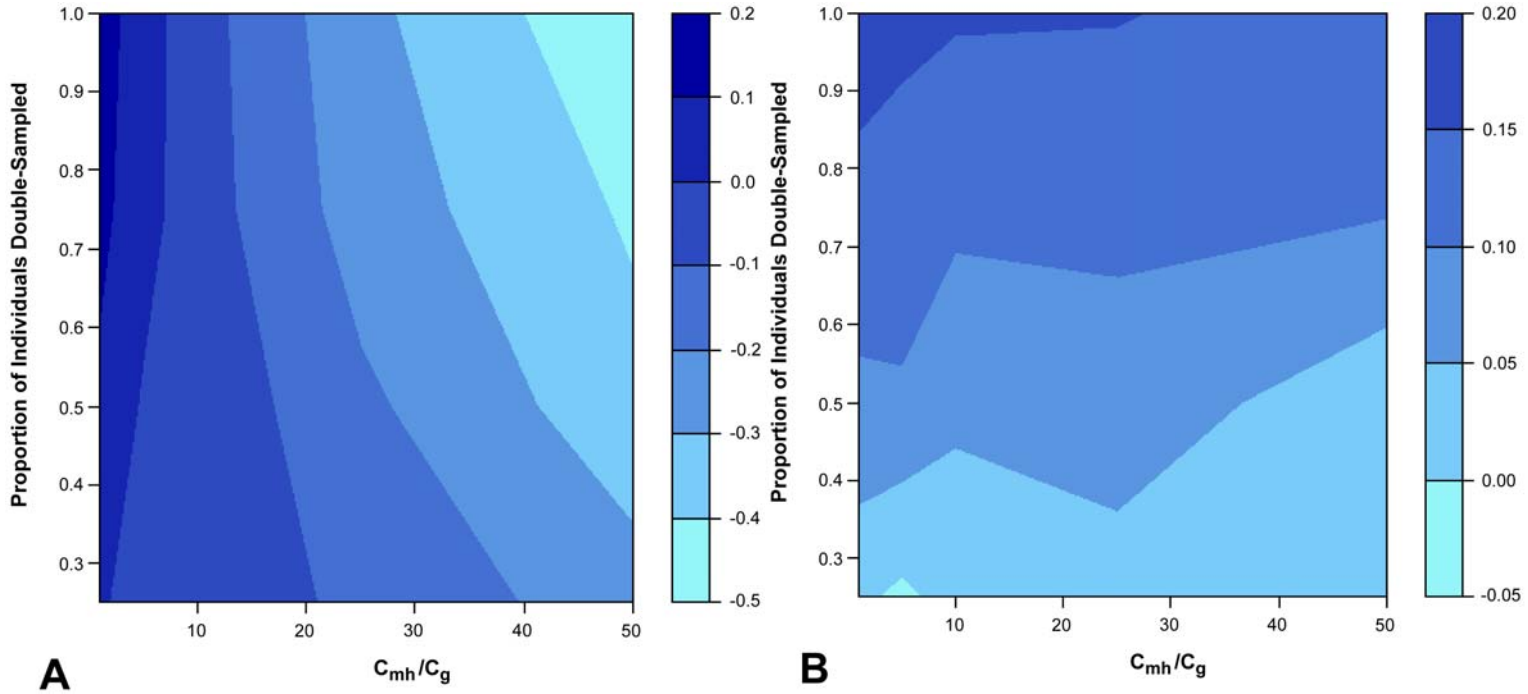
considered are 0.25, 0.50, 0.75, and 1.0. Simulations were performed with 1000 replicates and 1000 permutations for each combination of parameters, and SNPHAP v 1.3.1 was utilized for the haplotype inference. The sample size for LRT_{std}, N, was 1000 (equal numbers of cases and controls).

Figure 3.7 Contour plots of the power difference between LRT_{ae} and LRT_{std} at a significance level of 0.001 (multi-SNP scenario)



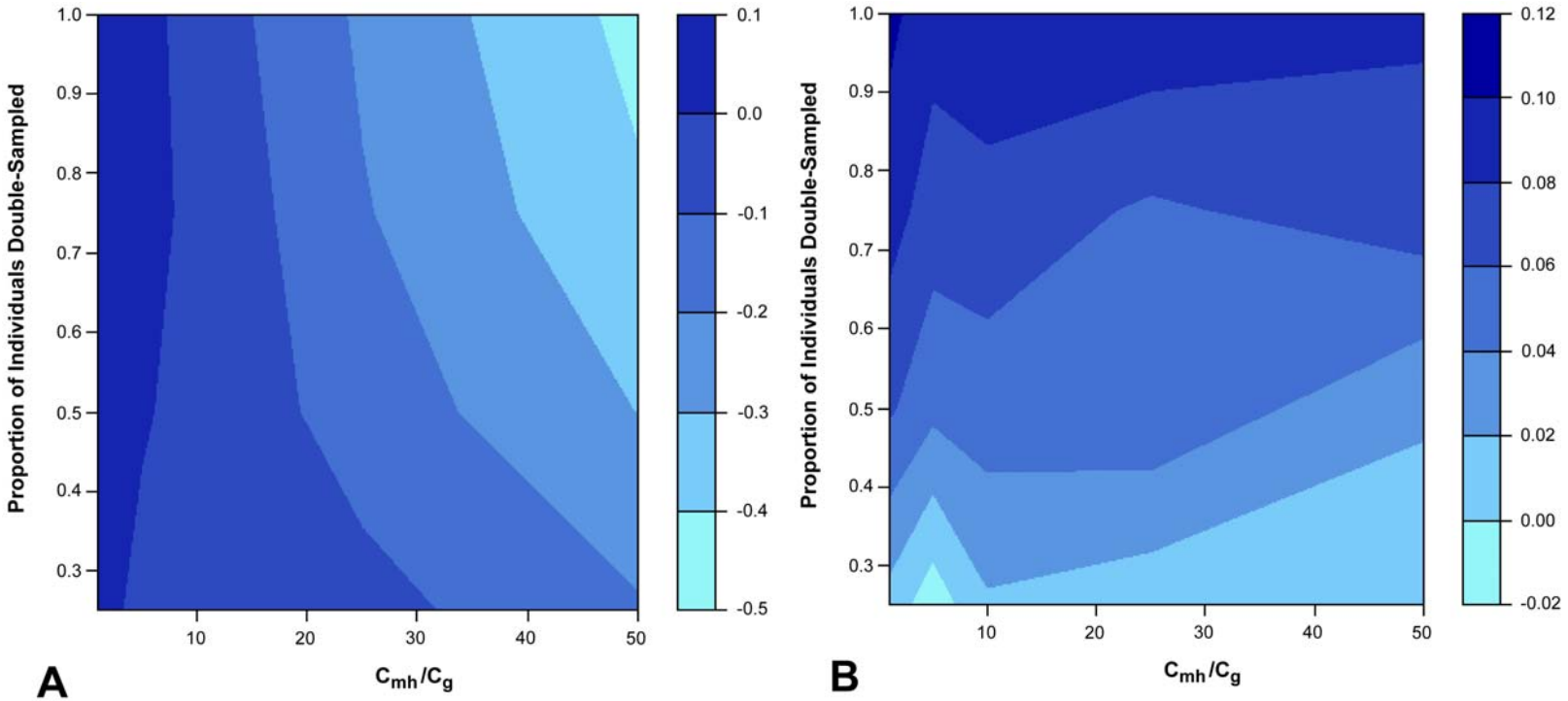
Legend for Figure 3.7: The contour plots display the power difference between LRT_{ae} and LRT_{std} at various settings for the cost ratio of molecular haplotyping to genotyping (r) and the proportion of individuals double-sampled (α). Power is compared at the 0.001 significance level. The cost ratio of phenotyping to genotyping for (A) is 25 while the cost ratio of phenotyping to genotyping for (B) is 1000. Generating haplotype frequencies for cases and controls were based on a dominant disease model with $\phi = 0.025$, $R_2 = 3.5$, and $DAF = 0.07$, as well as, population haplotype frequencies found from the HAPMAP TAP2 dataset (haplotype comprising 5 SNPs). The haplotype with a frequency closest to 0.05 was placed in LD ($D' = 0.9$) with the disease allele. LRT_{ae} was only computed with the random double-sample selection method. Haplotype pairs were inferred using SNPHAP v 1.3.1.

Figure 3.8 Contour plots of the power difference between LRT_{ae} and LRT_{std} at a significance level of 0.01 (multi-SNP scenario)



Legend for Figure 3.8: The contour plots display the power difference between LRT_{ae} and LRT_{std} at various settings for the cost ratio of molecular haplotyping to genotyping (r) and the proportion of individuals double-sampled (α). Power is compared at the 0.01 significance level. The cost ratio of phenotyping to genotyping for (A) is 25 while the cost ratio of phenotyping to genotyping for (B) is 1000. Generating haplotype frequencies for cases and controls were based on a dominant disease model with $\phi = 0.025$, $R_2 = 3.5$, and $DAF = 0.07$, as well as, population haplotype frequencies found from the HAPMAP TAP2 dataset (haplotype comprising 5 SNPs). The haplotype with a frequency closest to 0.05 was placed in LD ($D' = 0.9$) with the disease allele. LRT_{ae} was only computed with the random double-sample selection method. Haplotype pairs were inferred using SNPHAP v 1.3.1.

Figure 3.9 Contour plots of the power difference between LRT_{ae} and LRT_{std} at a significance level of 0.05 (multi-SNP scenario)



Legend for Figure 3.9: The contour plots display the power difference between LRT_{ae} and LRT_{std} at various settings for the cost ratio of molecular haplotyping to genotyping (r) and the proportion of individuals double-sampled (α). Power is compared at the 0.05 significance level. The cost ratio of phenotyping to genotyping for (A) is 25 while the cost ratio of phenotyping to genotyping for (B) is 1000. Generating haplotype frequencies for cases and controls were based on a dominant disease model with $\phi = 0.025$, $R_2 = 3.5$, and $DAF = 0.07$, as well as, population haplotype frequencies found from the HAPMAP TAP2 dataset (haplotype comprising 5 SNPs). The haplotype with a frequency closest to 0.05 was placed in LD ($D' = 0.9$) with the disease allele. LRT_{ae} was only computed with the random double-sample selection method. Haplotype pairs were inferred using SNPHAP v 1.3.

Like the two SNP scenario, Figures 3.7, 3.8, and 3.9 illustrate that for the multi-SNP scenario the power difference between LRT_{ae} and LRT_{std} increases as the significance threshold decreases. Figures 3.7A, 3.8A, and 3.9A show that LRT_{ae} provides a power advantage over LRT_{std} when r is less than 5 and α is greater than 0.5 when significance thresholds of 0.001, 0.01, and 0.05, respectively, are applied. In each case a maximum power gain (0.182 for the 0.001 significance level, 0.153 for the 0.01 significance level, and 0.084 for the 0.05 significance level) occurs when r and α are 1.0. Conversely, when the r is greater than 5, LRT_{ae} is less powerful than LRT_{std} for these parameter settings.

Figures 3.7B, 3.8B, and 3.9B show that LRT_{ae} is almost always at least as powerful as LRT_{std} when $C_p / C_g = 1000$ for the multi-SNP scenario. We observe a slight power loss of 0.02 at the 0.001 significance level when $\alpha = 0.25$ and $r = 10$ and of approximately 0.01 at the 0.01 and 0.05 significance levels when $\alpha = 0.25$ and $r = 5$. The maximum power gain of 0.217 occurs when r and α are 1.0 using a significance threshold of 0.001 (Figure 3.7B). As we observed with the two SNP scenario, Figures 3.7B, 3.8B, and 3.9B indicate that for any cost ratio, r, increasing the double-sampling proportion, α, always increases the power gain with the maximum power gain occurring when $\alpha = 1.0$.

Furthermore, comparing the multi-SNP scenario (Figures 3.7, 3.8, and 3.9) with the two SNP scenario (Figures 3.4, 3.5, and 3.6), we find the same fundamental trends for both $C_p / C_g = 25$ and $C_p / C_g = 1000$. However, the multi-SNP scenario generally displays a larger power advantage for LRT_{ae} over LRT_{std} due to the greater opportunity for misclassification of haplotypes composed of more double heterozygotes.

3.4 Discussion

In practice, few researchers employ molecular haplotyping techniques in genetic case-control studies. The absence of a high-throughput procedure relative to current SNP genotyping technologies is arguably the main reason that this methodology is not more widely used. Another related reason is the cost in terms of both time and money associated with employing this methodology. Our research suggests that the additional costs involved in molecular haplotyping may be worth the effort, especially if the cost of phenotyping is high relative to the cost of genotyping for a study. Ji *et al*. found analogous results for the effects of genotype misclassification on genotypic tests of association (Ji et al. 2005). Other research has shown that molecular haplotypes can greatly increase the power of family-based linkage studies for mapping complex diseases (Gillanders et al. 2006). In practice, the situation where the cost of phenotyping is high relative to the cost of genotyping arises for replication studies. A genome-wide scan involving thousands of SNP markers along with subsequent fine mapping in an initial set of case and control individuals may identify a number of promising regions for follow-up studies. These follow-up or replication studies involve recruiting an independent sample of cases and controls for which only SNPs in the promising regions will be genotyped (Skol et al. 2006). In replication studies for complex traits, the cost ratio of phenotyping to genotyping may be on the order of thousands. For these situations, the LRT_{ae} for testing haplotype association should provide the most utility. It is interesting to note, however, that applying the threshold double-sample selection method provided comparable powers for both high and low phenotyping to genotyping cost ratios. This

finding suggests that this selection strategy may provide additional power for an initial genome-wide association study as well as for a replication study.

One potential limitation of the test statistics that we selected is the increase in degrees of freedom associated with using haplotype pairs rather than individual haplotypes. In general, larger degrees of freedom may result in a loss of power. That is, methods that fully account for uncertainty in the phase assignment process (Schaid et al. 2002; Zaykin et al. 2002; Stram et al. 2003) may be more powerful than LRT_{ae} because the LRT_{ae} method examines haplotype pairs rather than single haplotypes and therefore has more degrees of freedom. We chose these statistics for the following reasons: 1) The most general misclassification model involves modeling errors in haplotype pairs rather than in individual haplotypes (Douglas et al. 2002; Sobel et al. 2002; Gordon et al. 2004). 2) When haplotype pair frequencies deviate from Hardy Weinberg Equilibrium in either case or control sample populations, test statistics that utilize single haplotype frequencies may increase false positive rates and/or lose power (Sasieni 1997; Czika and Weir 2004). 3) In contrast with methods that utilize single haplotype frequencies, the Cochran-Armitage Linear Test of Trend maintains the nominal false positive rate and does not lose power (Cochran 1954; Armitage 1955; Czika and Weir 2004). To our knowledge, a version of this test that incorporates double-sampling procedures to correct for haplotype miscalls does not currently exist.

A point for further research involves identifying the scenarios that produce differential and non-differential haplotype pair misclassification as well as the effects of each kind of misclassification on type I error and power. Under the null hypothesis that haplotype frequency distributions are equal in case and control populations, theoretical

and simulation studies (including the work presented in this chapter) suggest that misclassification is non-differential. Under the alternative hypothesis, it is conceivable that haplotype pair misclassification rates may be different in case and control populations. While recent research (Clayton et al. 2005; Moskvina et al. 2006) indicates that differential misclassification increases the type I error, the effects of differential misclassification on the power of these statistics remain unclear.

While the current perception may be that molecular haplotyping costs are not cost-effective, recent publications suggest that for relatively small regions of the genome accurate molecular haplotyping is no more expensive than performing fluorescent polymerase chain reactions (Proudnikov et al. 2004; Proudnikov et al. 2006). In addition, current techniques are able to provide molecular haplotypes for an entire chromosome at a cost ratio (C_{mh}/C_g) of approximately 5 (C. Ding; personal communication). Finally, as technology improves, the costs associated with molecular haplotyping will likely decrease, and the throughput will likely increase.

CHAPTER 4: ASCERTAINING THE DISTRIBUTION FOR THE LIKELIHOOD RATIO STATISTIC

4.1 Introduction

Although haplotype misclassification can decrease the power for a study, the issue can be avoided by applying an approach that does not infer haplotype pairs for each study participant. An alternative approach is to employ a test statistic that relies on haplotype frequency estimates rather than haplotype calls. Besides the consequences of estimates deviating from their true values, this alternative approach faces complications of its own. In some situations, the exact distribution of the test statistic under both the null and alternative hypotheses can be unclear.

Haplotype-based studies are often hindered by the fact that some haplotypes occur very rarely. The number of possible haplotypes grows exponentially as the number of component SNP loci increases. Consequently, the number of possible haplotypes is often quite large, and many of these possible haplotypes are rare or do not appear at all in the population. Recent studies have found that haplotypes appear in blocks such that there are several common variants while many other variants do not appear at all or are very rare (Daly et al. 2001; Patil et al. 2001; Stephens et al. 2001a; Subrahmanyan et al. 2001; Gabriel et al. 2002; International HapMap Consortium 2003; International HapMap Consortium 2005). As mentioned earlier, several strategies, such as clustering based on similarity (Hoehe et al. 2000), pooling rare haplotypes (Sham and Curtis 1995; Schaid et al. 2002; Zhao et al. 2003), and applying haplotype diversity criteria for SNP selection (Johnson et al. 2001; Jannot et al. 2004)

(http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/htsnp.pdf), have been utilized to reduce the number of haplotype categories and potentially to gain power. However, rare or non-existent haplotypes can have other effects on an analysis besides a reduction in power.

Since multilocus genotypes lack phase information, the testing situation for haplotype-based association studies is more complex than that for other genetic association studies where the variants under investigation are directly observed. In tests of haplotype-based association where haplotype frequencies are estimated from multilocus genotypes, estimation procedures may find a small frequency for some haplotypic variants. There is uncertainty whether haplotypes with small frequency estimates are present but rare in the sample or not present in the sample at all but merely compatible with the multilocus genotypes observed. The effect of this situation on the distribution of the resulting test statistic under both null and alternative hypotheses remains unclear. One still expects that the test statistic will follow a central χ^2 distribution under H_0 and a noncentral χ^2 distribution under H_1. However, the degrees of freedom associated with either χ^2 distribution are no longer well defined.

In this work, we investigate the distribution of a test statistic which relies on haplotype frequency estimates to detect an association between a haplotype and disease status. In particular, we are interested in the distribution of this statistic when some haplotypic variants are extremely rare or nonexistent. Furthermore, we apply a rule to predict the distribution of the statistic and evaluate the accuracy of its prediction.

4.2 Methods

Test statistic. We considered a likelihood ratio statistic for detecting haplotype-based association with disease. The null hypothesis we test is $H_0 : h_{0j} = h_{1j} = h_{*j}$ for all haplotypes j while the alternative hypothesis is $H_1 : h_{0j} \neq h_{1j}$ for at least one j where h_{0j}, h_{1j}, and h_{*j} are the haplotype frequencies for cases, controls, and the entire population, respectively, for the j^{th} haplotype. The statistic is computed using the equation

$$LRT = 2 \times ln\left[\frac{L_{H_1}}{L_{H_0}}\right] \tag{4.1}$$

where L_{H_1} and L_{H_0} are the likelihood of the data under the alternative and null hypotheses, respectively. Each likelihood (L) can be expressed as

$$L = \prod_{i=1}^{N}\left(\sum_{(i_1,i_2)\in H_i} h_{i_1} h_{i_2}\right) \tag{4.2}$$

where N is the number of individuals genotyped, H_i is the set of haplotype pairs compatible with the i^{th} multilocus genotype, and h_{i_1} and h_{i_2} are haplotype frequencies for a haplotype pair consistent with the i^{th} multilocus genotype. Expressing these likelihoods in terms of the haplotype frequencies (equation (4.2)), we have a missing data problem (since we do not observe phase directly). However, we can overcome this hurdle by applying the EM algorithm (Dempster et al. 1977) to find these likelihoods and estimates of the haplotype frequencies. We implement this strategy by employing the software package EHP (see Electronic Resource Information). This software was developed to compute haplotype frequency estimates for datasets where the DNA samples from several individuals have been pooled together. However, since our analysis did not

require any pooling of samples, we set our pool size to one. Equation (4.1) can be rewritten as:

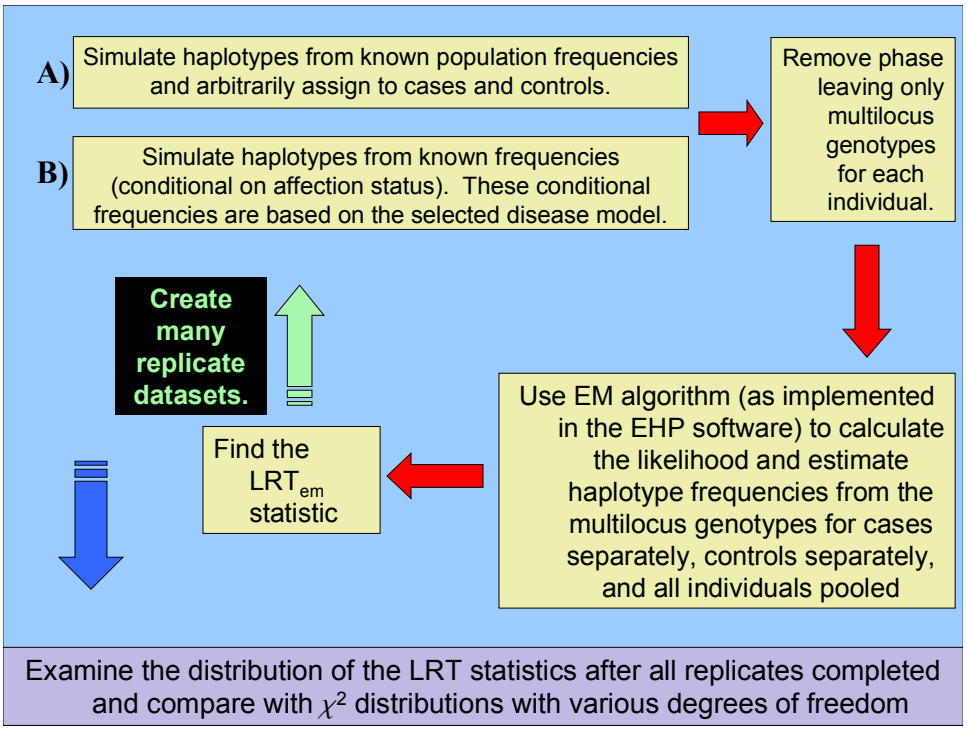
$$LRT = 2 \times \left[ln\left[L_0\right] + ln\left[L_1\right] - ln\left[L_*\right] \right] \tag{4.3}$$

where L_0, L_1, and L_* are the likelihoods computed from the multilocus genotypes from cases alone, controls alone, and all samples, respectively. The reader should note that L_* represents the likelihood under H_0 since the haplotype frequencies must be equal for cases and controls. In addition, the product of L_0 and L_1 represents the likelihood under H_1 since the haplotype frequencies for cases and controls are unconstrained. Applying the EM algorithm (Dempster et al. 1977) as implemented in EHP, we computed L_0, L_1, and L_* and utilized equation (4.3) to find the LRT statistic. To differentiate this LRT statistic from LRT_{ae} and LRT_{std} described in the previous chapter, we will refer to it as LRT_{em} from this point forward.

Description of data generation and analysis. To investigate the distribution of the *LRT* statistic for a variety of situations, we applied LRT_{em} to many simulated datasets. Figure 4.1 illustrates the procedure we used to simulate the data and compute the LRT_{em} statistic under the null and alternative hypotheses. For each replicate dataset, we simulated haplotype pairs for each individual from known frequencies (population haplotype frequencies under H_0 and conditional haplotype frequencies under H_1). Next we removed the phase information for each individual. We utilized the remaining multilocus genotypes to compute the LRT_{em} statistic using the EHP software as described above. After all replicate datasets had been simulated, we examined the distribution of the resulting LRT_{em} statistics. In order to find the distribution from which the LRT_{em}

statistics derive, we performed a goodness-of-fit test, the Kolmogorov-Smirnov (KS) test, for χ^2 distributions with various degrees of freedom. Since all KS tests in this chapter are for χ^2 distributions, we use the notation $KS_{v,j}$ to indicate a KS test for a χ^2 distribution with noncentral parameter, v, and degrees of freedom, j. In addition, we visually compared distributional plots of the LRT_{em} statistic with several χ^2 distributions with various degrees of freedom.

Figure 4.1 Schematic flow chart illustrating the procedure for data simulation and analysis



Legend for Figure 4.1: This schematic flow chart illustrates the procedure employed to create a distribution of LRT_{em} statistics under (A) the null hypothesis and (B) the alternative hypothesis by way of data simulation.

Two SNP scenario

Examination of the distributional properties of LRT_{em} under the null hypothesis. As in chapter 3, for the simplest non-trivial case, the scenario where the haplotype under evaluation includes two SNPs, we applied a factorial design (Box et al. 1978). Here we utilized the factorial design to perform a comprehensive study of the distributional properties of LRT_{em}. Table 4.1 contains the factorial design settings for the scenario involving two SNP markers. We consider a 2^g factorial design, where $g = 4$. We reduced the number of experimental runs from 16 to 12 due to redundancy. For example, a run with a haplotype comprised of a SNP with minor allele frequency 0.5 at the locus 1 and a SNP with minor allele frequency 0.01 at locus 2 is equivalent to a run with a haplotype comprised of a SNP with minor allele frequency 0.01 at locus 1 and a SNP with minor allele frequency 0.5 at locus 2 (with all other factors having equal settings to those for the first run). Since minor allele frequencies of 0.05 and 0.01 are commonly used thresholds for SNP selection in association studies, we chose 0.01 (the more extreme threshold) as the low setting for the minor allele frequency at loci 1 and 2. Within each replicate dataset, the number of cases and controls were equal. During each run, 10,000 replicate datasets were simulated.

Table 4.1 Factorial design parameter settings assuming the haplotype under investigation contains two SNP markers

Description of parameter	Low	High
Number of subjects (equal cases and controls)	500	2000
Minor allele frequency at locus 1 (MAF_1)	0.01	0.5
Minor allele frequency at locus 2 (MAF_2)	0.01	0.5
LD between locus 1 and 2 (measured by D')	0	0.9

Legend for Table 4.1: This table presents the settings for all parameters considered in the simulations to study the distribution of LRT_{em} under H_0 and H_1 assuming the haplotype under investigation contains two SNP markers. We consider a 2^g factorial design, where $g = 4$. The number of experimental runs was reduced from 16 to 12 due to redundancy. D' is the standardized linkage disequilibrium measure. The simulations included 10,000 replicates, and EHP was used to estimate haplotype frequencies and calculate likelihoods for LRT_{em}.

Examination of the distributional properties of LRT_{em} under the alternative hypothesis. We also examined the distribution of the LRT_{em} statistics under the hypothesis that a disease allele at an unobserved locus exists in linkage disequilibrium (LD) with the haplotype under study. Table 4.1 contains the factorial design settings for the study of the distribution of LRT_{em} under H_1. As for the study under H_0, the allele frequencies at each marker locus and the LD between marker loci were used to determine the population haplotype frequencies. For the study under H_0, these haplotype frequencies were used directly to simulate haplotypes. However, for the study under H_1, we used the population haplotype frequencies to compute the conditional (on case status) haplotype frequencies. These conditional haplotype frequencies were then used, in turn, to simulate haplotypes for case and control individuals. Conditional haplotype frequencies were found from population haplotype frequencies and specified disease model parameters by a method described by Sham and subsequently by De La Vega *et al.*

(Sham 1998; De La Vega et al. 2005) (also see the Power for Association with Error (PAWE) website at http://linkage.rockefeller.edu/derek/pawe1.html). However, we selected a specific haplotype to be in LD with the disease allele. For completeness, details regarding the conditional haplotype frequencies including notation and computation as described by De La Vega *et al*. (2005) are provided in subsection 3.2 of chapter 3. For all runs under H_1, the generating haplotype frequencies for cases and controls were based on a dominant disease model ($R_2 = R_1$) with $\phi = 0.025$, $R_2 = 3.5$, and $DAF = 0.07$. In addition, the marker haplotype with a frequency closest to 0.05 was placed in LD ($D' = 0.9$) with the disease allele. (In the previous chapter, we utilized these disease parameter settings for the in-depth power analysis for both the two SNP and multi-SNP scenarios.) Subsection 3.2 also provides an example of how the conditional frequencies are computed. As for the study under H_0, we reduced the number of experimental runs from 16 to 12 due to redundancy. Again, the number of cases and controls were equal within each replicate dataset, and 10,000 replicate datasets were simulated during each run.

Multi-SNP scenario

Examination of the distributional properties of LRT_{em} under the null and alternative hypotheses. We performed additional simulations to investigate the distributional properties of LRT_{em} when applied to haplotypes comprised of larger numbers of SNPs. Table 4.2 contains the factorial design settings for the study of the distribution of LRT_{em} under H_0 and H_1 when the haplotype under investigation contains many SNPs. Our simulations were based on haplotype frequencies from two datasets—

the Horan (Horan et al. 2003) and HapMap TAP2 (International HapMap Consortium 2003; International HapMap Consortium 2005) datasets. The datasets are described in subsection 3.2 of chapter 3 along with an explanation of how the generating population haplotype frequencies were attained from each dataset. Also in subsection 3.2, Figure 3.2 displays the inter-marker LD present in each dataset. For the experimental runs (both under H_0 and H_1) based on these datasets, we simulated haplotypes comprised of five and ten SNPs. In Figure 3.2, the five SNP markers comprising the five-SNP haplotype are indicated with an asterisk (*) for both Horan and HAPMAP TAP2 datasets. For the experimental runs with the ten-SNP haplotype, we used the last ten SNP markers appearing in Figure 3.2A for the Horan dataset, and all ten SNP markers appearing in Figure 3.2B for the HAPMAP TAP2 dataset. The number of cases and controls were equal within each replicate dataset, and 1000 replicate datasets were simulated during each run.

Table 4.2 Factorial design parameter settings assuming the haplotype under investigation contains many SNP markers

Description of parameter	Low	High
Inter-marker LD	HAPMAP TAP2	Horan
Number of SNPs comprising haplotype	5	10
Number of subjects (equal cases and controls)	500	2000

Legend for Table 4.2: This table presents the settings for all parameters considered in the simulations to study the distribution of LRT_{em} under H_0 and H_1 assuming the haplotype under investigation contains many (more than two) SNP markers. We consider a 2^g factorial design, where $g = 3$. Simulations were based on population haplotype frequencies from a dataset with low inter-marker LD, the Horan dataset (Horan et al. 2003), and on population haplotype frequencies from a dataset with high inter-marker LD, the HAPMAP TAP2 dataset (International HapMap Consortium 2003; International HapMap Consortium 2005). The simulations included 1000 replicates, and EHP was used to estimate haplotype frequencies and calculate likelihoods for LRT_{em}.

Predicting the degrees of freedom. One goal of this work is to establish a "rule of thumb" for predicting the degrees of freedom for the χ^2 distribution which most closely resembles the distribution of LRT_{em} for a set of simulation parameters. The rule that we test is that

$$df = \sum_{j=1}^{J} x_{0j} + \sum_{j=1}^{J} x_{1j} - \sum_{j=1}^{J} x_{*j} - 1 \tag{4.4}$$

where

$$x_{*j} = \begin{cases} 1 & \text{if } \hat{h}_{*j} \geq 1/2t \\ 0 & \text{if } \hat{h}_{*j} < 1/2t \end{cases}, \; x_{0j} = \begin{cases} 1 & \text{if } \hat{h}_{0j} \geq 1/2t_0 \\ 0 & \text{if } \hat{h}_{0j} < 1/2t_0 \end{cases}, \text{ and } \; x_{1j} = \begin{cases} 1 & \text{if } \hat{h}_{1j} \geq 1/2t_1 \\ 0 & \text{if } \hat{h}_{1j} < 1/2t_1 \end{cases}.$$

In equation (4.4), *df* is the predicted number of degrees of freedom for the χ^2 distribution; J is the total number of possible haplotypes; $\hat{h}_{0j}$, $\hat{h}_{1j}$, and $\hat{h}_{*j}$ are frequency estimates for

the j^{th} haplotype using cases alone, controls alone, and all samples, respectively; and t_0, t_1, and t are the number of cases, the number of controls, and the total number of samples, respectively.

According to statistical theory, for large sample sizes the LRT_{em} statistic asymptotically follows a central χ^2 distribution under H_0 and a noncentral χ^2 distribution under H_1 (Mitra 1958; Hogg and Craig 1995; Agresti 1996). The number of degrees of freedom associated with either χ^2 distribution equals the difference between the number of free parameters estimated under H_1 and H_0 in equation (4.1). For LRT_{em} in the context of haplotype-based association, this quantity can be expressed as

$$df = \eta_0 + \eta_1 - \eta_* - 1 \tag{4.5}$$

where η_0, η_1, and η_* are the number of haplotypes estimated using cases alone, controls alone, and all samples, respectively. The rule described above in equation (4.4) examines how η_0, η_1, and η_* should be found. Suppose we estimate haplotype frequencies from multilocus genotypes from *t* individuals. A single individual possessing one copy of the variant represents the minimum frequency of a haplotypic variant present in this sample. Thus, the rule described in equation (4.4) applies this minimum frequency ($1/2t$) as a threshold to distinguish haplotypes present in the sample from those that are not present.

To test the performance of this rule, we computed the average values for x_{*j}, x_{0j}, and x_{1j} over all replicate datasets for cases alone, controls alone, and all samples together and computed the predicted degrees of freedom using equation (4.4). We rounded the value computed for *df* and plotted the χ^2 distribution with *df* degrees of freedom (along with χ^2 distributions with $df - 1$ and $df + 1$) for comparison with the distribution LRT_{em}.

The software package R (see Electronic Resource Information) was used to create these plots. The noncentrality parameter (*ncp*) for the "predicted" χ^2 distribution was computed using

$$ncp = 2 \times t_1 \sum_{j=1}^{J} \frac{\left(h_{0j} - h_{1j}\right)^2}{\left(h_{0j} + \frac{t_1}{t_0} h_{1j}\right)} \tag{4.6}$$

as described by others (see http://linkage.rockefeller.edu/derek/pawe2.html) (Mitra 1958; Sham 1998; Gordon et al. 2002). Under the null hypothesis, $ncp = 0$ since h_{0j} and h_{1j} are equal for each haplotype j.

4.3 Results

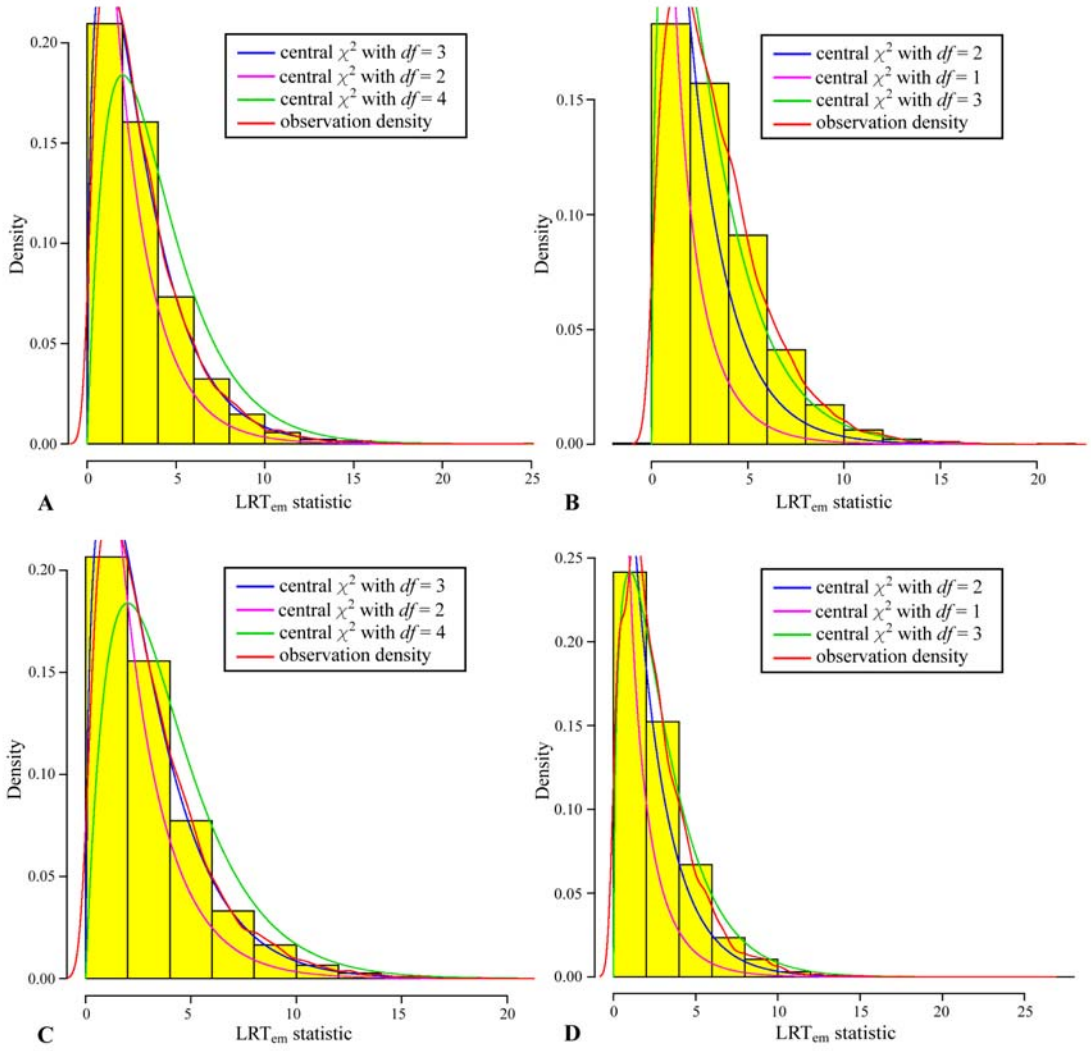
At the end of section 4.3, Table 4.3 summarizes the results for all experimental runs (both two SNP and multi-SNP scenarios under H_0 and H_1) presented in this chapter.

Two SNP scenario.

Examination of the distributional properties of LRT_{em} under the null hypothesis. Our simulation results under the null hypothesis can be classified into three categories—1) experimental runs where the rule described in equation (4.4) successfully predicts the correct distribution; 2) experimental runs where the rule described in equation (4.4) successfully predicts the correct distribution for larger sample sizes only; and 3) experimental runs where the rule described in equation (4.4) fails to predict the correct distribution regardless of sample size. Figure 4.2 displays the distribution of LRT_{em} for simulation runs that represent each of these categories. In our factorial design (Table 4.1), some experimental runs contain no rare haplotypes in the generating

haplotype frequencies. One example is the run in which $MAF_1 = 0.5$, $MAF_2 = 0.5$, and the LD between locus 1 and 2 (measured by D') is 0. These parameter settings result in four haplotypes with equal frequencies (0.25). These frequencies serve as the generating frequencies for the simulation. Figure 4.2A displays a histogram and density line for the LRT_{em} statistic computed from simulations utilizing these parameter settings under H_0 for 500 samples (equal numbers of cases and controls). Figure 4.2A shows that the distribution of LRT_{em} for this experimental run closely resembles a central χ^2 distribution with 3 degrees of freedom, the distribution predicted by the rule in equation (4.4). Since all the generating haplotype frequencies are large, we expected LRT_{em} for this run to exhibit this behavior. For this run, the $KS_{0,3}$ test (testing a central χ^2 distribution with $df = 3$) p-value = 0.248 indicating that the distribution of LRT_{em} is consistent with a central χ^2 distribution with 3 degrees of freedom. The experimental run with the same parameter settings and a sample size of 2000 showed similar results (results not shown).

Figure 4.2 Histograms displaying the distribution of LRT_{em} under H_0 for the two SNP scenario



Legend for Figure 4.2: The histograms display the distribution of LRT_{em} along with the density lines for several central χ^2 distributions for a number of experimental runs. The distribution of LRT_{em} was created by simulating haplotypes comprised of two SNPs under H_0. For (A), $MAF_1 = 0.5$ and $MAF_2 = 0.5$; for (B and C), $MAF_1 = 0.5$ and $MAF_2 = 0.01$; and for (D) $MAF_1 = 0.01$ and $MAF_2 = 0.01$. For all runs displayed, LD between SNP 1 and 2 = 0 (measured by D'). 10,000 replicate datasets comprised of 500 samples (A and B) and 2000 samples (C and D) were simulated. The graphs were scaled to the observed data, and density lines off the scale were truncated.

Other experimental runs required larger sample sizes for the rule described in equation (4.4) to predict the correct distribution. The experimental run in which $MAF_1 = 0.5$, $MAF_2 = 0.01$, and LD between locus 1 and 2 = 0 (measured by *D'*) exhibited this behavior. This run had a minimum generating haplotype frequency (0.005) that was substantially smaller than the minimum generating haplotype frequency for the run described above yet still greater than any of the thresholds established by the rule described in equation (4.4). Figures 4.2B and 4.2C display histograms for LRT_{em} computed from simulations utilizing these parameter settings. Figure 4.2B shows the distribution for simulated datasets containing 500 samples while Figure 4.2C shows the distribution for simulated datasets containing 2000 samples. In Figure 4.2B, the distribution of LRT_{em} does not resemble a central χ^2 distribution with 2 degrees of freedom, the distribution predicted by the rule in equation (4.4) for this run. Instead, it roughly resembles a central χ^2 distribution with 3 degrees of freedom. Figure 4.2C shows that increasing the sample size leads to a better fit with a central χ^2 distribution with 3 degrees of freedom, the distribution predicted with the increased sample size using the rule described in equation (4.4). Although the distribution of LRT_{em} visually matches the density plot for the central χ^2 distribution with 3 degrees of freedom in Figure 4.2C, the $KS_{0,3}$ *p*-values for both the 500 and 2000 sample size runs are approximately 0. Thus, the distribution of LRT_{em} for the 2000 sample run still deviates from a central χ^2 distribution with 3 degrees of freedom.

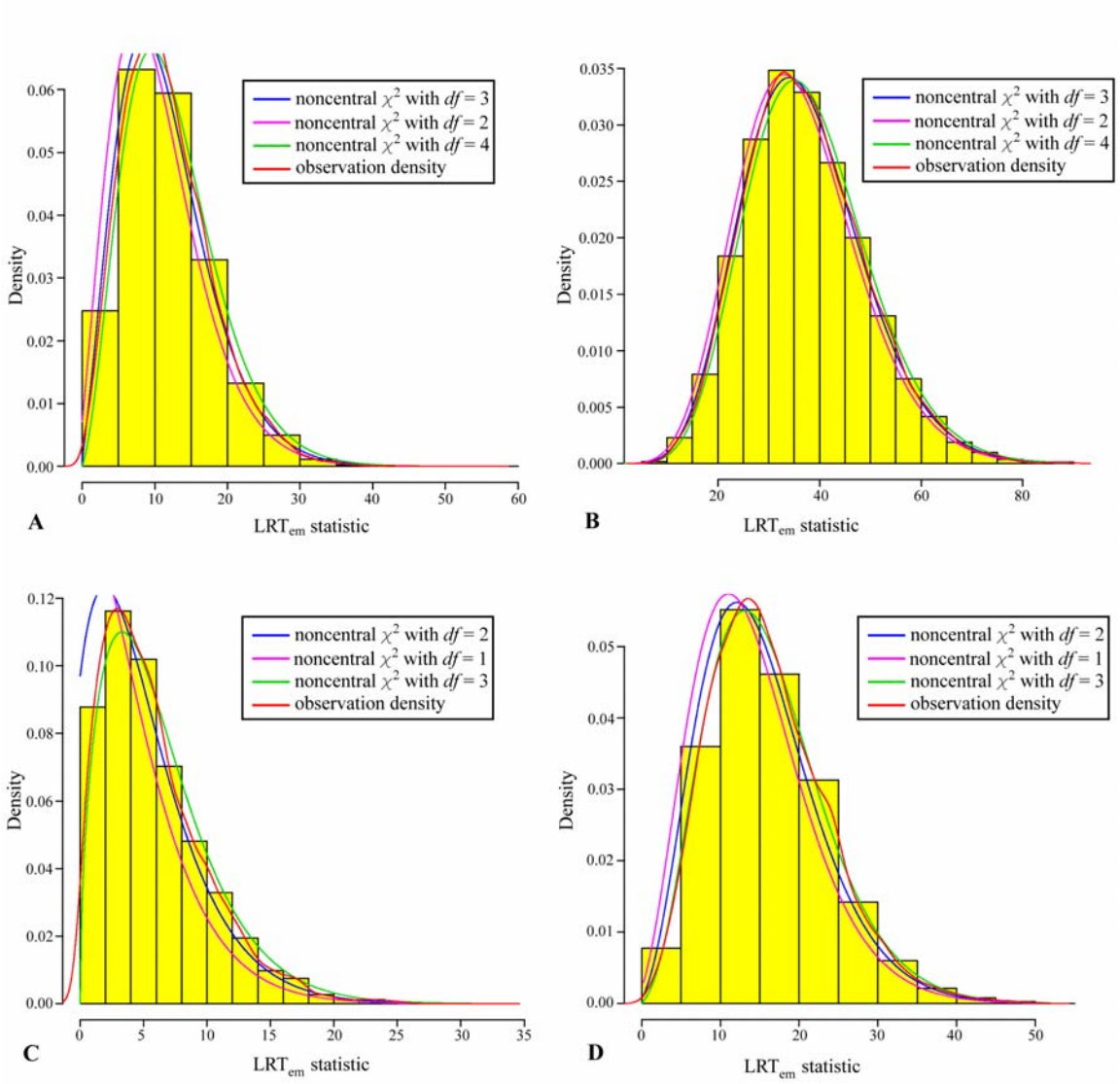
The rule described in equation (4.4) failed to predict the distribution for other experimental runs regardless of the sample size. The experimental run in which $MAF_1 = 0.01$, $MAF_2 = 0.01$, and LD between locus 1 and 2 = 0 (measured by *D'*) is in

this category. This run had a minimum generating haplotype frequency (0.0001) that was below all of the thresholds (for both a sample size of 500 and 2000) established by the rule described in equation (4.4). Figure 4.2D displays a histogram for LRT_{em} computed from simulations utilizing these parameter settings. Figure 4.2D shows the distribution for simulated datasets containing 2000 samples. In Figure 4.2D, the distribution of LRT_{em} does not resemble a central χ^2 distribution with 2 degrees of freedom, the distribution predicted by the rule in equation (4.4) for this experimental run. Instead, the distribution of LRT_{em} falls between central χ^2 distributions with 2 and 3 degrees of freedom. The distribution of LRT_{em} utilizing the same parameters for simulating datasets with 500 samples exhibited near identical behavior (results not shown). Thus, in this case, increasing the sample size did not increase the accuracy of the prediction rule described in equation (4.4).

Examination of the distributional properties of LRT_{em} under the alternative hypothesis. The prediction rule in equation (4.4) was not as successful for our simulations under the alternative hypothesis for the two SNP scenario. Although for the majority of cases the distribution of LRT_{em} did not resemble the distribution selected by the rule, in some situations increasing the sample size provided a distribution of LRT_{em} predicted by the rule (as we observed under H_0). Figures 4.3A and 4.3B display the distribution of LRT_{em} for one such set of experimental runs. Here $MAF_1 = 0.5$, $MAF_2 = 0.5$, and LD between SNP 1 and 2 = 0.9 (measured by D'). Figures 4.3A and 4.3B show the results from simulations of datasets with 500 and 2000 samples, respectively. Although the distribution of LRT_{em} appears to follow a noncentral χ^2

distribution with $df = 3$ (the distribution predicted for both runs) in Figure 4.3A, the fit is improved in Figure 4.3B. Furthermore, only the KS test results for the run with a sample size of 2000 support the idea that LRT_{em} follows the predicted distribution ($KS_{34.0,3}$ test p-value = 0.145 and $KS_{8.5,3}$ test p-value = 0). Figures 4.3C and 4.3D display the distribution of LRT_{em} for datasets of 500 and 2000 samples, respectively, simulated for haplotypes comprised of two SNPs where $MAF_1 = 0.01$, $MAF_2 = 0.01$, and LD between SNP 1 and 2 = 0 (measured by D'). The predicted distribution for both experimental runs is a noncentral χ^2 distribution with $df = 2$; however, the distribution of LRT_{em} in Figures 4.3C and 4.3D seems to bear a greater resemblance to a noncentral χ^2 distribution with $df = 3$. For the run with 2000 samples, the results of the KS test support the idea that LRT_{em} follows a noncentral χ^2 distribution with $df = 3$ ($KS_{13.1,3}$ test p-value = 0.286 and $KS_{13.1,2}$ test p-value = 0).

Figure 4.3 Histograms displaying the distribution of LRT_{em} under H_1 for the two SNP scenario



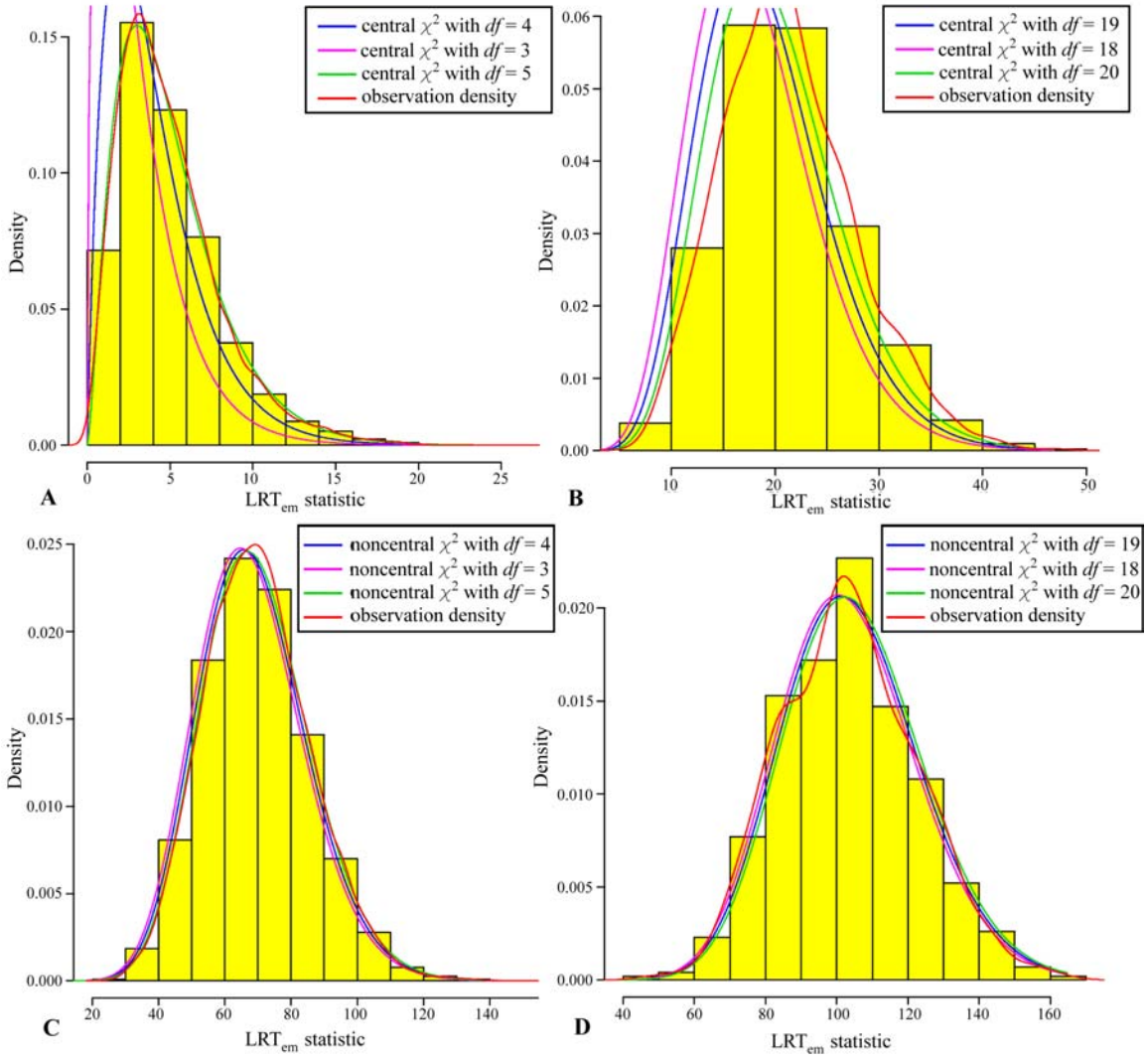
Legend for Figure 4.3: The histograms display the distribution of LRT_{em} computed from simulated datasets comprised of 500 samples (A and C) and 2000 samples (B and D) along with the density lines for several noncentral χ^2 distributions. The distribution of LRT_{em} was created by simulating haplotypes comprised of two SNPs under H_1. For (A) and (B), $MAF_1 = 0.5$, $MAF_2 = 0.5$, and LD between SNP 1 and $2 = 0.9$ (measured by D') while for (C) and (D), $MAF_1 = 0.01$, $MAF_2 = 0.01$, and LD between SNP 1 and $2 = 0$ (measured by D'). 10,000 replicate datasets containing equal numbers of cases and controls were simulated. The graphs were scaled to the observed data, and density lines off the scale were truncated.

Multi-SNP scenario

Examination of the distributional properties of LRT_{em} under the null and alternative hypotheses. For our simulations under the null and alternative hypotheses that rely on haplotype frequencies from the Horan dataset, LRT_{em} did not follow the distribution predicted by equation (4.4). Figure 4.4 displays histograms for LRT_{em} computed from simulations utilizing haplotype frequencies from the Horan dataset as the generating haplotype frequencies. Figures 4.4A and 4.4C show the distribution of LRT_{em} for a haplotype comprised of 5 SNP markers (for data simulated under H_0 and H_1, respectively) while Figures 4.4B and 4.4D show the distribution of LRT_{em} for a haplotype comprised of 10 SNP markers (for data simulated under H_0 and H_1, respectively). The simulations providing the data for Figures 4.4 created 1000 replicate datasets, each containing 2000 samples (equal numbers of cases and controls). For the haplotype simulations under H_0 or H_1 involving 5 SNP markers, the distribution predicted by equation (4.4) is a central or noncentral (*ncp* = 64.6), respectively, χ^2 distribution with 4 degrees of freedom. Figures 4.4A and 4.4C demonstrate that the distribution of LRT_{em} more closely approximates a central χ^2 distribution with 5 degrees of freedom. The $KS_{0,5}$ test (under H_0) *p*-value of 0.099 and the $KS_{64.6,5}$ test (under H_1) *p*-value of 0.163 confirm this similarity (while the $KS_{0,4}$ and $KS_{64.6,4}$ had *p*-values of 0). When the number of SNPs included in the haplotype is increased to ten for simulations under H_0 or H_1, the distribution predicted by equation (4.4) is a central or noncentral (*ncp* = 85.3), respectively, χ^2 distribution with 19 degrees of freedom. Figures 4.4B (under H_0) and 4.4D (under H_1) show that the distribution of LRT_{em} more closely approximates a central χ^2 distribution with 20 degrees of freedom and a noncentral χ^2 distribution with 18

degrees of freedom, respectively. The results from the KS tests indicate that there is the most evidence to support the idea that, under H_0, LRT_{em} is distributed as a central χ^2 distribution with 21 degrees of freedom ($KS_{0,21}$ p-value = 0.054) and, under H_1, LRT_{em} is distributed as a noncentral χ^2 distribution with 18 degrees of freedom ($KS_{85.3,18}$ p-value = 0.342). However, under H_1, a noncentral χ^2 distribution with 19 degrees of freedom is also consistent with the data ($KS_{85.3,19}$ p-value = 0.243).

Figure 4.4 Histograms displaying the distribution of LRT_{em} for simulations based on haplotype frequencies from the Horan dataset

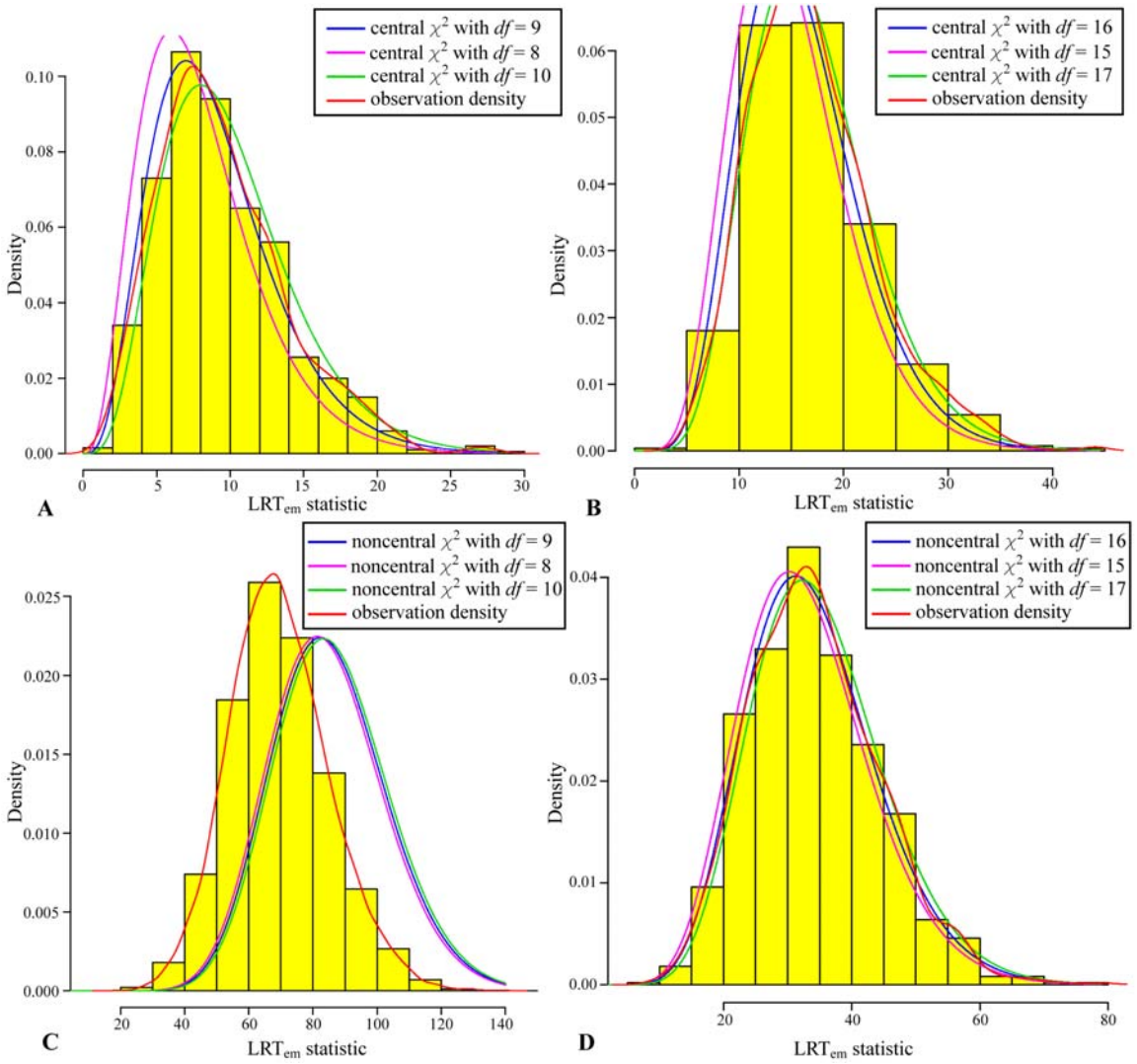


Legend for Figure 4.4: The histograms display the distribution of LRT_{em} computed from simulations based on haplotype frequencies from the Horan dataset along with the density lines for several central χ^2 distributions. The distribution of LRT_{em} was created by simulating haplotypes comprised of (A) 5 SNP markers and (B) 10 SNP markers under H_0 and haplotypes comprised of (C) 5 SNP markers and (D) 10 SNP markers under H_1. 1000 replicate datasets containing 2000 samples (equal numbers of cases and controls) were simulated. The graphs were scaled to the observed data, and density lines off the scale were truncated.

The rule described by equation (4.4) had some success in determining the distribution of LRT_{em} for the simulations under H_0 and H_1 based on haplotype frequencies from the HAPMAP TAP2 dataset. Figures 4.5A and 4.5C show the distribution under H_0 and H_1, respectively, for a haplotype comprised of 5 SNP markers while Figures 4.5B and 4.5D show the distribution under H_0 and H_1, respectively, for a haplotype comprised of 10 SNP markers. The simulations providing the data for Figures 4.5A and 4.5C created 1000 replicate datasets with 2000 samples (equal numbers of cases and controls) while the simulations providing the data for Figures 4.5B and 4.5D created 1000 replicate datasets with 500 samples (equal numbers of cases and controls). For the haplotype simulations involving 5 SNP markers, the distribution predicted by equation (4.4) under H_0 or H_1 is a central or noncentral ($ncp = 76.3$), respectively, χ^2 distribution with 9 degrees of freedom. In Figure 4.5A, the distribution of LRT_{em} under H_0 falls between central χ^2 distributions with 9 and 10 degrees of freedom. Although the *p*-values for the KS tests are small, they favor a central χ^2 distribution with 9 degrees of freedom ($KS_{0,9}$ *p*-value = 0.006). The distribution of LRT_{em} under H_1 presented in Figure 4.5C does not resemble the predicted noncentral χ^2 distribution but instead appears to be derived from a noncentral χ^2 distribution with many fewer degrees of freedom. When we increased the number of SNPs to ten, both under H_0 and H_1 equation (4.4) predicted that LRT_{em} would follow a central χ^2 distribution with 16 degrees of freedom. According to Figures 4.5B and 4.5D, the distribution of LRT_{em} falls between central χ^2 distributions with 16 and 17 degrees of freedom. The KS test results under H_0 indicate that LRT_{em} most likely follows a central χ^2 distribution with 17 degrees of freedom ($KS_{0,17}$ *p*-value = 0.750). (Interestingly, before rounding, equation (4.4)

predicted $df = 16.4$.) Figure 4.5D shows that, under H_1, LRT_{em} appears to follow a noncentral χ^2 distribution with $df = 16$ and $ncp = 18.0$ ($KS_{18.0,16}$ p-value = 0.628). Thus, under H_0 and H_1, equation (4.4) demonstrated an ability to predict the approximate correct degrees of freedom for the multi-marker haplotypes simulations although it lacked consistency for exacting precision.

Figure 4.5 Histograms displaying the distribution of LRT_{em} for simulations based on haplotype frequencies from the HAPMAP TAP2 dataset



Legend for Figure 4.5: The histograms display the distribution of LRT_{em} computed from simulations based on haplotype frequencies from the HAPMAP TAP2 dataset along with the density lines for several central χ^2 distributions. The distribution of LRT_{em} was created by simulating haplotypes comprised of (A) 5 SNP markers and (B) 10 SNP markers under H_0 and haplotypes comprised of (C) 5 SNP markers and (D) 10 SNP markers under H_1. 1000 replicate datasets containing (A and C) 2000 samples and (B and D) 500 samples (equal numbers of cases and controls) were simulated. The graphs were scaled to the observed data, and density lines off the scale were truncated.

Table 4.3 Summary table for the results from all experimental runs presented

Scenario	Hypothesis	MAF_1	MAF_2	LD between SNP 1 and SNP 2 (D')	Category Code
Two SNP	H_0	0.5	0.5	0.0	1
		0.5	0.01	0.0	2
		0.01	0.01	0.0	3
	H_1	0.5	0.5	0.9	2
		0.01	0.01	0.0	3
Scenario	Hypothesis	Dataset		Number of SNPs in Haplotype	Category Code
Multi-SNP	H_0	Horan		5	3
				10	3
	H_1			5	3
				10	3
	H_0	HAPMAP TAP2		5	3
				10	3
	H_1			5	3
				10	1

Legend for Table 4.3: This table summarizes the results for all experimental runs presented in section 4.3. The category codes are defined as: 1) experimental runs where the rule described in equation (4.4) successfully predicts the correct distribution; 2) experimental runs where the rule described in equation (4.4) successfully predicts the correct distribution for larger sample sizes only; and 3) experimental runs where the rule described in equation (4.4) fails to predict the correct distribution regardless of sample size.

4.4 Discussion

Even for the multi-SNP scenario where the range for the possible degrees of freedom of the χ^2 distribution is much wider (from 1 to 2^b, where b is the number of SNPs comprising the haplotype), the rule described in equation (4.4) was fairly consistent in predicting the χ^2 distribution closest to the distribution of LRT_{em} within a few degrees of freedom. However, while the rule sometimes predicted the correct distribution of the test statistic, it was not consistently accurate. Because of this inconsistency, we advocate

applying permutation and simulation methods to empirically generate the distribution of the test statistic under the null and alternative hypotheses, respectively, rather than applying the rule in equation (4.4). Future research is required to investigate alternative threshold settings and refine the prediction capability of this rule.

Knowing the precise distribution of a test statistic under the null and alternative hypotheses can be extremely practical. This knowledge allows researchers the freedom to employ the distribution to determine the statistical significance (distribution under H_0) and power (distribution under H_1) of the test rather than relying on more computationally intensive methods such as permutation and simulation to generate the null and alternative distributions empirically. Of course, reliance on a classically defined distribution (e.g. normal distribution, central χ^2 distribution, F distribution, etc.) that does not accurately describe the distribution of a statistic under the null and alternative hypotheses can lead to erroneous estimates of the type I error and power. In such cases, empirical techniques such as permutation and simulation are necessary even at the expense of computational resources. Often, this compromise is inconsequential when analyzing a real dataset. In fact, with modern computer processors and efficiently written code thousands of permutations can generally be performed in a reasonable timeframe. The limitation of this approach is often only apparent when many tests, all requiring a separate permutation procedure, are performed. Obviously, this situation arises for genome scans but can also be present for a haplotype-based association study that employs a sliding window approach across the SNPs in a single candidate gene.

Estimating low haplotype frequency estimates while computing LRT_{em} is somewhat analogous to constructing a sparse contingency table. However, methods that

utilize observations from a contingency table exhibit two qualities not available to likelihood-based methods that rely on haplotype frequency estimates – 1) a clear guideline defining when the central χ^2 distribution can be applied to determine the statistical significance and 2) the ability to combine categories containing rare observations. Unless Cochran's rule is violated (five or more observations in each cell of the contingency table), the central χ^2 distribution can be applied to determine the statistical significance for Pearson χ^2 or likelihood ratio statistics that utilize a contingency table (Cochran 1952). We have been unable to establish a parallel guideline for likelihood-based statistics that rely directly on haplotype frequency estimates. In addition, rare observations can be pooled (Sham and Curtis 1995; Schaid et al. 2002; Zhao et al. 2003) to produce a contingency table that is no longer sparse and contains a reduced number of categories. While frequency estimates for rare haplotypes can be pooled, for LRT_{em} the EM algorithm computes the likelihood during the haplotype frequency estimation step. Thus, pooling does not affect the computation of the statistic. The likelihood could be computed in a subsequent step using the multinomial distribution after haplotype frequencies were estimated and low haplotype frequency estimates were pooled. However, this approach is contrary to a key feature of LRT_{em} in that it treats expected counts from the estimates as observations rather than working directly with estimates. By working directly with haplotype frequency estimates in the expression for the likelihood, LRT_{em} avoids assumptions regarding the "observed" counts required for a contingency table.

CHAPTER 5: DISCUSSION

5.1 Synopsis

Although haplotypes can provide a powerful tool for gene mapping (Martin et al. 2000; Akey et al. 2001; Fallin et al. 2001; Morris and Kaplan 2002; Zaykin et al. 2002; Botstein and Risch 2003; Clark 2004), several factors add to the complexity of haplotype-based association studies relative to other forms of genetic association. First, in common practice, original observations are multilocus genotypes, which lack phase information. Consequently, estimation or inference procedures are required to apply a haplotype-based test. Second, haplotypes are a combination of alleles at multiple loci generally resulting in a large number of haplotypic variants. In the context of association studies, a large number of variants corresponds to many degrees of freedom and often a less powerful test. Third, as the number of marker loci comprising a haplotype grows, the number of possible haplotypic variants increases exponentially; however, many of these variants are not present in the population even though they may have positive frequency estimates. The complexity caused by these factors surfaces in several issues uniquely present in haplotype-base studies of association (as compared with other genetic association tests). For this thesis, we have developed work aimed at addressing several of these issues inherent in tests of haplotype-based association. Specifically, these issues include 1) the multiple testing problem introduced by employing hierarchical clustering to group similar haplotypes; 2) haplotype misclassification resulting from statistically inferring haplotype pairs from multilocus genotypes; and 3) uncertainty predicting the precise distribution of

the haplotype-based association test statistic when haplotype frequency estimates are very small or zero.

In the first part of this thesis, we examined the practice of applying a hierarchical clustering to haplotypes and then performing statistical tests at each step in the resulting hierarchy in the framework of multiple testing. To determine the empirical significance level or global *p*-value of the experiment, we proposed a method that takes into account the clustering process as well as the correlation structure of the tests performed. We applied our approach to datasets from haplotype association and microarray expression studies where hierarchical clustering has been used. In all of the cases we examined, we found that relying on one set of classes in the course of clustering leads to significance levels that are too small when compared with the significance level associated with an overall statistic that incorporates the process of clustering. In other words, relying on one step of clustering may furnish a formally significant result while the overall experiment is not significant.

In the second portion of this work, our simulations showed that the misclassification present in calling phased haplotypes from multilocus genotypes using statistical methods is complete. That is, each misclassified haplotype pair is consistently misclassified as the same incorrect haplotype pair throughout the entire dataset. In addition, our simulations under the null hypothesis of no association demonstrate that applying the central χ^2 distribution to evaluate the significance of test statistics produces conservative and anticonservative *p*-values while applying permutation methods consistently produces *p*-values that maintain the nominal false positive rate. Consequently, permutation methods should be exclusively used to determine statistical

significance for the tests we perform. As expected, the LRT_{ae} provides the greatest advantage in terms of power over the LRT_{std} in situations where more haplotype misclassification errors are present. These situations arise when the haplotype under investigation is comprised of many SNP markers with low pair-wise intermarker LD.

For fixed costs, the power gain of the LRT_{ae} over the LRT_{std} varied depending on the relative costs of genotyping, molecular haplotyping, and phenotyping. In general, the LRT_{ae} showed the greatest benefit over the LRT_{std} when the cost of phenotyping was very high relative to the cost of genotyping. This situation is likely to occur in a candidate gene replication study as opposed to a genome-wide association study. For intermediate phenotyping to genotyping cost ratios (e.g. $C_p / C_g = 25$), the LRT_{ae} may still provide a power advantage if the cost ratio of molecular haplotyping to genotyping is low ($C_{mh} / C_g < 10$ for $\alpha \geq 0.5$). Currently, inexpensive long-range PCR methods for molecular haplotyping are under development. As technology improves leading to less expensive molecular haplotyping methods, the LRT_{ae} will become applicable to a wider set of circumstances.

The final part of this thesis proposes a rule for predicting the distribution of a likelihood-based statistic that relies on haplotype frequency estimates. The rule consistently predicted the χ^2 distribution closest to the distribution of the statistic within a few degrees of freedom even for haplotypes containing many SNP markers. However, the rule did not consistently predict the distribution of the test statistic with pinpoint accuracy. Because of this inconsistent performance, we do not advocate applying the predicted distribution to determine statistical significance or power. Instead, permutation and simulation techniques should be employed to generate the distribution of the statistic

under the null hypothesis for determination of type I error and under the alternative hypotheses for determination of power, respectively.

5.2 Future Directions

This thesis introduces unique approaches for researchers utilizing haplotypes in case-control study designs to localize disease genes. The approaches proposed overcome pitfalls in analyzing datasets; however, they also have several limitations. One such limitation that is relevant for all three strategies described above is the means for computing type I error. In each case, permutation proves to be the most reliable method because of the possibility of sparse datasets. However, there are computational costs for this reliability. With modern processor speeds, analyses which utilize a large number of permutations can be performed in a practical amount of time. However, in the case of our computation of the global *p*-values for datasets where hierarchical clustering has been applied, the procedure is computationally more intensive. After permuting the data to compute null statistics, the procedure requires a myriad of comparisons between these null statistics (at the same step in the hierarchy) to compute null *p*-values. As a result, this procedure can be time-consuming, especially if the hierarchy created by clustering contains many steps. Similarly, the computational time required for permutation can be a factor when many association tests are performed at different locations in the genome, as is the case for a genome-wide scan. In addition, this situation arises for haplotype-based association studies within a single candidate gene that use a sliding window across the SNP markers in the gene. Permutation can be a valuable tool; however, the researcher

needs to be aware of the context of the application to plan for the time required for the procedure.

Aside from the computational issues, some limitations are inherent in the statistical methods themselves. For example, the LRT_{ae} procedure relies on haplotype pairs to detect association. As stated above, the number of haplotypes present can be quite large. Consequently, the number of inferred haplotype pairs can be very large ($\frac{w!}{2!(w-2)!}$, where w is the number of inferred haplotypes) resulting in many degrees of freedom for this test. Tests with larger degrees of freedom are generally equated with a loss in power. Thus, methods which examine single haplotypes (Schaid et al. 2002; Zaykin et al. 2002; Stram et al. 2003) rather than haplotype pairs may be more powerful than LRT_{ae}. Future research will need to compare the power for these approaches with that of the LRT_{ae}. Another option is to develop a version of the Cochran-Armitage Linear Test of Trend (Cochran 1954; Armitage 1955; Czika and Weir 2004) which incorporates a double-sampling procedure to correct for haplotype miscalls. Unlike the LRT_{ae} which makes no assumptions regarding a disease model, the Cochran-Armitage Linear Test of Trend relies on specific weights for each risk category and has only one degree of freedom. As a result, this test has the potential to be very powerful relative to other haplotype-based association tests, especially with the added capability of allowing for haplotype misclassification. However, specifying the incorrect disease model can negatively impact the power of the test (Freidlin et al. 2002). Future research will need to develop this test and assess its robustness to incorrect model selection.

Another potential limitation of LRT_{ae} is that the method assumes non-differential misclassification between cases and controls in estimating haplotype misclassification

rates. However, this assumption is not necessarily valid. A future research direction is to extend LRT_{ae} to estimate haplotype misclassification rates separately from cases and controls. Presumably, this feature will increase the effectiveness of the test.

In our power studies of LRT_{ae} and LRT_{std}, we used the entire dataset to infer haplotype pairs for each individual. We chose this approach because 1) it is conservative in terms of the power analysis (since differences between haplotype pair frequencies in cases and controls should not be as great); 2) the EM algorithm shows improved accuracy for haplotype frequency estimates when larger sample sizes are used (Fallin and Schork 2000); and 3) the EM algorithm assumes Hardy-Weinberg equilibrium, and one is more likely to violate this assumption when analyzing cases and controls separately. However, in practice researchers are more likely to examine cases and controls separately while inferring haplotype pairs. Presumably, the power will increase for both LRT_{ae} and LRT_{std} for an analysis conducted in this fashion; however, the relative power gain is not clear. Additional studies are required to assess the power of LRT_{ae} relative to LRT_{std} for data analyzed with this alternative inference scheme.

Finally, our rule for determining the distribution of LRT_{em} did not consistently provide a precisely accurate prediction. In some cases, a larger sample size improved the rule's accuracy. There are a number of possible explanations for this improvement. First, an increased sample size reduces the sparseness of the dataset. Second, an increased sample size improves the accuracy of the haplotype frequency estimates. Third, an increased sample size decreases the frequency threshold for distinguishing haplotypes present in the sample from those that are not present. Future research is

required to refine the prediction rule by investigating alternative algorithms for determining the thresholds for the estimated haplotype frequencies.

Technological advancements, in the form of SNP chips and online databases, have provided the capability to cost-effectively assay and manage hundreds of thousands of SNP markers throughout the genome (Smith 2005). With this explosion of genetic data, haplotype-based association studies have tremendous potential to localize disease genes. Specifically, genotypes are available genome-wide with an average density less than a kilobase. Prior to SNP chip technology which allows for this great density of genetic information, genome-scans were performed at substantially lower densities, such that the markers were in linkage equilibrium with one another and haplotype-based association analyses were less meaningful. Now the desire for molecular haplotypes presents a new technological frontier. Currently, the perception among molecular biologists appears to be that molecular haplotyping is too expensive to warrant widespread use. However, the cost of molecular haplotypes over small regions of the genome can be roughly equivalent to that for performing fluorescent polymerase chain reactions (Proudnikov et al. 2004; Proudnikov et al. 2006). In addition, industry has shown a serious interest in developing resources to reduce the cost of longer-range molecular haplotypes (Smith 2005). As molecular haplotyping becomes more affordable and hence more commonly used, the approaches explained in this thesis will continue to be relevant for identifying genes for complex traits.

NOTATION INDEX

Expression	Brief Description
C_g	cost of genotyping
C_{mh}	cost of molecular haplotyping
C_p	cost of phenotyping
C_p / C_g	cost ratio of phenotyping to genotyping
df	degrees of freedom for the (central or noncentral) χ^2 distribution
D'	standardized LD parameter, ($0 \le D' \le 1$)
D_{max}	maximum possible LD
DAF	disease allele frequency
f_i	penetrance associated with possessing i copies of the disease allele
g	number of variables in a (fractional) factorial design
h_j or h_{*j}	population haplotype frequency of the j^{th} haplotype (consisting of exclusively of marker loci)
h_{0j}	haplotype frequency in cases of the j^{th} haplotype
h_{1j}	haplotype frequency in controls of the j^{th} haplotype
$h_{+,j}$	frequency of disease-marker haplotype containing the wild-type allele (+) at the disease locus and the marker haplotype j
$h_{d,j}$	frequency of disease-marker haplotype containing the disease allele (d) at the disease locus and the marker haplotype j
h_{i_1} and h_{i_2}	pair of haplotype frequencies for a haplotype pair consistent with the i^{th} multilocus genotype
$\hat{h}_{*j}$	frequency estimates using all samples for the j^{th} haplotype
$\hat{h}_{0j}$	frequency estimates using cases alone for the j^{th} haplotype
$\hat{h}_{1j}$	frequency estimates using controls alone for the j^{th} haplotype
H_i	set of haplotype pairs compatible with the i^{th} multilocus genotype

I_{i,j_1,j_2}	conditional (on case status *i*) haplotype pair frequency for haplotype pair j_1, j_2
I_{i,j_3}	conditional (on case status *i*) haplotype frequency for haplotype j_3
J	number of total possible haplotypes
k	number of haplotype pairs
KS_{vj}	the Kolmogorov-Smirnov (KS) test for a χ^2 distribution with noncentrality parameter (*ncp*) of *v* and degrees of freedom (*df*) of *j*
$\ln(L_{1,ae})$	log-likelihood of data, where haplotype pair frequencies $p^t_{i'j'}$ are allowed to differ among different phenotype classes
$\ln(L_{0,ae})$	log-likelihood of data , where haplotype pair frequencies $p^t_{i'j'}$ are constrained to be equal among different phenotype classes
$\ln(L_{1,std})$	log-likelihood of data when not correcting for misclassification, where haplotype pair frequencies $p_{i'j}$ are allowed to differ among different phenotype classes
$\ln(L_{0,std})$	log-likelihood of data when not correcting for misclassification, where haplotype pair frequencies $p_{i'j}$ are constrained to be equal among different phenotype classes
L	likelihood of the data
L_{H_0}	likelihood of the data under the null hypothesis
L_{H_1}	likelihood of the data under the alternative hypothesis
L_*	likelihood computed from the multilocus genotypes from cases and controls together
L_0	likelihood computed from the multilocus genotypes from cases alone
L_1	likelihood computed from the multilocus genotypes from controls alone

LD(j)	amount of deviation from the equilibrium value for a disease-marker haplotype comprised of the j^{th} marker haplotype and either the wild type or disease allele.
LRT_{std}	standard likelihood ratio statistic (computed from contingency table)
LRT_{ae}	likelihood ratio statistic allowing for errors (computed from contingency table)
$m_{j'j}$	number of individuals that have been classified by the fallible method as haplotype pair j and by the infallible method as haplotype pair j', where $1 \le j, j' \le k$ (where k is the number of haplotype pairs)
$m_{j'+}$	number of individuals that have been classified by the infallible method as haplotype pair j', where $1 \le j' \le k$ (where k is the number of haplotype pairs)
m	number of permutations
$\min_i(p_i)$	minimum of local p-values
MAF_j	minor allele frequency at the j^{th} SNP locus
n	number of steps in hierarchy
$n^{(1)}_{i'j'j}$	number of individuals with (true) phenotype category i', true haplotype pair category j', and observed haplotype pair category j
$n^{(2)}_{i'j}$	number of individuals with (true) phenotype category i' and observed haplotype pair category j
ncp	noncentrality parameter for the noncentral χ^2 distribution
N	sample size for the LRT_{std}
N^{DS}	sample size for the LRT_{ae}
$N^{DS}*$	sample size for the LRT_{ae} determined from $\overline{\alpha}$
$\vec{p} = (p_1, p_2, ..., p_n)$	vector of local p-values
p_{min}	global p-value
p_d	allele frequency of disease-causing allele at the disease locus
p_+	allele frequency of the wild-type allele at the disease locus

$p_{i'j}$	observed population frequency of haplotype pair j for individuals with true phenotype i'
$p^t_{i'j'}$	true population frequency of haplotype pair j' for individuals with phenotype i'
$p^t_{*j'}$	true population frequency of haplotype pair j' under the null hypothesis that $p^t_{0j'} = p^t_{1j'} = p^t_{*j'}$
$q^t_{i'}$	true sampling frequency of phenotype i'
$r = C_{mh}/C_g$	cost ratio of molecular haplotyping to genotyping
R_1	genotype relative risk for the heterozygote
R_2	genotype relative risk for the homozygote
s	number of genetic variants
t	total number of individuals (used to determine a threshold from haplotype frequencies)
t_0	number of cases (used to determine a threshold from haplotype frequencies)
t_1	number of controls (used to determine threshold from haplotype frequencies)
$\vec{u}$	mean vector
$\mathbf{V}$	variance-covariance matrix
w	number of haplotypes (consisting of exclusively of marker loci)
x_{*j}	indicator function for the j^{th} haplotype (associated with frequency estimates using all samples)
x_{0j}	indicator function for the j^{th} haplotype (associated with frequency estimates using only cases)
x_{1j}	indicator function for the j^{th} haplotype (associated with frequency estimates using only controls)
$\vec{X} = (X_1, X_2, ..., X_n)$	vector of statistical values (generic)
$\mathbf{X_{null}}$	matrix of null statistics

X_j	event that an individual has observed haplotype pair $j, 1 \le j \le k$ (where k is the number of haplotype pairs)
$X^t_{j'}$	event that an individual has true haplotype pair $j', 1 \le j' \le k$ (where k is the number of haplotype pairs)
$X^t_{i'j'}$	event that an individual has phenotype $i', (i' = 0,1)$ and true haplotype pair $j', 1 \le j' \le k$ (where k is the number of haplotype pairs)
$Y^t_{i'}$	event that an individual has phenotype $i', (i' = 0,1)$
Y_i	multivariate normal random variable transformed from null local p-value at i^{th} step in hierarchy
α	double-sample proportion
$\overline{\alpha}$	mean double-sample proportion
δ	posterior probability threshold for the threshold double-sample selection method
η_*	number of haplotypes estimated using all samples
η_0	number of haplotypes estimated using controls alone
η_1	number of haplotypes estimated using cases alone
$\theta_{j'j}$	misclassification probability that the true haplotype pair j' will be misclassified as haplotype pair j
ϕ	disease prevalence

ELECTRONIC RESOURCE INFORMATION

The adenocarcinoma dataset published by Garber *et al.* (Garber et al. 2001) can be found at http://genome-www.stanford.edu/lung_cancer/adeno/index.shtml.

The B-cell lymphoma dataset published by Alizadeh *et al.* (Alizadeh et al. 2000) can be found at http://llmpp.nih.gov/lymphoma/.

The documentation for *StatXact 5* software can be found at http://www.cytel.com/.

The documentation for SNPHAP and PHASE can be found at http://www-gene.cimr.cam.ac.uk/clayton/software/ and http://www.stat.washington.edu/stephens/software.html, respectively.

The documentation for PAWE can be found at http://linkage.rockefeller.edu/derek/pawe1.html.

Data for the estimation of haplotype frequencies from SNP markers within the TAP2 gene were downloaded from http://www.hapmap.org/downloads/index.html.en (HapMap public release #16c.1).

LRT_{ae} software is available at ftp://linkage.rockefeller.edu/software/lrtae.

EHP software is available at http://linkage.rockefeller.edu/yyang/resources.html.

The documentation for the software package R is available at http://www.r-project.org/.

REFERENCES

Abecasis GR, Cookson WO (2000) GOLD—graphical overview of linkage disequilibrium. Bioinformatics 16:182-3

Abel L, Muller-Myhsok B (1998) Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. Am J Hum Genet 63:664-7

Adkins RM (2004) Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset. BMC Genet 5:22

Agresti A (1996) An Introduction to Categorical Data Analysis. John Wiley and Sons, NewYork

Akey J, Jin L, Xiong M (2001) Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291-300

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503-11

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sciences U S A 96:6745-50

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

Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933-44

Barral S, Haynes C, Levenstien MA, Gordon D (2005) Precision and type I error rate in the presence of genotype errors and missing parental data: a comparison between the original transmission disequilibrium test (TDT) and TDTae statistics. BMC Genet 6 Suppl 1:S150

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Statist Soc B 57:289-300

Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M (2001) Classification of human lung

carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 98:13790-5

Bolino A, Levy ER, Muglia M, Conforti FL, LeGuern E, Salih MA, Georgiou DM, Christodoulou RK, Hausmanowa-Petrusewicz I, Mandich P, Gambardella A, Quattrone A, Devoto M, Monaco AP (2000) Genetic refinement and physical mapping of the CMT4B gene on chromosome 11q22. Genomics 63:271-8

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl:228-37

Box GEP, Hunter WG, Hunter JS (1978) Statistic for Experimenters. John Wiley and Sons, New York

Brautbar C, Porat S, Nelken D, Gabriel KR, Cohen T (1977) HLA B27 and ankylosing spondylitis in the Israeli population. J Rheumatol Suppl 3:24-32

Brewerton DA, Hart FD, Nicholls A, Caffrey M, James DC, Sturrock RD (1973) Ankylosing spondylitis and HL-A 27. Lancet 1:904-7

Brumfield R, Beerli P, Nickerson D, Edwards S (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249-256

Burgtorf C, Kepper P, Hoehe M, Schmitt C, Reinhardt R, Lehrach H, Sauer S (2003) Clone-based systematic haplotyping (CSH): a procedure for physical haplotyping of whole genomes. Genome Res 13:2717-24

Chung CH, Bernard PS, Perou CM (2002) Molecular portraits and the family tree of cancer. Nat Genet Suppl 32:533-40

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111-22

Clark AG (2004) The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321-33

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595-612

Clark VJ, Metheny N, Dean M, Peterson RJ (2001) Statistical estimation and pedigree analysis of CCR2-CCR5 haplotypes. Hum Genet 108:484-93

Clayton D, Chapman J, Cooper J (2004) Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27:415-28

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, Nutland S, Howson JM, Faham M, Moorhead M, Jones HB, Falkowski M, Hardenbol P, Willis TD, Todd JA (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37:1243-6

Cochran WG (1952) The chi-square test of goodness of fit. Annals of Mathematical Statistics 23:315-345

Cochran WG (1954) Some methods for strengthening the common chi-squared tests. Biometrics 10:417-451

Collins FS, Brooks LD, Chakravarti A (1998) A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8:1229-31

Contu L, Capelli P, Sale S (1977) HLA B27 and ankylosing spondylitis: a population and family study in Sardinia. J Rheumatol Suppl 3:18-23

Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366:1121-31

Cox DR, Hinkley DV (1974) Theoretical Statistics. Chapman and Hall/CRC, Boca Raton

Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 56:811-2

Czika W, Weir BS (2004) Properties of the multiallelic trend test. Biometrics 60:69-74

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-32

De La Vega FM, Gordon D, Su X, Scafe C, Isaac H, Gilbert DA, Spier EG (2005) Power and sample size calculations for genetic case/control studies using gene-centric SNP maps: application to human chromosomes 6, 21, and 22 in three populations. Hum Hered 60:43-60

DeGroot MH (1991) Probability and Statistics. Addison-Wesley, Reading, MA

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc B 39:1-38

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

Devlin B, Roeder K, Wasserman L (2000) Genomic control for association studies: a semiparametric test to detect excess-haplotype sharing. Biostatistics 1:369-87

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

Devlin B, Roeder K, Wasserman L (2003) Analysis of multilocus models of association. Genet Epidemiol. 25:36-47

Ding C, Cantor CR (2003) Direct molecular haplotyping of long-range genomic DNA with M1-PCR. Proc Natl Acad Sci U S A 100:7449-53

Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 28:361-4

Douglas JA, Skol AD, Boehnke M (2002) Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. Am J Hum Genet 70:487-495

Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB (2000) Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci U S A 97:10483-8

Dudoit S, Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol 3:RESEARCH0036.1-0036.21

Dunn J (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3:32-57

Dupuis J, Brown PO, Siegmund D (1995) Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. Genetics 140:843-56

Edwards AWF (1992) Likelihood. The Johns Hopkins University Press, Baltimore

Eiberg H, Mohr J, Schmiegelow K, Nielsen LS, Williamson R (1985) Linkage relationships of paraoxonase (PON) with other markers: indication of PON-cystic fibrosis synteny. Clin Genet 28:265-71

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863-8

Ellis NA, Kirchhoff T, Mitra N, Ye TZ, Chuai S, Huang H, Nafa K, Norton L, Neuhausen S, Gordon D, Struewing JP, Narod S, Offit K (2005) Localization of

breast cancer susceptibility loci by genome-wide SNP linkage disequilibrium mapping. Genet Epidemiol

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523-42

Ewens WJ, Spielman RS (2005) What is the significance of a significant TDT? Hum Hered 60:206-10

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-7

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227-33

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143-51

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947-59

Fisher R (1922a) On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85:87-94

Fisher R (1922b) On the mathematical foundations of theoretical statistics. Philos Trans Roy Soc Lond A 222:309-368

Fisher R (1925) Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh
Fisher RA (1935) The Design of Experiments. Hafner, New York

Freidlin B, Zheng G, Li Z, Gastwirth JL (2002) Trend tests for case-control studies of genetic markers: power, sample size and robustness. Hum Hered 53:146-52

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225-9

Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I (2001) Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 98:13784-9

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Bio Cell 11:4241-57

Gillanders EM, Pearson JV, Sorant AJ, Trent JM, O'Connell J R, Bailey-Wilson JE (2006) The value of molecular haplotypes in a family-based linkage study. Am J Hum Genet 79:458-68

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-7

Gordon D, Finch SJ, Nothnagel M, Ott J (2002) Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered 54:22-33

Gordon D, Haynes C, Johnnidis C, Patel SB, Bowcock AM, Ott J (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. Eur J Hum Genet 12:752-61

Gorroochurn P, Hodge SE, Heiman G, Greenberg DA (2004) Effect of population stratification on case-control association studies. II. False-positive rates and their limiting behavior as number of subpopulations increases. Hum Hered 58:40-8

Guo L, Ma Y, Ward R, Castranova V, Shi X, Qian Y (2006) Constructing molecular classifiers for the accurate prognosis of lung adenocarcinoma. Clin Cancer Res 12:3344-54

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234-8

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684-9

Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409-11

Heid IM, Lamina C, Bongardt F, Fischer G, Klopp N, Huth C, Kuchenhoff H, Kronenberg F, Wichmann HE, Illig T (2005) [How about the uncertainty in the

haplotypes in the population-based KORA studies?]. Gesundheitswesen 67 Suppl 1:S132-6

Heiman GA, Hodge SE, Gorroochurn P, Zhang J, Greenberg DA (2004) Effect of population stratification on case-control association studies. I. Elevation in false positive rates and comparison to confounding risk ratios (a simulation study). Hum Hered 58:30-9

Helms C, Cao L, Krueger JG, Wijsman EM, Chamian F, Gordon D, Heffernan M, Daw JA, Robarge J, Ott J, Kwok PY, Menter A, Bowcock AM (2003) A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. Nat Genet 35:349-56

Helms C, Saccone NL, Cao L, Daw JA, Cao K, Hsu TM, Taillon-Miller P, Duan S, Gordon D, Pierce B, Ott J, Rice J, Fernandez-Vina MA, Kwok PY, Menter A, Bowcock AM (2005) Localization of PSORS1 to a haplotype block harboring HLA-C and distinct from corneodesmosin and HCR. Hum Genet 118:466-76

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

Heyer LJ, Kruglyak S, Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Res 9:1106-15

Hindorff LA, Psaty BM, Carlson CS, Heckbert SR, Lumley T, Smith NL, Lemaitre RN, Rieder MJ, Nickerson DA, Reiner AP (2006) Common Genetic Variation in the Prothrombin Gene, Hormone Therapy, and Incident Nonfatal Myocardial Infarction in Postmenopausal Women. Am J Epidemiol

Hoehe MR, Kopke K, Wendel B, Rohde K, Flachmeier C, Kidd KK, Berrettini WH, Church GM (2000) Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence. Hum Mol Genet 9:2895-908

Hogg R, Craig A (1995) Introduction to Mathematical Statistics. Prentice Hall, Upper Saddle River, NJ

Hoh J, Wille A, Ott J (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. Genome Res 11:2115-9

Hoppe B, Haupl T, Gruber R, Kiesewetter H, Burmester GR, Salama A, Dorner T (2006) Detailed analysis of the variability of peptidylarginine deiminase type 4 in

German patients with rheumatoid arthritis: a case-control study. Arthritis Res Ther 8:R34

Hoppe B, Heymann GA, Tolou F, Kiesewetter H, Doerner T, Salama A (2004) High variability of peptidylarginine deiminase 4 (PADI4) in a healthy white population: characterization of six new variants of PADI4 exons 2-4 by a novel haplotype-specific sequencing-based approach. J Mol Med 82:762-7

Horan M, Millar DS, Hedderich J, Lewis G, Newsway V, Mo N, Fryklund L, Procter AM, Krawczak M, Cooper DN (2003) Human growth hormone 1 (GH1) gene expression: complex haplotype-dependent influence of polymorphic variation in the proximal promoter and locus control region. Hum Mutat 21:408-23

Horimoto K, Toh H (2001) Statistical estimation of cluster boundaries in gene expression profile data. Bioinformatics 17:1143-51

Huang J, Vieland VJ (2001) The null distribution of the heterogeneity lod score does depend on the assumed model for the trait. Hum Hered 52:217-222

International HapMap Consortium (2003) The International HapMap Project. Nature 426:789-96

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299-320

Jannot AS, Essioux L, Clerget-Darpoux F (2004) Association in multifactorial traits: how to deal with rare observations? Hum Hered 58:73-81

Ji F, Yang Y, Haynes C, Finch SJ, Gordon D (2005) Computing asymptotic power and sample size for case-control genetic association studies in the presence of phenotype and/or genotype misclassification errors. Stat Appl Genet Mol Biol 4:Article 37

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233-7

Johnson S (1967) Hierarchical clustering schemes. Psychometrika 2:241-254

Joosten PH, Toepoel M, Mariman EC, Van Zoelen EJ (2001) Promoter haplotype combinations of the platelet-derived growth factor alpha-receptor gene predispose to human neural tube defects. Nat Genet 27:215-7

Kalbfleisch JD, Prentice RL (1980) The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York

Kang H, Qin ZS, Niu T, Liu JS (2004) Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. Am J Hum Genet 74:495-510

Kendall M, Stuart A, Ord JK (1994) The Advanced Theory of Statistics. Vol I. Oxford University Press, New York

Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073-80

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385-9

Knowlton RG, Cohen-Haguenauer O, Van Cong N, Frezal J, Brown VA, Barker D, Braman JC, Schumm JW, Tsui LC, Buchwald M, et al. (1985) A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. Nature 318:380-2

Kohler K, Bickeboller H (2006) Case-control association tests correcting for population stratification. Ann Hum Genet 70:98-115

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241-7

Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48:67-81

Lee MH, Gordon D, Ott J, Lu K, Ose L, Miettinen T, Gylling H, Stalenhoef AF, Pandya A, Hidaka H, Brewer B, Jr., Kojima H, Sakuma N, Pegoraro R, Salen G, Patel SB (2001) Fine mapping of a gene responsible for regulating dietary cholesterol absorption; founder effects underlie cases of phytosterolaemia in multiple communities. Eur J Hum Genet 9:375-84

Levitin PM, Gough WW, Davis JSt (1976) HLA-B27 antigen in women with ankylosing spondylitis. JAMA 235:2621-2

Levy-Lahad E, Bird TD (1996) Genetic factors in Alzheimer's disease: a review of recent advances. Ann Neurol 40:829-40

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49-67

Li CC (1955) Population Genetics. University of Chicago Press, Chicago

Little RJA, Rubin DB (1987) Statistical Analysis with Missing Data. John Wiley, New York

Lo Nigro C, Cusano R, Scaranari M, Cinti R, Forabosco P, Morra VB, De Michele G, Santoro L, Davies S, Hurst J, Devoto M, Ravazzolo R, Seri M (2000) A refined physical and transcriptional map of the SPG9 locus on 10q23.3-q24.2. Eur J Hum Genet 8:777-82

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799-810

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Paper presented at Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley

Maksymowych WP, Reeve JP, Reveille JD, Akey JM, Buenviaje H, O'Brien L, Peloso PM, Thomson GT, Jin L, Russell AS (2003) High-throughput single-nucleotide polymorphism analysis of the IL1RN locus in patients with ankylosing spondylitis by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry. Arthritis Rheum 48:2011-8

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, Qin ZS, Munro HM, Abecasis GR, Donnelly P (2006) A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 78:437-50

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

Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res 24:4841-3

Miller MB (1997) Genomic scanning and the transmission/disequilibrium test: analysis of error rates. Genet Epidemiol 14:851-6

Mitchell AA, Cutler DJ, Chakravarti A (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. Am J Hum Genet 72:598-610

Mitra SK (1958) On the limiting power function of the frequency chi-square test. Annals of Mathematical Statistics 29:1221-1233

Monaco AP, Kunkel LM (1988) Cloning of the Duchenne/Becker muscular dystrophy locus. Adv Hum Genet 17:61-98

Morris RW, Kaplan NL (2002) On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221-33

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277-318

Morton NE, Collins A (1998) Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci U S A 95:11389-93

Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC (2006) Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered 61:55-64

Mote VL, Anderson RL (1965) An investigation of the effect of misclassification on the properties of chisquare-tests in the analysis of categorical data. Biometrika 52:95-109

Niu T (2004) Algorithms for inferring haplotypes. Genet Epidemiol 27:334-47

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet 70:157-69

O'Hely M, Slatkin M (2003) The loss of statistical power to distinguish populations when certain samples are ambiguous. Theor Popul Biol 64:177-92

Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588-97

Ott J (1999) Analysis of Human Genetic Linkage. The Johns Hopkins University Press, Baltimore

Ouvrier R (1996) Correlation between the histopathologic, genotypic, and phenotypic features of hereditary peripheral neuropathies in childhood. J Child Neurol 11:133-46

Paluru P, Ronan SM, Heon E, Devoto M, Wildenberg SC, Scavello G, Holleschau A, Makitie O, Cole WG, King RA, Young TL (2003) New locus for autosomal dominant high myopia maps to the long arm of chromosome 17. Invest Ophthalmol Vis Sci 44:1830-6

Papadopoulos N, Leach FS, Kinzler KW, Vogelstein B (1995) Monoallelic mutation analysis (MAMA) for identifying germline mutations. Nat Genet 11:99-102

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719-23

Pearson K (1900) On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. Ser. 50:157-175

Perreard L, Fan C, Quackenbush JF, Mullins M, Gauthier NP, Nelson E, Mone M, Hansen H, Buys SS, Rasmussen K, Orrico AR, Dreher D, Walters R, Parker J, Hu Z, He X, Palazzo JP, Olopade OI, Szabo A, Perou CM, Bernard PS (2006) Classification and risk stratification of invasive breast carcinomas using a real-time quantitative RT-PCR assay. Breast Cancer Res 8:R23

Pfeiffer RM, Gail MH (2003) Sample size calculations for population- and family-based case-control association studies on marker genotypes. Genet Epidemiol 25:136-48

Pitman E (1937) Significance tests which may be applied to samples from any population. Supp Roy Stat Soc 4:119-130, 225-32

Pitman E (1938) Significance tests which may be applied to samples from any population. . Part III. The analysis of variance test. Biometrika 29:322-35

Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. Theor Popul Biol 60:227-37

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genetics 155:945-59

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67:170-81

Proudnikov D, Laforge KS, Hofflich H, Levenstien M, Gordon D, Barral S, Ott J, Kreek MJ (2006) Association analysis of polymorphisms in serotonin 1B receptor (HTR1B) gene with heroin addiction: a comparison of molecular and statistically estimated haplotypes. Pharmacogenet Genomics 16:25-36

Proudnikov D, LaForge KS, Kreek MJ (2004) High-throughput molecular haplotype analysis (allelic assignment) of single-nucleotide polymorphisms by fluorescent polymerase chain reaction. Anal Biochem 335:165-7

Qin ZS, Niu T, Liu JS (2002) Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am J Hum Genet 71:1242-7

Rademakers R, Cruts M, Sleegers K, Dermaut B, Theuns J, Aulchenko Y, Weckx S, De Pooter T, Van den Broeck M, Corsmit E, De Rijk P, Del-Favero J, van Swieten J, van Duijn CM, Van Broeckhoven C (2005) Linkage and association studies identify a novel locus for Alzheimer disease at 7q36 in a Dutch population-based sample. Am J Hum Genet 77:643-52

Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19:368-75

Rice KM, Holmans P (2003) Allowing for genotyping error in analysis of unmatched cases and controls. Ann Hum Genet 67:165-174

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066-73

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-7

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405:847-56

Ross S (2002) A First Course in Probability. Prentice Hall, Upper Saddle River, NJ

Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25:127-41

Rubinstein P, Walker M, Carpenter C, Carrier C, Krassner J, Falk CT, Ginsburg F (1981) Genetics of HLA disease associations. The use of the haplotype relative risk (HRR) and the "haplo-delta" (Dh) estimates in juvenile diabetes from three radical groups. Hum Immunol 3:384 (Abstract)

Sabatti C, Service S, Freimer N (2003) False Discovery Rate in Linkage and Association Genome Screens for Complex Disorders. Genetics 164:829-833

Sabbagh A, Darlu P (2005) Inferring haplotypes at the NAT2 locus: the computational approach. BMC Genet 6:30

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-33

Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53:1253-61

Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, Gusella JF, Crapper-MacLachlan DR, Alberts MJ (1993)

Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease.[comment]. Neurology 43:1467-72

Schaid DJ (2002) Relative efficiency of ambiguous vs. directly measured haplotype frequencies. Genet Epidemiol 23:426-43

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425-34

Schaid DJ, Sommer SS (1993) Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 53:1114-26

Schlosstein L, Terasaki PI, Bluestone R, Pearson CM (1973) High association of an HL-A antigen, W27, with ankylosing spondylitis. N Engl J Med 288:704-6

Schmiegelow K, Eiberg H, Tsui LC, Buchwald M, Phelan PD, Williamson R, Warwick W, Niebuhr E, Mohr J, Schwartz M, et al. (1986) Linkage between the loci for cystic fibrosis and paraoxonase. Clin Genet 29:374-7

Seri M, Cusano R, Forabosco P, Cinti R, Caroli F, Picco P, Bini R, Morra VB, De Michele G, Lerone M, Silengo M, Pela I, Borrone C, Romeo G, Devoto M (1999) Genetic mapping to 10q23.3-q24.2, in a large Italian pedigree, of a new syndrome showing bilateral cataracts, gastroesophageal reflux, and spastic paraparesis with amyotrophy. Am J Hum Genet 64:586-93

Sham P (1998) Statistics in Human Genetics. J. Wiley and Sons, Inc., New York

Sham PC, Curtis D (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Human Genet 59:97-105

Shmulewitz D, Zhang J, Greenberg DA (2004) Case-control association studies in mixed populations: correcting using genomic control. Hum Hered 58:145-53

Šidák Z (1967) Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association 62:626-633

Simpson EH (1951) The interpretation of interaction in contingency tables. J Roy Statist Soc B:238-241

Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38:209-13

Smith C (2005) Genomics: Getting down to details. Nature 435:991-94

Sobel E, Papp JC, Lange K (2002) Detection and integration of genotyping errors in statistical genetics. Am J Hum Genet 70:496-508

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 98:10869-74

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983-9

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-16

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al. (2001a) Haplotype variation and linkage disequilibrium in 313 human genes. Science 293:489-93

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162-9

Stephens M, Smith NJ, Donnelly P (2001b) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-89

Storey JD, Tibshirani R (2001) Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays. Stanford University, Stanford, CA

Stram DO, Leigh Pearce C, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Thomas DC (2003) Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered 55:179-90

Subrahmanyan L, Eberle MA, Clark AG, Kruglyak L, Nickerson DA (2001) Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. Am J Hum Genet 69:381-95

Sullivan LS, Daiger SP (1996) Inherited retinal degeneration: exceptional genetic and clinical heterogeneity. Mol Med Today 2:380-6

Tenenbein A (1970) A double sampling scheme for estimating from binomial data with misclassifications. Journal of the American Statistical Association 65:1350-1361

Tenenbein A (1972) A double sampling scheme for estimating from misclassified multinomial data with applications to sampling inspection. Technometrics 14:187-202

Terwilliger JD, Ott J (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Hum Hered 42:337-46

Terwilliger JD, Ott J (1994) Handbook of Human Genetic Linkage. Johns Hopkins, Baltimore

Thomas S, Porteous D, Visscher PM (2004) Power of direct vs. indirect haplotyping in association studies. Genet Epidemiol 26:116-24

Thomson G, Robinson WP, Kuhner MK, Joe S, Klitz W (1989) HLA and insulin gene associations with IDDM. Genet Epidemiol 6:155-60

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518-22

Tu IP, Whittemore AS (1999) Power of association and linkage tests when the disease alleles are unobserved. Am J Hum Genet 64:641-9

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, et al. (2001) The sequence of the human genome.[comment][erratum appears in Science 2001 Jun 5;292(5523):1838]. Science 291:1304-51

Wainwright BJ, Scambler PJ, Schmidtke J, Watson EA, Law HY, Farrall M, Cooke HJ, Eiberg H, Williamson R (1985) Localization of cystic fibrosis locus to human chromosome 7cen-q22. Nature 318:384-5

Wang L, Xu Y (2003) Haplotype inference by maximum parsimony. Bioinformatics 19:1773-80

Weller JI, Song JZ, Heyen DW, Lewin HA, Ron M (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics 150:1699-706

Westfall PH, Young SS (1993) Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. John Wiley & Sons, New York

Xie X, Ott J (1993) Testing linkage disequilibrium between a disease gene and marker loci. Am J Hum Genet 53:1107 (Abstract)

Xu CF, Lewis K, Cantone KL, Khan P, Donnelly C, White N, Crocker N, Boyd PR, Zaykin DV, Purvis IJ (2002) Effectiveness of computational methods in haplotype prediction. Hum Genet 110:148-56

Xu H, Wu X, Spitz MR, Shete S (2004) Comparison of haplotype inference methods using genotypic data from unrelated individuals. Hum Hered 58:63-8

Yan H, Papadopoulos N, Marra G, Perrera C, Jiricny J, Boland CR, Lynch HT, Chadwick RB, de la Chapelle A, Berg K, Eshleman JR, Yuan W, Markowitz S, Laken SJ, Lengauer C, Kinzler KW, Vogelstein B (2000) Conversion of diploidy to haploidy. Nature 403:723-4

Yang Y, Hoh J, Broger C, Neeb M, Edington J, Lindpaintner K, Ott J (2003) Statistical methods for analyzing microarray feature data with replications. J of Comput Biol 10:157-69

Yu CE, Devlin B, Galloway N, Loomis E, Schellenberg GD (2004) ADLAPH: A molecular haplotyping method based on allele-discriminating long-range PCR. Genomics 84:600-12

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79-91

Zhang J, Vingron M, Hoehe MR (2005) Haplotype reconstruction for diploid populations. Hum Hered 59:144-56

Zhao JH, Curtis D, Sham PC (2000) Model-free analysis and permutation tests for allelic associations. Hum Hered 50:133-9

Zhao JH, Sham P (2002) Faster haplotype frequency estimation using unrelated subjects. Hum Hered 53:36-41

Zhao LP, Li SS, Khalid N (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231-50

Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89-100