2008

Comparisons of Sequential Testing Approaches for Detection of Association Between Disease and Haplotype Blocks

Andres Azuero

University of Alabama at Birmingham

COMPARISONS OF SEQUENTIAL TESTING APPROACHES FOR DETECTION OF ASSOCIATION BETWEEN DISEASE AND HAPLOTYPE BLOCKS

by

ANDRES AZUERO

DAVID T. REDDEN, COMMITTEE CHAIR
CHARLES R. KATHOLI
SHARINA D. PERSON
JOSHUA S. RICHMAN
HEMANT K. TIWARI

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2008

July 28, 2008

COMPARISONS OF SEQUENTIAL TESTING APPROACHES FOR DETECTION OF ASSOCIATION BETWEEN DISEASE AND HAPLOTYPE BLOCKS

ANDRES AZUERO

BIOSTATISTICS

ABSTRACT

Since Wald's development of the Sequential Probability Ratio Test (SPRT) in 1945, sequential analysis has evolved into a rich and well-developed field. Sequential methods are commonly used in industrial applications, clinical trials, and genetic association studies. Genetic association studies are typically conducted using a case-control design, where deoxyribonucleic acid (DNA) samples from affected cases and unaffected controls, unrelated to each other, are collected. The distribution of alleles for the groups is compared at a set of genetic markers. Substantial between-group differences in allele frequencies are indicative of association with susceptibility to the disease. The most common variation in the human genome is a single nucleotide polymorphism (SNP). SNPs are currently the most common markers in association studies. For a particular disease, testing all 10 million common SNPs in the human genome for association would be extremely expensive. However, adjacent SNP alleles tend to be inherited together. Haplotype blocks are inferred locations along the genome where sequences of linked alleles are likely to be inherited as units. Once there is more clarity about the haplotype block structure in the human genome, using blocks as markers could represent a drastic dimension reduction for association testing. In spite of this reduction, multiple testing will remain an obstacle in association studies. Ignoring multiplicity results in substantial numbers of false positive detections. Controlling for multiplicity results in a decrease of statistical power; and increasing the sample size to maintain power translates into higher costs. Sequential procedures have been proposed as a solution to these problems. This dissertation modifies a fully sequential design for application in association studies; it develops an algorithm for simulation of markers in Linkage Disequilibrium; and through simulations of an association study using haplotype blocks, this dissertation compares the modified sequential procedure against *ad hoc* sequential procedures published during the past 15 years, as well as common fixed sample size approaches, when applied to the problem of testing a relatively large number of markers in the same case-control cohort. Comparisons are made in terms of observed experiment-wise type I error rate, false positive rate, experiment-wise power, and a proposed measure of penalized power.

DEDICATION

To my family and wife.

ACKNOWLEDGEMENTS

To my professor David Redden, PhD, for helping me all along the way these past 5 years; for his support, mentoring and friendship.

To my family and wife, for being always patient and supportive.

To the faculty and staff of the UAB Departments of Biostatistics and Preventive Medicine, for giving me a chance and for providing a great learning environment.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure *Page*

LIST OF ABBREVIATIONS

ALR	Adaptive likelihood ratio test
DNA	Deoxyribonucleic acid
FDR	False discover rate
FPR	False positive rate
FTR	Fail to reject
FWER	Family-wise error rate
GLR	Generalized likelihood ratio
LD	Linkage disequilibrium
MLE	Maximum likelihood estimator
SGLR	Sequential generalized likelihood ratio test
SNP	Single nucleotide polymorphism
SPRT	Sequential probability ratio test

1. INTRODUCTION AND BACKGROUND

1.1 Introduction to Sequential Analysis

Formal sequential testing procedures have existed since the publication of Wald's (1945) seminal paper on sequential tests of hypotheses. However, the history of sequential procedures dates back to the 18th and 19th centuries when the rudiments of sequential analysis were introduced by famous mathematicians such as Bernoulli, DeMoivre, Lagrange and Laplace while studying the problem known as the *Duration of Play,* a problem commonly referred to currently as the *Gambler's Ruin* (Ghosh, 1991). In the early 20th century, sequential *ad hoc* procedures were developed for industrial applications in manufacturing. Research into sequential methods advanced again in 1942, when the United States Department of Defense formed a statistical research group at Columbia University in New York. The purpose of this research group was to advise the Department of Defense on the use of statistical methods for large scale experiments which were being conducted due to the country's World War II involvement. Wald, along with contributions from other famous statisticians such as Friedman, Wallis, Wolfowitz and Hotelling, developed the Sequential Probability Ratio Test (SPRT) in 1943 for military applications. The SPRT's objective was to minimize sample size when the cost of each observation was high (i.e., firing an experimental rocket or testing anti-craft gunnery (Lai, 2001)). The original technical reports on the SPRT were considered military secrets; however, at the end of World War II, the Department of Defense approved public release of Wald's SPRT research. Shortly after the release, Wald published the seminal paper on sequential tests of hypotheses.

A sequential test is a statistical testing procedure which at any stage of the experiment gives a specific rule for making one of the following three decisions:

a) Fail to Reject (FTR) the null hypothesis, i.e. 'accept' the null hypothesis, and stop collecting observations,

b) reject the null hypothesis, i.e. accept the alternative hypothesis, and stop collecting observations,

c) continue the experiment by collecting an additional observation.

The original SPRT is a sequential observation-by-observation Likelihood Ratio Test for simple hypotheses based on the Neyman-Pearson Lemma. The number of observations for a fully sequential test is not predetermined, but is a random variable that depends on the true parameter value, the hypothesized null and alternative parameter values, desired type I and type II error probabilities, and the previous observations. Wald derived approximations to the decision boundaries in terms of type I and type II error probabilities. When the true parameter value is close to either the null or the alternative hypothesized parameter values, the SPRT is optimal in the sense that it minimizes the sample size required to make a decision.

Since the development of the SPRT, sequential analysis has evolved into a well-developed field (Province, 2000). For instance, in statistical genetics, a major and direct application of the SPRT is the LOD score developed by Morton (1955) for detection of linkage between genetic loci. And in group sequential methods in clinical trials, O'Brien and Fleming (1979) proposed a class of group sequential tests based on an adaptation of a truncated SPRT. The paper by O'Brien and Fleming, along with papers by Pocock (1977) and Lan & DeMets (1983) have been particularly influential and together form the starting point for recent methodological research and the basis of current practice in clinical trial design (Jennison & Turnbull, 2000).

Throughout the years, different versions of sequential likelihood ratio tests have been proposed that address some of the shortcomings of the original SPRT procedure both in the frequentist and Bayesian settings. Within the next section, the statistical issues facing genetic association studies that motivate the application of SPRT methodologies to genetic studies will be presented.

1.2 The Problem of Multiple Testing in Genetic Association Studies

Genetic association studies are typically conducted using a case-control design, where deoxyribonucleic acid (DNA) samples from affected cases and unaffected controls, unrelated to each other, are collected. The distribution of alleles/genotypes for the two groups is compared at a set of genetic markers. The two common genetic association study designs are candidate gene association studies and genome-wide association scans. A candidate gene association study tests for association between disease status and a prespecified subset of candidate genetic markers from a few regions of the genome. In contrast, testing markers from all regions of the genome for association with disease status is referred to as a genome-wide association scan. Markers that show substantial between-group differences in allele frequencies are taken to be associated with susceptibility to the disease under study (Siegmund & Yakir, 2007).

The ability to obtain a large number of markers, such as single-nucleotide polymorphisms (SNPs, pronounced 'snips'), has accelerated research interests in association studies (Satagopan & Elston, 2003). More than 10 million SNPs are estimated to occur commonly in the human genome (www.hapmap.org accessed on April 15th, 2008). Since 2002, the International HapMap Project, a multi-country effort to identify and catalog genetic similarities and differences in four populations representing different parts of the world, has identified approximately 3.8 million SNPs in each of the four populations. For a particular disease, testing all 10 million common SNPs in the human genome for association would be extremely expensive and redundant. However, examination of high-density SNP markers over contiguous regions suggested a surprisingly simple pattern. Adjacent markers form blocks of variable length that tend to be inherited as units. These blocks are delimited by recombination ‘hotspots’. And within each block only a few common sequences are observed (Daly et al., 2001; Gabriel et al., 2002). These sequences of linked markers are known as haplotypes. Consequently, the International HapMap Project has as goals to catalog the regions that contain haplotypes (referred to as ‘haplotype blocks’) and to determine 'tag' SNPs that would identify the haplotypes within these blocks. It is believed that by identifying an individual's tag SNPs, researchers will be able to determine the collection of haplotypes in a person's DNA. It is anticipated that 300,000 to 600,000 tag SNPs will summarize most of the genetic variability within human populations (www.hapmap.org accessed on April 15th, 2008). Although the vast majority of association studies are currently conducted using SNP markers, once there is more clarity about the haplotype structure in the human genome, the haplotype mapping approach could potentially be more efficient than single SNP mapping.

In spite of the potential drastic dimension reduction that using haplotype blocks represents, compared to single SNPs, the issue of multiple testing will still remain a substantial statistical problem in genetic association studies. If multiplicity is not accounted for, an inappropriately large number of false detections will be observed. To illustrate this point, assume an association study that tests 2000 approximately independent markers, of which none are truly associated to the disease. At a significance level of 0.05 for each individual test, approximately 100 false detections are expected, and further examination and study of these false leads might be very costly. For a fixed sample size, usual multiple testing correction methods such as Bonferroni will greatly decrease power. However, large samples needed to maintain power after correcting for multiplicity translate into substantially higher costs. Sequential testing procedures have been proposed as potential solutions to these problems (Sobel et al., 1993; Sham, 1994; Satagopan et al., 2002; Satagopan & Elston, 2003; Satagopan et al., 2004; Thomas et al., 2005; Skol et al., 2006; Wang et al., 2006).

1.3 Dissertation Objectives

The purpose of this dissertation is to 1) modify the Sequential Generalized Likelihood Ratio test (SGLR) for application in haplotype studies, and 2) compare and contrast the properties of the SGLR and four other sequential testing procedures proposed and published in the literature during the past 15 years, when applied to the problem of testing a relatively large number of haplotype blocks in the same case-control cohort. Each sequential procedure is to be compared against each other as well as with the standard fixed-sample-size chi-square test. The comparisons are to be made in terms of observed experiment-wise type I error rate, False Positive Rate (FPR), observed experiment-wise statistical power, and a measure of penalized power. Table 1 summarizes the testing methods and multiplicity adjustments examined in this dissertation.

Table 1. *Testing methods and multiplicity adjustments examined in this dissertation.*

Testing Procedure	Multiplicity Adjustment
1) Fixed sample size Pearson's χ^2 tests	a) Uncorrected b) Bonferroni c) Holm d) Benjamini and Hochberg e) Benjamini and Yekutieli
2) Sobel et al., 1993 (2 and 3 stages)	a) Holm
3) Sham, 1994 (2 and 3 stages)	
4) Satagopan et al., 2004 (2 stages)	
5) Skol et al., 2006 (2 stages)	
6) Modified SGLR, Chan & Lai, 2005	a) Uncorrected b) Bonferroni

Each simulated experiment consists of applying the testing methods in Table 1 to detect 100 blocks from a total of 2150 blocks (following a simulated block structure of the human chromosome 22), and with 400 subjects (200 per group). Since chromosome 22 is small in length compared to other chromosomes, screening this chromosome should not require a very large sample size. Consequently, the specific aims of this dissertation are the following:

a) Modify the SGLR for application in haplotype-based genetic association studies.
b) Develop a novel simulation approach for generation of correlated multinomial variables (representing haplotype blocks).

c) Using the novel simulation approach, generate 1000 simulated datasets with the following characteristics:

- Each dataset consists of 2150 correlated multinomial variables representing haplotype blocks. The underlying uniform (0,1) variables used to generate the multinomial variables have an approximate autoregressive structure with ρ=0.8 to simulate Linkage Disequilibrium (LD) across blocks.
- Of the 2150 multinomial variables, 100 variables differ in proportions of realizations of multinomial outcomes (representing haplotypes) between case and control groups, according to predetermined chi-square effect sizes. The remaining 2050 multinomial variables have equal distribution of multinomial outcomes between case and control groups.
- Each dataset has 800 observations (each of the 200 subjects per group contributes two independent chromosomes, for a total of 400 observations per group)

d) Using the simulated datasets from (c), compare and contrast the methods in Table 1, in terms of observed experiment-wise type I error rate, FPR, observed experiment-wise power, and a measure of penalized power

e) Propose improvements to the methods in Table 1.

2. LITERATURE REVIEW

2.1 Sequential Analysis

The history of sequential procedures dates back to the 18th and 19th centuries when the rudiments of sequential analysis were introduced by famous mathematicians such as Bernoulli, DeMoivre, Lagrange and Laplace while studying the problem known as the *Duration of Play,* a problem commonly referred to currently as the *Gambler's Ruin* (Ghosh, 1991). In the early 20th century, sequential *ad hoc* procedures were developed for industrial applications in manufacturing. However, with the publication of Wald's (1945) seminal paper on sequential tests of hypotheses using the Sequential Probability Ratio Test (SPRT), formal statistical research into the properties of sequential testing procedures was introduced. Wald began this paper by defining a sequential test as a statistical testing procedure which at any stage of the experiment gives a specific rule for making one of the following three decisions:

a) Fail to Reject (FTR) the null hypothesis, i.e. 'accept' the null hypothesis and collect no more observations,
b) reject the null hypothesis, i.e. accept the alternative hypothesis and collect no more observations,
c) continue the experiment by collecting an additional observation.

The number of observations for a sequential test is not predetermined, but is a random variable that depends on the true parameter value, the hypothesized null and alternative parameter values, desired type I and type II error probabilities, and the previous observations. Wald's idea was to use the 'current most powerful procedure' as he called it, in a sequential manner. He described the 'current most powerful procedure' as follows:

The test is H$_0$: the distribution of X is $f_0(x)$ versus H$_A$: the distribution of X is $f_1(x)$.

Determine the critical region by:

$$\frac{\prod_{i=1}^{N} f_1(x_i)}{\prod_{i=1}^{N} f_0(x_i)} \geq k \quad \text{i.e.} \quad \frac{L_{1N}}{L_{0N}} \geq k$$

where x_i, i=1..N, are random observations from a variable X; $f_0(x)$ is the distribution of X under the null hypothesis H$_0$; $f_1(x)$ is the distribution of X under the alternative hypothesis H$_A$; L_{0N} is the likelihood or joint probability of the N sample observations under H$_0$, and L_{1N} is the likelihood of the sample under H$_A$. The critical value k is determined so that the probability of type I error is a pre-assigned value α, and N is equal to the smallest sample size for which the probability of type II error does not exceed a pre-assigned value β. If H$_0$ and H$_A$ are simple hypotheses, then this likelihood ratio test is an application of the Neyman-Pearson Lemma (Casella & Berger, 2002), which defines the Uniformly Most Powerful Test for simple hypotheses.

Thus, Wald developed the SPRT as a sequential, observation by observation, Neyman-Pearson-based test. H$_0$ and H$_A$ are simple hypotheses, i.e. H$_0$: $\theta = \theta_0$ vs. H$_A$: $\theta = \theta_1$, where θ is the parameter of interest and θ_0 and θ_1 are the values of the parameter under H$_0$ and H$_A$, respectively. The procedure is as follows:

Let $f(x|\theta)$ be the distribution function of a random variable X with parameter θ.

The hypothesis of interest is: H$_0$: the distribution of X is $f(x|\theta_0)$ vs. H$_A$: the distribution of X is $f(x|\theta_1)$. At each observation (at the m$^{\text{th}}$ observation), calculate

$$\frac{L_{1m}}{L_{0m}} = \frac{\prod_{i=1}^{m} f(x_i|\theta_1)}{\prod_{i=1}^{m} f(x_i|\theta_0)},$$

Reject H_0 (conclude H_A) if: $\frac{L_{1m}}{L_{0m}} \geq A$,

FTR H_0 (conclude H_0) if: $\frac{L_{1m}}{L_{0m}} \leq B$,

Collect an additional observation if: $B < \frac{L_{1m}}{L_{0m}} < A$,

or alternatively, $\log(B) < \log\big(L_{1m}(x_i)\big) - \log\big(L_{0m}(x_i)\big) < \log\,(A)$,

since it is easier to work with logarithms in practice.

The solutions to the quantities A and B are not trivial (Bain and Engelhardt, 1992), but Wald was able to cleverly derive practical approximations in terms of the type I error probability, α, and the type II error probability, β: $A = \frac{1-\beta}{\alpha}$, and $B = \frac{\beta}{1-\alpha}$. Hence for a test with α=0.05 and β=0.2, the boundaries A and B are equal to 16 and 0.21, respectively.

Wald noted that the repeated assessment after each observation allows for termination immediately after the decision boundaries are crossed. However, the procedure can be used as a group-sequential test. The key feature of a group-sequential test is that the accumulating observations are analyzed in groups rather than after every new observation. In the case of the SPRT, since the likelihood ratios are calculated for each observation, the only effect of taking groups of observations at a time instead of a single observation is that more observations will be taken, approximately enough to fill the last group required to make a decision. But the benefit of such an approach is that the probability of making an incorrect decision will be somewhat smaller by having more observations (Wald, 1945).

Wald derived approximations to the expected number of observations required for the SPRT to reach a decision given a true value of a location parameter for a normal distribution with known variance, and for the case of a binomial proportion.

The original SPRT has three major shortcomings:

1) the procedure is restricted to simple null and simple alternative hypotheses,
2) the procedure is only optimal, in the sense of the sample required to make a terminal decision, when the true value of the parameter of interest is either the null or the alternative hypothesized parameter,
3) the SPRT is not be robust to distributional misspecifications.

Soon after the publication of Wald’s paper, probabilists/statisticians recognized the SPRT as a stochastic process, and as consequence, the mathematical statistics journals now contain a vast literature on the subject of sequential analysis (Jennison & Turnbull, 2000). Throughout the years, numerous versions of sequential likelihood ratio tests have been proposed that address some of the aforementioned shortcomings of the original SPRT both in the frequentist and Bayesian settings. Ghosh (1991) provides a detailed chronologic survey of statistical methods for sequential analysis up to 1990. Lai (2001) discusses more recent developments, including the use of Generalized Likelihood Ratios (GLRs), where the parameters of interest are replaced by their Maximum Likelihood Estimates (MLEs).

In the frequentist paradigm, Chan & Lai (2005) propose a sequential test for composite hypotheses based on GLRs. The Sequential Generalized Likelihood Ratio test (SGLR) is as follows:

Let *f(x|θ)* be the distribution function of a random variable *X* with parameter *θ*. Let Θ be the parameter space of θ.

The hypothesis of interest is: H_0: $\theta \in \Theta_0$ vs. H_1: $\theta \in \Theta_1$, where Θ_0 is the parameter space restricted to the null hypothesis, and Θ_1 is the parameter space restricted to the alternative hypothesis. Let $\Theta_1 \cap \Theta_0 = \emptyset$, and $\Theta_1 \cup \Theta_0 = \Theta$.

At each observation (at the m$^{\text{th}}$ observation), calculate

$$\lambda_{0m} = \frac{\prod_{i=1}^{m} f(x_i|\hat{\theta})}{\prod_{i=1}^{m} f(x_i|\hat{\theta}_0)} \quad \text{and} \quad \lambda_{1m} = \frac{\prod_{i=1}^{m} f(x_i|\hat{\theta})}{\prod_{i=1}^{m} f(x_i|\hat{\theta}_1)}$$

where $\hat{\theta}, \hat{\theta}_0$and $\hat{\theta}_1$ are respectively the unrestricted MLE of *θ*, the restricted MLE of *θ* under the parameter space defined by H_0, and the restricted MLE of *θ* under the parameter space defined by H_1. The decision rules are defined as:

Reject H_0 (conclude H_1) if $\lambda_{om} \geq B_{GLR}$,

FTR H_0 (conclude H_0) if $\lambda_{1m} \geq B_{GLR}$,

Otherwise, collect an additional observation.

Chan and Lai (2005) propose using Monte Carlo methods in order to calculate the decision boundary B_{GLR}. In the direct Monte Carlo approach, *K* simulations of the experiment under H_0 are generated until a terminal decision is reached in each of the *K* experiments. For a desired type I error probability, *α*, the boundary B_{GLR} is calculated numerically by:

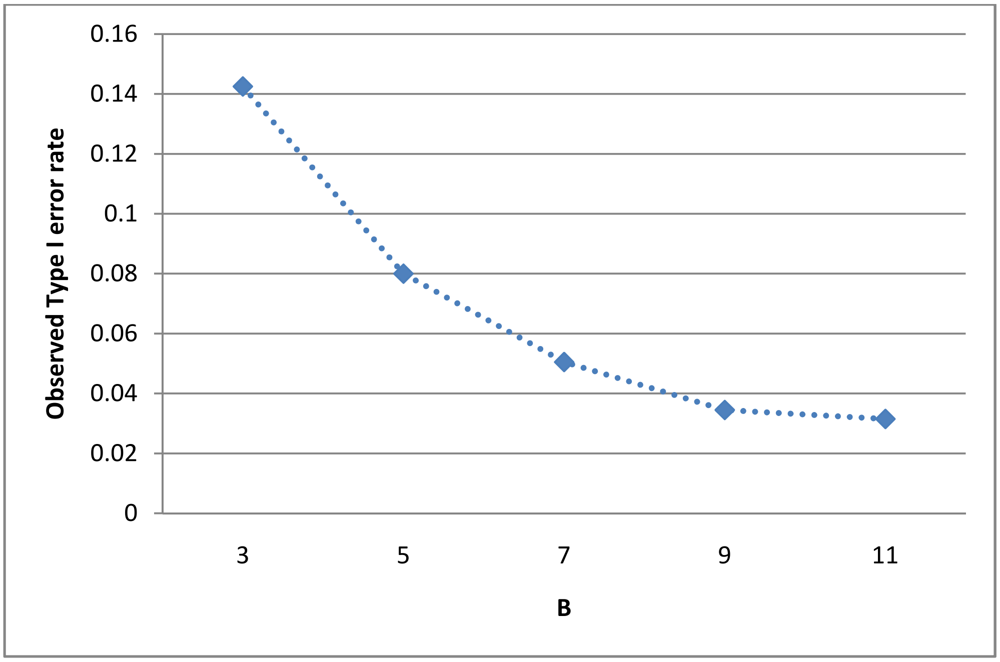
$$B_{GLR} = B\text{: } \frac{\sum(I_{(\lambda_m \geq B)})}{K} \leq \alpha \text{ , where } I_{(\lambda_{m \geq B})} = \begin{cases} 0 \text{ if } \lambda_{1m} \geq B, \text{i.e. FTR } H_0 \\ 1 \text{ if } \lambda_{0m} \geq B, \text{i.e. } H_0 \text{ is rejected} \end{cases}$$

To illustrate this approach consider an experiment where independent observations x_i come from a normal distribution *f(x|μ,σ)* with known scale parameter *σ*=1, and location parameter μ such that $0 \leq \mu < \infty$. Suppose the hypothesis test of interest is of the form:

$H_0: 0 \leq \mu \leq 0.2$ vs. $H_1: \mu > 0.2$. Assume the true value of μ=0.1.

Also suppose that this test is truncated at m=1000, i.e. H_0 is not rejected, if after 1000 collected observations, no terminal decision has been made. For this example, Figure 1 shows calculated type I error rates for different values of B, for K=2000 simulations under H_0.

Figure 1. *Calculated type I error rates by the direct Monte Carlo method, for K=2000 simulations under H_0 of the SGLR test of $H_0: 0 \leq \mu \leq 0.2$ vs. $H_1: \mu > 0.2$, where $x_i \sim N(\mu=0.1, \sigma=1)$, with Boundaries B=3, 5, 7, 9 and 11.*



In the example in Figure 1, for an approximate type I error rate of α=0.035, the boundary B_{GLR} should be set at the value B=9. As in fixed sample size tests, lower error rates require larger sample sizes. Table 2 shows the average sample size used to reach a decision for the K=2000 simulations under H_0 for different values of B.

Table 2. *Observed type I error rate and average sample size used by the direct Monte Carlo method, for K=2000 simulations under H_0 of the SGLR test of* $H_0: 0 \leq \mu \leq 0.2$ *vs.* $H_1: \mu > 0.2$*, where* $x_i \sim N(\mu=0.1, \sigma=1)$*, with Boundaries B=3, 5, 7, 9 and 11.*

B	Observed Type I error rate	Sample size used, m	
		Average	Std. Dev.
3	0.143	112.32	174.04
5	0.080	197.55	245.34
7	0.051	265.06	278.62
9	0.035	315.33	299.19
11	0.032	345.61	311.54

Chan & Lai (2005) also propose the use of importance sampling for calculating the decision boundary B_{GLR}. This is a technique that consists of giving weights (i.e. importance) to different values of the parameter space. Thus, if there is prior knowledge of values most likely to be the true values of the parameter, then greater weights are given to these values. The parameter weights with greater values have higher probability of being used in generating the K simulated samples. In Bayesian analysis, importance sampling is often used to estimate posterior densities or posterior expectations in probabilistic models that are difficult to treat analytically. In the Bayesian setting, specified prior distributions of the parameters are used as weight functions.

Chan & Lai (2005) also compare through simulations their proposed SGLR procedure to the Adaptive Likelihood Ratio test (ALR), an older frequentist extension of the

SPRT for composite hypotheses, proposed by Pavlov (1990). The ALR procedure has the advantage of using a simple decision boundary in terms of the type I error probability, α. The ALR procedure is as follows:

As with the SGLR, let *f(x|θ)* be the distribution function of a random variable *X* with parameter *θ*. Let Θ be the parameter space of θ.

The hypothesis of interest is: H_0: $\theta \in \Theta_0$ vs. H_1: $\theta \in \Theta_1$, where Θ_0 is the parameter space restricted to the null hypothesis, and Θ_1 is the parameter space restricted to the alternative hypothesis. Let $\Theta_1 \cap \Theta_0 = \emptyset$, and $\Theta_1 \cup \Theta_0 = \Theta$.

At each observation (at the m^th observation), calculate

$$\lambda_{0m}^* = \frac{\prod_{i=2}^{m} f(x_i|\theta_{i-1})}{\prod_{i=2}^{m} f(x_i|\hat{\theta}_0)} \quad \text{and} \quad \lambda_{1m}^* = \frac{\prod_{i=2}^{m} f(x_i|\theta_{i-1})}{\prod_{i=2}^{m} f(x_i|\hat{\theta}_1)}$$

where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the restricted MLEs of θ under the parameter spaces defined by H_0 and H_1, respectively, and θ_{i-1} is an estimate of θ calculated with the x_1,.., x_{i-1} for each *i*. The decision rules are defined as:

Reject H_0 (conclude H_1) if $\lambda_{0m}^* \geq B_{ALR}$,

FTR H_0 (accept H_0) if $\lambda_{1m}^* \geq B_{ALR}$.

The decision boundary B_{ALR} is simply $1/\alpha$, where α is the type I error probability. Pavlov (1990) shows that λ_{0m}^* under H_0 is a stochastic process known as a non-negative submartingale with expected value under H_0 of 1. Due to a property of non-negative submartingales known as Doob's inequality we have:

$P_{\Theta_0}(\lambda_{0m}^* > B) \leq \frac{E_{\Theta_0}(\lambda_{0m}^*)}{B} = \frac{1}{B}$. Thus setting $\frac{1}{B} = \alpha$ provides a bound in terms of the desired type I error rate.

The simple decision boundary $B_{ALR} = 1/\alpha$, of the ALR test makes it somewhat easier to apply than the SGLR procedure. However, Chan & Lai (2005) criticize the ALR procedure for restricting each x_i in the numerator of λ_{0m}^* and λ_{1m}^* to be associated only with θ_{i-1} even though m observations have already been collected. Chan & Lai (2005) describe the ALR procedure as “contrived and inefficient”. In addition, Chan & Lai’s (2005) simulation results suggest that the SGLR procedure results in considerable savings in sample size compared to the ALR procedure.

Casella & Berger (2000) provide an asymptotic approximation for a fixed sample size test of the form H$_0$: $\theta = \theta_0$ vs. H$_1$: $\theta \neq \theta_0$, using a Generalized Likelihood Ratio (GLR) statistic. This two-sided test can be approximated asymptotically relying on the fact that this test statistic converges asymptotically in distribution to a chi-square distribution with 1 degree of freedom. This can be symbolically represented by:

$$-2\log\lambda(\underline{x}) \xrightarrow{D} \chi_{(1)}^2 \,;\ \lambda(\underline{x}) = \frac{\prod_{i=1}^{m} f(x_i|\theta_0)}{\prod_{i=1}^{m} f(x_i|\hat{\theta})} \,;\ \hat{\theta} = \mathrm{MLE}(\theta).$$

This asymptotic approximation is useful because when a closed form of the GLR statistic $\lambda(\underline{x})$ cannot be analytically obtained, the MLE of θ can usually be computed numerically and thus the test statistic $\lambda(\underline{x})$ can be obtained from the observed data. For a significance level α, the asymptotic approximation rejects H$_0$ if $\lambda(\underline{x}) \leq \exp\left(\frac{\chi_{(1)1-\alpha}^2}{-2}\right)$, where $\chi_{(1)1-\alpha}^2$ is the (1-α)$^{\text{th}}$ percentile of the chi-square distribution with 1 degree of freedom. For example, at a significance level $\alpha \approx 0.05$, H$_0$ is rejected if $\lambda(\underline{x}) \leq \frac{1}{7}$. However, for the asymptotic approximation to be valid, four regularity conditions must hold. The first condition is that the x_i be iid, i.e. independent and identically-distributed. The second condition is that range of the x_i must not depend upon θ. The third condition is that θ_0 is an interior point of the parameter space of θ. The fourth condition is that $f(x|\theta)$ is differentiable in θ.

Apart from the asymptotic approximation to the GLR statistic discussed immediately above, up to this point this review has discussed an area of sequential analysis referred to as 'fully sequential procedures', i.e. sequential procedures based on likelihood ratios calculated on accumulating independent observations. In the paragraphs below, this review focuses on a different area of sequential analysis referred to as 'repeated significance tests'. In a repeated significance test, at each analysis, a non-sequential test is applied to the data collected up to that point. The null hypothesis is rejected if the non-sequential test statistic is significant at a modified significance level that accounts for multiple looks at the data.

Through the widespread use of repeated significance tests, one of the areas where sequential analysis methods have been most influential is clinical trials. In clinical trials, repeated significance tests are often referred to as 'group sequential' tests. Jennison & Turnbull (2000) provide a short history of sequential analysis in clinical trials. Fully sequential plans in the medical field were pioneered in the 1950's. However, these methods did not receive widespread acceptance, perhaps because the assessment of study results after each collected observation was considered impractical. Later in the 1970's group sequential designs with small numbers of interim analyses were introduced using repeated significance tests.

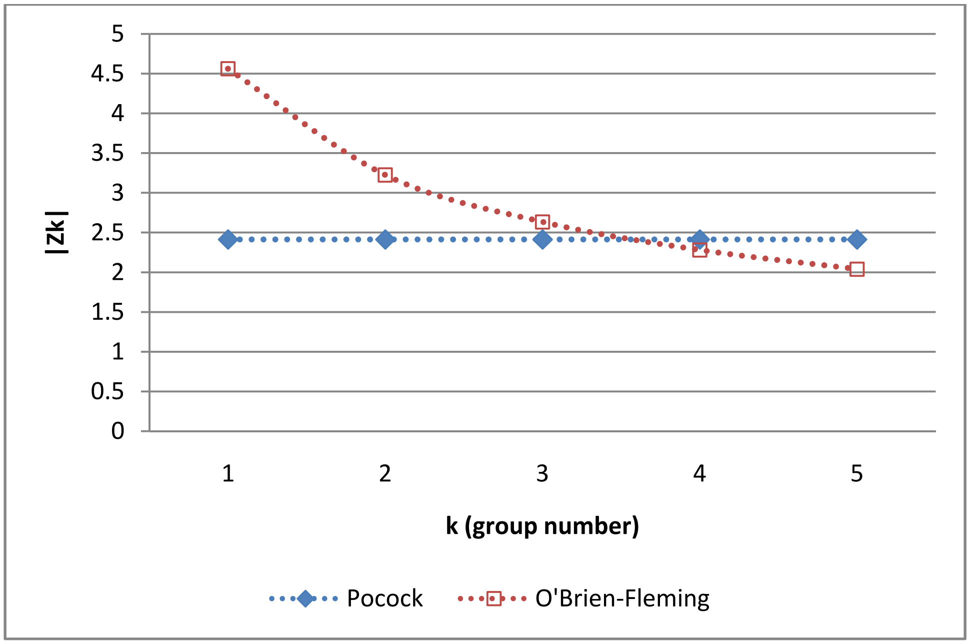
Major impetus for group sequential methods came after Pocock (1977) provided clear and easy guidelines for the implementation of group sequential experimental designs that attained approximate type I error rates and statistical power requirements similar to the fixed-sample size approach. In Pocock's approach, patient entry is divided into *k*=1, 2, .., *K* equally sized groups containing m subjects on each treatment. Assuming that the responses of subjects allocated to two treatments, A and B, are distributed $x_{Ai} \sim N(\mu_A, \sigma)$ and $x_{Bi} \sim N(\mu_B, \sigma)$, for testing the null hypothesis of no treatment difference H_0: $\mu_A = \mu_B$ vs. H_1: $\mu_A \neq \mu_B$, Pocock's test rejects H_0 if the absolute value of the standardized test statistic $Z_k = \frac{1}{\sqrt{(2mk\sigma^2)}}(\sum_{i=1}^{mk} x_{Ai} - \sum_{i=1}^{mk} x_{Bi})$, calculated after the observations on each k group of patients have been collected, is greater than a constant $C_P(K, \alpha)$, such that $P_{\mu_A=\mu_B}\{\cup_{k=1}^{K}(|z_k| > \ C_P(K,\alpha))\} = \alpha$. The constant $C_P(K, \alpha)$, is calculated numerically, using the joint distribution of the sequence of z_k. The numerical calculations are described in detail by Jennison & Turnbull (2000).

Shortly after Pocock's paper, O'Brien & Fleming (1979) propose a class of group sequential tests with conservative stopping significance levels at early analyses and a decision rule similar to the fixed sample size test if the last stage is reached. Thus, with the O'Brien-Fleming test it is more difficult to reject H_0 at the earliest analyses, but easier later on, a feature that turned out to be very appealing to practitioners (Jennison & Turnbull, 2000). For a testing problem similar to that described for Pocock's test, the O'Brien-Fleming test rejects H_0 if the absolute value of the standardized test statistic Z_k is greater than a critical value $C_{O-B}(K,\alpha)\sqrt{K/k}$. The value of $C_{O-B}(K,\alpha)$ is calculated numerically, such that:

$$P_{\mu_A=\mu_B}\{\cup_{k=1}^{K}(|z_k| > \ C_{O-B}(K,\alpha)\sqrt{K/k})\} = \alpha.$$

The numerical calculations are described in detail by Jennison & Turnbull (2000). Figure 2 shows an example of critical values for Pocock's and O'Brien-Fleming's tests, with *K*=5 equally-sized groups of observations, and *α*=0.05.

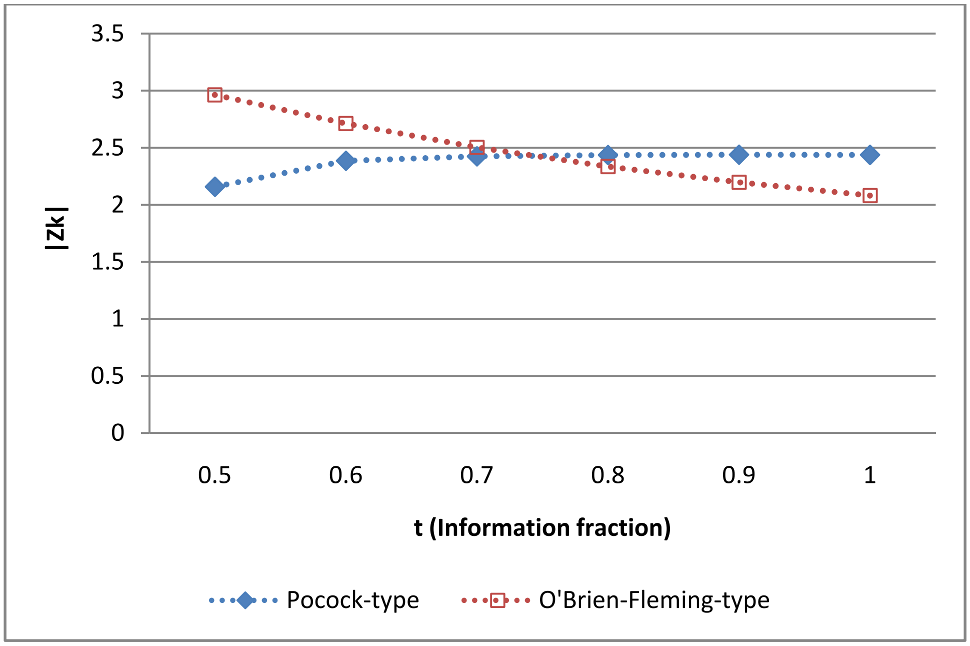
Figure 2. *Critical values for Pocock's and O'Brien-Fleming's tests, with K=5 equally-sized groups of observations, and α=0.05.*



Both Pocock's and O'Brien-Fleming's approaches require the interim analyses to follow a pre-specified schedule so that the calculated significance levels accurately protect for inflation of the type I error rate due to the multiple looks. A later paper by Lan & DeMets (1983) introduces type I error spending functions. A type I error spending function is a method for allocating the type I error probability based on the proportion of the total planned sample collected at each analysis. The error spending approach is based on modeling the sequence of test statistics as a stochastic process. This approach has the advantage of not requiring the timing of the analyses to be pre-specified, which adds flexibility to the timing and number of interim looks. In addition, the type I error spending function can take shapes similar to those of the Pocock or the O'Brien-Fleming tests.

There have been a number of suggestions for the form of the alpha spending function. Lan & DeMets (1983) show that the function $f(t) = \min\{2 - 2\Phi(\mathrm{z}_{1-\alpha/2}/\sqrt{\mathrm{t}}),\ \alpha\}$, yields critical values similar to those of the O’Brien-Fleming test when group sizes are equal (Here, t is the ‘information fraction’, i.e. the proportion of the target sample size collected; Φ is the standard normal cumulative distribution function; and $z_{1-\alpha/2}$ denotes the $(1- \alpha/2)^{\text{th}}$ percentile of the standard normal distribution). For a close analogue to the Pocock critical values, Lan & DeMets (1983) suggest $f(t) = \min\{\alpha\, log[1 + (e - 1)t], \alpha\}$. Kim & DeMets (1987) propose a family of spending functions indexed by a parameter ρ>0. The function $f(t) = \min\{at^{\rho}, \alpha\}$ with ρ=1 and ρ=3 yields tests with properties similar to those of the Pocock and O’Brien-Fleming tests, respectively. Reboussin et al. (2000) developed a software application that calculates spending functions similar to the Pocock and O’Brien-Fleming tests, among others, for user-specific information fractions. Figure 3 shows an example of critical values for Pocock-type and O’Brien-Fleming-type spending functions with information fractions *t*=0.5, 0.6, 0.7, 0.8, 0.9, 1, produced with Reboussin et al.’s (2000) software application.

Figure 3. *Pocock-type and O’Brien-Fleming-type spending functions with information fractions t=0.5, 0.6, 0.7, 0.8, 0.9, 1, produced with the software application by Reboussin et al. (2000).*



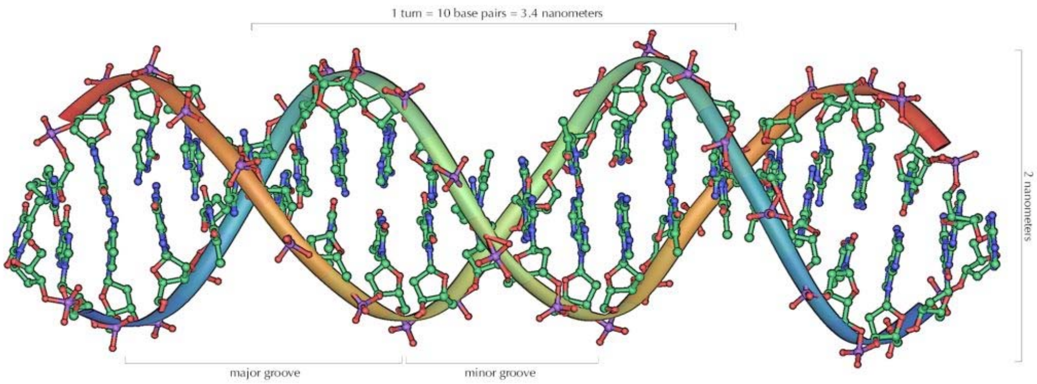
The aforementioned papers by Pocock, O'Brien and Fleming, and Lan & DeMets have been particularly influential and together form the starting point for recent methodological research and the basis of current practice in clinical trial design (Jennison & Turnbull, 2000). The properties of Pocock's and O'Brien-Fleming tests, among other recent developments, along with the error spending approach are studied in depth by Jennison & Turnbull (2000) and Proschan et al. (2006). This concludes the review of sequential statistical methodologies. However, given that the application of these methods will be in the context of genetic association studies, it is imperative that genetic terms and concepts that will be important in this dissertation are reviewed and carefully defined. This review of genetic terms is presented in the following section.

2.2 Genetic Association Studies, SNPs and Haplotype Blocks

Genetic association studies are typically conducted using a case-control design, where deoxyribonucleic acid (DNA) samples from affected cases and unaffected controls, unrelated to each other, are collected. The distribution of alleles/genotypes for the groups is compared at a set of genetic markers. Testing only a subset of candidate genetic markers from a few regions of the genome for association with disease is referred to as a candidate gene association study. In contrast, testing markers from all regions of the genome for association to a particular disease is referred to as a genome-wide association scan. Assuming that the racial admixture for each group is similar, markers that show substantial between-group differences in allele frequencies are taken to be associated with susceptibility to the disease under study (Siegmund & Yakir, 2007). The ability to obtain a large number of markers, such as single-nucleotide polymorphisms (SNPs, pronounced 'snips') on the human genome has accelerated research interests in association studies (Satagopan & Elston, 2003).

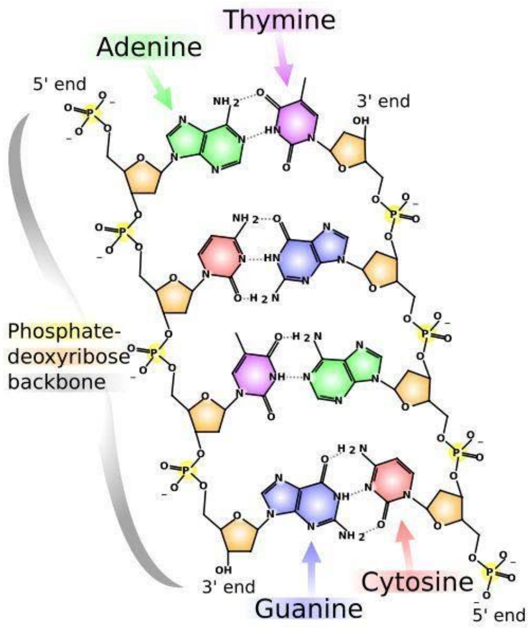
DNA is a double-helix polymer consisting of two strands wound around each other. Each strand is composed of a long chain of nucleotides. Figure 4 depicts the three-dimensional structure of DNA. A nucleotide of DNA consists of a deoxyribose sugar, a phosphate group and one of four nitrogenous bases: Adenine, Guanine, Cytosine, and Thymine (Speer, 1988). Figure 5 depicts the chemical structure of DNA.

Figure 4. *Depiction of the three-dimensional structure of DNA.*



Note: From Wikipedia.org. Permission granted to copy, distribute and/or modify this image under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation.

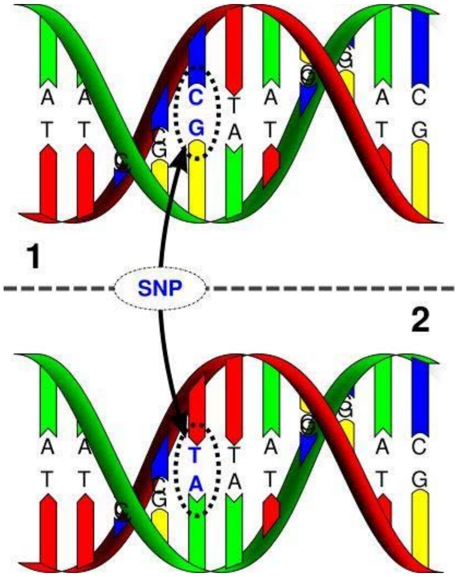
Figure 5. *Depiction of the chemical structure of DNA.*



Note: From Wikipedia.org. Permission granted to copy, distribute and/or modify this image under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation.

More than 6 billion of these bases, strung together in 23 pairs of chromosomes, exist in a human cell. However, the genetic sequences of two different individuals are remarkably similar. On average, a difference in 2 unrelated individuals' genetic sequences is observed in every 1200 bases (www.hapmap.org accessed on April 15th, 2008). A SNP involves a variation in the nucleotide at a specific location. For example, while some subjects in a population may have the base Cytosine at a given location on a particular strand of DNA, other subjects may have Thymine at the very same location in their DNA sequence. Figure 6 provides a depiction of a SNP.

Figure 6. *Depiction of SNP. A SNP is a change of a nucleotide at specific location in the DNA sequence.*



Note: From Wikipedia.org. Permission granted to copy, distribute and/or modify this image under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation.

Although in principle, SNPs may have four distinct alternative forms (one for each base) they are typically bi-allelic, i.e. they typically present only two alternatives (Siegmund & Yakir, 2007). Approximately 10 million SNPs are estimated to occur commonly in the human genome (www.hapmap.org accessed on April 15th, 2008). Since 2002, the International HapMap Project, a multi-country effort to identify and catalog genetic similarities and differences in four populations from different parts of the world, has identified approximately 3.8 million SNPs in each of these four populations (www.hapmap.org accessed on April 15th, 2008). For a particular disease, testing all 10 million common SNPs in the human genome for association with a particular disease would be extremely expensive.

Homologous recombination is the process during meiosis (cell division resulting in egg or sperm) by which an individual's homologous chromosomes (similar chromosomes in length and structure, but inherited separately from each parent) exchange DNA and form new combinations of DNA sequences to be transmitted to the individual's offspring (Speer, 1988). Examination of high-density SNP markers over contiguous regions suggested a surprisingly simple pattern. Recombination rates are not uniform over the genome, and 'cold' and 'hot' spots of recombination cause the genome to appear partitioned into blocks that are inherited as units (Daly et al. 2001; Schaid, 2004). Within each block only a few common sequences are observed (Gabriel, et al., 2002). These sequences of linked markers are known as haplotypes. Consequently, the International HapMap Project has as goals to catalog the regions that contain haplotypes (referred to as haplotype blocks) and to determine 'tag' SNPs that would identify the haplotypes within these blocks. It is believed that by identifying an individual's tag SNPs, researchers will be able to determine the collection of haplotypes in a person's DNA. It is anticipated that 300,000 to 600,000 tag SNPs will summarize most of the genetic variability within human populations (www.hapmap.org accessed on April 15th, 2008). Although the majority of current association studies are conducted using SNPs, once there is more clarity about the haplotype structure of the human genome, the haplotype mapping approach is anticipated to be more efficient than SNP mapping, which could require 10 million common SNPs.

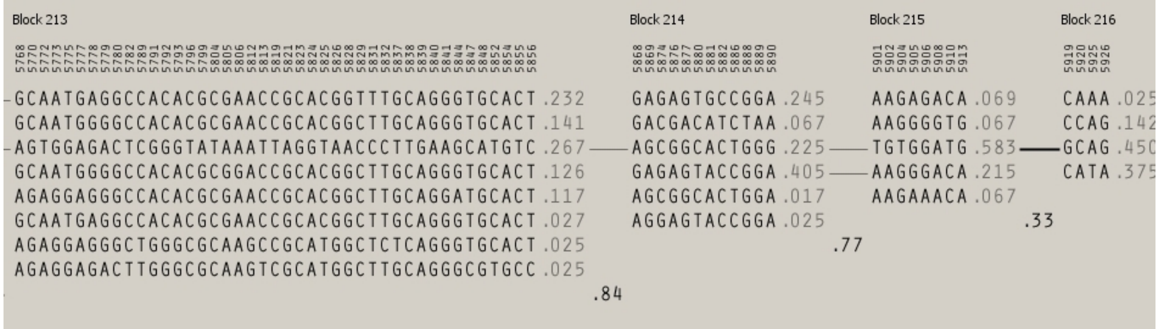
It is now anticipated that haplotypes will play a key role in the discovery and mapping of common human disorders, yet debate still remains on the likely success of haplotype-based association studies (Schaid, 2004; Terwilliger & Hiekkalinna, 2006). The focus of this debate is whether common diseases are caused by common genetic variants, and whether haplotype-block structure is a general feature of the human genome. The true test of the haplotype-map approach will come from application of a completed map to a variety of common diseases (Schaid, 2004). Another issue is that so far there is no universally accepted method to define blocks. Zhao et al. (2003) provide some examples:

1. A contiguous set of markers in which the average allelic association measure, known as D’, is greater than some threshold.
2. Regions with limited haplotype diversity and strong Linkage Disequilibrium (LD) except for a few markers.
3. Regions with absolutely no evidence for historical recombination between any pair of SNPs.

The first definition is a common definition of a haplotype block proposed by Gabriel et al. (2002) and is one of the methods implemented in the software package HaploView (Barrett et al. 2005), which allows visualization of LD and haplotype blocks from SNP marker data publicly available at the HapMap project website. In this case, LD refers to the situation in which some combinations of alleles from adjacent markers occur more or less frequently than what would be expected if the combinations were formed randomly.

For a particular set of data visualized with HaploView, the number and boundaries of the blocks inferred would depend on the threshold for allelic association measure, D', entered by the user. Figure 7 shows inferred haplotype blocks by HaploView, from a region in chromosome 22, in the sample of subjects of northern European descent collected by the HapMap Project.

Figure 7. *Screenshot of some inferred haplotype blocks in a region on chromosome 22, from the sample of subjects of northern European descent collected by the HapMap Project. The haplotype blocks are defined according to the definition by Gabriel et al. (2002) and implemented in the software package HaploView.*



```
Block 213                                                Block 214                 Block 215            Block 216
5768 5770 5772 5773 5775 5777 5778 5779 5780 5782 5789 5791 5792 5793 5796 5799 5804 5805 5806 5812 5813 5819 5821 5823 5824 5825 5826 5828 5829 5831 5832 5837 5838 5839 5840 5841 5844 5847 5848 5852 5854 5855 5856
5868 5869 5874 5876 5877 5880 5881 5882 5886 5888 5889 5890
5901 5902 5904 5905 5906 5908 5910 5913
5919 5920 5925 5926

-GCAATGAGGCCACACGCGAACCGCACGGTTTGCAGGGTGCACT .232    GAGAGTGCCGGA .245    AAGAGACA .069    CAAA .025
 GCAATGGGGCCACACGCGAACCGCACGGCTTGCAGGGTGCACT .141    GACGACATCTAA .067    AAGGGGTG .067    CCAG .142
-AGTGGAGACTCGGGTATAAATTAGGTAACCCTTGAAGCATGTC .267 ── AGCGGCACTGGG .225 ── TGTGGATG .583 ━━ GCAG .450
 GCAATGGGGCCACACGCGGACCGCACGGCTTGCAGGGTGCACT .126    GAGAGTACCGGA .405 ── AAGGGACA .215    CATA .375
 AGAGGAGGGCCACACGCGAACCGCACGGCTTGCAGGATGCACT .117    AGCGGCACTGGA .017    AAGAAACA .067
 GCAATGAGGCCACACGCGAACCGCACGGCTTGCAGGGTGCACT .027    AGGAGTACCGGA .025                  .33
 AGAGGAGGGCTGGGCGCAAGCCGCATGGCTCTCAGGGTGCACT .025                       .77
 AGAGGAGACTTGGGCGCAAGTCGCATGGCTTGCAGGGCGTGCC .025
                                                    .84
```

It is foreseeable that once large projects such as the HapMap are completed and there is more clarity about the haplotype block structure in the human genome, using blocks as markers for association scans would provide higher statistical power than scans with individual SNPs due to a drastic dimension reduction. However, about 6% of the genome sequence falls within the recombination hotspots (The International HapMap Consortium, 2007). Thus, SNPs within the hotspots will not be included in any haplotype blocks. The implication for association studies is that thorough scans will have to include individual SNPs from recombination hotspots along with haplotype blocks. Furthermore, if association is detected between a haplotype block and a disease, not only the haplotypes within the block should be further examined, but also the SNPs comprised in the block, since it is possible that either haplotypes or SNPs could be causative for the disease. This concludes the review of genetic terms and concepts used in this dissertation. Having discussed both sequential analysis concepts and genetics concepts, the following section reviews *ad hoc* sequential designs that have been proposed and published in the literature specifically for genetic association studies.

2.3 Sequential Designs in Genetic Association Studies

During the past 2 decades, the goals of sequential designs in genetic association studies have been twofold: first, to minimize genotyping costs, and second, to screen large numbers of markers. As it is further discussed in the paragraphs below, it is pertinent to note that rapid advances in 'high throughput' technologies have made genotyping costs less of a problem. The following paragraphs elaborate on sequential designs proposed during the last 15 years in the context of genetic association studies.

Sobel et al. (1993) propose a case-control sequential testing scheme with the objective of detecting association between a number candidate genetic markers and disease.

Their proposed testing scheme is as follows: calculate the sample size required to detect a target difference between cases and controls with a desired power and significance level α. This sample size of cases and controls is divided in two or three sub-samples to be used in two or three independent testing stages, respectively. After each stage, i.e. after each subsample test, two groups are selected: a group of definitive significant markers, with p-values $\leq \alpha/i$, where i is the marker number; and a group of suggestive markers, with p-values between α/i and 0.1. The suggestive genetic marker candidates are retained and then retested on the next subsample and, according to the authors, after the last stage only the true associations are likely to be retained. Sobel et al. (1993) state that their approach controls for false positive associations, while not seriously affecting power. An evident downside of their proposed step-wise multiplicity correction is that it depends on the order of the hypotheses tested. As a commentary to the sequential scheme proposed above, Sham (1994) considers that Sobel et al.’s (1993) proposed testing procedure controls for false positives, but on the other hand, decreases the overall power, since it divides the sample. However, Sham acknowledges that the advantage of Sobel et al.'s (1993) procedure is the fact that it decreases the amount of genotyping, because only a portion of the candidate markers are retained after each stage. Sham states that one way to overcome the decrease in power, under a sequential setting, is to use all existing data at the end of each stage in an overall test for association in order to reduce the effect of chance fluctuations on each independent stage. However the issue of dependency among test statistics calculated on accumulating data is not addressed by Sham. In the conclusion of his article, Sham (1994) suggests the study and adaptation of more formal sequential test procedures; in particular the works of Wald (1945) and Morton (1955) to case-control association studies.

Schaid & Sommer (1994) who collaborated with Sobel et al. (1993), as a response to Sham, justify Sobel et al.'s (1993) method by the need to control for false positives. They acknowledge that there is greater power in the combined-sample approach, but state that the probability of false positive findings is greater when combining the subsamples than when analyzing them separately. Schaid & Sommer (1994) explain that the focus of Sobel et al.'s approach is the need to control for false positive detections and that their procedure requires each stage to be adequately powered. To this date, neither Sobel, Schaid, Sommer nor Sham has published further comments regarding their respective proposed procedures. Sham did not publish any further research from his idea of using Wald's SPRT in genetic association studies, nor is he aware of other authors who have done so (Sham P., personal communication, January 23rd, 2007).

Mitchell (1995) uses simulations to examine a two-stage approach based on Sobel et al.'s (1993) approach. The objective is to find association between 360 markers and a 'rare oligogenic disease' i.e. a disease that is produced by two or more genes working together. There are two true associations by design. The simulated cohort consists of 200 cases and 100 controls, and the significance level is set at α/i, where i is the marker number and α=0.05. The results are mixed. The observed power is low (0.5), yet the observed type I error rate is relatively low also (0.01). In agreement with Sham (1994), Mitchell concludes that Sobel et al.'s (1993) approach requires large samples to attain adequate power, but that the approach could provide a reasonable strategy for screening a large number of marker-disease associations. However, Mitchell does not consider the false positive rate, i.e. the fraction of erroneous rejections among all hypotheses rejected, which, calculated from the results in that paper, is very high (0.83).

Province (2000) proposes a Sequential Selection and Ranking test for linkage, based on sequential methods published in the monograph by Bechhoffer et al. (1968). Province suggests that the methodology could be extended also to genetic association studies. He proposes using Sequential Selection and Ranking methods as an analysis technique for data that have been collected with 'more practical fixed sample designs', and then using the data not used due to early termination of the sequential analysis procedure to confirm (or refute) the positive regions detected by the sequential analysis methods. For the particular situation of a genome-wide linkage scan using the Haseman-Elston regression method, Province uses simulations of sib-pair data to compare fixed sample analyses versus a Bonferroni Corrected SPRT and a Sequential Ranking and Selection procedure. His simulation results suggest that both sequential methods on the average outperform the fixed sample designs in regards to number of the sib-pairs used to reach a decision within the specified error rates.

Boddeker & Ziegler (2001) publish a review of the literature on sequential designs in the context of genetic linkage and association studies. They begin by discussing Wald's SPRT and the linkage detection method by Morton (1955) and state that the assessment required after each observation is taken, as well as the a-priori specification of a recombination fraction for linkage, rendered these methods impractical. The fact that group sequential methods have been developed to overcome the impractical issues of the fully sequential methods is discussed. Boddeker & Ziegler (2001) refer their readers to the textbook by Jennison & Turnbull (2000). The problem of multiplicity is acknowledged in the case of two-stage genomic screenings (which will be discussed further below). Next, Boddeker & Ziegler (2001) review a group of 32 articles that they have classified as 'heuristic sequential designs', i.e. designs developed for a specific study, that are based on practical grounds but lack theoretical justification as well as conclusive evaluation of type I and type II errors. It is noted that although lacking strong theoretical basis, these 'heuristic' designs were successful in stimulating further research and confirmation studies. A second group of 19 articles are classified as 'procedures based on computer simulations'. It is commented that these studies allow evaluation of error rates within the specific simulated scenarios, yet they cannot be generalized to other settings, thus requiring simulations for every study. Finally Boddeker & Ziegler (2001) review a third group of 10 articles that they have classified as 'theoretically based procedures'. These articles propose formal sequential strategies in the context of genetic linkage or association studies. The discussion of this last group begins with the papers (cited above) by Sobel et al. (1993), Sham (1994), Schaid & Sommer (1994), and Mitchell (1995). It is commented that some studies introduce specific adjustments for type I error rates, such as Sobel et al. (1993); others aim to adapt group sequential designs for application in genetic epidemiological studies (Chotai, 1984; Muller & Ziegler, 1998) and a third approach (Elston et al., 1996; Guo & Elston, 2000) is specific for screening a large amount of markers in a first stage and then re-test in a second stage while minimizing a cost function. Boddeker & Ziegler (2001) conclude that the approaches, although promising, still require more development: the procedure by Chotai (1984) does not consider power calculations, that of Muller & Ziegler (1998) is based on erroneous sample size calculations, and the procedure by Elston et al.(1996) and Guo & Elston (2000) uses an 'inadequate cost function and power definition'. Finally, 'more formal' sequential methods are suggested, including adaptations of Wald's SPRT, as was also suggested by Sham (1994). Considering these suggestions, one of the objectives of this dissertation is to apply modern fully sequential procedures such as the SGLR test (Chan & Lai, 2005), which is a recent adaptation of Wald's SPRT.

Satagopan et al. (2002) propose a two-stage design for marker-disease association studies, that aims to minimize the amount of genotyping in a study by screening, in the first stage of testing, all markers under evaluation with a proportion of the individuals in the sample, and then in a second stage using the rest of the individuals, to validate the promising markers from stage one. Satagopan et al. (2002) avoid using significance tests and propose using absolute values of test statistics as measure of association. In their approach, the markers with the highest absolute value of test statistics at stage two would be selected as the markers most likely to be truly associated with the disease. However, the number of markers after stage two to be selected is an arbitrary number decided by the investigators conducting the association study. Assuming an asymptotically normally distributed test statistic, and a small correlation between markers, the proportions of subjects to be used and markers to be tested on each of the two stages that minimize a cost function under this model are calculated under a variety of conditions. In order to obtain 'near optimal' power compared to a one-stage design, those authors propose as a general rule using 75% of the genotyping resources in stage one to screen all markers and then using the remaining 25% genotyping resources to validate the top 10% markers ranked by magnitude of the test statistic from stage one.

It is pertinent to note that the method by Satagopan et al. (2002) aims to minimize the cost of genotyping in a study. At that time, 6 years ago, genotyping large numbers of markers on each study participant required the use of several different biological assays, since a single assay could only be used to genotype a relatively small number of markers. In addition, each assay was relatively expensive. New 'high throughput' technologies have made genotyping costs less of a problem. For instance, in 2000, a 'genome-wide' study was attempted with 600 SNPs (Mei, et al., 2000). In contrast, one of the latest SNP genotyping 'chips' produced by the company Affymetrix called "Genome-Wide Human SNP Array 6.0", can genotype over 900,000 SNPs (www.affymetrix.com, accessed on April 16th, 2008). With respect to the rapid advance of genotyping technology, Siegmund & Yakir (2007) comment: "New technologies emerge almost daily, pushing down the price and increasing the rate of the genotyping of SNPs". Although genotyping costs do not have the same importance that they had in the recent past, the procedure proposed by Satagopan et al. (2002), as well as some other procedures that will be mentioned below, have the advantage of allowing screening of large numbers of markers at stage one, which could be an advantage in regards to controlling the number of false positive detections.

Satagopan & Elston (2003) recommend the procedure by Satagopan et al. (2002) for large numbers of markers, as in genome-wide scans, to obtain similar power compared to a one-stage study but with a 45% decrease in genotyping costs. Also, Satagopan et al. (2004) consider the two-stage approach in a scenario where the number of subjects is fixed, and propose as general rule using 50% of the available subjects in stage one to screen all markers and then using the remaining 50% of the sample to validate the top 10% markers ranked by magnitude of the test statistic from stage one.

Aplenc et al. (2003) suggests the use of Group Sequential Methods in association studies due to the fact that 'early termination' may result in significant cost and sample size savings, however the approach is shown for only two markers, and the issue of multiplicity was not discussed.

Thomas et al. (2005) summarize the discussions of an international group of 165 investigators at the University of Southern California on how best to design and analyze association studies with ultra-high genotyping volume: “A broad consensus emerged that the time was now ripe for launching such studies”, and several common themes are identified such as the problem of stratification in the samples, how to incorporate environmental exposure information in the study, how to biologically validate positive association detections, and the potential efficiency gains of multistage sampling designs, specifically the two-stage approach in which only a portion of the subjects are screened with a high-density genome-wide technology, followed by testing the promising SNPs identified by the first scan on the additional subjects in the sample. Considering these common themes, one of the objectives of this dissertation is to compare and contrast the properties of the aforementioned sequential testing approaches, including the two-stage procedures, when applied to the problem of testing a relatively large number of markers in the same case-control cohort.

Writing for less sophisticated statistical readers, Cordell and Clayton (2005) provide a very readable general overview of the methods for design and analysis of genetic association studies, and compare similarities with classic epidemiological studies of environmental risk factors, but point out the importance of the design, statistical analysis and interpretation of such studies.

Wang et al. (2006) propose the two-stage approach by Satagopan & Elston (2003) as an efficient alternative to the typical one-stage approach in the context of high-dimension data from genome-wide association studies. In the same context, Skol et al. (2006) suggest a 'joint analysis' at stage two, instead of a 'replication-based' analysis (as Skol et al. termed the two-stage approach originally proposed by Satagopan et al., 2004) in order to attain greater statistical power by using the pooled sample at stage two, in the same way that Sham (1994) suggests a joint approach as an improvement to the sequential design by Sobel et al. (1993) to attain higher power. Skol et al.'s (2006) 'joint analysis' consists of using 50% of the available subjects in stage one to screen all markers and then adding the remaining 50% of the sample to validate the top 10% markers ranked by magnitude of the test statistic from stage one. Then, significance tests are to be conducted at the second stage, controlling for multiplicity with a Bonferroni correction.

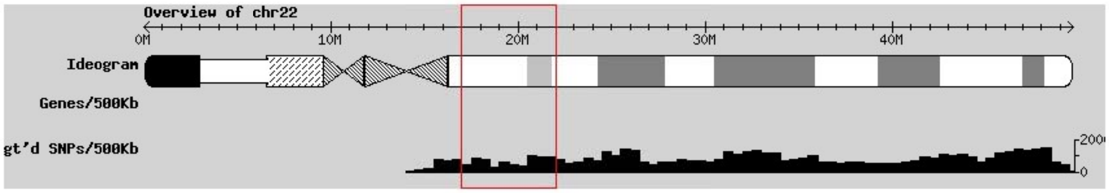
Elston & Spence (2006) discuss advances in statistical human genetics over the last 25 years. In regards to association studies, the authors optimistically note: “we are in the middle of an explosion in the development of statistical methods to detect genetic associations, just as there is currently an explosion in the molecular methods to measure genetic markers”. However, some caution is suggested: “There is presently a rush towards genome-wide association analyses, but we believe this is being driven more by the technology that is available than by any scientific rationale [...] Although genome-wide association analysis will perhaps one day be the method of choice for gene-finding, issues remain [...] This is an area in which new statistical research on both design and analysis will continue unabated because we do not yet know the best statistical framework for such studies”.

3. METHODS

3.1 An Example of Haplotype Blocks

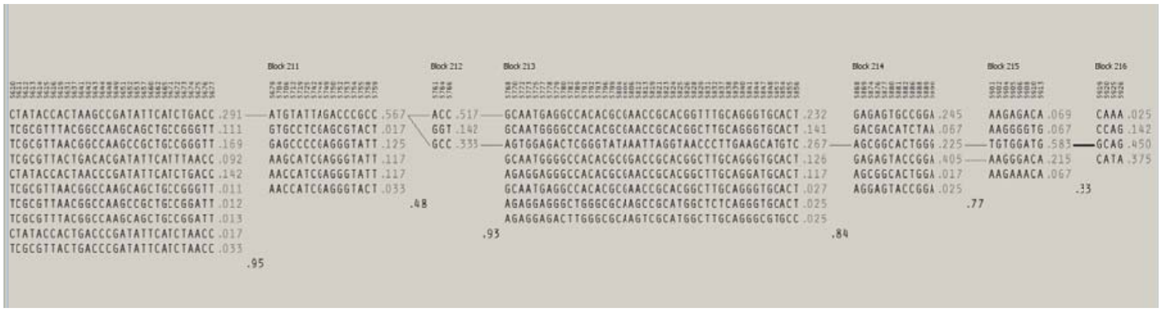
In order to examine the haplotype block structure of the human chromosome 22 in one of the populations studied by the HapMap project, a 5-million-base (5Mb) sample from a region on chromosome 22, in the population of northern European descent collected by the HapMap project, is obtained. Figure 8 shows the selected sample region on chromosome 22. This 5Mb region contains 5,929 SNPs.

Figure 8. *Genetic region from positions 17M to 22M (5Mb) on chromosome 22, in the sample of subjects of northern European descent collected by the HapMap project.*



With these data from HapMap as input, the algorithm developed by Gabriel et al. (2002) and implemented in the software HaploView (Barrett et al., 2005) identifies 216 haplotype blocks in the selected sample region (chromosome 22, positions 17M to 22M). The haplotypes in each block are also identified and their population frequencies estimated. Figure 9 shows a screen shot from HaploView with blocks 211 to 216.

Figure 9. *Haplotype blocks 211 to 216 from the region within positions 17M to 22M (5Mb) on chromosome 22, in the sample of subjects of northern European descent collected by the HapMap project. The haplotype blocks are defined according to the definition by Gabriel et al. (2002) and implemented in the software package HaploView (Barrett et al., 2005). The estimated haplotype population proportions are the numbers on the right hand side of each haplotype. The LD measure, known as D', is shown at the bottom between adjacent blocks.*

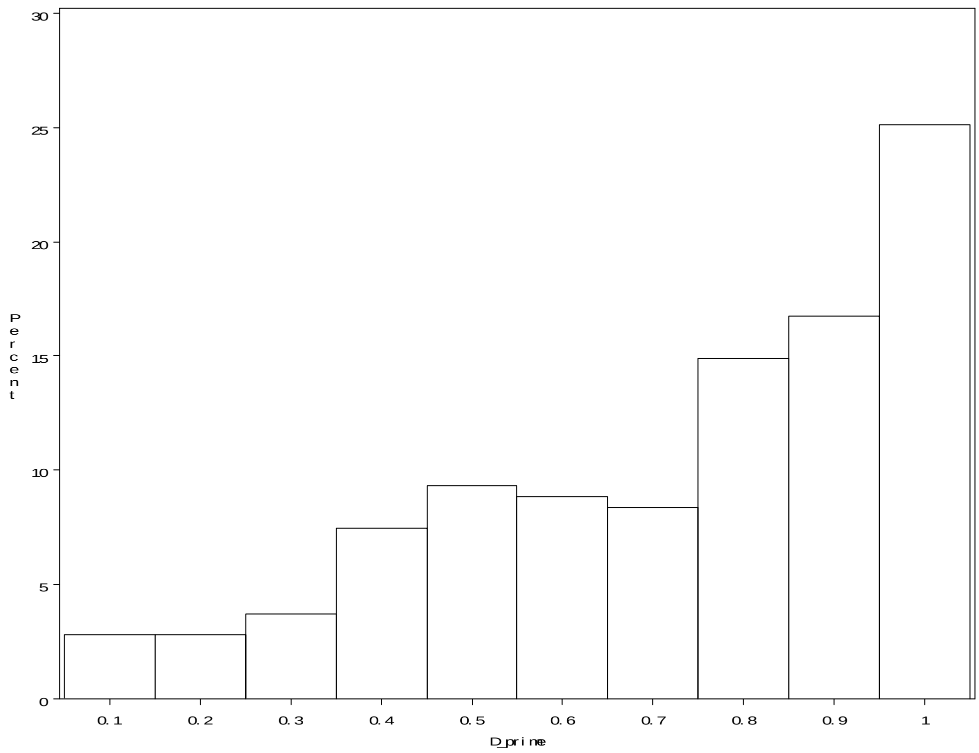


```
                                     Block 211              Block 212 Block 213                                         Block 214          Block 215      Block 216
CTATACCACTAAGCCGATATTCATCTGACC .291  ATGTATTAGACCCGCC .567  ACC .517  GCAATGAGGCCACACGCGAACCGCACGGTTTGCAGGGTGCACT .232  GAGAGTGCCGGA .245  AAGAGACA .069  CAAA .025
TCGCGTTTACGGCCAAGCAGCTGCCGGGTT .111  GTGCCTCGAGCGTACT .017  GGT .142  GCAATGGGGCCACACGCGAACCGCACGGCTTGCAGGGTGCACT .141  GACGACATCTAA .067  AAGGGGTG .067  CCAG .142
TCGCGTTAACGGCCAAGCCGCTGCCGGGTT .169  GAGCCCCGAGGGTATT .125  GCC .333  AGTGGAGACTCGGGTATAAATTAGGTAACCCTTGAAGCATGTC .267  AGCGGCACTGGG .225  TGTGGATG .583  GCAG .450
TCGCGTTACTGACACGATATTCATTTAACC .092  AAGCATCGAGGGTATT .117            GCAATGGGGCCACACGCGGACCGCACGGCTTGCAGGGTGCACT .126  GAGAGTACCGGA .405  AAGGGACA .215  CATA .375
CTATACCACTAACCCGATATTCATCTGACC .142  AACCATCGAGGGTATT .117            AGAGGAGGGCCACACGCGAACCGCACGGCTTGCAGGATGCACT .117  AGCGGCACTGGA .017  AAGAAACA .067
TCGCGTTAACGGCCAAGCAGCTGCCGGGTT .011  AACCATCGAGGGTACT .033            GCAATGAGGCCACACGCGAACCGCACGGCTTGCAGGGTGCACT .027  AGGAGTACCGGA .025             .33
TCGCGTTAACGGCCAAGCCGCTGCCGGATT .012                     .48         AGAGGAGGGCTGGGCGCAAGCCGCATGGCTCTCAGGGTGCACT .025                .77
TCGCGTTTACGGCCAAGCAGCTGCCGGATT .013                                 AGAGGAGACTTGGGCGCAAGTCGCATGGCTTGCAGGGCGTGCC .025
CTATACCACTGACCCGATATTCATCTAACC .017                               .93                                                .84
TCGCGTTACTGACCCGATATTCATCTAACC .033
                                 .95
```

Note that the estimated haplotype population proportions might or might not add to 1. This happens because haplotypes with small frequencies (≤0.01) are not shown. Also in Figure 9, connecting lines from a haplotype in one block to a haplotype in an adjacent block are shown. These connecting lines represent the percentage of times the two connected haplotypes are inherited together. The user can select the percentages to be shown.

In the crossing areas between blocks, a value of multiallelic D' is shown. D' is a measure of linkage disequilibrium (LD) between two adjacent loci (0≤D'≤1). In this case, LD refers to the situation in which some combinations of haplotypes from adjacent blocks occur more or less frequently than what would be expected if the combinations were formed randomly. A value of D'=0 means linkage equilibrium (no association), however a value of D'=1 does not necessarily mean perfect LD, but higher values imply higher LD (non-random association). Figure 10 shows a histogram of the D' values calculated by HaploView in the 5Mb sample.

Figure 10. *Histogram of D' measures between adjacent haplotype blocks from the 216 blocks detected by HaploView, contained in 5Mb on chromosome 22, locations 17M-22M, from the HapMap northern European ancestry sample.*



The histogram in Figure 10, as well as an average D' of 0.72 (std. dev.= 0.25) calculated among the 216 blocks from the sample, suggest that adjacent haplotype blocks are in high LD. This means that there is association between adjacent blocks, which must be included in the simulations.

Table 3 shows the number of haplotypes per block in the sample of 216 blocks.

Table 3. *Tabulated number of haplotypes per block, from the 216 blocks detected by HaploView, contained in 5Mb on chromosome 22, locations 17M-22M, from the HapMap northern European ancestry sample.*

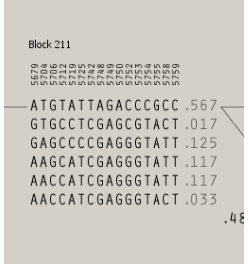
Number of Haplotypes per block	Frequency	Percent
2	29	13.43
3	49	22.69
4	35	16.20
5	38	17.59
6 or more	65	30.09
Total	216	100.0

The frequencies shown in table 3 are followed in the simulations generated for this dissertation. Further details for the simulation setup are discussed in section 3.4.

3.2 Example of a Simulated Haplotype Block

In this dissertation it is proposed to model, at a population level, a haplotype block as a multinomial random variable, where each haplotype in the block is a multinomial outcome. Figure 11 shows a parallel between block 211 from the sample of 216 blocks, and a hypothetical multinomial variable called B211.

Figure 11. *A parallel between block 211 from the 216 blocks detected by HaploView, contained in 5Mb on chromosome 22, locations 17M-22M, from the HapMap northern European ancestry sample, and a hypothetical multinomial variable called B211.*

 Block 211 5679 5704 5706 5712 5719 5725 5742 5748 5749 5750 5752 5753 5754 5755 5758 5759 ATGTATTAGACCCGCC .567 GTGCCTCGAGCGTACT .017 GAGCCCCGAGGGTATT .125 AAGCATCGAGGGTATT .117 AACCATCGAGGGTATT .117 AACCATCGAGGGTACT .033 .48	Variable B211 (7 outcomes) P(B211=”1”) = 0.567 P(B211=”2”) = 0.017 P(B211=”3”) = 0.125 P(B211=”4”) = 0.117 P(B211=”5”) = 0.117 P(B211=”6”) = 0.033 P(B211=”7”) = 0.024

In Figure 11, note that the frequencies in block 211 do not add up to 1 because haplotypes with frequencies ≤0.01 are not shown. Thus, for the hypothetical multinomial variable B211, “7” is the outcome for all the haplotypes with frequencies ≤0.01 in block 211.

In the same way, if a chromosome can be seen as a sequence of blocks in LD, then in this dissertation it is proposed to model a chromosome as a sequence of ‘correlated’ multinomial variables.

3.3 Algorithm for Simulating ‘Correlated’ Multinomial Variables

Sabatti et al. (2003) and Satagopan et al. (2004) use autoregressive structures as simplified models of dependency in order to account for LD-induced correlation between adjacent genetic markers. Satagopan et al. (2004) specifies an autoregressive structure directly on simulated test statistics, whereas Sabatti et al. (2003) specifies an autoregressive structure on the joint distribution between each pair of adjacent bi-allelic markers. In Sabatti et al.’s (2003) approach, a value of ρ=0 indicates independent markers, i.e. markers in linkage equilibrium; values of 0<ρ≤0.1 indicate a scenario where markers are in low LD; values of 0.1<ρ≤0.4 characterize a scenario where markers are in low to medium

LD; values of $0.4<\rho\leq 0.8$ indicate a scenario where markers are in medium to high LD; and values of $\rho>0.8$ characterize a scenario where markers are in high LD.

A common method to generate random multinomial realizations from a multinomial variable consist of generating uniform(0,1) variables and then transforming them into multinomial realizations by dividing the range [0,1] according to the respective probabilities of each multinomial outcome. For simulating high LD between adjacent haplotype blocks (represented by multinomial variables) in this dissertation it is proposed to specify an autoregressive structure ($\rho=0.8$) on a matrix of randomly generated uniform(0,1) variables and then transform these highly correlated uniform(0,1) variables into multinomial observations. As a consequence of the correlation among the underlying uniform(0,1) variables, some combinations of the resulting multinomial outcomes (representing haplotypes) from adjacent multinomial variables (representing blocks) occur more or less frequently than what would be expected if the combinations were formed randomly, thus providing a simplified model for LD. The proposed algorithm is explained in detail in the following paragraphs.

A common method to generate *k* correlated random standard normal variables, with *n* observations for each variable, given a symmetric moment correlation matrix $\mathbf{C}_{k*k}$, consists of finding an upper triangular matrix $\mathbf{D}_{k*k}$ such that $\mathbf{D}^T\mathbf{D}=\mathbf{C}$, where $\mathbf{D}$ is calculated by an Eigenvalue decomposition or a Cholesky decomposition (for positive definite $\mathbf{C}$). Then, after generating a matrix of uncorrelated random normal standard variables $\mathbf{R}_{n*k}$, the matrix $\mathbf{(RD)}_{n*k}$ yields a matrix of *k* standard normal variables with *n* observations, having the specified moment correlation structure among its *k* columns. In order to extend this method to non-normal variables, Phoon et al. (2004) consider the fact that for uniform(0,1) variables the moment correlation is equal to the fractile or quantile correlation. In Phoon et al's (2004) approach the initial step is to select a fractile or quantile correlation matrix $\mathbf{F}_{k*k}$ and then transform this matrix to a moment correlation matrix $\mathbf{C}_{k*k}$ by $c_{ij} = 2\sin(\frac{\pi}{6} f_{ij})$. This transformation, derived by Hotelling & Pabst (1936), applies only to standard normal variables. Next, the matrix **D** is calculated using an Eigenvalue decomposition. Then, after generating a matrix of uncorrelated random normal standard variables $\mathbf{R}_{n*k}$ and calculating the matrix of correlated standard normal variables $\mathbf{(RD)}_{n*k}$, a matrix $\mathbf{U}_{n*k}$ of uniform (0,1) variables is obtained by applying the probability integral transformation to **(RD)**, i.e. applying the standard normal cumulative distribution function (CDF) to each of the elements of **(RD)**. This matrix **U** has the specified quantile correlation structure among its *k* columns. Then the elements of **U** can be transformed from uniform(0,1) into a different distribution by an inverse CDF transformation. The quantile correlation is invariant to monotone transformations. Thus, as long as the inverse CDF transformation is monotone, the quantile correlation is not affected and therefore the new variables will retain the initial quantile correlation structure $\mathbf{F}_{k*k}$, regardless of their final distribution function.

The aforementioned Choleski decomposition is a matrix factorization for symmetric positive definite matrices (Bock, 1998). It results in an upper triangular matrix and lower triangular matrix that is the transpose of the upper triangular matrix. Consider the Cholesky decomposition $\mathbf{C}=\mathbf{D}^T\mathbf{D}$, where $\mathbf{C}_{k*k}$ is a symmetric positive definite matrix, $\mathbf{D}^T$ is the lower triangular matrix and **D** is the upper triangular matrix. Since $\mathbf{D}^T$ and **D** are triangular matrices, one of the features of this decomposition is that the element in row 1 column 1 of **C**, c_{11}, is factored into $\sqrt{c_{11}}$, that is:

$$\begin{bmatrix} c_{11} & \cdots & c_{1k} \\ \vdots & \ddots & \vdots \\ c_{k1} & \cdots & c_{kk} \end{bmatrix} = \begin{bmatrix} d_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ d_{k1} & \cdots & d_{kk} \end{bmatrix} \begin{bmatrix} d_{11} & \cdots & d_{1k} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_{kk} \end{bmatrix},$$

where $c_{11} = d_{11} * d_{11} + 0 * 0 + \cdots + 0 * 0 = (d_{11})^2 \Rightarrow \sqrt{c_{11}} = d_{11}$

If **C** is a symmetric positive definite correlation matrix then all the diagonal elements of **C** are equal to 1 and since c_{11}=1 then $d_{11}=\sqrt{1}$ =1. Thus the first column of **D** is made of d_{11}=1 and the remaining elements are equal to zero. The consequence of this feature when generating random variables is that if **C**=**D**^{T}**D** is a Cholesky decomposition then after generating a matrix of uncorrelated random normal standard variables $\mathbf{R}_{n*k}$, the matrix $\mathbf{(RD)}_{n*k}$ yields a matrix of *k* standard normal variables with *n* observations, having the specified moment correlation structure among its *k* columns, and the first column of $\mathbf{(RD)}_{n*k}$ is equal to the first column of $\mathbf{R}_{n*k}$. Based on this consideration, in this dissertation an algorithm is proposed that generates a long sequence of uniform(0,1) variables with an approximate autoregressive structure (ρ>0). The proposed algorithm breaks the sequence in small parts and avoids having to define one large rank correlation matrix **F** for the whole sequence. Then, this sequence of correlated uniform(0,1) variables is used to generate the multinomial variables that represent haplotype blocks. The proposed algorithm is as follows:

1. Generate a matrix $\mathbf{R}_{n*k}$ of k=5 random standard normal variables (columns) with n rows. Each column will be independent of the other columns.
2. Input the desired 5 by 5 autoregressive fractile correlation matrix $\mathbf{F}_{k*k}$ (must be symmetric, positive, definite).
3. Transform the fractile correlation matrix $\mathbf{F}_{k*k}$ into a moment correlation matrix $\mathbf{C}_{k*k}$ by $c_{ij} = 2\sin\left(\frac{\pi}{6} f_{ij}\right)$.

4. Calculate $\mathbf{D}_{k*k}$, the upper triangular Choleski decomposition matrix of the moment correlation matrix **C**, where $\mathbf{C}=\mathbf{D}^T\mathbf{D}$.
5. Postmultiply the matrix of independent standard normal variables $\mathbf{R}_{n*k}$ by the upper triangular Cholesky decomposition matrix **D**, i.e. $\mathbf{(RD)}_{n*k}=\mathbf{R}^{\mathbf{(1)}}{}_{n*k}$. The transformed set of standard normal variables will have the desired moment correlation structure, yet the first column remains unchanged, i.e. column 1 of $\mathbf{R}^{\mathbf{(1)}}{}_{n*k}$ is equal to column 1 of $\mathbf{R}_{n*k}$.

To generate the next set:

6. Generate another matrix of 5 independent random standard normal variables (columns) with n rows.
7. Take the last column (column 5) of the previous set $\mathbf{R}^{\mathbf{(1)}}{}_{n*k}$ and make it the first column of the new set, resulting in a matrix of n by 6, $\mathbf{S}_{n*k}$, k=6.
8. Input the 6 by 6 autoregressive fractile correlation matrix $\mathbf{F}_{k*k}$.
9. Transform the fractile correlation matrix $\mathbf{F}_{k*k}$ into a moment correlation matrix $\mathbf{C}_{k*k}$.
10. Calculate the upper triangular Choleski decomposition matrix $\mathbf{D}_{k*k}$ of the moment correlation matrix **C**.
11. Postmultiply the matrix of independent standard normal variables $\mathbf{S}_{n*k}$ by the upper triangular Cholesky decomposition matrix **D**, i.e. $\mathbf{(SD)}_{n*k}$ The transformed set of standard normal variables will have the desired moment correlation structure, yet the first column remains unchanged, which is the last column (column 5) of the previous set $\mathbf{R}^{\mathbf{(1)}}{}_{n*k}$.

12. Remove the first column of the new correlated set $\mathbf{(SD)}_{n*k}$ (which is the same last column of the previous set $\mathbf{R}^{\mathbf{(1)}}{}_{n*k}$), resulting in a second n by 5 matrix $\mathbf{R}^{\mathbf{(2)}}{}_{n*k}$.
13. Join both n by 5 correlated sets, resulting in a n by 10 matrix of correlated standard normal variables $\mathbf{Q}_{n*2k} = [\mathbf{R}^{\mathbf{(1)}}{}_{n*k} \mid \mathbf{R}^{\mathbf{(2)}}{}_{n*k}]$ having an approximate autoregressive correlation structure.

To generate a long sequence:

14. Repeat steps 6 to 13 as needed in order to generate the desired number of correlated standard normal variables $\mathbf{Q}_{n*Kk} = [\mathbf{R}^{\mathbf{(1)}}{}_{n*k} \mid \mathbf{R}^{\mathbf{(2)}}{}_{n*k} \mid \ldots \mid \mathbf{R}^{\mathbf{(K)}}{}_{n*k}]$.
15. Transform the correlated standard normal variables $\mathbf{Q}$ into correlated uniform(0,1) columns using the probability integral transformation, i.e. $\mathbf{U}=\Phi(\mathbf{Q})$, where Φ is the standard normal CDF. The uniform(0,1) columns of $\mathbf{U}$ will have approximately the desired autoregressive fractile correlation structure, and can be used to generate correlated multinomial variables.

The simulations conducted for testing this algorithm resulted in a minimal decrease in correlation values when compared to an approach that models the correlation structure for all the variables in a single matrix.

3.4 Simulation Set-up: Number of Blocks, Sample Size, and Differences to Detect

The simulations for this dissertation are based upon the following design. Assume a hypothetical case-control association study where the researchers seek to screen a specific human chromosome for association with a disease status in the context of a limited fixed amount of funds as well as a limited fixed number of cases and controls. Specifically assume chromosome 22 is under investigation with 400 total cases and controls being available for investigation. Under this context, the aim is to determine the effect sizes that can be detected with adequate (80%) power after correction for multiple hypothesis tests. In order to estimate the effect sizes, the first step is to determine the total number of markers to test for association; the next step is to set a significance level that accounts for the multiple tests; and the third step is to determine the minimum detectable differences that can be declared significant with adequate power (80%).

The simulation set-up for this dissertation is based on the steps above. The following paragraphs elaborate on each these steps.

3.4.1 Number of Markers to Test

In the sample of subjects of northern European descent collected by the HapMap Project, the region from positions 17M to 22M (5Mb) in chromosome 22, discussed in section 3.1, yielded 216 haplotype blocks using the algorithm by Gabriel et al. (2002) implemented in HaploView (Barrett et al., 2005). This region includes 5,292 SNPs. Assuming that this region is representative of the whole chromosome in terms of number of SNPs and number of haplotype blocks, these numbers extrapolated to the length of the whole chromosome (49.69Mb) result in approximately 58,922 SNPs and 2150 haplotype blocks for the entire chromosome. Due to the drastic dimension reduction that haplotype blocks represent compared to SNPs, blocks are selected as the 'markers' for this hypothetical study. Thus the total number of haplotypes to test in this hypothetical study is 2150. Accordingly, 2150 blocks are simulated, each block represented as a multinomial variable. For simulating high LD across blocks, for each simulation, a sequence of 2150 uniform(0,1) variables is generated using the algorithm described in section 3.3 with an autoregressive structure (ρ=0.8). Then this correlated uniform(0,1) variables are transformed into ‘correlated’ multinomial variables. Of the 2150 multinomial variables, 100 are simulated under the alternative hypothesis of different proportions of multinomial outcomes (representing haplotypes) between cases and controls. The remaining 2050 simulated blocks are simulated under the null hypothesis with equal proportions of haplotypes between cases and controls by design. It is acknowledged that in a realistic association study it is not likely to observe 100 haplotype blocks associated with disease status within a single chromosome. This relatively large number of associations is designed with the objective of obtaining precise power estimates.

3.4.2 Number of Participants and Sample Size

For a sample size of 200 cases and 200 controls, each participant contributes 2 versions of chromosome 22. Each version is inherited independently from each parent for a total sample size of 800 chromosomes (400 in the case group, 400 in the control group).

3.4.3 Significance Level

Under the null hypothesis of no association, at the traditional significance level of 0.05 for each independent hypothesis, 2150*0.05= 108 hypotheses are expected to be declared significant without any multiplicity correction. With a Bonferroni correction, the significance level goes down to 0.05/2150 $=2.32*10^{-5}$ and thus 2150*0.05/2150= 0.05 hypotheses are expected to be declared significant assuming all null hypotheses are true. However, if true associations are present, the downside of setting such strict significance level is a substantial decrease in statistical power to detect any true associations. In order to maintain power at the expense of type I errors, the hypothesis specific significance level is set to 5/2150=2.32*10^{-3} (i.e. with tolerance for 5 experiment-wise type I errors if all hypotheses tested are true null hypotheses). Thus, if all 2150 hypotheses were actually independent null hypotheses, at this significance level, 5 false positive rejections would be expected.

3.4.4 Detectable Differences

Consider a haplotype block for which at a population level 5 haplotypes have been observed. The haplotypes observed for this hypothetical block in a case-control group can be summarized in a 2 by 5 table of frequencies such as Table 4 below. In Table 4 the total number of observations N_{Total} is equal to two times the number of subjects, since each subject contributes 2 independent haplotypes to the 2 by 5 contingency table. Likewise, the row marginal totals N_{Cases} and $N_{Controls}$ are equal to two times the number of subjects in the case and control groups, respectively. The column marginal totals N_1, N_2, N_3, N_4, and N_5 are the number of haplotypes type 1, 2, 3, 4, and 5, respectively, observed in the combined case-control cohort.

Table 4. *Observed haplotype frequencies in a haplotype block.*

Disease Status	Haplotype1	Haplotype2	Haplotype3	Haplotype4	Haplotype5	Total
Case	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	$n_{1,5}$	N_{Cases}
Control	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2,4}$	$n_{2,5}$	$N_{Controls}$
Total	N_1	N_2	N_3	N_4	N_5	N_{Total}

With the data tabulated as a contingency table, a standard tool for association testing is Pearson's χ^2 test, which compares the observed table of frequencies with the table of frequencies that would be expected under the assumption of independence (i.e. no association). The general form of the test statistic is: $X^2 = \sum_i \sum_j \frac{(Observed_{i,j}-Expected_{i,j})^2}{Expected_{i,j}}$, where Observed$_{i,j}$ is the frequency at the cell in the ith row and jth column from the table of observed frequencies. In this hypothetical example, Observed$_{i,j}$ are equal to the n$_{i,j}$ in Table 4. The Expected$_{i,j}$ are calculated using the marginal totals, as shown in Table 5.

Table 5. *Expected haplotype frequencies in a haplotype block.*

Disease Status	Haplotype1	Haplotype2	Haplotype3	Haplotype4	Haplotype5	Total
Case	$\frac{N_1 * N_{Cases}}{N_{Total}}$	$\frac{N_2 * N_{Cases}}{N_{Total}}$	$\frac{N_3 * N_{Cases}}{N_{Total}}$	$\frac{N_4 * N_{Cases}}{N_{Total}}$	$\frac{N_5 * N_{Cases}}{N_{Total}}$	N_{Cases}
Control	$\frac{N_1 * N_{Controls}}{N_{Total}}$	$\frac{N_1 * N_{Controls}}{N_{Total}}$	$\frac{N_1 * N_{Controls}}{N_{Total}}$	$\frac{N_1 * N_{Controls}}{N_{Total}}$	$\frac{N_1 * N_{Controls}}{N_{Total}}$	$N_{Controls}$
Total	N_1	N_2	N_3	N_4	N_5	N_{Total}

Under the null hypothesis of no association, the distribution of the test statistic X^2 has approximately a χ^2 distribution with degrees of freedom equal to the product (I-1)(J-1), where I is the number of rows and J is the number of columns in the table, thus in this example the degrees of freedom are equal to (2-1)(5-1)=4.

For an individual Pearson's χ^2 test of hypothesis, the sample size necessary to detect significant association depends on the following quantities: a determined χ^2 effect size, the degrees of freedom, a target statistical power, and a target type I error rate. In this case, since the sample size is fixed at N=800, the significance level is previously set at 2.32*10^{-3}, and the desired power is 80%, the only quantities left to determine are the minimum χ^2 effect sizes that can be detected under these conditions. These χ^2 effect sizes for blocks having 2 to 6 haplotypes, are given in table 6:

Table 6. *Minimum χ^2 effect sizes that can be detected with N=800 observations, for blocks having 2 to 6 haplotypes.*

Haplotypes per Block	Degrees of freedom	Significance level	Power	N	Effect size, w	χ^2 value
2	1	0.00232	0.8	800	0.1374	15.1127
3	2	0.00232	0.8	800	0.1482	17.5774
4	3	0.00232	0.8	800	0.1556	19.3607
5	4	0.00232	0.8	800	0.1614	20.8294
6	5	0.00232	0.8	800	0.1662	22.1074

In table 6, the effect size, w, is related to the χ^2 value of the test statistic, and the sample size N by $w = \sqrt{\frac{\chi^2}{N}}$.

For the simulations in this dissertation, the 100 multinomial variables representing haplotype blocks associated with disease status are set up so that the differences in proportions of multinomial outcome realizations between case and control groups, when tested in contingency tables with Pearson's χ^2 tests, result in the effect sizes shown in Table 6.

For any one multinomial variable, the total sum of the probabilities of its outcome realizations must be equal to 1. This implies that when comparing frequencies of multinomial outcome realizations between groups, for a variable with 2 outcomes, a change in the proportion of one outcome results in a change of the same magnitude in the opposite direction for the other outcome's proportion. However, for a multinomial variable with 3 or more outcomes, a change in the proportion of one outcome results in a change in the proportion of at least one of the other outcomes; the magnitude of these changes depending on how many outcomes' proportions change. The more outcomes present in a variable, the more possible combinations of proportions that when tested in contingency tables with Pearson's χ^2 test would result in the same effect size. For N=800 (400 chromosomes per group), Tables 7 to 11 each show five examples of combinations of multinomial outcome proportions for 2, 3, 4, 5, and 6 haplotypes per block, respectively, that when tested in contingency tables with Pearson's χ^2 test result in the effect sizes shown in Table 6. These examples are a subset of the combinations that are used for the simulations in this dissertation.

Table 7. *Five combinations of case and control haplotype proportions for a Pearson's χ^2 test with 1 degree of freedom that result in an effect size of approximately 0.1374, with N=800 (400 cases and 400 controls).*

Combination	Control		Case	
	Hap1	**Hap2**	**Hap1**	**Hap2**
1	0.5	0.5	0.364	0.636
2	0.55	0.45	0.413	0.587
3	0.6	0.4	0.463	0.537
4	0.65	0.35	0.514	0.486
5	0.7	0.3	0.568	0.432

Table 8. *Five combinations of case and control haplotype proportions for a Pearson's χ^2 test with 2 degrees of freedom that result in an effect size of approximately 0.1482, with N=800 (400 cases and 400 controls).*

Combination	Control			Case		
	Hap1	**hap2**	**Hap3**	**Hap1**	**Hap2**	**Hap3**
1	0.4	0.36	0.24	0.282	0.36	0.358
2	0.45	0.33	0.22	0.330	0.33	0.340
3	0.5	0.30	0.20	0.380	0.30	0.320
4	0.55	0.27	0.18	0.431	0.27	0.299

5	0.6	0.24	0.16	0.483	0.24	0.277

Table 9. *Five combinations of case and control haplotype proportions for a Pearson's χ^2 test with 3 degrees of freedom that result in an effect size of approximately 0.1556, with N=800 (400 cases and 400 controls).*

Combination	Control				Case			
	Hap1	**Hap2**	**Hap3**	**Hap4**	**Hap1**	**Hap2**	**Hap3**	**Hap4**
1	0.35	0.33	0.23	0.10	0.258	0.33	0.23	0.189
2	0.4	0.30	0.21	0.09	0.308	0.30	0.21	0.182
3	0.45	0.28	0.19	0.08	0.358	0.28	0.19	0.174
4	0.5	0.25	0.18	0.08	0.409	0.25	0.18	0.166
5	0.55	0.23	0.16	0.07	0.461	0.23	0.16	0.157

Table 10. *Five combinations of case and control haplotype proportions for a Pearson's χ^2 test with 4 degrees of freedom that result in an effect size of approximately 0.1614, with N=800 (400 cases and 400 controls).*

Combination	Control					Case				
	Hap1	**Hap2**	**Hap3**	**Hap4**	**Hap5**	**Hap1**	**Hap2**	**Hap3**	**Hap4**	**Hap5**
1	0.3	0.28	0.21	0.14	0.07	0.215	0.28	0.21	0.14	0.155
2	0.35	0.26	0.195	0.13	0.065	0.264	0.26	0.195	0.13	0.151
3	0.4	0.24	0.18	0.12	0.06	0.314	0.24	0.18	0.12	0.146
4	0.45	0.22	0.165	0.11	0.055	0.365	0.22	0.165	0.11	0.140
5	0.5	0.2	0.15	0.1	0.05	0.416	0.2	0.15	0.1	0.134

Table 11. *Five combinations of case and control haplotype proportions for a Pearson's χ^2 test with 5 degrees of freedom that result in an effect size of approximately 0.1662, with N=800 (400 cases and 400 controls).*

Combination	Case						Control					
	Hap1	**Hap2**	**Hap3**	**Hap4**	**Hap5**	**Hap6**	**Hap1**	**Hap2**	**Hap3**	**Hap4**	**Hap5**	**Hap6**
1	0.25	0.263	0.188	0.15	0.113	0.038	0.177	0.263	0.188	0.15	0.113	0.111
2	0.3	0.245	0.175	0.14	0.105	0.035	0.226	0.245	0.175	0.14	0.105	0.109
3	0.35	0.228	0.163	0.13	0.098	0.033	0.275	0.228	0.163	0.13	0.098	0.107

4	0.4	0.210	0.150	0.12	0.090	0.030	0.326	0.210	0.150	0.12	0.090	0.104
5	0.45	0.193	0.138	0.11	0.083	0.028	0.376	0.193	0.138	0.11	0.083	0.101

3.5 Testing Procedures

The testing Procedures examined in this dissertation are listed in Table 1 (section 1.3). These methods include the standard Pearson's χ^2 test; the sequential methods proposed by Sobel et al. (1993); Sham (1994); Satagopan et al. (2004); Skol et al. (2006); and a modification of the SGLR test (original procedure proposed by Chan & Lai, 2005). Some adaptations are applied to these tests in order to utilize them in 1000 simulations of an association scan of chromosome 22, using haplotype blocks. The details for the aforementioned tests and the modifications applied are discussed in sections 3.5.1 to 3.5.6. The testing results from each sequential procedure are compared against each other as well as with the standard fixed-sample-size Pearson's χ^2 tests. Comparisons among procedures are made in terms of observed experiment-wise type I error rate, FPR (False Positive Rate), observed experiment-wise statistical power, and a measure of experiment-wise penalized power.

The observed experiment-wise type I error rate is calculated by the number of true null hypotheses rejected divided by the total number of true null hypotheses by design on each replication of the simulated chromosome scan. The range of the observed experiment-wise type I error rate is [0,1].

The observed experiment-wise power is calculated by the number of true alternative hypotheses rejected divided by the total number of true alternative hypotheses by design. The range of observed experiment-wise power is [0,1].

The FPR is calculated by the number of true null hypotheses rejected divided by the total number of rejections (Zakharkin et al., 2006). The range of observed FPR is [0,1].

The proposed measure of experiment-wise penalized power is calculated by subtracting the observed FPR from the observed experiment-wise power. The range for this measure is [-1,1], where the value of -1 corresponds to testing results where only null hypotheses were incorrectly rejected and no true alternative hypotheses were rejected. In contrast, the value of 1 corresponds to testing results where all the true alternative hypotheses were correctly rejected and no null hypotheses were incorrectly rejected. The purpose of the proposed measure of penalized power is to capture the overall usefulness of a procedure by incorporating into one measure both statistical power and the rate of false positive detections.

Details of the testing procedures examined in this dissertation are discussed below.

3.5.1 Pearson's χ^2 test

This test, described in detail in section 3.4, is a standard tool for association studies when the allele/genotype frequencies for case and control groups are tabulated in a contingency table. In each of the simulated case-control cohorts, for each of the 2150 multinomial variables representing haplotype blocks, a contingency table is tabulated and a Pearson's χ^2 test of association is conducted. The first applied approach is the naïve approach in which there is no correction for multiple tests, at the customary significance level of α=0.05 for each individual hypothesis. Next, well-known multiplicity adjustments are applied including: Bonferroni correction, Holm correction (Holm, 1979), Benjamini-Hochberg ‘control of False Discovery Rate’ (FDR) (Benjamini & Hochberg, 1995), and Benjamini-Yekutieli control of FDR under dependency (Benjamini & Yekutieli, 2001).

The Bonferroni and Holm corrections control a measure of error known as the family-wise error rate (FWER). The FWER is defined as the probability of incorrectly rejecting at least 1 null hypothesis. The FWER is an appropriate measure of error when there is an overriding reason to not make any incorrect rejections of null hypotheses (Sabatti, 2006). Control of the FWER is based on setting stringent significance levels on the hypotheses. The downside of setting such stringent significance levels is a substantial decrease in statistical power.

In the well-known Bonferroni correction, each p-value is compared to a significance level of α divided by the total number of hypotheses tested. In this case the Bonferroni significance level is $\alpha/2150$.

The Holm correction is a step-wise procedure in which the p-values are sorted from smallest to largest before comparing them to step-wise significance levels. To illustrate this approach, assume a study that involves testing n hypotheses. The n p-values are sorted from smallest to largest and then each ordered p-value, p-value number i, $p_{(i)}$, is compared to $\alpha/(n-i+1)$ in a step-wise manner, where i=1 corresponds to the smallest p-value and $i=n$ corresponds to the largest p-value. Beginning with i=1, if $p_{(i)}\leq\alpha/(n-i+1)$ then the corresponding null hypothesis is rejected and the procedure continues with the next p-value. The procedure stops if $p_{(i)}>\alpha/(n-i+1)$ and all null hypotheses corresponding to p-values i to n are not rejected. This approach provides a less conservative correction than the Bonferroni correction.

The Benjamini-Hochberg and Benjamini-Yekutieli procedures control a measure of error known as the False Discovery Rate (FDR). The FDR is a less stringent criterion than the FWER and is defined as the expected fraction of erroneous rejections among all hypotheses rejected i.e. $FDR = E(\frac{number\ of\ true\ null\ hypotheses\ rejected}{number\ of\ rejections})$. The FDR criterion captures the idea that if in an experiment there are a number of true alternative hypotheses present, we become more lenient toward committing a small fraction of false rejections when detecting the true alternative hypotheses, because the error from a single erroneous rejection (i.e. the FWER criterion) is not considered as crucial as the detection of true alternative hypotheses. Thus the proportion of incorrect rejections is controlled instead of the probability of a single incorrect rejection. It has been shown that adjusting for multiplicity with the FDR criterion substantially increases power compared to controlling the FWER. Another benefit of the FDR criterion is that if all hypotheses being tested are true null hypotheses, controlling the FDR is equivalent to controlling the FWER (Sabatti et al., 2003).

In Benjamini & Hochberg's approach, the first step is sorting the p-values from smallest to largest. Assuming a number of n hypotheses being tested, beginning with the smallest of the n p-values, each $p_{(i)}$ is compared to $\alpha_{(i)}=qi/n$ where the quantity q is the target FDR. If $p_{(i)}\leq\alpha_{(i)}$ the corresponding null hypothesis is rejected and the procedure continues with the next p-value. The procedure stops if $p_{(i)}>\alpha_{(i)}$ and all remaining null hypotheses from i to n are not rejected. It has been shown that the Benjamini & Hochberg procedure controls the FDR when the hypothesis tests are independent as well as when the tests are under a form of dependency technically referred to as ‘Positive Regression Dependency on each one from a Subset’ (PRDS). The formal definition of PRDS is rather technical and may appear quite arcane. Formally, the definition of PRDS is as follows:

The set D is called increasing if $x \in D$ and $y \geq x$ imply that $y \in D$ as well. The random variables X_1, …, X_n are PRDS on I_0 if, for any increasing set D, and for each $i \in I_0$:

$P(X_1, \dots, X_n \in D | X_i = x)$ is non-decreasing in x.

However, Sabatti et al. (2003) note that PRDS is nothing other than a formal requirement for what it is less formally referred to as ‘positive dependence’. A simple example of positive dependence is a group of test statistics distributed multivariate normal with all correlations greater or equal than zero (Benjamini & Yekutieli, 2001). In the context of genetic association studies, Sabatti et al. (2003) interpret positive dependence as follows: if two markers are in LD (i.e. non random association), and neither is related to the disease, the p-values of the tests conducted at each marker tend to be positively correlated.

The Benjamini & Yekutieli (2001) control of FDR under dependency is a modification to the Benjamini and Hochberg procedure that takes into account dependency types between tests other than PRDS. In this approach, the first step is sorting the p-values from smallest to largest. Beginning with the smallest of the n p-values, each $p_{(i)}$ is compared to $\alpha_{(i)} = \frac{qi}{n\sum_{i=1}^{n} 1/i} \approx \frac{qi}{n[\log(n)+\gamma]}$ where the quantity q is the target FDR and γ is a constant known as the Euler-Mascheroni constant ($\gamma \approx 0.5772$). If $p_{(i)} \leq \alpha_{(i)}$ the corresponding null hypothesis is rejected and the procedure continues with the next p-value. The procedure stops if $p_{(i)} > \alpha_{(i)}$ and all remaining null hypotheses from i to n are not rejected.

This approach provides a more conservative control of FDR than the Benjamini and Hochberg procedure, and it may lead to a substantial loss of power. However, it has been shown to control the FDR under any kind of dependency (Sabatti, 2006).

For both Benjamini & Hochberg and Benjamini & Yekutieli procedures, the target FDR level is set at q=0.05 for the simulations in this dissertation.

3.5.2 Sequential procedure proposed by Sobel et al. (1993)

Sobel et al. (1993) proposed the sample size of cases and controls be divided in two or three sub-samples to be used in two or three independent testing stages, respectively. The tests conducted at each stage for each simulated haplotype block are Pearson's χ^2 tests. In the original procedure after each stage, i.e. after each subsample test, two groups are selected: a group of definitive significant markers, with p-values $\leq \alpha/i$, where i is the marker number; and a group of suggestive markers, with p-values between α/i and 0.1. The suggestive genetic markers are then retested on the next subsample. At the last stage, after the group of definitive significant markers with p-values $\leq \alpha/i$ is selected, testing is stopped. An evident downside of this procedure's proposed step-wise multiplicity correction is that it depends on the order of the hypotheses tested. Thus a Holm-type correction is applied instead of the original proposed correction. At each stage the n p-values are sorted from smallest to largest and then each ordered p-value, p-value number i, $p_{(i)}$, is compared to $\alpha/(n\text{-}i+1)$ in a step-wise manner. A marker is selected as definitive significant if its p-value is $\leq \alpha/(n\text{-}i+1)$ and it is not tested any further. If a marker's p-value falls between $\alpha/(n\text{-}i+1)$ and 0.1, then the marker is selected as suggestive and retested on the next stage. At the last stage, after the group of definitive significant markers with p-values $\leq \alpha/(n\text{-}i+1)$ is selected, testing is stopped. For this procedure, the significance level is set at the customary α=0.05 level.

3.5.3 Sequential procedure proposed by Sham (1994)

As in Sobel et al.'s (1993) approach, the sample size of cases and controls is divided in two or three sub-samples. The tests conducted at each stage are Pearson's χ^2 tests. However, the data is used cumulatively instead of independently in the two or three testing stages. A Holm-type correction is also applied here. At each stage the n p-values are sorted from smallest to largest and then each ordered p-value, p-value number i, $p_{(i)}$, is compared to $\alpha/(n\text{-}i+1)$ in a step-wise manner. A marker is selected as definitive significant if its p-value is $\leq \alpha/(n\text{-}i+1)$ and it is not tested any further. If a marker's p-value falls between $\alpha/(n\text{-}i+1)$ and 0.1, then the marker is selected as suggestive and retested on the next stage. At the last stage, after the group of definitive significant markers with p-values $\leq \alpha/(n\text{-}i+1)$ is selected, testing is stopped. For this procedure, the significance level is set at the customary α=0.05 level.

3.5.4 Sequential procedure proposed by Satagopan et al. (2004)

In this approach 50% of the available subjects are used in stage one to screen all markers and then, at stage two, the remaining 50% of the sample is used to validate, independently, the top 10% markers ranked by magnitude of the test statistic from stage one. The tests conducted at each stage are Pearson's χ^2 tests. In the original procedure, the markers with the highest absolute value of test statistics at stage two would be selected as the markers most likely to be truly associated with the disease. However, the number of markers at the end of stage two to be selected is an arbitrary number decided by the investigators conducting the association study. Since the magnitude of the test statistic depends also on the degrees of freedom of the χ^2 distribution, in this dissertation the procedure is modified by grouping the test statistics by their degrees of freedom, and then selecting the top 10% markers ranked my magnitude of test statistic within each group. Then at stage two significance tests are used, with a Holm correction, in order to detect significant associations. For this procedure, the significance level is set at the customary α=0.05 level.

3.5.5 Sequential procedure proposed by Skol et al. (2006)

This approach consists of using 50% of the available subjects in stage one to screen all markers and then, at stage two, using the full sample to validate the top 10% markers ranked by magnitude of the absolute value of the test statistic from stage one. Significance tests are to be conducted at the second stage, controlling for multiplicity with a Bonferroni correction. In this dissertation two modifications are made to this design. First, since the magnitude of the test statistic depends also on the degrees of freedom of the χ^2 distribution, the first modification consists of grouping the test statistics by their degrees of freedom, and then selecting the top 10% markers ranked my magnitude of the test statistic within each group. Second, a Holm correction is applied instead of Bonferroni in order to attain higher power as the Bonferroni correction is too conservative.

This procedure is also considered as a screening tool with 100% of the sample on stage one. Since the main motivation for this approach, as well as Satagaopan et al.'s

(2004) procedure, is to minimize the cost of genotyping, it is assumed a highly likely future scenario where 'high throughput' technologies have made genotyping costs less of a problem. The main interest then is accurate detection of associations. Thus, tests using the full sample are conducted in order to calculate test statistics for all markers, next the top 10% markers ranked by magnitude of the test statistic are selected, and then significance testing is conducted with a Holm correction. For this procedure, the significance level is set at the customary α=0.05 level.

3.5.6 Modified version of SGLR test (Chan & Lai, 2005)

The original SGLR test is described in detail in section 2.1. Briefly, the procedure is as follows: Let $f(x|\theta)$ be the distribution function of a random variable X with parameter θ. Let Θ be the parameter space of θ. The hypothesis of interest is: H_0: $\theta \in \Theta_0$ vs. H_1: $\theta \in \Theta_1$, where Θ_0 is the parameter space restricted to the null hypothesis, and Θ_1 is the parameter space restricted to the alternative hypothesis. Let $\Theta_1 \cap \Theta_0 = \emptyset$, and $\Theta_1 \cup \Theta_0 = \Theta$.

At each observation (at the mth observation), calculate

$$\lambda_{0m} = \frac{\prod_{i=1}^{m} f(x_i|\hat{\theta})}{\prod_{i=1}^{m} f(x_i|\hat{\theta}_0)} \quad \text{and} \quad \lambda_{1m} = \frac{\prod_{i=1}^{m} f(x_i|\hat{\theta})}{\prod_{i=1}^{m} f(x_i|\hat{\theta}_1)}$$

Where $\hat{\theta}, \hat{\theta}_0$and $\hat{\theta}_1$ are respectively the unrestricted MLE of θ, the restricted MLE of θ under H_0, and the restricted MLE of θ under H_1.

Reject H_0 (conclude H_1) if $\lambda_{om} \geq B_{GLR}$;

FTR H_0 (conclude H_0) if $\lambda_{1m} \geq B_{GLR}$;

Otherwise, collect an additional observation.

The decision boundary B_{GLR} is to be calculated by a Monte Carlo approach. If after certain number of observations a terminal decision has not been reached, then the test is truncated and H_0 is not rejected.

As shown through the SGLR example discussed in section 2.1, it is straightforward to apply the SGLR test to situations involving a single stream of observed data. In a case-control scenario, as long as there is a form for the distribution of the difference between each pair of case and control observations, then the SGLR test can be easily applied. For instance, the difference between two normally distributed variables is a normal variable. In the case considered in this dissertation, using the SGLR to test a difference in proportions in case-control multinomial data presents major difficulties. First, each pair of case-control multinomial outcome realizations constitutes nominal-type data and thus it makes no sense to 'subtract' one from the other in the sense of continuous data, and then use the result of the subtraction individually in a likelihood function. Second, if the focus is on the difference in accumulating proportions instead of the outcome realizations themselves, then the proportions, calculated after each pair of multinomial outcome realizations is observed, are not independent from the previously observed proportions, a situation more suited for repeated significance testing methods than for fully sequential procedures, which assume independence among accumulating observations. Third, apart from the aforementioned issue of non-independence between accumulating proportions, as discussed in section 3.4.4 when comparing frequencies of multinomial outcome realizations between groups, for a multinomial variable with 3 or more outcomes, a change in the proportion of one outcome results in a change in the proportion of at least one of the other outcomes. Therefore, for multinomial variables with 3 or more outcomes, to test H_0:

the proportions between case and control groups are equal, vs. H_1: at least one proportion differs between case and control groups, there is no clear way to partition the parameter spaces between null and alternative spaces in a single SGLR test for all the proportions involved. Thus, since there are many combinations of proportions that can cause a significant change, more than one test would be required, which would result in multiplicity within a single test of hypothesis. In fact, to this date, the author of this dissertation has been unable to find in the literature a formal fully sequential test of association for contingency tables.

After careful consideration of the problems involved in conducting a SGLR test with case-control multinomial data, an adaptation to the procedure is proposed that consists of using the Pearson's χ^2 tests statistic to calculate the likelihood ratios required for this test, and an error spending function to account for multiple looks at accumulating data. The proposed adaptation combines two ideas proposed in the literature separately. First, in a Bayesian setting, Johnson (2005) proposes likelihood ratio tests based on common test statistics such as Pearson's χ^2 and Student's *t*. Second, Pearson's χ^2 test statistics have been used successfully in repeated significance testing since O'Brien and Fleming's (1979) seminal paper, which directly deals with repeated looks at a Pearson's χ^2 test statistic resulting from accumulating case-control dichotomous data. The proposed adaptation is not free of problems, however, as it is discussed in the following paragraphs.

In Pearson's χ^2 test, under the null hypothesis of no association, the test statistic is approximately distributed as a (central) χ^2 variable with degrees of freedom depending on the number of categories in the contingency table. Under the alternative hypothesis, the test statistic is approximately distributed as a non-central χ^2 variable with non-centrality parameter $\delta \geq 0$. Since the non-central χ^2 distribution with non-centrality parameter $\delta=0$ reduces to the (central) χ^2 distribution, the SGLR test can be set in terms of the non-centrality parameter δ. The test requires the MLE of the non-centrality parameter. Let x_i, $i=1..n$ be iid from a non-central χ^2 distribution with ν degrees of freedom and non-centrality parameter δ, the MLE of δ, $\hat{\delta}$, is given by:

$\hat{\delta} = \max\{(\sum_{i=1}^{n} x_i) - n\nu, 0\} = \max\{\bar{x} - \upsilon, 0\}$ (Saxena & Alam, 1982).

According to the simulation design for this dissertation, 100 multinomial variables representing haplotype blocks associated with disease status are set up so that the differences in proportions of multinomial outcome realizations between case and control groups, when tested in contingency tables with Pearson's χ^2 tests, result in the χ^2 values shown in Table 6. For these χ^2 values, the non-centrality parameters δ that can be observed under the alternative hypothesis of different proportions of haplotype realizations between case and control groups, for blocks having 2 to 6 haplotypes, are given in Table 12.

Table 12. *Non-centrality parameters δ that can be observed with N=800 observations, for blocks having 2 to 6 haplotypes, under the alternative hypothesis of different proportions of haplotype realizations between case and control groups.*

Haplotypes per Block	Degrees of freedom	χ^2 value	δ
2	1	15.1127	14.1127
3	2	17.5774	15.5774
4	3	19.3607	16.3607
5	4	20.8294	16.8294
6	5	22.1074	17.1074

It is assumed that a non-centrality parameter <5 is considered noise, and therefore the minimum value of a non-centrality parameter to be considered relevant is 5, which is an interior point of the parameter space of δ. Thus, the SGLR hypothesis test can be formulated as H_0: $\delta \leq \delta_0$ vs. H_1: $\delta > \delta_0$, where $\delta_0=5$.

At each observation (at the m^{th} observation), calculate

$$\lambda_{0m} = \frac{\prod_{i=1}^{m} f(x_i|\hat{\delta})}{\prod_{i=1}^{m} f_{\delta\leq\delta_0}(x_i|\hat{\delta})} \quad \text{and} \quad \lambda_{1m} = \frac{\prod_{i=1}^{m} f(x_i|\hat{\delta})}{\prod_{i=1}^{m} f_{\delta>\delta_0}(x_i|\hat{\delta})}$$

Reject H_0 (conclude H_1) if $\lambda_{om} \geq B_{GLR}$;

FTR H_0 (conclude H_0) if $\lambda_{1m} \geq B_{GLR}$;

Otherwise, collect an additional observation. If after certain maximum number of observations a terminal decision has not been reached, then the test is truncated and H_0 is not rejected.

Ideally, each 'observation' of the Pearson's χ^2 statistic is calculated sequentially after each pair of case and control multinomial observations is tabulated in an accumulating contingency table. However, for a fixed number of cells in a contingency table, the true sampling distribution of the Pearson's χ^2 statistic only converges to the chi-square distribution as the sample size increases. Low sample sizes and sparse tables result in test statistics poorly approximated to the chi-square distribution (Agresti, 2002). Therefore, the first problem of the proposed modified SGLR is convergence of the test statistic. Thus, as there is only assurance of good convergence with contingency tables including as many counts as possible, for this version of the SGLR test, the Pearson's χ^2 statistics are calculated beginning with 80% of the sample, and then at 1% increases until 100% of the sample has been utilized, for a maximum number of 21 χ^2 'observations'.

In the original SGLR procedure, the decision boundary B_{GLR} is calculated by a Monte Carlo approach. However in this case, the sequence of 21 Pearson's χ^2 'observations' is not independent, and the autocorrelation in the sequence is unknown beforehand. Thus, an accurate Monte Carlo estimate of B_{GLR} cannot be obtained with an acceptable degree of certainty. Therefore, the second problem of the proposed modified SGLR is obtaining the decision boundary B_{GLR}. Instead of a Monte Carlo approach, the approach taken in this dissertation uses two tools to obtain an estimate of the boundary B_{GLR}. The first tool is the asymptotic distribution of the GLR statistic. The second tool is an O'Brien-Fleming-type error spending function that accounts for repeated looks at a test statistic calculated on accumulating data.

Under the regularity conditions stated in section 2.1, for a fixed sample size test of the form H$_0$: $\theta = \theta_0$ vs. H$_1$: $\theta \neq \theta_0$, the GLR statistic can be approximated asymptotically by:

$$-2\log\lambda(\underline{x}) \xrightarrow{D} \chi^2_{(1)}\ ;\ \lambda(\underline{x}) = \frac{\prod_{i=1}^{m} f(x_i|\theta_0)}{\prod_{i=1}^{m} f(x_i|\hat{\theta})}\ ;\ \hat{\theta} = \mathrm{MLE}(\theta)$$

Thus, for a significance level $\alpha \approx 0.05$ using the asymptotic approximation, H$_0$ is rejected if $\lambda(\underline{x}) \leq \frac{1}{7}$. In this case, since the x_i in $\underline{x}$ are not independent, one of the regularity conditions is violated and an error spending function is needed to modify the significance level in order to account for the χ^2 'observations' arising from multiple looks at a test statistic calculated on accumulating data. Table 13 shows the nominal significance levels and respective $\lambda(\underline{x})$ values that would cross the boundary resulting from an

O'Brien-Fleming-type error spending function with 21 information fractions, beginning with 80%, calculated with the software application by Reboussin et al. (2000).

Table 13. *Nominal significance levels and respective λ(x̲) values resulting from an O'Brien-Fleming-type error spending function with 21 information fractions, beginning with 80%, calculated with the software application by Reboussin et al. (2000)*

χ^2 number	Accumulated sample per group	Information fraction	Nominal significance level	*λ(x̲)*
1	320	0.8	0.00978	1/28
2	324	0.81	0.0098	1/28
3	328	0.82	0.00983	1/28
4	332	0.83	0.00987	1/27
5	336	0.84	0.00992	1/27
6	340	0.85	0.01008	1/27
7	344	0.86	0.01029	1/26
8	348	0.87	0.01053	1/26
9	352	0.88	0.01079	1/25
10	356	0.89	0.01107	1/25
11	360	0.9	0.01137	1/24
12	364	0.91	0.01168	1/24
13	368	0.92	0.012	1/23
14	372	0.93	0.01234	1/22
15	376	0.94	0.01268	1/22
16	380	0.95	0.01303	1/21
17	384	0.96	0.01338	1/21
18	388	0.97	0.01374	1/20
19	392	0.98	0.01411	1/20
20	396	0.99	0.01449	1/19
21	400	1	0.01487	1/19

For a naïve approach in which there is no correction for multiplicity at the customary α=0.05 significance level, for each hypothesis test the boundaries B_{GLR} are set equal to the column $\lambda(\underline{x})$ in Table 13, at each of the 21 χ^2 'observations' accordingly.

For an approach with a Bonferroni correction for multiplicity, Table 14 shows the nominal significance levels and respective $\lambda(\underline{x})$ values that would cross the boundary resulting from an O'Brien-Fleming-type error spending function with 21 information fractions, beginning with 80%, calculated with the software application by Reboussin et al. (2000), with a Bonferroni correction for 2150 hypothesis tests.

Table 14. *Nominal significance levels and respective $\lambda(\underline{x})$ values resulting from an O'Brien-Fleming-type error spending function with 21 information fractions, beginning with 80%, with a Bonferroni correction.*

χ^2 number	Accumulated sample per group	Information fraction	Nominal significance level	$\lambda(\underline{x})$
1	320	0.8	$4.549*10^{-6}$	1/36651
2	324	0.81	$4.558*10^{-6}$	1/36580
3	328	0.82	$4.572*10^{-6}$	1/36473
4	332	0.83	$4.591*10^{-6}$	1/36331
5	336	0.84	$4.614*10^{-6}$	1/36156
6	340	0.85	$4.688*10^{-6}$	1/35606
7	344	0.86	$4.786*10^{-6}$	1/34909
8	348	0.87	$4.898*10^{-6}$	1/34147
9	352	0.88	$5.019*10^{-6}$	1/33359
10	356	0.89	$5.149*10^{-6}$	1/32550
11	360	0.9	$5.288*10^{-6}$	1/31727
12	364	0.91	$5.433*10^{-6}$	1/30921
13	368	0.92	$5.581*10^{-6}$	1/30131
14	372	0.93	$5.74*10^{-6}$	1/29336
15	376	0.94	$5.898*10^{-6}$	1/28582
16	380	0.95	$6.06*10^{-6}$	1/27847
17	384	0.96	$6.223*10^{-6}$	1/27150
18	388	0.97	$6.391*10^{-6}$	1/26469

19	392	0.98	$6.563*10^{-6}$	1/25804
20	396	0.99	$6.74*10^{-6}$	1/25156
21	400	1	$6.916*10^{-6}$	1/24541

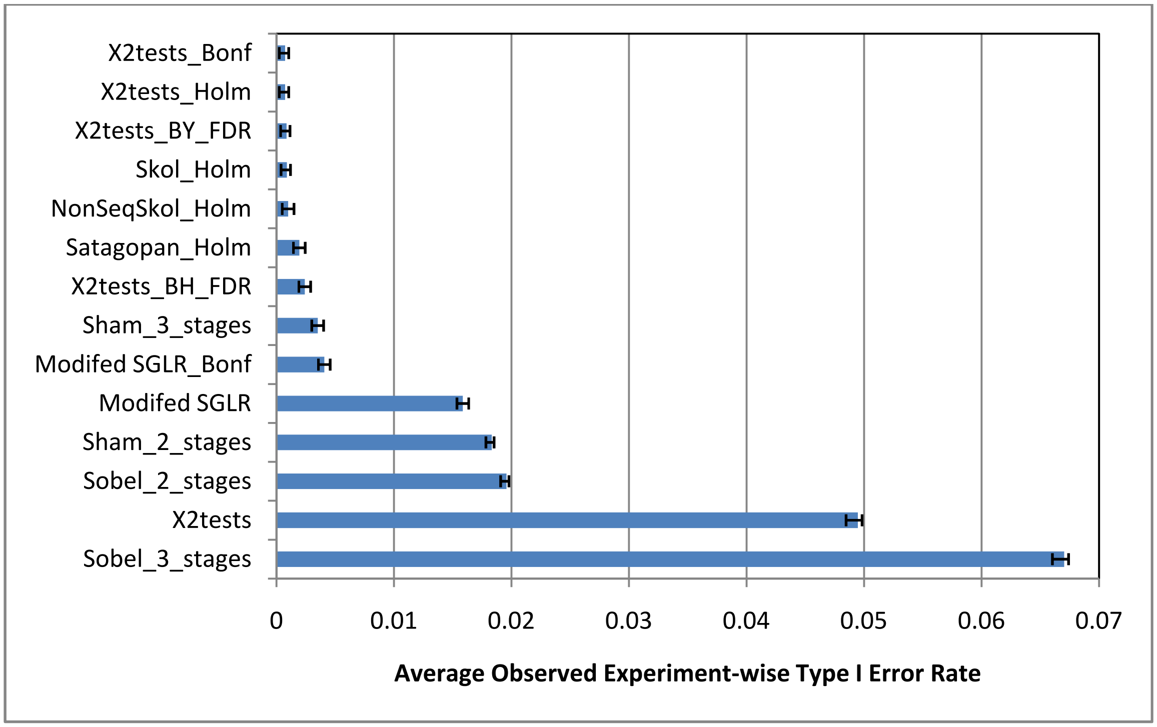
For an approach with a Bonferroni correction for multiplicity, for each hypothesis test the boundaries B_{GLR} are set equal to the column $\lambda(\underline{x})$ in Table 14, at each of the 21 χ^2 'observations' accordingly.

4. RESULTS

4.1 Graphical Comparison of Simulation Testing Results

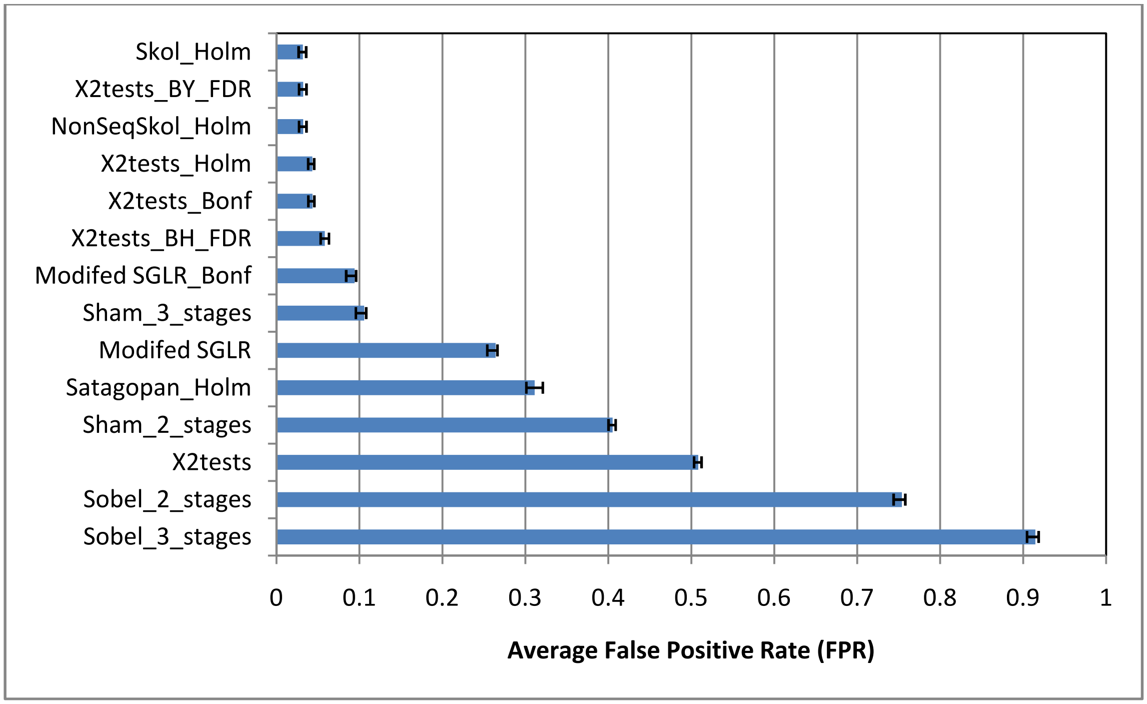
Figures 12, 13, 14, and 15 summarize graphically the simulation results in terms of observed experiment-wise type I error rate, FPR, observed experiment-wise power, and a proposed measure of penalized power, respectively.

Figure 12. *Observed experiment-wise type I error rate from 1000 simulations. For each procedure, the length of the solid bar represents the estimated mean; and the short lines extending from the solid bar represent a 95% C.I. for the mean.*



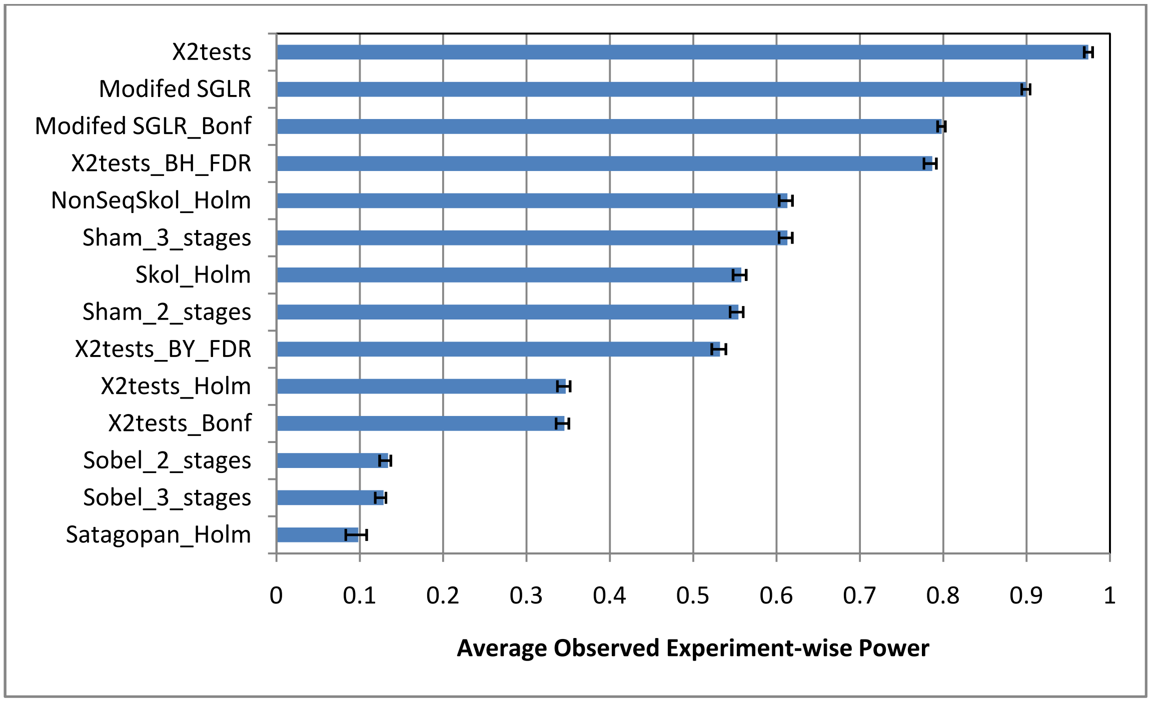
In terms of the mean observed experiment-wise type I error rate, Figure 12 shows that on the average, almost all methods resulted in low experiment-wise type I error rates. If the criterion for selecting the best procedure was experiment-wise type I error rate, then the procedure with the lowest experiment-wise type I error rate and therefore best result is Pearson's χ^2 test with a Bonferroni correction, closely followed by Pearson's χ^2 test with a Holm correction, and Pearson's χ^2 with Benjamini-Yekutieli control of FDR under dependency. As expected, control of the FWER and strong control of the FDR result in very low type I error rates. Among the sequential approaches, it is noticeable that the sequential procedure proposed by Skol et al. modified with a Holm correction produces results comparable to those of the non-sequential approaches. The mean observed experiment-wise type I error rate for this sequential procedure is slightly higher than that of the three 'best' procedures according to this performance criterion. A surprising result is the mean observed experiment-wise type I error rate of about 0.015 obtained for the modified SGLR test uncorrected for multiplicity, which is substantially lower than the expected level of 0.05. On the other hand, Pearson's χ^2 test with no multiplicity adjustment resulted almost exactly in the expected level of 0.05.

Figure 13. *False Positive Rate (FPR) from 1000 simulations. For each procedure, the length of the solid bar represents the estimated mean; and the short lines extending from the solid bar represent a 95% C.I. for the mean.*



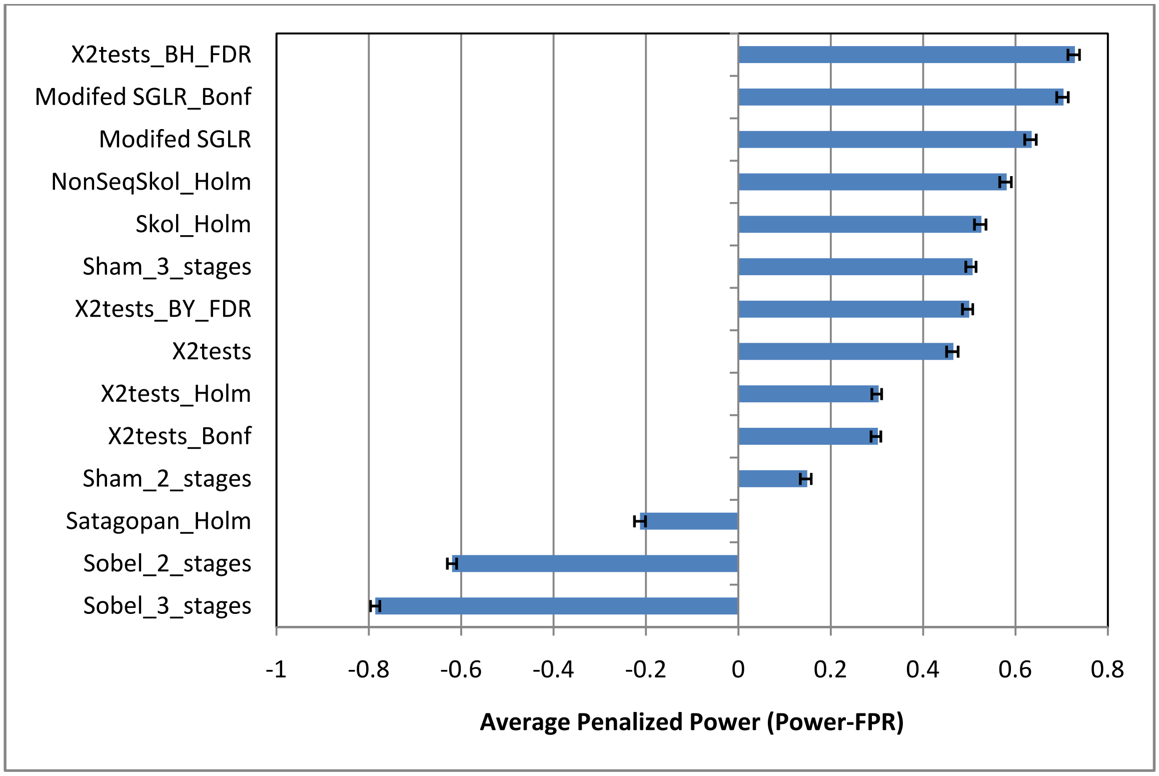
In terms of the mean FPR, Figure 13 shows that on the average, the fixed-sample-size methods with multiplicity adjustments resulted in the expected low levels, whereas the sequential procedures produced mixed results. A surprising result is the mean FPR obtained for the sequential procedure proposed by Skol et al. modified with a Holm correction. If the criterion for selecting the best procedure was FPR, then this sequential procedure resulted in the lowest FPR and therefore best result, closely followed by the non-sequential Pearson's χ^2 tests with multiplicity adjustments, thus suggesting the usefulness of this sequential method to filter out false positive detections. On the other hand, the sequential procedure proposed by Sobel et al. produced very poor results in regards to the FPR performance criterion. For instance, on the average, over 90% of the detections with Sobel et al.'s procedure with 3 stages were false detections.

Figure 14. *Observed experiment-wise power from 1000 simulations. For each procedure, the length of the solid bar represents the estimated mean; and the short lines extending from the solid bar represent a 95% C.I. for the mean.*



In terms of the mean observed experiment-wise power, Figure 14 shows that on the average, the two methods uncorrected for multiplicity, Pearson's χ^2 tests and the modified SGLR, obtained the highest power, as expected. If the criterion for selecting the best procedure was experiment-wise power, then the procedure with the highest and therefore best result is the fixed-sample-size Pearson's χ^2 uncorrected, followed by the SGLR uncorrected. A surprising result is the mean observed experiment-wise power of about 80% obtained for the modified SGLR test with a Bonferroni correction. On the other hand, Pearson's χ^2 test with Benjamini & Hochberg control of FDR adjustment resulted in substantially higher power than Pearson's χ^2 tests with FWER corrections as expected. Pearson's χ^2 tests with FWER corrections resulted in low power, as expected. It is noticeable that the sequential methods that divide the sample size in independent subsamples produced the poorest results in regards of the experiment-wise power criterion.

Figure 15. *Penalized Power (experiment-wise power-FPR) from 1000 simulations. For each procedure, the length of the solid bar represents the estimated mean; and the short lines extending from the solid bar represent a 95% C.I. for the mean.*



The purpose of the proposed measure of penalized power is to capture the overall usefulness of a procedure by incorporating into one measure the procedure's statistical power counterbalanced by its observed rate of false positive detections. In terms of this measure, Figure 15 shows that on the average, the non-sequential Pearson's χ^2 test with Benjamini & Hochberg control of FDR adjustment provided the best balance between power and false positive detection, closely followed by the modified SGLR test with a Bonferroni correction. Although this is not a completely surprising result, it is noticeable that neither of these two procedures obtained the 'best' results in the other three performance criteria examined; that is, neither of these two procedures obtained the 'best' mean observed experiment-wise type I error rate, FPR, or observed experiment-wise power. Results for the other sequential procedures were mixed. The sequential methods that divide the sample size into independent subsamples produced the poorest results in regards of the penalized power criterion. However, the sequential methods that pool the sample obtained better results in this criterion than the non-sequential procedures, with the exception of the non-sequential modification to the procedure by Skol et al. with a Holm correction. The high power obtained by Pearson's χ^2 tests uncorrected for multiplicity was counterbalanced by high FPR as expected, whereas Pearson's χ^2 tests with FWER corrections although resulted in very low FPR, also resulted in low power, thus yielding a somewhat low performance in terms of the penalized power criterion.

4.2 Detailed Simulation Results by Testing Procedure

4.2.1 Pearson's χ^2 tests, uncorrected for multiplicity

The simulation results for this testing procedure are tabulated in Table 15.

Table 15. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using Pearson's χ^2 tests with no multiplicity correction.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0.03073	0.04537	0.04927	0.04947	0.05366	0.06878	0.005778026
FPR	0.3889	0.4894	0.51	0.5084	0.5298	0.5966	0.03003555
Power	0.89	0.96	0.98	0.9741	0.99	1	0.02102801

Penalized Power	0.3263	0.44	0.466	0.4657	0.4907	0.6011	0.03976779

Each row in Table 15 shows descriptive statistics for the measure in the leftmost column. For this procedure, at a significance level of 0.05 for each individual test, the mean observed experiment-wise type I error rate of 0.0495 (Table 15, row 1) is very close, as expected, to the significance level of 0.05. The average FPR of 0.508 (Table 15, row 2) is somewhat high and indicates that on the average, approximately 50.8% of the hypotheses rejected were true null hypotheses. The average observed experiment-wise power of 0.974 (Table 15, row 3) is high as expected and indicates that on the average the procedure detected 97.4% of the simulated true alternative hypotheses. The average penalized power of 0.466 (Table 15, row 4) in this case indicates that the usefulness of the procedure to detect associations, resulting from high statistical power, is greatly counterbalanced by a high FPR. The ideal procedure would have a Penalized Power close to 1 indicating highly desirable characteristics such as high statistical power and low FPR.

4.2.2 Pearson's χ^2 tests, Bonferroni correction

The simulation results for this testing procedure are tabulated in Table 16.

Table 16. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using Pearson's χ^2 tests with a Bonferroni correction.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0	0.0004878	0.0004878	0.0007415	0.0009756	0.003415	0.000614554
FPR	0	0.02083	0.03704	0.04338	0.06522	0.2258	0.03642932
Power	0.12	0.28	0.34	0.3455	0.4	0.61	0.08504818
Penalized Power	-0.02048	0.2347	0.3068	0.3021	0.37	0.5733	0.1014953

Each row in Table 16 shows descriptive statistics for the measure in the leftmost column. For this procedure the mean experiment-wise type I error rate of $7.41*10^{-4}$ (Table 16, row 1) is very low as expected. The average FPR of 0.043 (Table 16, row 2) is low and indicates that on the average, approximately 4.3% of the hypotheses rejected were true null hypotheses. However, the mean experiment-wise power of 0.345 (Table 16, row 3) is also low, and indicates that on the average the procedure only detected 34.5% of the true alternative hypotheses. The average penalized power of 0.302 (Table 16, row 4) in this case indicates that the usefulness of the procedure to detect a high fraction of true associations is somewhat low, resulting from low statistical power, even though the procedure has the advantage of a low FPR resulting from small numbers of false detections.

4.2.3 Pearson's χ^2 tests, Holm correction

The simulation results from this testing procedure are tabulated in Table 17.

Table 17. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using Pearson's χ^2 tests with a Holm correction.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0	0.0004878	0.0004878	0.000742	0.0009756	0.003415	0.000614546
FPR	0	0.02083	0.03704	0.04321	0.06522	0.2188	0.03621273
Power	0.12	0.28	0.35	0.347	0.4	0.61	0.0852938
Penalized Power	-0.02048	0.2375	0.3074	0.3038	0.3745	0.5733	0.1015424

Each row in Table 17 shows descriptive statistics for the measure in the leftmost column. As expected, for this procedure the average observed experiment-wise type I error rate of $7.42*10^{-4}$ (Table 17, row 1) is low and slightly larger than the observed average experiment-wise type I error level of $7.41*10^{-4}$ for the Bonferroni-corrected procedure. The average FPR of 0.0432 (Table 17, row 2) is low and indicates that on the average, approximately 4.32% of the hypotheses rejected were true null hypotheses; it is approximately the same average FPR of the Bonferroni-corrected procedure. However, as with the Bonferroni-corrected procedure, the mean experiment-wise power of 0.347 (Table 17, row 3) is low and indicates that on the average the procedure only detected 34.7% of the simulated true alternative hypotheses. The average penalized power of 0.304 (Table 17, row 4) in this case indicates that the usefulness of the procedure to detect a high fraction of true associations is somewhat low, resulting from low statistical power, even though the procedure has the advantage of a low FPR resulting from small numbers of false detections.

4.2.4 Pearson's χ^2 tests with Benjamini and Hochberg's control of FDR procedure

The simulation results for this testing procedure are tabulated in Table 18.

Table 18. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using Pearson's χ^2 tests with Benjamini and Hochberg's control of FDR procedure.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0	0.001463	0.002439	0.002407	0.003415	0.006341	0.001186348
FPR	0	0.03797	0.05682	0.05826	0.07692	0.1446	0.02661439
Power	0.5	0.74	0.79	0.7868	0.84	0.98	0.07923096
Penalized Power	0.3929	0.6767	0.7344	0.7285	0.7858	0.9297	0.08344801

Each row in Table 18 shows descriptive statistics for the measure in the leftmost column. For this procedure the average observed experiment-wise type I error rate of 0.0024 (Table 18, row 1) is low but, but as expected it is larger than that for both the Bonferroni-corrected and Holm-corrected procedures. The average FPR of 0.058 (Table 18, row 2) is slightly larger than the set q-level of 0.05 expected under independence or positive dependence of the hypotheses. This average FPR of 0.058 indicates that on the average, approximately 5.8% of the hypotheses rejected were true null hypotheses. The mean observed experiment-wise power of 0.79 (Table 18, row 3) is over two times higher than that of the Bonferroni-corrected and Holm-corrected procedures and indicates that on the average the procedure detected 79% of the true alternative hypotheses. The average penalized power of 0.728 (Table 18, row 4) in this case indicates that the usefulness of the detections is not greatly counterbalanced by large numbers of false detections. Thus, in this case, this procedure results in an improvement in the balance between statistical power and false detections compared to the non-sequential uncorrected, Bonferroni-corrected, and Holm-corrected procedures. This procedure obtained the best average penalized power among all non-sequential and sequential procedures examined.

4.2.5 Pearson's χ^2 tests with Benjamini and Yekutieli control of FDR under dependency procedure

The simulation results for this testing procedure are tabulated in Table 19.

Table 19. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using Pearson's χ^2 tests with Benjamini and Yekutieli control of FDR under dependency procedure.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0	0.0004878	0.0009756	0.0008527	0.001463	0.003415	0.000671754
FPR	0	0.01587	0.02985	0.03218	0.04762	0.1613	0.02504971
Power	0.17	0.46	0.54	0.5322	0.61	0.82	0.1116895
Penalized Power	0.06474	0.4247	0.5044	0.5	0.5783	0.7847	0.1189962

Each row in Table 19 shows descriptive statistics for the measure in the leftmost column. For this procedure the average observed experiment-wise type I error rate of $8.5*10^{-4}$ (Table 19, row 1) is low but as expected, it is larger than that for the Bonferroni-corrected and Holm-corrected procedures. The average FPR of 0.032 (Table 19, row 2) is lower than the set q-level of 0.05 expected under independence or positive dependence of the hypotheses. This observed average FPR of 0.032 indicates that on the average, approximately 3.2% of the hypotheses rejected were true null hypotheses. The mean observed experiment-wise power of 0.53 (Table 19, row 3) is somewhat low and indicates that on the average the procedure detected 53% of the true alternative hypotheses. The average Penalized Power of 0.5 (Table 19, row 4) in this case indicates that the usefulness of the procedure to detect associations is somewhat low, resulting from somewhat low statistical power, even though the procedure has the advantage of a low FPR resulting from small numbers of false detections. Thus, in this case, this procedure does not result in an improvement in the balance between statistical power and false detections compared to the non-sequential procedure with Benjamini-Hochberg adjustment.

4.2.6 Sequential procedure proposed by Sobel et al. (1993), two stages

The simulation results for this testing procedure are tabulated in Table 20.

Table 20. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a modification of the sequential procedure proposed by Sobel et al. (1993) with two stages.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0.00976	0.01707	0.01951	0.01958	0.02195	0.0322	0.00331401
FPR	0.4546	0.6964	0.76	0.7539	0.8113	0.9546	0.08099078
Power	0.02	0.1	0.13	0.1338	0.17	0.39	0.05441359
Penalized Power	-0.9345	-0.713	-0.6305	-0.62	-0.5317	-0.1401	0.132517

Each row in Table 20 shows descriptive statistics for the measure in the leftmost column. For this procedure the average observed experiment-wise type I error rate of 0.019 (Table 20, row 1) is lower than the uncorrected significance level of 0.05. However, the average observed FPR of 0.754 (Table 20, row 2) is very high and indicates that on the average, approximately 75.4% of the hypotheses rejected were true null hypotheses. The average observed experiment-wise power of 0.17 (Table 20, row 3) is very low and indicates that on the average the procedure detected only 17% of the true alternative hypotheses. The average Penalized Power of -0.62 (Table 20, row 4) indicates a procedure with very low usefulness, resulting from highly undesirable properties such as low power and high FPR.

4.2.7 Sequential procedure proposed by Sobel et al. (1993), three stages

The simulation results for this testing procedure are tabulated in Table 21.

Table 21. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a modification of the sequential procedure proposed by Sobel et al. (1993) with three stages.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0.05024	0.06293	0.06732	0.06703	0.07122	0.08927	0.005997684
FPR	0.8079	0.8978	0.9156	0.9147	0.9353	0.9864	0.02868389
Power	0.02	0.1	0.13	0.1285	0.15	0.3	0.04572199
Penalized Power	-0.9664	-0.8371	-0.79	-0.7862	-0.7429	-0.518	0.07378462

Each row in Table 21 shows descriptive statistics for the measure in the leftmost column. For this procedure the average observed experiment-wise type I error rate of 0.067 (Table 21, row 1) is higher than the uncorrected significance level of 0.05. The average FPR of 0.914 (Table 21, row 2) is very high and indicates that on the average, approximately 91.4% of the hypotheses rejected were true null hypotheses. The average observed experiment-wise power of 0.13 (Table 21, row 3) is very low and indicates that on the average the procedure detected only 13% of the true alternative hypotheses. The average Penalized Power of -0.78 (Table 21, row 4) indicates a procedure with very low usefulness, resulting from highly undesirable properties such as low power and high FPR.

4.2.8 Sequential procedure proposed by Sham (1994), two stages

The simulation results for this testing procedure are tabulated in Table 22.

Table 22. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a modification to the sequential procedure proposed by Sham (1994) with two stages.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0.00829	0.0161	0.01805	0.01832	0.02049	0.03024	0.003184611
FPR	0.2208	0.3656	0.4043	0.4052	0.4444	0.5875	0.05835
Power	0.25	0.49	0.56	0.5542	0.61	0.83	0.09302497
Penalized Power	-0.2954	0.05608	0.1576	0.149	0.2452	0.5587	0.1399902

Each row in Table 22 shows descriptive statistics for the measure in the leftmost column. For this procedure the average observed experiment-wise type I error rate of 0.0183 (Table 22, row 1) is lower than the uncorrected significance level of 0.05. The average FPR of 0.405 (Table 22, row 2) is somewhat high and indicates that on the average, approximately 40.5% of the hypotheses rejected were true null hypotheses. The average observed experiment-wise power of 0.55 (Table 22, row 3) is somewhat low and indicates that on the average the procedure detected 55% of the true alternative hypotheses. The average Penalized Power of 0.15 (Table 22, row 4) in this case indicates that the usefulness of the procedure to detect true associations is low due to somewhat low power counterbalanced by a high FPR.

4.2.9 Sequential procedure proposed by Sham (1994), three stages

The simulation results for this testing procedure are tabulated in Table 23.

Table 23. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a modification to the sequential procedure proposed by Sham (1994) with three stages.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0.00049	0.00244	0.00341	0.003508	0.00439	0.0078	0.001310606
FPR	0.01333	0.07812	0.1	0.1057	0.1292	0.2388	0.03858013
Power	0.28	0.55	0.62	0.613	0.68	0.88	0.09326274
Penalized Power	0.1285	0.4336	0.5167	0.5073	0.5866	0.8055	0.1140816

Each row in Table 23 shows descriptive statistics for the measure in the leftmost column. For this procedure the average observed experiment-wise type I error rate of 0.0035 (Table 23, row 1) is lower than the uncorrected significance level of 0.05. Also, the average FPR of 0.105 (Table 23, row 2) is somewhat low and indicates that on the average, approximately 10.5% of the hypotheses rejected were true null hypotheses. The average observed experiment-wise power of 0.61 (Table 23, row 3) is however somewhat low and indicates that on the average the procedure detected 61% of the simulated true alternative hypotheses. The average Penalized Power of 0.507 (Table 23, row 4) indicates that the usefulness of the procedure to is somewhat limited, resulting from somewhat low statistical power, even though the procedure has the advantage of a low FPR resulting from low to moderate numbers of false detections.

4.2.10 Sequential procedure proposed by Satagopan et al. (2004), Holm correction

The simulation results for this testing procedure are tabulated in Table 24.

Table 24. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a modification to the sequential procedure proposed by Satagopan et al. (2004).*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0	0.00146	0.00195	0.001943	0.00244	0.00634	0.0009468
FPR	0	0.2	0.2941	0.3112	0.4	1	0.1583019
Power	0	0.06	0.09	0.09818	0.13	0.32	0.04875926
Penalized Power	-1	-0.325	-0.19	-0.213	-0.07	0.25	0.1933043

Each row in Table 24 shows descriptive statistics for the measure in the leftmost column. For this procedure the average observed experiment-wise type I error rate of 0.002 (Table 24, row 1) is lower than the uncorrected significance level of 0.05. The average FPR of 0.31 (Table 24, row 2) is somewhat high and indicates that on the average, approximately 31% of the hypotheses rejected were true null hypotheses. The average observed experiment-wise power of 0.098 (Table 24, row 3) is very low and indicates that on the average the procedure detected only 9.8% of the true alternative hypotheses. The average Penalized Power of -0.21 (Table 24, row 4) indicates a procedure with undesirable properties such as low power and high FPR.

4.2.11 Sequential procedure proposed by Skol et al. (2006), Holm correction

The simulation results for this testing procedure are tabulated in Table 25.

Table 25. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a modification to the sequential procedure proposed by Skol et al. (2006)*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0	0.00049	0.00098	0.0008894	0.00146	0.00341	0.000673201
FPR	0	0.01587	0.02985	0.03176	0.04762	0.1346	0.02383731

Power	0.28	0.4975	0.56	0.5578	0.62	0.85	0.09095859
Penalized Power	0.1957	0.46	0.5308	0.5261	0.59	0.8051	0.09823053

Each row in Table 25 shows descriptive statistics for the measure in the leftmost column. For this procedure the observed experiment-wise type I error rate of $8.8*10^{-4}$ (Table 25, row 1) is very low, about 1.17 times the observed experiment-wise type I error rate of $7.41*10^{-4}$ of the Pearson's χ^2 test with a Bonferroni correction. The average FPR of 0.031 (Table 25, row 2) is very low and indicates that on the average, approximately 3.1% of the hypotheses rejected were true null hypotheses. However, the observed experiment-wise power of 0.557 (Table 25, row 3) is somewhat low and indicates that on the average the procedure detected only 55.7% of the simulated true alternative hypotheses. The average Penalized Power of 0.526 (Table 25, row 4) is almost two times that of the non-sequential Bonferroni-corrected, and Holm-corrected procedures, and in this case indicates that even though the procedure does not attain very high power, the usefulness of the true detections is not greatly counterbalanced by large numbers of false detections. Thus, this procedure results in an improvement in the balance between statistical power and false detections compared to the non-sequential Bonferroni-corrected, and Holm-corrected procedures, and the sequential procedures by Sobel et al. (1993), Sham (1994), and Satagopan et al. (2004).

4.2.12 Non-sequential modification to the procedure proposed by Skol et al. (2006), Holm correction

The results for this testing procedure are tabulated in Table 26.

Table 26. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a non-sequential modification to the procedure proposed by Skol et al. (2006)*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0	0.0004878	0.0009756	0.0009927	0.001463	0.003902	0.000717245
FPR	0	0.01538	0.0303	0.03222	0.04688	0.122	0.02283374
Power	0.28	0.55	0.62	0.6131	0.68	0.87	0.09386667
Penalized Power	0.2133	0.5149	0.5856	0.5809	0.6504	0.85	0.1005349

Each row in Table 26 shows descriptive statistics for the measure in the leftmost column. For this procedure the observed experiment-wise type I error rate of $9.9*10^{-4}$ (Table 26, row 1) is very low, about 1.3 times the observed type I error rate of $7.41*10^{-4}$ of the Pearson's χ^2 tests with a Bonferroni correction. The average FPR of 0.032 (Table 26, row 2) is very low and indicates that on the average, approximately only 3.2% of the hypotheses rejected were true null hypotheses. The observed experiment-wise power of 0.61 (Table 26, row 3) is somewhat low and indicates that on the average the procedure detected 61% of the simulated true alternative hypotheses. The average Penalized Power of 0.458 (Table 26, row 4) is two times that of the non-sequential Bonferroni-corrected, and Holm-corrected procedures, and in this case indicates that even though the procedure does not attain very high power, the usefulness of the true detections is not greatly counterbalanced by large numbers of false detections. Thus, this modified procedure results in an improvement in the balance between statistical power and false detections compared to the non-sequential Bonferroni-corrected, and Holm-corrected procedures, and the sequential procedures by Sobel et al. (1993), Sham (1994), Satagopan et al. (2004), and the original sequential procedure proposed by Skol et al. (2006) with a Holm correction.

4.2.13 Modified SGLR test, uncorrected for multiplicity

The simulation results for this testing procedure are tabulated in Table 27.

Table 27. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a modification to the SGLR test, uncorrected for multiplicity.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0.00732	0.01366	0.01561	0.01586	0.01805	0.02634	0.002951244
FPR	0.14	0.24	0.27	0.2641	0.29	0.37	0.03732244
Power	0.72	0.87	0.9	0.8991	0.93	1	0.04648739
Penalized Power	0.4022	0.595	0.638	0.6349	0.6802	0.8136	0.06602789

Each row in Table 27 shows descriptive statistics for the measure in the leftmost column. For this procedure the observed experiment-wise type I error rate of 0.015 (Table 27, row 1) is lower than the uncorrected significance level of 0.05. The average FPR of 0.264 (Table 27, row 2) is moderate and indicates that on the average, approximately 26.4% of the hypotheses rejected were true null hypotheses. The observed experiment-wise power of 0.899 (Table 27, row 3) is high and indicates that on the average the procedure detected 89.9% of the simulated true alternative hypotheses. The average Penalized Power of 0.63 (Table 27, row 4) is somewhat high and indicates that the high power is somewhat counterbalanced by the false positive detections. Thus, this procedure results in an improvement in the balance between statistical power and false detections compared to the non-sequential Bonferroni-corrected, and Holm-corrected procedures, and the sequential procedures by Sobel et al. (1993), Sham (1994), Satagopan et al. (2004), and the sequential procedure proposed by Skol et al. (2006).

4.2.14 Modified SGLR test, Bonferroni correction

The simulation results for this testing procedure are tabulated in Table 28.

Table 28. *Testing results from 1000 simulated association scans of 2150 haplotype blocks using a modification to the SGLR test proposed, with a Bonferroni correction for multiplicity.*

Measure	Min	1st_Q	Median	Mean	3rd_Q	Max	Std_Dev
Type I error	0.00049	0.00293	0.0039	0.00407	0.00488	0.01073	0.001478181
FPR	0.01	0.07	0.09	0.09407	0.11	0.21	0.0317355
Power	0.53	0.76	0.8	0.7982	0.84	0.97	0.06767282
Penalized Power	0.44	0.6551	0.7091	0.7041	0.7625	0.9194	0.08039072

Each row in Table 28 shows descriptive statistics for the measure in the leftmost column. For this procedure the observed experiment-wise type I error rate of 0.0048 (Table 28, row 1) is low. The average FPR of 0.094 (Table 28, row 2) is somewhat low and indicates that on the average, approximately only 9.4% of the hypotheses rejected were true null hypotheses. The observed experiment-wise power of 0.798 (Table 28, row 3) is somewhat high and indicates that on the average the procedure detected 79.8% of the simulated true alternative hypotheses. The average penalized power of 0.7041 (Table 28, row 4) is somewhat high and indicates that the somewhat high power is counterbalanced to a small extent by the false positive detections. Thus, this procedure results in an improvement in the balance between statistical power and false detections compared to the non-sequential Bonferroni-corrected, and Holm-corrected procedures, and the sequential procedures by Sobel et al. (1993), Sham (1994), Satagopan et al. (2004), the sequential procedure proposed by Skol et al. (2006), and the uncorrected modified SGLR test.

5. DISCUSSION AND CONCLUSIONS

5.1 Summary

In order to fulfill the main objective of this dissertation, which is to 1) modify the Sequential Generalized Likelihood Ratio test (SGLR) for application in haplotype studies, and 2) compare and contrast the properties of the SGLR and four other sequential testing procedures proposed during the past 15 years, when applied to the problem of testing a relatively large number of haplotype blocks in the same case-control cohort, the following steps are taken:

1. Although current genetic association scans typically use bi-allelic SNPs as markers, it is foreseeable that once there is more clarity about the haplotype block structure in the human genome, using a haplotype mapping approach for association scans would provide higher statistical power compared to scans with individual SNPs, due to a drastic dimension reduction. It is noted that haplotype blocks can be modeled as multinomial variables at a population level, and thus chromosomes can be modeled as long sequences of multinomial variables. Therefore, assuming a case-control study, this dissertation research seeks to simulate 1000 association scans of the human chromosome 22 using haplotype blocks as genetic markers. In order to have an idea of the number of haplotype blocks found in chromosome 22 for a determined human population, a sample of SNP marker data made publicly available by the HapMap project from the human chromosome 22 is obtained. This sample encompasses approximately 10% of the length of

the chromosome and includes 5,292 SNPs. Next, one of the proposed algorithms in the literature that identify haplotype blocks within the SNP marker sample is applied. The algorithm yields 216 blocks in the sample. Then the results from the sample are extrapolated to the length of the chromosome yielding 2150 haplotype blocks (containing 58,922 SNPs) for the whole chromosome. This number of haplotype blocks is used for the simulated chromosomes. Also other characteristics of the haplotype blocks identified in the sample are tabulated, such as the numbers of haplotypes per block and a measure of LD between adjacent blocks. In this case, LD refers to the situation in which some combinations of haplotypes from adjacent blocks occur more or less frequently than what would be expected if the combinations were formed randomly. It is noted that adjacent haplotype blocks are in moderate to high LD on the average, a feature that is included in the design of the simulations. Also, frequencies of haplotypes within blocks are designed following those found in the HapMap sample.

2. To simulate observed haplotype data, this dissertation research uses a novel algorithm to generate long sequences of correlated uniform(0,1) variables with an approximate autoregressive correlation structure (ρ=0.8) and then transform the highly correlated uniform(0,1) variables into realizations of multinomial outcomes. As a consequence of the high correlation among the underlying uniform(0,1) variables, some combinations of the resulting multinomial outcomes (representing haplotypes) from adjacent multinomial variables (representing blocks) occur more or less frequently than what would be expected if the combinations were formed randomly, thus providing a simplified model for LD.

3. From the 2150 simulated multinomial variables representing contiguous blocks in chromosome 22, 100 variables are designed with differences in proportions between case and control groups that would result in a rejection of the null hypothesis of no association, when the multinomial observations from the case-control cohort are tabulated in a contingency table and tested with Pearson's χ^2 test, with a sample size of 200 individuals per group, 80% power, and at significance level of $5/2150=2.32*10^{-3}$. The remaining 2050 simulated multinomial variables are designed with the same proportions for case and control groups.

4. During the past 2 decades, the goals of sequential designs in genetic association studies have been twofold: first, to minimize genotyping costs, and second, to screen large numbers of markers. Since rapid advances in 'high throughput' technologies have made genotyping costs less of a problem, this dissertation focuses on the latter objective of sequential designs in genetic association studies. Within this context, this dissertation examines sequential designs introduced in the last 15 years, including the procedures proposed by Sobel et al. (1993); Sham (1994); Satagopan et al. (2004); Skol et al. (2006); and the SGLR test by Chan & Lai (2005). The SGLR test is modified for application in haplotype studies and some adaptations are applied to the other sequential procedures in order to utilize them in the simulations.

5. 1000 simulations of a hypothetical association scan are programmed, run and the results are tabulated. The testing results from each sequential procedure are compared against each other as well as with the standard fixed-sample-size Pearson's χ^2 tests with common multiplicity adjustments. Comparisons among procedures are made in terms of observed experiment-wise type I error rate, FPR (False Positive Rate), observed experi-

ment-wise power, and a measure of experiment-wise penalized power, defined as the subtraction of FPR from observed experiment-wise power. The purpose of the proposed measure of penalized power is to capture the overall usefulness of a procedure by incorporating into one measure the procedure's statistical power counterbalanced by its observed rate of false positive detections.

6. The comparisons among the examined methods indicated that, under the assumptions of the simulations, on the average, the non-sequential Pearson's χ^2 test with Benjamini & Hochberg control of FDR provides the best balance between power and false positive detections, closely followed by the modified version of the SGLR test with a Bonferroni correction.

5.2 Strengths and limitations of this dissertation research

5.2.1 Strengths

1. Testing methods examined are applicable to both haplotype blocks and SNP markers

As previously mentioned, in this dissertation haplotype blocks are used as genetic markers instead of the more commonly used SNPs. Based on empirical observation reported by Daly et al. (2001) and Gabriel et al. (2002), among others, of what appears as a surprisingly simple haplotype structure of the human genome, in this dissertation it is assumed that regions of contiguous SNP markers can be 'grouped' into haplotype blocks, since adjacent SNPs in the genome tend to be in high LD, except for the 'hot-spot' regions were genetic recombination occurs. If this assumption holds for the majority of the genome, then for testing purposes, the dimension reduction from SNP markers alone to haplotype blocks is drastic. It is recognized, however, that there is still much to learn about the haplotype structure of the human genome, and debate still remains on the likely success of haplotype-based association studies (Schaid, 2004; Terwilliger & Hiekkalinna, 2006). The focus of this debate is whether common diseases are caused by common genetic variants, and whether the haplotype-block structure is a general feature of the human genome. The true test of the haplotype-map approach will come from application of a completed map to a variety of common diseases (Schaid, 2004). As a consequence the great majority of current genetic association scans typically use bi-allelic SNPs as markers. In spite of a degree of uncertainty on whether haplotype-based association studies will be viable in the future, it is noted that at a population level, observed bi-allelic SNPs from a case-control cohort can be seen as binomial variables, whereas observed haplotypes within a block can be seen as multinomial variables. Since a binomial variable is equivalent to multinomial variable with only two outcomes, all the testing methods examined in this dissertation are applicable to SNP markers as well as haplotype blocks.

2. Feasible sample sizes

Elston & Spence (2006) comment on the so-called genome-wide association scans: "There is presently a rush towards genome-wide association analyses, but we believe this is being driven more by the technology that is available than by any scientific rationale". The author of this dissertation concurs with Elton & Spence's remark. The fact that hundreds of thousands of SNPs can be genotyped does not mean that they all should be tested simultaneously. Assuming that a particular disease is associated to the combined small effects of many variants, the consequence of such brute-force approach is a potential data deluge where the true signals are diluted by thousands of false positive detections. It is sensible to first examine LD patterns among SNPs from data made available by consortiums such as the HapMap project. This would allow a reduction in the number of hypothesis tests by either applying a haplotype approach or, if testing SNPs directly, avoid testing thousands of redundant SNPs. Another consequence of conducting hundreds of thousands of tests simultaneously is that after adjusting for multiple testing, the sample size required for detection of association with an acceptable power can be so large that it might be economically unfeasible to obtain for a single study. Even if sample sizes of several thousand subjects would be feasible to obtain for common conditions (e.g. obesity, hypertension), for other less common conditions small prevalence of the condition would make the recruitment of large sample sizes unfeasible for a single study. An alternative approach is considered in this dissertation, by targeting a specific region of the genome (i.e. chromosome 22), as in candidate gene approaches, and thus in this dissertation a relatively small attainable sample size is considered. For diseases with very small prevalence within large geographical areas, the author of this dissertation envisions networks of multi-country studies each targeting specific areas of the genome and recruiting small attainable sample sizes. This, of course, brings other issues, such as population stratification and admixture, or environmental exposure differences.

3. Simple and fast algorithm for simulating markers in LD

In this dissertation a novel, simple and fast-to-implement algorithm is proposed to model LD among long sequences of variables. This algorithm breaks the sequence in small parts and avoids having to define one large correlation matrix for the whole sequence. The proposed algorithm is of easy implementation and has no limitation with respect to the desired length of the sequence of variables to generate. It also avoids the need to specify the joint distribution of each pair of adjacent bi-allelic markers (in the case of SNPs), or multi-allelic markers (in the case of haplotype blocks), as it is the case with the common 'moving window' algorithms, such as the one used by Sabatti (2006).

4. No assumptions with respect to a genetic model

As it would be the case in a realistic association study, a genetic model (dominant, co-dominant, or recessive) as well as penetrance (i.e. the probability of disease status given a particular genotype) are unknown. Thus, in the simulation design for this dissertation, no assumption is made in regards of whether certain alleles (haplotypes) within the simulated haplotype blocks associated to the disease follow a dominant or recessive model, or whether individuals with certain combinations of alleles (haplotypes) have a higher probability of disease status. In this regard only two assumptions are made. The first assumption is that for an individual, the haplotypes observed at any given haplotype block, are selected randomly from a 'population' of haplotypes, such as under random mating. The second assumption is that when comparing the haplotype frequencies between case and control groups, substantial divergence in frequencies between the groups is taken to be associated with disease status, as defined by Siegmund & Yakir (2007).

5. Results in terms of one comprehensive performance measure

In this dissertation, the testing results from each sequential procedure are compared against each other as well as with those from the standard fixed-sample-size Pearson's χ^2 tests with common multiplicity adjustments. Comparisons among procedures are made in terms of usual measures such as observed experiment-wise type I error rate, FPR (False Positive Rate), and observed experiment-wise power. In this dissertation a measure of experiment-wise penalized power is proposed, defined as the subtraction of FPR from observed experiment-wise power. The purpose of the proposed measure of penalized power is to capture the overall usefulness of a procedure by incorporating into one measure the procedure's statistical power counterbalanced by its observed rate of false positive detections.

5.2.2 Limitations

1. Assumption of random mating, but no formal testing of Hardy-Weinberg equilibrium (HWE)

Under random mating, mating takes place at random with respect to the genotypes under consideration. Mating can be random with respect to some traits, but non-random with respect to others in the same population. In human populations, for example, mating seems to be random with respect to blood groups, and many other characteristics, but mating is nonrandom with respect to other traits such as skin color and height (Hartl, 2000). Genotype frequencies are also influenced by various evolutionary forces including mutation, migration and natural selection. Under a random-mating model, the random mating of individuals is equivalent to random union of gametes (i.e. sperm or ovum) and as a consequence, the allele frequencies remain the same generation after generation, and so do genotype frequencies, a concept often called HWE. In order to illustrate the idea of HWE consider an autosomal chromosome with a bi-allelic marker whose alleles are called 'A' and 'a'. The genotype of a randomly selected individual in a population can be

AA, Aa, or aa. Under HWE the frequencies of these genotypes are $p_A^2, 2p_A(1 - p_A)$, and $(1 - p_A)^2$, respectively, where p_A denotes the population proportion of allele A, and $(1 - p_A)$ the population proportion of allele a. These frequencies of genotypes are expected in a random sample from a population that is assumed to be in HWE (and thus under random mating). In this dissertation, it is assumed that for an individual, the haplotypes observed at any given haplotype block, are selected randomly from a 'population' of haplotypes, such as under random mating. In a realistic association study, however, this assumption should be tested especially among the controls (Siegmund & Yakir, 2007). A violation of the HWE assumption in a sample may indicate divergence in the equilibrium of the population, errors in the determination of the genotypes, or a sample not representative of the population that can potentially cause spurious results. A common method to test for HWE is the χ^2 goodness-of-fit test. The observed genotype frequencies in the sample are tested against the expected genotype frequencies under HWE, calculated with the allele population proportions.

2. Simplistic model for LD

In this dissertation, for simulating high LD between adjacent haplotype blocks (represented by multinomial variables) an autoregressive structure (ρ=0.8) is specified on a matrix of randomly generated uniform(0,1) variables and then these highly correlated uniform(0,1) variables are transformed into multinomial observations. As a consequence of the high correlation among the underlying uniform(0,1) variables, some combinations of the resulting multinomial outcomes (representing haplotypes) from adjacent multinomial variables (representing blocks) occur more or less frequently than what would be expected if the combinations were formed randomly, thus providing a simplified model for LD. Sabatti et al. (2003) and Satagopan et al. (2004) use autoregressive structures as simplified models of dependency in order to account for LD-induced correlation between adjacent genetic markers. Satagopan et al. (2004) specifies an autoregressive structure directly on simulated test statistics, whereas Sabatti et al. (2003) specifies an autoregressive structure on the joint distribution between each pair of adjacent bi-allelic markers. However, concurring with Satagopan et al. (2004), the author of this dissertation acknowledges that an autoregressive structure is, at best, a crude approximation. With the sequencing of the human genome and development of high-throughput genomic methods, it has become clear that the human genome generally displays more LD than under simple population genetic models, and that LD is more varied across regions, and more segmentally structured, than had previously been supposed (The International HapMap Consortium, 2005). For instance, as shown on the histogram in Figure 10, LD between adjacent haplotype blocks appears to be present, and to be moderate to high on the average, but it is not constant, suggesting that a more complex model is required to attain more realism in simulations.

3. Limited number of multiplicity adjustments examined

In this dissertation four common multiplicity adjustment procedures are examined. Two of the examined procedures, the Bonferroni and Holm corrections, control a measure of error known as the family-wise error rate (FWER). The FWER is defined as the probability of incorrectly rejecting at least 1 null hypothesis. The two other procedures examined, Benjamini-Hochberg and Benjamini-Yekutieli, control a measure of error known as the False Discovery Rate (FDR). The FDR is a less stringent criterion than the FWER and is defined as the expected fraction of erroneous rejections among all hypotheses. Other recently-developed procedures based for instance on resampling or permutation techniques that control either the FWER or the FDR, such as those discussed by Sabatti (2006) in the context of multiple testing in genomics, are not examined in this dissertation.

4. Non-applicability of the SGLR procedure

As discussed in section 3.5.6, it is straightforward to apply the SGLR test to situations involving a single stream of observed data. In a case-control scenario, as long as there is a form for the distribution of the difference between each pair of case and control observations, then the SGLR test can be easily applied. However, using the SGLR to test a difference in proportions in case-control multinomial data presents major difficulties. First, each pair of case-control multinomial outcome realizations constitutes nominal-type data and thus it makes no sense to 'subtract' one from the other in the sense of continuous data, and then use the result of the subtraction individually in a likelihood function. Second, if the test is focused on the difference in accumulating proportions instead of the outcome realizations themselves, then the proportions, calculated after each pair of multinomial outcome realizations is observed, are not independent from the previously observed proportions, a situation more suited for repeated significance testing methods than for fully sequential procedures, which assume independence among accumulating observations. Third, when comparing frequencies of multinomial outcome realizations between groups, for a multinomial variable with 3 or more outcomes, a change in the proportion of one outcome results in a change in the proportion of at least one of the other outcomes. Therefore, for multinomial variables with 3 or more outcomes, to test H_0: the proportions between case and control groups are equal, vs. H_1: at least one proportion differs between case and control groups, there is no clear way to partition the parameter spaces between null and alternative spaces in a single SGLR test for all the proportions involved. Thus, since there are many combinations of proportions that can cause a significant change, more than one test would be required, which would result in multiplicity within a single test of hypothesis. To this date, the author of this dissertation has been unable to find in the literature a formal fully sequential test of association for contingency tables. After careful consideration of the problems involved in conducting a SGLR test with case-control multinomial data, in this dissertation a modification to the procedure is proposed by means of using the Pearson's χ^2 tests statistic to calculate the likelihood ratios required for this test, and an error spending function to account for multiple looks at accumulating data. However, this adaptation is an application of repeated significance testing methodologies, such as those used in clinical trials, as opposed to a fully sequential testing method as the original SGLR test is proposed by Chan & Lai (2005).

5. Population stratification might cause spurious results

In the context of genetics, population stratification or population substructure refers to the presence of a systematic difference in allele frequencies between subpopulations comprised within a larger population, possibly due to different ancestry (Siegmund & Yakir, 2007). Inspection of the haplotype block structure of the human genome by the HapMap consortium has shown substantially similarity in the haplotype patterns for the four populations examined by that consortium. However, the frequencies of haplotypes across populations often differ (The International HapMap Consortium, 2003 and 2005). Failure to take into account population substructure, if present, may produce spurious associations between markers (blocks) and disease status due to natural differences in allele (haplotype) frequencies between subpopulations. The approach taken in this dissertation assumes that all subjects belong to the same population, or that differences due to population substructure are negligible. A more complex model including population substructure would be needed to attain more realism in simulations.

5.3 Lessons learned

1. Performance of a procedure: balance between power and false positive detection rate

As shown in Figure 15, under the assumptions of this dissertation, on the average, the non-sequential Pearson's χ^2 test with Benjamini & Hochberg control of FDR adjustment, closely followed by a proposed modification of the SGLR test with a Bonferroni correction, provided the best balance between power and false positive detection, as measured by the proposed measure of penalized power. It is noticeable that neither of these two procedures obtained the 'best' results in the other three performance criteria examined; that is, neither of these two procedures obtained the 'best' mean observed experiment-wise type I error rate, FPR, or observed experiment-wise power. As previously mentioned, the purpose of the proposed measure of penalized power is to capture the overall usefulness of a procedure by incorporating into one measure the procedure's statistical power counterbalanced by its observed rate of false positive detections.

2. FWER vs. FDR

The FWER is defined as the probability of incorrectly rejecting at least 1 null hypothesis. The FWER is an appropriate measure of error when there is an overriding reason to not make any incorrect rejections of null hypotheses (Sabatti, False Discovery Rate and Multiple Comparison Procedures, 2006). Control of the FWER is based on setting stringent significance levels on the hypotheses. The downside of setting such stringent significance levels is a substantial decrease in statistical power. On the other hand, the FDR is a less stringent criterion than the FWER and is defined as the expected fraction of erroneous rejections among all hypotheses rejected. The FDR criterion captures the idea that if in an experiment there are a number of true alternative hypotheses present, we become more lenient toward committing a small fraction of false rejections when detecting the true alternative hypotheses, because the error from a single erroneous rejection (i.e. the FWER criterion) is not considered as crucial as the detection of true alternative hypotheses. Thus the proportion of incorrect rejections is controlled instead of the probability of a single incorrect rejection. Concurring with the results of this dissertation, it has been shown that with fixed sample size tests, adjusting for multiplicity with the FDR criterion substantially increases power compared to controlling the FWER. Another benefit of the FDR criterion is that if all hypotheses being tested are true null hypotheses, controlling the FDR is equivalent to controlling the FWER (Sabatti et al., 2003). Based on the considerations discussed above, and the simulation results from this dissertation, it is the opinion of the author of this dissertation that the definition of error provided by the FDR criterion is much more appropriate for large dimensional hypothesis-generating exploratory investigations such as genetic association studies. The behavior of sequential tests with control of FDR remains as an issue requiring further research.

3. Fully sequential methods and repeated significance tests

Currently available fully sequential methods (i.e. methods based on likelihood ratios on accumulating independent observations) are not well suited for genetic association studies due to the nominal nature of the allele data. As previously mentioned, to this date, the author of this dissertation has been unable to find in the literature a formal fully sequential test of association for contingency tables. For this dissertation, using the SGLR procedure to test a difference in proportions in case-control multinomial data presents major difficulties. After careful consideration of the problems involved in conducting a SGLR test with case-control multinomial data, in this dissertation a modification to the procedure is proposed that consists of using Pearson's χ^2 test statistics to calculate the likelihood ratios required for the SGLR test, and an error spending function to account for multiple looks at accumulating data. However, this adaptation is an application of repeated significance testing methodologies, such as those used in clinical trials. Furthermore, in order to avoid convergence problems of the Pearson's χ^2 test statistic, the proposed adaptation of the SGLR procedure requires the use of as much of the sample as possible, thus rendering the procedure unable to provide substantial savings in sample size, which is one of its objectives. However, the testing results from modified SGLR procedure with a FWER correction are very close, though not superior, to the non-sequential test with the Benjamini-Hochgberg control of FDR (the test whose results are considered 'best' in this dissertation). Thus the adaptation of the SGLR procedure did not result in a substantial improvement compared to the available non-sequential methods. Based on the considerations discussed above, and the simulation results from this dissertation, it is the opinion of the author of this dissertation that until a formal sequential test of association for contingency tables is developed, the currently-available fully sequential procedures are not only more difficult to apply but also do not constitute an improvement from the non-sequential methods in the context of genetic association studies.

5. *Ad hoc* sequential procedures

As shown in Figures 13, 14 and 15, under the assumptions of this dissertation, on the average, the sequential procedures that divide the sample size in independent subsamples (Sobel et al., 1993; Satagopan et al., 2004) produced poor results in regards to all the performance measures considered in this dissertation. These procedures should be avoided. The procedure proposed by Skol et al. (2006), while yielding the best performance of all procedures examined in regards to the FPR performance measure, resulted in substantially lower power than the non-sequential and modified SGLR procedures. While Skol et al.'s procedure is a cost-effective approach when the cost of genotyping is considerable; it is pertinent to note that rapid advances in 'high throughput' technologies have made genotyping costs less of a problem. It is foreseeable that, in the context of genetic association studies, if sequential procedures are used at all, their use would depend on how these procedures would help screening large amounts of markers or increasing power; but the focus of their use would not be on cost reductions due to a decrease in genotyping.

6. High performance computing

The simulations for this dissertation were run in a computing cluster. Each computing node in the cluster is equivalent to a single PC with a 2.4 GHz processor and 2GB of RAM memory. For the 1000 simulations in this dissertation, the code was divided in 25 programs, which were submitted simultaneously to the cluster. The cluster allocated each program to a node, and the 25 nodes ran the programs simultaneously. The average running time for each program was about 4½ hours. If the programs were run by a single PC, it would have taken approximately 112.5 hours (4.7 days) to finish the 1000 simulations. The advantage of using high performance computing clusters for simulations and analyses of large dimensional data are evident. However, computing clusters do not have graphic user interfaces, so commands in Unix/Linux are required to submit the programs. In addition, if programming from a Windows-based PC, the programs themselves need to be converted from DOS file format to Unix/Linux file format. Finally, statistical software available in the clusters is currently limited to the R system.

5.4. Directions for future research

1. A fully sequential test of association/independence for contingency tables

This dissertation showed the limitations of currently available fully sequential methods when applied to contingency tables. A challenging problem to be resolved is the development of a fully sequential test of association/independence for contingency tables.

2. Sequential Ranking and Selection (SRS) procedures

The issue of multiple testing is not commonly addressed in the literature of formal sequential methods. However, a related topic that is commonly addressed is referred to as Sequential Ranking and Selection (SRS). The objective of SRS procedures is to select a subset of variables from a given set of variables (Lai, 2001). For instance, Gupta & Liang (1988) proposed a SRS procedure to select a subset of random variables with the largest location parameters from a given set of variables. The subset includes the variable with the individually largest location parameter. In this approach, as data accumulates, at each stage the variables with smaller location parameters are eliminated, while the variables with larger location parameters are labeled as 'good'. The procedure reaches a terminal decision when there is only one variable left, or when all variables left are labeled as 'good'. An interesting and challenging problem is to determine whether it is possible to adapt an SRS procedure to genetic association studies, and if so, how.

3. Comparison of association detection between haplotype-block and tag-SNP based approaches

This dissertation uses haplotype blocks as markers for association testing. In this approach, ‘tag-SNPs’ would be used exclusively for identification of a person’s collection of haplotypes in each haplotype block. The author of this dissertation acknowledges that this is an alternative approach. Currently, the main approach for association testing is the use of tag-SNPs directly as markers. In the main approach, ‘tag-SNPs’ are defined as representative SNPs in a region of the genome in high LD. The great majority of current association studies use tag-SNPs as markers. An interesting and challenging problem is to determine in what situations a haplotype-based association testing approach is more efficient than the common tag-SNP association testing approach.

4. Algorithm for simulation of complex LD models

In this dissertation a simple, easy-to-implement algorithm to simulate LD is developed. The downside of this algorithm is that the resulting autoregressive LD model is too simplistic. A useful contribution to the literature would consist of a fast and easy-to-implement algorithm, but allowing complex LD structures, perhaps based on LD structures observed from real data available for instance from the HapMap project.

5. Bayesian methods

Finally, this dissertation has only dealt with procedures within the frequentist paradigm. It would be interesting to compare the performance of frequentist methods vs. Bayesian alternatives. For instance Bayesian sequential and non-sequential testing approaches have been proposed in the literature in terms of Bayes factors (Kass & Raferty, 1995; Berger et al. 1999; Johnson, 2005). In a Bayesian setting, let $x \sim f(x, \theta)$ where θ is the parameter of interest. For two competing hypotheses H_1 and H_2 in terms of the parameter of interest θ, let $\pi(\theta_1)$ and $\pi(\theta_2)$ be the prior distribution of θ under H_1 and H_2 respectively. The Bayes factor is defined as the ratio:

$$\frac{\int f(\underline{x}|\theta_1) \cdot \pi(\theta_1) d\theta_1}{\int f(\underline{x}|\theta_2) \cdot \pi(\theta_2) d\theta_2}.$$

Kass & Raferty (1995) provide guidelines for evaluation of Bayes factors as well as procedures for approximating the integrals, and methods for handling the problem of model uncertainity; Sabatti (2006) discusses Bayesian approaches to the multiplicity problem; and Agresti & Hitchcock (2005) provide a survey of Bayesian inference for categorical data.

6. REFERENCES

Agresti, A. (2002). *Categorical data analysis, 2nd Ed.* Hoboken, NJ: Wiley.

Agresti, A., & Hitchcock, D. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, (14):297-330.

Aplenc, R., Zhao, H., Rebbeck, T. R., & Propert, K. J. (2003). Group Sequential Methods and Sample Size Savings in Biomarker-Disease Association Studies. *Genetics*, (163):1215-1219.

Bain, L., & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics 2nd Ed.* Pacific Grove, CA: Duxbury.

Barrett, J. C., Fry, B., Maller, J., & Daly, J. M. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, (21):263-265.

Bechhofer, R. E., Kiefer, J., & Sobel, M. (1968). *Sequential Identification and Ranking Procedures.* Chicago, IL: The University of Chicago Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, (57):289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, (29):1165-1188.

Berger, J., Boukai, B., & Wang, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses . *Biometrika*, (86):79-92.

Bock, R. (1998). *The data analysis briefbook.* New York, NY: Springer.

Boddeker, I. R., & Ziegler, A. (2001). Sequential Designs for Genetic Epidemiological Linkage. *Biometrical Journal*, (43):501-525.

Casella, G., & Berger, R. L. (2002). *Statistical Inference - 2nd ed.* Pacific Grove, CA: Duxbury.

Chan, H., & Lai, T. (2005). Importance sampling for generalized likelihood ratio procedures in sequential analysis. *Sequential Analysis*, (24):259-278.

Chotai, J. (1984). On the LOD Score Method in Linkage Analysis. *Annals of Human Genetics*, 359-378.

Cordell, H. J., & Clayton, D. G. (2005). Genetic Association Studies. *The Lancet*, (366):1121-1131.

Daly, M., Schaffner, S., Hudson, T., & Lander, E. (2001). High-resolution haplotype structure in the human genome . *Nature Genetics*, (29):229-232.

Elston, R. C., & Spence, A. (2006). Advances in statistical human genetics over the last 25 years. *Statistics in Medicine*, (25):3049-3080.

Elston, R. C., Guo, X., & Williams, L. V. (1996). Two-stage global search designs for linkage analysis using pairs of affected relatives. *Genetic Epidemiology*, 535-558.

Gabriel, S., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science*, (296):2225-2229.

Ghosh, B. (1991). A Brief History of Sequential Analysis. In B. Ghosh, & P. K. Sen, *Handbook of Sequential Analysis* (pp. 1-19). New York, NY: Marcel Dekker.

Guo, X., & Elston, R. C. (2000). Two-stage global search designs for linkage analysis I: use of the mean statistic for affected sib pairs. *Genetic Epidemiology*, 97-110.

Gupta, S., & Liang, T. (1988). *On a sequential subset selection procedure, Technical Report No. 88-23.* West Lafayette, IN: Purdue Univ. Dept. of Statistics.

Hartl, D. (2000). *A primer of population genetics, 3rd Ed* . Sunderland, MA: Sinauer.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65-70.

Hotelling, P., & Pabst, M. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics*, (7):29-43.

Jennison, C., & Turnbull, B. W. (2000). *Group Sequential Methods with Applications in Clinical Trials.* Boca Raton, FL: Chapman & Hall/CRC Press.

Johnson, V. (2005). Bayes factors based on test statistcs. *Journal of the Royal Statistical Society - B*, (67):689-701.

Kass, R., & Raferty, A. (1995). Bayes factors. *Journal of the American Statistical Association*, (90): 773-795.

Kim, K., & DeMets, D. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*, (74), 149-154.

Lai, T. (2001). Sequential Analysis: Some Classical Problems and New Challenges. *Statistica Sinica*, (11):303-408.

Lan, K., & DeMets, D. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, (70), 659-663.

Mei, R., Galipeau, P. C., Prass, C., Berno, A., Ghandour, G., Patil, N., et al. (2000). Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Research*, 1126-1137.

Mitchell, L. (1995). Sequential Analysis of Marker Data for a Rare Oligogenic Disease. *Genetic Epidemiology*, (12):647-651.

Morton, N. E. (1955). Sequential Tests for Detection of Linkage. *American Journal of Human Genetics*, (7):277-318.

Muller, H. H., & Ziegler, A. (1998). Sequential testing using the transmission-disequilibrium test. In E. Greiser, & M. Wischnewsky, *Medizinische Informatik, Biometrie und Epidemiologie* (pp. 467-470). Munich, Germany: MMV Medien & Medizin Verlag.

O'Brien, P., & Fleming, T. (1979). A multiple testing procedure for clinical trials. *Biometrics*, (40), 1079-1087.

Pavlov, I. (1990). Sequential procedure for testing composite hypotheses with applications to the Kiefer-Weiss problem . *Theory of Probability and its Applications*, (35):280-292.

Phoon, K., Quek, S., & Huang, H. (2004). Simulation of non-Gaussian processes using fractile correlation. *Probabilistic Engineering Mechanics*, (19):287-292.
Pocock, S. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, (64), 191-199.

Proschan, M., Lan, G., & Wittes, J. (2006). *Statistical monitoring of clinical trials: a unified approach.* New york, NY: Springer.

Province, M. A. (2000). A single, Sequential, Genome-Wide Test to Identify Simultaneously All Promising Areas in a Linkage Scan. *Genetic Epidemiology*, (19):301-322.

Reboussin, D., DeMets, D., Kim, K., & Lan, K. (2000). Computations for group sequential boundaries using the Lan-DeMets spending function . *Controlled Clinical Trials*, (21):190-207.

Sabatti, C. (2006). False Discovery Rate and Multiple Comparison Procedures. In D. B. Allison, G. P. Page, T. M. Beasley, & J. W. Edwards, *DNA Microarrays and Related Genomic Techniques: design, analysis, and interpretation of experiments* (pp. 289-304). Boca Raton, FL: Chapman & Hall/CRC.

Sabatti, C., Service, S., & Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics*, (164):829-833.

Satagopan, J., & Elston, R. (2003). Optimal Two-Stage Genotyping in Population-Based Association Studies. *Genetic Epidemiology*, (25):149-157.

Satagopan, J., Venkatraman, E., & Begg, C. (2004). Two-Stage Designs for Gene-Disease Association Studies with Sample Size Constrains. *Biometrics*, (60):589-597.

Satagopan, J., Verbel, D., Venkatraman, E., Offit, K., & Begg, C. (2002). Two-Stage Designs for Gene-Disease Association Studies. *Biometrics*, (58):163-170.

Saxena, K., & Alam, K. (1982). Estimation of the non-centrality parameter of a chi-squared distribution. *The Annals of Statistics*, 10(3):1012-1016.

Schaid, D. J. (2004). Genetic Epidemiology and Haplotypes. *Genetic Epidemiology*, (27):317-320.

Schaid, D., & Sommer, S. (1994). Need to Confirm Promising Case-Control Association Studies: Reply to Sham. *American Journal of Medical Genetics*, (54):156-157.

Sham, P. (1994). Sequential Analysis and Case-Control Candidate Gene Association Studies: Reply to Sobell et al. *American Journal of Medical Genetics*, (54):154-155.

Siegmund, D., & Yakir, B. (2007). *The Statistics of Gene Mapping.* New York, NY: Springer.

Skol, A., Scott, J., Abecasis, G., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature Genetics*, (38):209-213.

Sobel, J., Heston, L., & Sommer, S. (1993). Novel Association Approach for Determining the Genetic Predisposition to Schizophrenia: Case-Control Resource and Testing of a Candidate Gene. *American Journal of Medical Genetics*, (48):28-35.

Speer, M. (1988). Basic Concepts in Genetics. In J. Haines, & M. Pericak-Vance, *Approaches to Gene Mapping in Complex Human Diseases* (pp. 17-52). New York, NY: Wiley-Liss.

Terwilliger, D., & Hiekkalina, T. (2006). An utter refutation of the 'Fundamental Theorem of the HapMap'. *European Journal of Human Genetics,* (14):426-437.

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

The International HapMap Consortium. (2007). A second generation human haplotype map over 3.1 million SNPs. *Nature*, (449): 851-861.

The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, 789-795.

Thomas, D. C., Haile, R., & Duggan, D. (2005). Recent Developments in Genomewide Association Scans: A Workshop Summary and Review. *American Journal of Human Genetics*, (77):337-345.

Thomas, D., Xie, R., & Gebegziabher, M. (2004). Two-Stage Sampling Designs for Gene Association Studies. *Genetic Epidemiology*, (27):401-414.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, (16):117-186.

Wang, H., Thomas, C., Pe'er, I., & Stram, D. (2006). Optimal Two-Stage Genotyping Designs for Genome-Wide Association Scans. *Genetic Epidemiology*, (30):356-368.

Zakharkin, S., Mehta, T., Tanik, M., & Allison, D. B. (2006). Epistemological Foundations of Statistical Methods for High-Dimensional Biology. In D. B. Allison, G. P. Page, T. M. Beasley, & J. W. Edwards, *DNA Microarrays and Related Genomic Techniques: design, analysis, and interpretation of experiments* (pp. 57-75). Boca Raton, FL: Chapman & Hall/CRC.

Zhao, H., Pfeiffer, R., & Gail, M. H. (2003). Haplotype analysis in population genetics and association studies. *Pharmacogenomics*, (171):171-178.