METHODS OF ASSOCIATION FOR GENOME DATA

WITH RARE VARIANTS AND

A MULTINOMIAL RESPONSE

By

JANAE ELIZABETH NICHOLSON

Bachelor of Science in Mathematics
University of Kansas
Lawrence, Kansas
2000

Master of Science in Statistics
Oklahoma State University
Stillwater, Oklahoma
2005

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2013

METHODS OF ASSOCIATION FOR GENOME DATA

WITH RARE VARIANTS AND

A MULTINOMIAL RESPONSE

Dissertation Approved:

Dr. Lan Zhu

Dissertation Adviser

Dr. Carla Goad

Dr. Ibrahim Ahmad

Dr. Udaya DeSilva

Outside Committee Member

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my husband, Randy, who has been supportive though this long and difficult process. I would also like to acknowledge my parents, Leo and Janelle Caspar, who have helped make my pursuit of a doctoral degree possible.

I would also like to thank my advisor, Dr. Lan Zhu, for her assistance in preparing this dissertation and two posters. Her understanding helped me though this difficult and at times frustrating process. I would also like to thank my committee members Dr. Carla Goad, Dr. Ibrahim Ahmad, and Dr. Udaya DeSilva. Their comments and guidance helped me make this work complete. I would also like to acknowledge Dr. William Warde who was an original committee member until his unexpected death in 2010. Dr. Warde was a friend who helped me through the toughest of times and convinced me to finish the doctoral program at OSU. I would like to thank Dr. Dana Brunson, Will Flanery, and Jesse Schafer of the High Performance Computing Center at OSU. These three people made it possible for me to run my simulations in a reasonable amount of time. Finally, I would like to thank Dr. Jonathan C. Cohen for sharing his resequencing data from the Dallas Heart Study.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF ABREVIATIONS

AMELIA	allele matching empirical locus-specific integrated association
BMI	body mass index
CDCV	common disease common variant
CDRV	common disease rare variant
CMAT	cumulative minor allele test
CMC	combined multivariate and collapsing
DNA	deoxyribonucleic acid
FDR	false discovery rate
FWER	family wise error rate
GWAS	genome wide association studies
KBAC	kernel based adaptive cluster
KBAT	kernel based association test
LD	linkage disequilibrium
MAF	minor allele frequency
MLOGIT	multinomial logistic regression
MLOGITC	multinomial logistic regression with collapsing
MNWSS	multinomial weighted sum statistic
MNWSSP	multinomial weighted sum statistic with a permutation test

QT-KBAT	quantitative trait kernel based association test
SNP	single nucleotide polymorphism
SMA	single marker analysis
WSS	weighted sum statistic

CHAPTER I

INTRODUCTION

1.1 Motivation of Research

A rare variant is a Single Nucleotide Polymorphism (SNP) with a minor allele frequency (MAF) of 5% or less. Approximately 60% of human SNPs are rare variants (Gorlov, Gorlova, Sunyaev, Spitz, & Amos, 2008). A debate is playing out as to whether these low frequency mutations are important in disease susceptibility (Schork, Murray, Frazer, & Topol, 2009). Including rare variants will cost more in terms of money and time as well as make the analysis more complex (Hirschhorn & Daly, 2005). New rapid genotyping technologies now make it possible to efficiently survey these rare variants. Many new statistical methods (Asimit & Zeggini, 2010; Bansal, Ondrej, Torkamani, & Schork, 2010) are being developed to analyze the associations between rare variants and phenotypes. Current methods have focused on dichotomous phenotypes such as case/control status or quantitative phenotypes such as weight or cholesterol level. The power of these methods depends on the underlying genetic model (Basu & Pan, 2011). Rare variant association methods for multinomial phenotypes, or categorical outcomes with more than two possibilities, have not been adequately addressed. There is one published method that can be used for rare variant association when the phenotype is multinomial. This method is the Allele Matching Empirical Locus-specific Integrated

Association (AMELIA) method (Zeggini & Asimit, 2010). However this method has not been evaluated using multinomial phenotypes. It is a modification of the Kernel Based Association Test (KBAT) (Mukhopadhyay, Feingold, Weeks, & Thalamuthu, 2010). The limited simulations on the AMELIA test using case control data showed it had a lower power than the KBAT.

Multinomial phenotypes have occurred in an association study where rare variants were included (Sulem, et al., 2007). Examples of these phenotypes would be hair color, eye color, schizophrenia sub-classification, and treatment outcome. New statistical methods of association need to be developed for rare variant association analysis when the phenotype is multinomial. Such a method could also be extremely useful in testing for population stratification of rare variants.

This dissertation proposes and investigates several methods for rare variant association analysis when the phenotype is multinomial. The recommendations contained in a later chapter aid geneticists in planning studies and analyzing this type of data. This work also provides a starting point for future researchers to build their own rare variant association methods for multinomial phenotypes.

Sections 1.1.1 through 1.1.3 begins by introducing the background in genetics needed to understand the work. Section 1.2 describes the statistical problems encountered when including rare variants in an association study. Section 1.3 outlines the relevant statistical methods currently available. Section 1.4 reveals the proposed work. Chapter 2 details the proposed methods. Chapter 3 lays out the simulation study of the proposed methods. Chapter 4 discusses the results of the simulation study and makes recommendations on the methods. Chapter 5 presents the results of applying the proposed methods to resequencing data from the Dallas Heart Study. Chapter 6 provides concluding remarks.

1.1.1 Common and Complex Disease

The human genome project finished sequencing the roughly three billion base pairs of the human genome in 2003. Since then there has been a considerable amount of work trying to decode what those base pairs do. Discovering the gene responsible for a simple, or Mendelian, disease is straightforward since there is a clear relationship between having the mutated gene responsible for the disease and having the disease (Mackay, 2009). However this relationship does not exist for what are called complex diseases. Here the term penetrance is used to specify the proportion of individuals with the gene that exhibit the disease. Not every individual with the gene will develop the disease. Rather an individual with the gene will have a greater probability of developing the disease. Further convoluting the ability to detect genes responsible for complex diseases is allelic heterogeneity, or the presence of different mutations at a single locus that produce the same phenotype. A common disease is a disease that occurs frequently in the population. It is usually assumed that common diseases have a complex genetic structure and hence fall under the domain of complex diseases. These common diseases include cancer, diabetes, and schizophrenia as well as many others. Common diseases are the focus of the rest of this paper.

1.1.2 The Common Disease Common Variant Hypothesis

There are two separate hypotheses used by current researchers trying to discover the genetic components of common diseases. The first is the Common Disease Common Variant (CDCV) hypothesis. It states that a common disease is the result of one or a few common variants with high penetrances in the genome (Reich & Lander, 2001; Pritchard & Cox, 2002; Schork, Murray, Frazer, & Topol, 2009). Supporters of this line of thinking argue that interactions between a small number of alleles produce the disease prevalence seen in the population (Smith & Lusis, 2002). Many early studies adopted this hypothesis because it required a smaller number of observations and markers to detect associations. If the CDCV hypothesis is true then with a reasonable sample size current

methods of association and linkage analysis should have adequate power to detect a locus responsible for a disease (Risch & Merikangas, 1996). The power of linkage disequilibrium mapping is low when common alleles have low penetrance (Hirschhorn & Daly, 2005).

1.1.3 The Common Disease Rare Variant Hypothesis

The second hypothesis is the Common Disease Rare Variant (CDRV) hypothesis. For this study a rare variant is defined as an allele that has a MAF of 5% or less. Some authors define a rare variant as an allele that has a MAF of 1% or less (Li & Leal, 2008; Morris & Zeggini, 2010). The CDRV hypothesis states that a number of rare variants and possibly several common variants are responsible for the disease (Schork, Murray, Frazer, & Topol, 2009). These variants have moderate to high penetrances (Li & Leal, 2008). Some researchers assume the variants are independent in the CDRV hypothesis (Li & Leal, 2008). While others assume a dependency called linkage disequilibrium (LD), or the phenomenon where a string of DNA tends to be inherited together, is present (Madsen & Browning, 2009; Basu & Pan, 2011).

The difference between two above hypotheses amounts to whether to include or exclude rare variants. A large multi-database study concluded that approximately 60% of human SNPs have a MAF of less than 5% (Gorlov, Gorlova, Sunyaev, Spitz, & Amos, 2008). Yet most studies using association methods or linkage mapping are not adequately powered to detect associations with variants with a low MAF (Risch & Merikangas, 1996). Commercially available SNP platforms exclude most rare variants (Zeggini & Asimit, 2010). It is common for studies to not follow up on significant SNPs when the MAF is less than 5% (Asimit & Zeggini, 2010, p. 294). Pritchard showed at least theoretically through simulations that rare and common variants could contribute to the genetic variation of a complex disease (2001). The combined results of several simulation studies show that for the low penetrance and high allelic heterogeneity seen in many common diseases, the

frequency of alleles responsible for that common disease will be near zero or one (Pritchard & Cox, 2002).

Some researchers are beginning to adopt the CDRV hypothesis. A review of all human genome wide association studies (GWAS) published up to December 2009 found 43 significant associations involving a rare variant with a p-value of 10^{-7} or less in 28 different studies (Panagiotou, Evangelou, & Ioannidis, 2010). It is important to note that the average sample size of these studies was 10,647 individuals.

1.2 Difficulties with Including Rare Variants

Since including rare variants means adding many more markers to an analysis, false discovery rates, degrees of freedom, and power become problematic. It has been shown by many researchers that single marker analysis adjusted to control the family-wise error rate (FWER) suffers from extremely low power to detect a true association (Li & Leal, 2008; Madsen & Browning, 2009; Basu & Pan, 2011). Even with a reasonable FWER control many false positives may still occur. Multivariate analysis such as a Hotelling's T^2 test or multiple logistic regression has slightly higher power in these studies but it is still not adequate (Li & Leal, 2008). Since the degrees of freedom in the likelihood ratio test of the multiple logistic regression and the numerator degrees of freedom in the Hotelling's T^2 test statistic are equal to the number of markers, a large number of markers results in a large number degrees of freedom in these tests. This makes it difficult to detect a true susceptibility allele hence reducing the power in these tests.

Sample size also becomes a huge consideration when including rare variants. As previously mentioned Panagiotou, Evangelou, and Ioannidis (2010) found that published studies that discovered a significant rare variant had an average sample size of 10,647. It has also been shown that for case control studies the sample size necessary to achieve a fixed power increases as MAF decreases

(Gorlov, Gorlova, Sunyaev, Spitz, & Amos, 2008). These increases are more dramatic for rare variants with marginal effects.

It has been argued that current methods based on LD are inadequate for rare variant analysis (Asimit & Zeggini, 2010). This stems from the fact that for a strong correlation to exist the MAFs of the two variants must be roughly equal (VanLiere & Rosenberg, 2008). Commercially available SNP platforms rely on LD to balance the number of SNPs with the amount of genetic variation captured. Since these panels include very few rare variants the ability to detect a causal rare variant is low using them.

1.3 Methods for Detecting Associations with Rare Variants

Including rare variants in the search for an association has necessitated the development of new statistical methods. While most of the methods available for rare variant association analysis are for case control data or a dichotomous response there are some methods for quantitative phenotypes (Asimit & Zeggini, 2010). To date there is only one published method for rare variant analysis when the phenotype is multinomial. Methods for quantitative phenotypes will be discussed first, followed by case control data, and lastly multinomial phenotypes.

1.3.1 Quantitative Phenotypes

A few methods exist for rare variant association analysis when the phenotype is quantitative. Morris and Zeggini (2010) present two methods based on linear regression. Price and authors (2010) propose a variable threshold approach that considers multiple threshold cutoffs at once. Hoffman, Marini, and Witte (2010) illustrate a step-up procedure to build a model relating the quantitative phenotype to the variants. Thalamuthu, Zhao, Keong, Kondragunta, and Mukhopadhyay (2011) have extended their previous Kernel-Based Association Test (KBAT) (Mukhopadhyay, Feingold, Weeks, & Thalamuthu, 2010) so that quantitative phenotypes could be analyzed. The new Quantitative Trait Kernel Based Association Test (QT-KBAT) and the original KBAT are also modified to better handle

rare variants included with common variants in the test. The simulations for Morris and Zeggini's methods, the variable threshold approach, and the step-up procedure only considered a normally distributed phenotype. The QT-KBAT method was only applied to a single generated data set. Hence the behavior of these tests is unknown when the phenotype is non-normal. Also the variable threshold approach, step-up procedure, and quantitative kernel-based association test are all very computationally intensive.

1.3.2 Case Control Designs

The majority of the research in rare variant association published to date is for case control designs (Asimit & Zeggini, 2010). In general there is not one method that performs best under all situations (Basu & Pan, 2011). Rather the performance of each of the methods depends on the underlying genetic model. The following pages will lay out some of the available association methods for case control designs.

The first group of methods considered here are the model type methods. These methods usually have many degrees of freedom in the test or require an adjustment for multiple tests. A majority of the tests are based on the generalized linear model for case control data via multiple logistic regression. Since the degrees of freedom is equal to the number of variants in the model, this test can have a large number of degrees of freedom if a large number of variants are used. Li and Leal (2008) found that multiple logistic regression had an inflated Type I Error rate as a method for rare variant association. Han and Pan (2010) proposed the adaptive sum test to strike a balance between large number of degrees of freedom in the multiple logistic regression test and large adjustment for multiple tests. The step-up procedure (Hoffmann, Marini, & Witte, 2010) mentioned above can also be used for a dichotomous phenotype. This procedure builds a logit model using a step-up algorithm. Both the adaptive sum test and step-up procedure aim to limit the number of variants used in the logit model but still require a large number of degrees of freedom. The univariate

minP test which tests a logit model with each variant individually was considered in simulations by Basu and Pan (2011). This procedure requires a large adjustment for multiple tests.

Several methods seek to reduce the dimensionality of the data and hence preserve the degrees of freedom in the test (Asimit & Zeggini, 2010). This should theoretically increase the power of the tests at the one locus where the combining of the data is made. Since these methods collapse or pool rare variants together they are often called collapsing or pooling methods.

The Combined Multivariate and Collapsing (CMC) method proposed by Li and Leal groups markers by MAF at a locus and collapses within groups by creating a single dummy variable for the group (2008). A Hotellings T^2 test is then performed on the collapsed data and any other high frequency variants included. The authors show through simulations that the CMC method has higher power than a single marker analysis and the Hotelling's T^2 test without collapsing. Additionally the CMC method controls the Type I error at the desired level.

A weighted sum statistic (WSS) presented by Madsen and Browning (2009) also combines markers within a locus to test for association with rare variants. In this method the number of rare alleles is weighted by the inverse of the standard deviation of the number of rare alleles. A "genetic score" is created for each individual by summing up the weighted number of rare alleles. The sum of the ranks of the genetic scores in cases is then used in a permutation test. The researchers show through simulations that their test has improved power over the CMC method and a single marker analysis.

Although the WSS and Li and Leal's CMC method have been the benchmark methods that most researchers compare their proposed case control methods to, many other methods also use collapsing or pooling to achieve better power. Feng, Elson, and Zhu (2011) propose a modification to Madsen and Browning's WSS. They adjust the weight for both sib-pair and case control designs. The variable threshold test for quantitative phenotypes previously mentioned can also be applied to

dichotomous phenotypes (Price, et al., 2010). This method considers multiple thresholds for collapsing rare variants. Ionita-Laza, Buxbaum, Laird, and Lange (2011) propose the Replication Based Test that groups variants together by minor allele counts. The Kernel Based Adaptive Cluster (KBAC) test (Liu & Leal, 2010) pools individuals together with the same rare variant haplotype and looks for differences in the proportions in cases and controls. The Cumulative Minor-Allele Test (CMAT) (Zawistowski, Gopalakrishnan, Ding, Li, Grimm, & Zollner, 2010) pools allele counts for cases and controls. The Kernel Based Association Test (KBAT) for case control data is modified to better handle rare variants (Thalamuthu, Zhao, Keong, Kondragunta, & Mukhopadhyay, 2011). Rare variants are pooled to increase power and decrease computation time. The C-alpha test (Neale, et al., 2011) uses the distribution between cases and controls of individuals with a rare variant. Markers are pooled by summing over the markers in the locus.

A number of methods for analysis of common variants are suggested for rare variant analysis. Some of these methods include the kernel-machine test (Wu, et al., 2010), Hotelling's T^2 test (Li & Leal, 2008), the SSU test (Basu & Pan, 2011), the Sum test (Basu & Pan, 2011), multivariate distance matrix regression (MDMR) (Wessel & Schork, 2006), the ZGlobal Statistic (Schaid, McDonnell, Hebbring, Cunningham, & Thibodeau, 2005), Logic regression, ridge regression, and LASSO. Most but not all of these methods are evaluated in simulation studies where rare variants are included as both causal and non-causal variants.

1.3.3 Multinomial Methods

To date there is only one method as proposed available for rare variant association testing when more than two phenotypic categories are present. This method is called the Allele Matching Empirical Locus-specific Integrated Association (AMELIA) test (Zeggini & Asimit, 2010). It is a modification of the original KBAT method by Mukhopadhyay, Feingold, Weeks, and Thalamughu (2010) mentioned above. The AMELIA method includes genotype quality scores and allows for

more than two categories in the phenotype. The researchers only compare their AMELIA test to the original KBAT method on case control data. Multinomial phenotypes are not used in any of the simulations. Both methods have very low power to detect a true association in the case control data. When a region of 342 simulated SNPs is included in the test, the AMELIA test has a power of 8.71% and the KBAT has a power of 9.53%. When a neighborhood of 11 SNPs around the one causal SNP is considered the AMELIA test has a power of 17.31% and the KBAT method has a power of 21.61%. These limited simulations show that the KBAT test has higher power than the AMELIA test. Additionally the AMELIA test requires SNP quality scores which may not be available. Thus the AMELIA test is not further considered.

The KBAT method for case control data described above is laid out generally so that more than two categories in the phenotype are possible. However the test statistic used forces two categories and an equal number of observations in each category. The test statistic is easily modified to allow for more than two categories and an unbalanced design as was done for the AMELIA test. There are no power estimates available for rare variant analysis with multinomial phenotypes in the published literature. Early work on this dissertation considered this modified KBAT method. An attempt to analyze one data set with 320 variants and 1000 individuals on one node of the super computer Pistol Pete timed out after 24 hours. Since 1,000 iterations would need to be run, it was decided the method was too computationally intensive to include in the simulation study. This method could be considered in a future simulation study with a much smaller number of individuals and SNPs.

In the study of more than two phenotypic categories, it is possible to collapse to two categories and proceed with a case control method. However information is lost when this is done and the results depend on which categories are collapsed. Another possibility is using a stratified approach. A stratified single marker analysis is used by Sulem et al. (2007) for an analysis of hair and eye color. However the power of these types of approaches to detect a true association has not

been evaluated. Morris and Zeggini's methods, the threshold tests, the step-up procedure, and the QT-KBAT method detailed in section 1.3.1 for quantitative phenotypes can also be used to test for an association when the phenotype is categorical. All of these methods can be applied to the multinomial phenotype data as long as the categories are coded as a number. The methods provided by Morris and Zeggini (2010) assume a normal distribution which is violated in the case of categorical phenotypes. The threshold tests, QT-KBAT, and the step-up test use permutations so the departure from normality should not be a problem. However none of these methods have been evaluated when the phenotype is a non-normal distribution such as the multinomial distribution. Additionally each of these methods is extremely computationally extensive. In the interest of time and brevity none of the quantitative phenotype methods are considered as a method of association with a multinomial phenotype. Proposed methods of association are laid out in the next section.

1.4 Scope of Study

It is plausible that rare variants are responsible in some part for common diseases (Panagiotou, Evangelou, & Ioannidis, 2010; Pritchard, 2001; Pritchard & Cox, 2002; Schork, Murray, Frazer, & Topol, 2009). Current methods for testing for an association with a nominal response are lacking. Therefore three novel methods of rare variant analysis for genetic data with a multinomial response are investigated. The first method is an extension of the weighted sum statistic by Madsen and Browning (2009). A test statistic for the weighted sum statistic procedure is chosen that can incorporate more than two outcomes in the phenotype. A single marker analysis (SMA) is created to work with multinomial data. A test procedure is developed and the appropriate test statistic is determined for each test of association at each marker. Then a multiple testing procedure is necessary to adjust for the large number of tests being run. Finally the results of all of these tests are put together for a single decision about the association at the locus. Finally multinomial logistic regression is investigated as a method of association. This method is the generalized linear model approach so it should be a good baseline for comparison. The appropriate test statistic is taken from

the literature as multinomial logistic regression is a published procedure. All of the methods are developed and evaluated so that recommendations can be made on how to run a rare variant analysis when the phenotype is multinomial.

Simulations are run to assess the performance of each of the methods. First genetic data is generated. To determine the Type I Error rate, data is produced under the null hypothesis of no association. To determine the power data is simulated under the alternative hypothesis of an association between the phenotype and SNPs. The proposed methods are applied to these data sets to detect an association between the markers and phenotypes. The results are recorded and presented in later chapters of this document. Recommendations based on the simulations are presented.

CHAPTER II

METHODOLOGY

This chapter outlines the three proposed methods for testing for an association between a multinomial phenotype and multiple rare variants. In general it is assumed that the Common Disease Rare Variant hypothesis detailed in Section 1.1.3 holds. The first method considered is the weighted sum statistic for multinomial data (see Section 2.1). This is followed by a single marker analysis (see Section 2.2). A false discovery rate controlling method is included for the single marker analysis (see Section 2.2.1). Multinomial logistic regression is considered as a method to test for association (see Section 2.3). The multinomial logistic regression routine fails to fit the model at times. Due to these failures, collapsing of variants in the multinomial logistic regression procedure is also considered (see Section 2.4). The terms variant and marker are used interchangeably in this document.

2.1 A Weighted Sum Statistic for Multinomial Data

A weighted sum statistic is proposed to test for an association between a multinomial response and collectively all rare variants at a locus. Markers along a chromosome must first be grouped into a locus or gene where the test is to be conducted. Rare variants within the locus will be pooled together in the test. Following the lead of other researchers it is assumed that this grouping can be done in a meaningful way (Li & Leal, 2008; Madsen & Browning, 2009). Since multiple loci or genes are tested in a GWAS, a multiple testing control must be used to control

either the familywise experiment rate or the false discovery rate. The hypotheses to test for an association at one locus are:

H_0: The mutation frequency is the same for all response groups ($k = 1, \ldots, K$).

H_a: The mutation frequency is different for at least one of the response groups.

These hypotheses are equivalent to testing for association between the multinomial response and the variants at a locus since the frequency of the alleles are the same for each response group if there is no association. The test consists of the following steps:

1. For each variant or marker, $j = 1, \ldots, J$, identify the mutant allele that is thought to be a susceptibility allele. If it is not known which allele is the mutant one then the rarer allele will be used.
2. Next a weight is calculated for each genotype. Define m_j as the number of mutant alleles in individuals at variant j and $n_{\cdot j}$ as the number of individuals genotyped for variant j. Since each individual is genotyped on a pair of chromosomes there are $2 \cdot n_{\cdot j}$ alleles for each marker. The weight is

$$w_j = \sqrt{n_{\cdot j} q_j (1 - q_j)}$$

where

$$q_j = \frac{m_j + 1}{2n_{\cdot j} + 2}.$$

Thus q_j is the adjusted frequency of mutant alleles at marker j. The weight is the standard deviation of the number of mutant alleles using pseudo counts. The adjustment in q_j is necessary since the frequency of mutant alleles is expected to be small possibly zero. In a future step a quantity is calculated where w_j is used in the denominator. Hence it is undesirable for w_j to be zero. Individuals missing the phenotype but possessing genotypic information are allowed to contribute to the weights.

3. Define I_{ijk} as the number of mutant alleles in variant *j* for individual *i* in group *k*. Since each individual has two copies of each chromosome for an additive model $I_{ijk} \in \{0, 1, 2\}$. However if a recessive or dominant model is expected then restrict $I_{ijk} \in \{0, 1\}$. For the recessive model only the homozygous mutants receive a 1. For the dominant model both the heterozygous and homozygous mutants receive a 1. The additive model is the default model used. Any individual which is missing the multinomial response or all genetic markers must be eliminated from further calculations. For the remaining individuals any missing I_{ijk} is assumed to be zero. For each individual calculate the genetic score following Madsen and Browning (2009) as

$$\gamma_{ik} = \sum_{j=1}^{J} \frac{I_{ijk}}{w_j}.$$

Hence the genetic score is the weighted sum of the number of mutations in each individual. The lower the mutant allele frequency in the sample, the more a mutation at that variant contributes to the genetic score. This allows a mutation at a variant with an allele frequency of 1% to contribute more to the genetic score than a mutation at a variant with a 5% allele frequency. It also allows a mutation at a rare variant to contribute more than a mutation at a common variant if common variants are included in the test.

4. Using the genetic scores, γ_{ik}'s, conduct a Kruskal-Wallis test to detect a difference in the distribution of the genetic scores between response groups. A test statistic assuming ties is necessary since ties in the genetic scores are possible. For a tie the average of the ranks is assigned.
5. An observed significance level is obtained either using the asymptotic chi-square distribution of the test statistic or through a permutation test. The method used depends on whether it is assumed that the observations are independent or not. If it can be assumed that the observations are independent, then the observed significance level is the

probability that a chi-square random variable with K-1 degrees of freedom is greater than the test statistic, T. If it is assumed that there is dependence among the observations, an empirical distribution for T is found by permuting the response group among individuals and recalculating the test statistic at least 1000 times. The observed significance level of the test statistic is the percentile of the observed test statistic in the empirical distribution.

2.1.1 The Distribution of the Weights and Genetic Scores

The proposed Multinomial Weighted Sum Statistic utilizes the Kruskal-Wallis test statistic to test for an association between all rare variants collectively and the phenotype. This is not necessary if the distribution of the genetic scores is known or the central limit theorem applies to the sum in the genetic score. Since the genetic score is a sum of ratios it is not unreasonable to expect the central limit theorem to apply. An empirical study of the genetic scores showed that the data is very far from normal. One thousand data sets were generated using the procedure described in chapter 3. The genetic score was calculated for each individual in each of the data sets. Lilliefors test for Normality (Conover, 1999, p. 443) and the Shapiro-Wilk test for Normality (Conover, 1999, p. 450) were both run on each data set to test the null hypothesis that the data is normally distributed. All one thousand times the null hypothesis was rejected for both tests. This provides strong evidence that the central limit theorem does not apply. Inspection of the generated genetic scores revealed that the distribution is skewed and large outliers are possible. Feng, Elson, and Zhu (2011) also reported finding the distribution of the genetic scores skewed with possible outliers.

Additionally the distribution of the genetic scores is currently unknown. However it is possible to derive the distributions of some of the quantities input into the genetic score. These derivations are produced as part of the work of this dissertation. For all derivations shown here individuals and markers are considered independent. This may not be the case if pedigree

structure or linkage disequilibrium exists. Let $n_{\cdot j}$ be the number of individuals genotyped for variant *j*. Also let N_k be the number of individuals in response group *k*. Assuming an additive model, use I_{ijk} as the number of mutant alleles at variant *j* for individual *i* of group *k* as before. Also assume a fixed probability of a mutant allele at variant *j*, p_j, in the population. Then $I_{ijk} \sim Binomial(n = 2, p = p_j)$. Now fix *j* and consider the mutations at variant *j*. Since the individuals are assumed independent the joint distribution of the I_{ijk} 's at variant *j* is

$$f(I_{1j1}, I_{2j1}, \dots, I_{N_K jK}) = \left[\prod_{i,k} \binom{2}{I_{ijk}}\right] \cdot p_j^{\sum_{i,k} I_{ijk}} \cdot (1 - p_j)^{2n_{\cdot j} - \sum_{i,k} I_{ijk}}.$$

Consider the joint transformation

$$y_1 = \sum_{i,k} I_{ijk},$$

$$y_2 = I_{1j1},$$

$$y_3 = I_{2j1}, \dots,$$

$$y_{n_{\cdot j}} = I_{(N_K - 1)jK}.$$

Then the joint distribution of the y_i's is

$$f\left(y_1, y_2, \dots, y_{n_{\cdot j}}\right) = \binom{2}{y_2} \cdot \binom{2}{y_3} \cdot \dots \cdot \binom{2}{y_{n_{\cdot j}}} \cdot \binom{2 n_{\cdot j}}{y_1 - \sum_{a=2} y_a} \cdot p_j^{y_1} \cdot (1 - p_j)^{2n_{\cdot j} - y_1}$$

where $y_a \in \{0,1,2\}$ for $a = 2, 3, \dots, n_{\cdot j}$ and $y_1 - \sum_{a=2}^{n_{\cdot j}} y_a \in \{0,1,2\}$. Notice that y_1 is dependent on the y_a's since there is no way to factor the joint distribution. Therefore the weights in the genetic score which are a function of y_1 will also be dependent on the I_{ijk}'s. Randomly choose one of the y_a's $a = 2, 3, \dots, n_{\cdot j}$ to keep. For convenience y_2 is used in the following derivations

but any y_a, $a = 2, 3, \dots, n_{\cdot j}$ can be chosen. It is desired to find the joint distribution of y_1, y_2. First find the marginal joint distribution of $y_1, y_2, \dots, y_{n_{\cdot j}-1}$.

$$f\left(y_1, y_2, \dots, y_{n_{\cdot j}-1}\right)$$

$$= \binom{2}{y_2} \cdot \binom{2}{y_3} \cdot \dots \cdot \binom{2}{y_{n_{\cdot j}-1}} \cdot p_j^{y_1} \cdot (1-p_j)^{2n_{\cdot j}-y_1} \cdot \binom{4}{y_1 - \sum_{a=2}^{n_{\cdot j}-1} y_a}$$

$$\cdot \sum_{y_{n_{\cdot j}}=max\left(0, y_1-\sum_{a=2}^{n_{\cdot j}-1} y_a\right)}^{min\left(2, y_1-\sum_{a=2}^{n_{\cdot j}-1} y_a\right)} \frac{\binom{2}{y_{n_{\cdot j}}} \cdot \binom{2}{y_1 - \left(\sum_{a=2}^{n_{\cdot j}-1} y_a\right) - y_{n_{\cdot j}}}}{\binom{4}{y_1 - \sum_{a=2}^{n_{\cdot j}-1} y_a}}$$

$$= \binom{2}{y_2} \cdot \binom{2}{y_3} \cdot \dots \cdot \binom{2}{y_{n_{\cdot j}-1}} \cdot p_j^{y_1} \cdot (1-p_j)^{2n_{\cdot j}-y_1} \cdot \binom{4}{y_1 - \sum_{a=2}^{n_{\cdot j}-1} y_a}$$

where $y_a \in \{0,1,2\}$ for $a = 2, 3, \dots, n_{\cdot j} - 1$ and $y_1 - \sum_{a=2}^{n_{\cdot j}-1} y_a \in \{0,1,2,3,4\}$ since

$$\frac{\binom{2}{y_{n_{\cdot j}}} \cdot \binom{2}{y_1 - \sum_{a=2}^{n_{\cdot j}-1} y_a - y_{n_{\cdot j}}}}{\binom{4}{y_1 - \sum_{a=2}^{n_{\cdot j}-1} y_a}} \sim Hypergeometric(n, M, N)$$

with $n = y_1 - \sum_{a=2}^{n_{\cdot j}-1} y_a$, $M = 2$, and $N = 4$ (Bain & Engelhardt, 1992, p. 96). Continuing find the joint distribution of $y_1, y_2, \dots, y_{n_{\cdot j}-2}$.

$$f\left(y_1, y_2, \dots, y_{n_{\cdot j}-2}\right)$$

$$= \binom{2}{y_2} \cdot \binom{2}{y_3} \cdot \dots \cdot \binom{2}{y_{n_{\cdot j}-2}} \cdot p_j^{y_1} \cdot \left(1-p_j\right)^{2n_{\cdot j}-y_1} \cdot \binom{6}{y_1 - \sum_{a=2}^{n_{\cdot j}-2} y_a}$$

$$\cdot \sum_{y_{n_{\cdot j}-1}=max\left(0, y_1-\sum_{a=2}^{n_{\cdot j}-2} y_a\right)}^{min\left(2, y_1-\sum_{a=2}^{n_{\cdot j}-2} y_a\right)} \frac{\binom{2}{y_{n_{\cdot j}-1}} \cdot \binom{4}{y_1 - \left(\sum_{a=2}^{n_{\cdot j}-2} y_a\right) - y_{n_{\cdot j}-1}}}{\binom{6}{y_1 - \sum_{a=2}^{n_{\cdot j}-2} y_a}}$$

$$= \binom{2}{y_2} \cdot \binom{2}{y_3} \cdot \dots \cdot \binom{2}{y_{n_{\cdot j}-2}} \cdot p_j^{y_1} \cdot \left(1-p_j\right)^{2n_{\cdot j}-y_1} \cdot \binom{6}{y_1 - \sum_{a=2}^{n_{\cdot j}-2} y_a}$$

where $y_a \in \{0,1,2\}$ for $a = 2, 3, \dots, n_{\cdot j} - 2$ and $y_1 - \sum_{a=2}^{n_{\cdot j}-2} y_a \in \{0, 1, 2, 3, 4, 5, 6\}$. By induction continuing this process until only the joint distribution of y_1 and y_2 remains finds

$$f(y_1, y_2) = \binom{2}{y_2} \cdot \binom{2\left(n_{\cdot j} - 1\right)}{y_1 - y_2} \cdot p_j^{y_1} \cdot \left(1-p_j\right)^{2n_{\cdot j}-y_1}$$

where $y_2 \in \{0, 1, 2\}$ and $y_1 \in \left\{y_2, \dots, 2\left(n_{\cdot j} - 1\right) + y_2\right\}$. Notice that if the process had also summed out y_2 the resulting marginal distribution of y_1 would be

$$f(y_1) = \binom{2n_{\cdot j}}{y_1} \cdot p_j^{y_1} \cdot \left(1-p_j\right)^{2n_{\cdot j}-y_1}.$$

Hence the distribution of the sum of the mutant alleles at variant j is $Binomial\left(n = 2n_{\cdot j}, p = p_j\right)$. Now consider the following joint transformation of $f(y_1, y_2)$:

$$q_{j1} = \frac{y_1 + 1}{2n_{\cdot j} + 2},$$

$$q_{j2} = y_2.$$

Then the joint distribution of q_{j1} and q_{j2} is

$$f(q_{j1}, q_{j2}) = \binom{2}{q_{j2}} \cdot \binom{2(n_{\cdot j} - 1)}{q_{j1}(2n_{\cdot j} + 1) - 1 - q_{j2}} \cdot p_j^{q_{j1}(2n_{\cdot j}+2)-1} \cdot (1 - p_j)^{2n_{\cdot j}+1-q_{j1}(2n_{\cdot j}+2)}$$

Where $q_{j2} \in \{0, 1, 2\}$ and $q_{j1} \in \left\{\frac{q_{j2}+1}{2n_{\cdot j}+2}, \frac{q_{j2}+2}{2n_{\cdot j}+2}, \dots, \frac{2(n_{\cdot j}-1)+q_{j2}+1}{2n_{\cdot j}+2}\right\}$. Now consider the joint transformation

$$w_{j1} = \sqrt{n_{\cdot j} q_{j1}(1 - q_{j1})}$$

$$w_{j2} = q_{j2}.$$

The equation for w_{j1} is not one to one and has a maximum at $1/2$. Using the quadratic formula to solve for q_{j1} yields

$$q_{j1} = \frac{1}{2}\left[1 \pm \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right].$$

Now define $A_1 = \left\{\frac{q_2+1}{2n_{\cdot j}+2}, \frac{q_2+2}{2n_{\cdot j}+2}, \dots, \frac{1}{2}\right\}$ and $A_2 = \left\{\frac{n_{\cdot j}+2}{2n_{\cdot j}+2}, \frac{q_2+3}{2n_{\cdot j}+2}, \dots, \frac{2(n_{\cdot j}-1)+q_{j2}+1}{2n_{\cdot j}+2}\right\}$. Then w_{j1} is one to one on the sets A_1 and A_2. On A_1 the inverse is $q_{j1}^- = \frac{1}{2}\left[1 - \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right]$ and on A_2 the inverse is $q_{j1}^+ = \frac{1}{2}\left[1 + \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right]$. Define the following sets

$$(w_{j1}, w_{j2}) \in B_1 = \left\{\left(\sqrt{n_{\cdot j}\left(\frac{1}{2n_{\cdot j}+2}\right)\left(1 - \frac{1}{2n_{\cdot j}+2}\right)}, 0\right), \left(\sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j}+2}\right)\left(1 - \frac{2}{2n_{\cdot j}+2}\right)}, 0\right)\right\},$$

$$B_2 = \left\{\left(\sqrt{n_{\cdot j}\left(\frac{2n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1 - \frac{2n_{\cdot j}}{2n_{\cdot j}+2}\right)}, 2\right), \left(\sqrt{n_{\cdot j}\left(\frac{2n_{\cdot j}+1}{2n_{\cdot j}+2}\right)\left(1 - \frac{2n_{\cdot j}+1}{2n_{\cdot j}+2}\right)}, 2\right)\right\}$$

$$B_3 = \left\{\left(\sqrt{n_{\cdot j}\left(\frac{3}{2n_{\cdot j}+2}\right)\left(1-\frac{3}{2n_{\cdot j}+2}\right)}, 0\right), \left(\sqrt{n_{\cdot j}\left(\frac{4}{2n_{\cdot j}+2}\right)\left(1-\frac{4}{2n_{\cdot j}+2}\right)}, 0\right), \ldots, \left(\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}, 0\right)\right\}$$

$$B_4 = \left\{\left(\sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j}+2}\right)\left(1-\frac{2}{2n_{\cdot j}+2}\right)}, 1\right), \ldots, \left(\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}, 1\right)\right\}$$

$$B_5 = \left\{\left(\sqrt{n_{\cdot j}\left(\frac{3}{2n_{\cdot j}+2}\right)\left(1-\frac{3}{2n_{\cdot j}+2}\right)}, 2\right), \left(\sqrt{n_{\cdot j}\left(\frac{4}{2n_{\cdot j}+2}\right)\left(1-\frac{4}{2n_{\cdot j}+2}\right)}, 2\right), \ldots, \left(\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}, 2\right)\right\}.$$

$$B_6 = \left\{\left(\frac{\sqrt{n_{\cdot j}}}{2}, 0\right), \left(\frac{\sqrt{n_{\cdot j}}}{2}, 1\right), \left(\frac{\sqrt{n_{\cdot j}}}{2}, 2\right)\right\}$$

If $(w_{j1}, w_{j2}) \in B_1$ then the joint distribution of w_{j1} and w_{j2} is

$$f(w_{j1},\, w_{j2}) = \binom{2}{w_{j2}} \cdot \left(\frac{2(n_{\cdot j}-1)}{\left(1-\sqrt{1-4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1)-1-w_{j2}}\right) \cdot p_j^{\left(1-\sqrt{1-4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1)-1} \cdot (1-p_j)^{2n_{\cdot j}+1-\left(1-\sqrt{1-4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1)}.$$

If $(w_{j1}, w_{j2}) \in B_2$ then the joint distribution of w_{j1} and w_{j2} is

$$f(w_{j1},\, w_{j2}) = \binom{2}{w_{j2}} \cdot \left(\frac{2(n_{\cdot j}-1)}{\left(1+\sqrt{1-4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1)-1-w_{j2}}\right) \cdot p_j^{\left(1+\sqrt{1-4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1)-1} \cdot (1-p_j)^{2n_{\cdot j}+1-\left(1+\sqrt{1-4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1)}.$$

If $(w_{j1}, w_{j2}) \in B_3, B_4,$ or B_5 then the joint distribution of w_{j1} and w_{j2} is

$$f(w_{j1},\, w_{j2}) = \binom{2}{w_{j2}} \cdot \left(\frac{2(n_{\cdot j} - 1)}{\left(1 - \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j} + 1) - 1 - w_{j2}} \right) \cdot p_j^{\left(1 - \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1) - 1}$$

$$\cdot (1 - p_j)^{2n_{\cdot j} + 1 - \left(1 - \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1)}$$

$$+ \binom{2}{w_{j2}} \cdot \left(\frac{2(n_{\cdot j} - 1)}{\left(1 + \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j} + 1) - 1 - w_{j2}} \right) \cdot p_j^{\left(1 + \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1) - 1}$$

$$\cdot (1 - p_j)^{2n_{\cdot j} + 1 - \left(1 + \sqrt{1 - 4w_{j1}^2/n_{\cdot j}}\right)(n_{\cdot j}+1)}.$$

If $(w_{j1}, w_{j2}) \in B_6$ then the joint distribution of w_{j1} and w_{j2} is

$$f(w_{j1}, w_{j2}) = \binom{2}{w_{j2}} \cdot \binom{2(n_{\cdot j} - 1)}{n_{\cdot j} - w_{j2}} \cdot p_j^{n_{\cdot j}} \cdot (1 - p_j)^{n_{\cdot j}}.$$

Now consider the joint transformation

$$r_{j1} = \frac{w_{j2}}{w_{j1}} = \frac{I_{1j1}}{w_j}$$

$$r_{j2} = w_{j1}.$$

Define the sets

$$(r_{j1}, r_{j2}) \in C_1$$

$$= \left\{ \left(0, \sqrt{n_{\cdot j}\left(\frac{1}{2n_{\cdot j} + 2}\right)\left(1 - \frac{1}{2n_{\cdot j} + 2}\right)}\right), \left(0, \sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j} + 2}\right)\left(1 - \frac{2}{2n_{\cdot j} + 2}\right)}\right) \right\}$$

$$C_2 = \left\{ \left(2 \Big/ \sqrt{n_{\cdot j}\left(\frac{1}{2n_{\cdot j} + 2}\right)\left(1 - \frac{1}{2n_{\cdot j} + 2}\right)}, \sqrt{n_{\cdot j}\left(\frac{1}{2n_{\cdot j} + 2}\right)\left(1 - \frac{1}{2n_{\cdot j} + 2}\right)} \right), \right.$$

$$\left(2\Big/\sqrt{n_{.j}\left(\frac{2}{2n_{.j}+2}\right)\left(1-\frac{2}{2n_{.j}+2}\right)},\sqrt{n_{.j}\left(\frac{2}{2n_{.j}+2}\right)\left(1-\frac{2}{2n_{.j}+2}\right)}\right)\right\}$$

$$C_3=\left\{\left(0,\sqrt{n_{.j}\left(\frac{3}{2n_{.j}+2}\right)\left(1-\frac{3}{2n_{.j}+2}\right)}\right),\left(0,\sqrt{n_{.j}\left(\frac{4}{2n_{.j}+2}\right)\left(1-\frac{4}{2n_{.j}+2}\right)}\right),\dots,\right.$$

$$\left.\left(0,\sqrt{n_{.j}\left(\frac{n_{.j}}{2n_{.j}+2}\right)\left(1-\frac{n_{.j}}{2n_{.j}+2}\right)}\right)\right\}$$

$$C_4=\left\{\left(2\Big/\sqrt{n_{.j}\left(\frac{3}{2n_{.j}+2}\right)\left(1-\frac{3}{2n_{.j}+2}\right)},\sqrt{n_{.j}\left(\frac{3}{2n_{.j}+2}\right)\left(1-\frac{3}{2n_{.j}+2}\right)}\right),,\dots,\right.$$

$$\left.\left(2\Big/\sqrt{n_{.j}\left(\frac{n_{.j}}{2n_{.j}+2}\right)\left(1-\frac{n_{.j}}{2n_{.j}+2}\right)},\sqrt{n_{.j}\left(\frac{n_{.j}}{2n_{.j}+2}\right)\left(1-\frac{n_{.j}}{2n_{.j}+2}\right)}\right)\right\}$$

$$C_5=\left\{\left(1\Big/\sqrt{n_{.j}\left(\frac{2}{2n_{.j}+2}\right)\left(1-\frac{2}{2n_{.j}+2}\right)},\sqrt{n_{.j}\left(\frac{2}{2n_{.j}+2}\right)\left(1-\frac{2}{2n_{.j}+2}\right)}\right),,\dots,\right.$$

$$\left.\left(1\Big/\sqrt{n_{.j}\left(\frac{n_{.j}}{2n_{.j}+2}\right)\left(1-\frac{n_{.j}}{2n_{.j}+2}\right)},\sqrt{n_{.j}\left(\frac{n_{.j}}{2n_{.j}+2}\right)\left(1-\frac{n_{.j}}{2n_{.j}+2}\right)}\right)\right\}$$

$$C_6=\left\{\left(0,\frac{\sqrt{n_{.j}}}{2}\right),\left(\frac{2}{\sqrt{n_{.j}}},\frac{\sqrt{n_{.j}}}{2}\right),\left(\frac{4}{\sqrt{n_{.j}}},\frac{\sqrt{n_{.j}}}{2}\right)\right\}$$

If $(r_{j1},r_{j2})\in C_1$ then the joint distribution of r_{j1} and r_{j2} is

$$f(r_{j1},r_{j2})=\binom{2}{r_{j1}r_{j2}}\cdot\begin{pmatrix}2(n_{.j}-1)\\ \left(1-\sqrt{1-4r_{j2}^2/n_{.j}}\right)(n_{.j}+1)-1-r_{j1}r_{j2}\end{pmatrix}\cdot p_j^{\left(1-\sqrt{1-4r_{j2}^2/n_{.j}}\right)(n_{.j}+1)-1}$$

$$\cdot\left(1-p_j\right)^{2n_{.j}+1-\left(1-\sqrt{1-4r_{j2}^2/n_{.j}}\right)(n_{.j}+1)}.$$

If $(r_{j1}, r_{j2}) \in C_2$ then the joint distribution of r_{j1} and r_{j2} is

$$f(r_{j1}, r_{j2}) = \binom{2}{r_{j1} r_{j2}} \cdot \binom{2(n_{\cdot j} - 1)}{\left(1 + \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1) - 1 - r_{j1} r_{j2}} \cdot p_j^{\left(1 + \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1) - 1}$$

$$\cdot (1 - p_j)^{2n_{\cdot j} + 1 - \left(1 + \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1)}.$$

If $(r_{j1}, r_{j2}) \in C_3, C_4$ or C_5 then the joint distribution of r_{j1} and r_{j2} is

$$f(r_{j1}, r_{j2}) = \binom{2}{r_{j1} r_{j2}} \cdot \binom{2(n_{\cdot j} - 1)}{\left(1 - \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1) - 1 - r_{j1} r_{j2}} \cdot p_j^{\left(1 - \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1) - 1}$$

$$\cdot (1 - p_j)^{2n_{\cdot j} + 1 - \left(1 - \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1)}$$

$$+ \binom{2}{r_{j1} r_{j2}} \cdot \binom{2(n_{\cdot j} - 1)}{\left(1 + \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1) - 1 - r_{j1} r_{j2}} \cdot p_j^{\left(1 + \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1) - 1}$$

$$\cdot (1 - p_j)^{2n_{\cdot j} + 1 - \left(1 + \sqrt{1 - 4r_{j2}^2 / n_{\cdot j}}\right)(n_{\cdot j} + 1)}.$$

If $(r_{j1}, r_{j2}) \in C_6$ then the joint distribution of r_{j1} and r_{j2} is

$$f(r_{j1}, r_{j2}) = \binom{2}{r_{j1} r_{j2}} \cdot \binom{2(n_{\cdot j} - 1)}{n_{\cdot j} - r_{j1} r_{j2}} \cdot p_j^{n_{\cdot j}} \cdot (1 - p_j)^{n_{\cdot j}}.$$

Now find the marginal distribution of r_{j1}. If $r_{j1} = 0$ then

$$f_1(r_{j1}) = \sum_{t=3}^{n_{\cdot j}} \left[\binom{2(n_{\cdot j} - 1)}{t - 1} \cdot p_j^{t-1} \cdot (1 - p_j)^{2n_{\cdot j} + 1 - t} + \binom{2(n_{\cdot j} - 1)}{2n_{\cdot j} + 1 - t} \cdot p_j^{2n_{\cdot j} + 1 - t} \right.$$

$$\left. \cdot (1 - p_j)^{t-1} \right] + \sum_{t=1}^{2} \left[\binom{2(n_{\cdot j} - 1)}{t - 1} \cdot p_j^{t-1} \cdot (1 - p_j)^{2n_{\cdot j} + 1 - t} \right]$$

$$+\binom{2(n_{\cdot j}-1)}{n_{\cdot j}}\cdot p_j^{n_{\cdot j}}\cdot(1-p_j)^{n_{\cdot j}}$$

If $r_{j1}\in\left\{2/\sqrt{n_{\cdot j}\left(\frac{1}{2n_{\cdot j}+2}\right)\left(1-\frac{1}{2n_{\cdot j}+2}\right)},2/\sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j}+2}\right)\left(1-\frac{2}{2n_{\cdot j}+2}\right)}\right\}$ then the distribution of r_{j1} is

$$f_2(r_{j1})=\left(\frac{2(n_{\cdot j}-1)}{\left(1+\sqrt{1-16/(r_{j1}^2n_{\cdot j})}\right)(n_{\cdot j}+1)-3}\right)\cdot p_j^{\left(1+\sqrt{1-16/\left(r_{j1}^2n_{\cdot j}\right)}\right)(n_{\cdot j}+1)-1}$$

$$\cdot(1-p_j)^{2n_{\cdot j}+1-\left(1+\sqrt{1-16/\left(r_{j1}^2n_{\cdot j}\right)}\right)(n_{\cdot j}+1)}.$$

If $r_{j1}\in\left\{1/\sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j}+2}\right)\left(1-\frac{2}{2n_{\cdot j}+2}\right)},\dots,1/\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}\right\}$ then the distribution of r_{j1} is

$$f_3(r_{j1})=2\cdot\left(\frac{2(n_{\cdot j}-1)}{\left(1-\sqrt{1-4/(r_{j1}^2n_{\cdot j})}\right)(n_{\cdot j}+1)-2}\right)\cdot p_j^{\left(1-\sqrt{1-4/\left(r_{j1}^2n_{\cdot j}\right)}\right)(n_{\cdot j}+1)-1}$$

$$\cdot(1-p_j)^{2n_{\cdot j}+1-\left(1-\sqrt{1-4/\left(r_{j1}^2n_{\cdot j}\right)}\right)(n_{\cdot j}+1)}+2$$

$$\cdot\left(\frac{2(n_{\cdot j}-1)}{\left(1+\sqrt{1-4/(r_{j1}^2n_{\cdot j})}\right)(n_{\cdot j}+1)-2}\right)\cdot p_j^{\left(1+\sqrt{1-4/\left(r_{j1}^2n_{\cdot j}\right)}\right)(n_{\cdot j}+1)-1}$$

$$\cdot(1-p_j)^{2n_{\cdot j}+1-\left(1+\sqrt{1-4/\left(r_{j1}^2n_{\cdot j}\right)}\right)(n_{\cdot j}+1)}.$$

If $r_{j1}\in\left\{2/\sqrt{n_{\cdot j}\left(\frac{3}{2n_{\cdot j}+2}\right)\left(1-\frac{3}{2n_{\cdot j}+2}\right)},\dots,2/\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}\right\}$ then the distribution of r_{j1} is

$$f_4(r_{j1}) = \left(\frac{2(n_{\cdot j}-1)}{\left(1-\sqrt{1-16/(r_{j1}^2 n_{\cdot j})}\right)(n_{\cdot j}+1)-3} \right) \cdot p_j^{\left(1-\sqrt{1-16/(r_{j1}^2 n_{\cdot j})}\right)(n_{\cdot j}+1)-1}$$

$$\cdot (1-p_j)^{2n_{\cdot j}+1-\left(1-\sqrt{1-16/(r_{j1}^2 n_{\cdot j})}\right)(n_{\cdot j}+1)}$$

$$+ \left(\frac{2(n_{\cdot j}-1)}{\left(1+\sqrt{1-16/(r_{j1}^2 n_{\cdot j})}\right)(n_{\cdot j}+1)-3} \right) \cdot p_j^{\left(1+\sqrt{1-16/(r_{j1}^2 n_{\cdot j})}\right)(n_{\cdot j}+1)-1}$$

$$\cdot (1-p_j)^{2n_{\cdot j}+1-\left(1+\sqrt{1-16/(r_{j1}^2 n_{\cdot j})}\right)(n_{\cdot j}+1)}.$$

If $r_{j1} \in \{2/\sqrt{n_{\cdot j}}, 4/\sqrt{n_{\cdot j}}\}$ then the distribution of r_{j1} is

$$f_5(r_{j1}) = \binom{2}{r_{j1} \cdot \sqrt{n_{\cdot j}}/2} \binom{2(n_{\cdot j}-1)}{n_{\cdot j} - r_{j1} \cdot \sqrt{n_{\cdot j}}/2} \cdot p_j^{n_{\cdot j}} \cdot (1-p_j)^{n_{\cdot j}}.$$

In order to find the distribution of the genetic score the joint distribution of $r_{11}, r_{21}, \ldots, r_{J1}$ is needed. Since it is assumed that the variants are independent this joint distribution is

$$f(r_{11}, r_{21}, \ldots, r_{J1}) = \prod_{j=1}^{J} f(r_{j1}).$$

However there are five different formulas for $f(r_{j1})$ depending on the value of r_{j1}. Hence there are 5^J different formulas for the joint distribution of $f(r_{11}, r_{21}, \ldots, r_{J1})$ depending on the values of $r_{11}, r_{21}, \ldots, r_{J1}$. It is possible to classify these distributions into five different cases. Let there be a_1 of the r_{j1}'s such that $r_{j1} \in \{2/\sqrt{n_{\cdot j}}, 4/\sqrt{n_{\cdot j}}\}$, a_2 of the r_{j1}'s such that $r_{j1} \in \left\{2/\sqrt{n_{\cdot j}\left(\frac{3}{2n_{\cdot j}+2}\right)\left(1-\frac{3}{2n_{\cdot j}+2}\right)}, \ldots, 2/\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}\right\}$, a_3 of the r_{j1}'s such that $r_{j1} \in \left\{1/\sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j}+2}\right)\left(1-\frac{2}{2n_{\cdot j}+2}\right)}, \ldots, 1/\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}\right\}$, a_4 of the r_{j1}'s such that

$r_{j1} \in \left\{2/\sqrt{n_{\cdot j}\left(\frac{1}{2n_{\cdot j}+2}\right)\left(1-\frac{1}{2n_{\cdot j}+2}\right)}, 2/\sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j}+2}\right)\left(1-\frac{2}{2n_{\cdot j}+2}\right)}\right\}$, and $a_5 = J - a_1 - a_2 - a_3 - a_4$ of the r_{j1}'s such that $r_{j1} = 0$. Let j' be an index for reordering the r_{j1}'s such that $r_{j'1} \in \{2/\sqrt{n_{\cdot j}}, 4/\sqrt{n_{\cdot j}}\}$ for $j' = 1, \ldots, a_1$,

$r_{j'1} \in \left\{2/\sqrt{n_{\cdot j}\left(\frac{3}{2n_{\cdot j}+2}\right)\left(1-\frac{3}{2n_{\cdot j}+2}\right)}, \ldots, 2/\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}\right\}$ for $j' = a_1 + 1, \ldots, a_1 + a_2$, $r_{j'1} \in \left\{1/\sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j}+2}\right)\left(1-\frac{2}{2n_{\cdot j}+2}\right)}, \ldots, 1/\sqrt{n_{\cdot j}\left(\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)\left(1-\frac{n_{\cdot j}}{2n_{\cdot j}+2}\right)}\right\}$ for $j' = a_1 + a_2 + 1, \ldots, a_1 + a_2 + a_3$, $r_{j'1} \in \left\{2/\sqrt{n_{\cdot j}\left(\frac{1}{2n_{\cdot j}+2}\right)\left(1-\frac{1}{2n_{\cdot j}+2}\right)}, 2/\sqrt{n_{\cdot j}\left(\frac{2}{2n_{\cdot j}+2}\right)\left(1-\frac{2}{2n_{\cdot j}+2}\right)}\right\}$ for $j' = a_1 + a_2 + a_3 + 1, \ldots, a_1 + a_2 + a_3 + a_4$, and $r_{j'1} = 0$ for $j' = a_1 + a_2 + a_3 + a_4 + 1, \ldots, J$. Extending the notation, let $n_{\cdot j'}$ be the number of individuals genotyped for variant j' and $p_{j'}$ be the probability of a mutant allele at variant j'. The general form of the joint distribution of $r_{11}, r_{21}, \ldots, r_{J1}$ can be rewritten as

$$f(r_{11}, r_{21}, \ldots, r_{J1}) = \prod_{j'=1}^{a_1} f_5(r_{j'1}) \cdot \prod_{j'=a_1+1}^{a_1+a_2} f_4(r_{j'1}) \cdot \prod_{j'=a_1+a_2+1}^{a_1+a_2+a_3} f_3(r_{j'1}) \cdot \prod_{j'=a_1+a_2+a_3+1}^{a_1+a_2+a_3+a_4} f_2(r_{j'1}) \cdot \prod_{j'=a_1+a_2+a_3+a_4+1}^{J} f_1(r_{j'1}).$$

Now consider the joint transformation $g_1 = \sum_{j'=1}^{J} r_{j'1}$, $g_2 = r_{2'1}, g_3 = r_{3'1}, \ldots, g_J = r_{J'1}$. Clearly the formulas for most of the variables will be similar since all but one of the $r_{j'1}$'s is directly transformed to a $g_{j'}$. However $r_{1'1} = g_1 - \sum_{j'=2}^{J} r_{j'1}$ thus the distribution of $r_{1'1}$ must be considered when finding the joint distribution of $g_1, g_2, \ldots, g_J$. First consider a case 1 where at least one $r_{j'1} \in \{2/\sqrt{n_{\cdot j}}, 4/\sqrt{n_{\cdot j}}\}$. Then the joint distribution of $g_1, g_2, \ldots, g_J$ is

$$f(g_1, g_2, \dots, g_J)$$

$$= \left(\left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)\left(g_1 - \sum_{j'=2}^{J} g_{j'}\right)^2\right)\left(n_{\cdot 1'} - \left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)\left(g_1 - \sum_{j'=2}^{J} g_{j'}\right)^{2(n_{\cdot 1'}-1)}\right) \cdot p_{1'}^{n_{\cdot 1'}}$$

$$\cdot (1 - p_{1'})^{n_{\cdot 1'}} \cdot \prod_{j'=2}^{a_1} f_5(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_1+1}^{a_1+a_2} f_4(r_{j'1} = g_{j'})$$

$$\cdot \prod_{j'=a_1+a_2+1}^{a_1+a_2+a_3} f_3(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_1+a_2+a_3+1}^{a_1+a_2+a_3+a_4} f_2(r_{j'1} = g_{j'})$$

$$\cdot \prod_{j'=a_1+a_2+a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

where $g_1 - \sum_{j'=2}^{J} g_{j'} \in \{2/\sqrt{n_{\cdot 1'}}, 4/\sqrt{n_{\cdot 1'}}\}$. Now sum over $g_{a_1+a_2+a_3+a_4+1}, \dots, g_J$. Since each of these values is zero and the $f_1(g_{j'})$'s are constants the joint distribution of $g_1, \dots, g_{a_1+a_2+a_3+a_4}$ is

$$f(g_1, g_2, \dots, g_{a_1+a_2+a_3+a_4})$$

$$= \left(\left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)\left(g_1 - \sum_{j'=2}^{a_1+a_2+a_3+a_4} g_{j'}\right)^2\right)\left(n_{\cdot 1'} - \left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)\left(g_1 - \sum_{j'=2}^{a_1+a_2+a_3+a_4} g_{j'}\right)^{2(n_{\cdot 1'}-1)}\right) \cdot p_{1'}^{n_{\cdot 1'}}$$

$$\cdot (1 - p_{1'})^{n_{\cdot 1'}} \cdot \prod_{j'=2}^{a_1} f_5(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_1+1}^{a_1+a_2} f_4(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_1+a_2+1}^{a_1+a_2+a_3} f_3(r_{j'1} = g_{j'})$$

$$\cdot \prod_{j'=a_1+a_2+a_3+1}^{a_1+a_2+a_3+a_4} f_2(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_1+a_2+a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

where $g_1 - \sum_{j'=2}^{a_1+a_2+a_3+a_4} g_{j'} \in \{2/\sqrt{n_{\cdot 1'}}, 4/\sqrt{n_{\cdot 1'}}\}$. Now sum over $g_{a_1+a_2+a_3+1}, \dots, g_{a_1+a_2+a_3+a_4}$. Let

$diff_1 = g_1 - \sum_{j'=2}^{a_1+a_2+a_3} g_{j'} - \sum_{i=1}^{a_4} \frac{2}{\sqrt{n_{\cdot(a_1+a_2+a_3+i)'}\left(\frac{u_i}{2n_{\cdot(a_1+a_2+a_3+i)'+2}}\right)\left(1-\frac{u_i}{2n_{\cdot(a_1+a_2+a_3+i)'+2}}\right)}}$ for $u_i \in \{1,2\}$. Then the joint distribution of $g_1, \dots, g_{a_1+a_2+a_3}$ is

$$f\left(g_1, g_2, \dots, g_{a_1+a_2+a_3}\right)$$

$$= p_{1'}^{n_{\cdot 1'}} \cdot (1-p_{1'})^{n_{\cdot 1'}} \cdot \prod_{j'=2}^{a_1} f_5\left(r_{j'1} = g_{j'}\right) \cdot \prod_{j'=a_1+1}^{a_1+a_2} f_4\left(r_{j'1} = g_{j'}\right)$$

$$\cdot \prod_{j'=a_1+a_2+1}^{a_1+a_2+a_3} f_3\left(r_{j'1} = g_{j'}\right) \cdot \prod_{j'=a_1+a_2+a_3+a_4+1}^{J} f_1\left(r_{j'1} = g_{j'}\right)$$

$$\cdot \left[\sum_{u_{a_4}=1}^{2} \dots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2\left(n_{\cdot(a_1+a_2+a_3+i)'} - 1\right)}{2n_{\cdot(a_1+a_2+a_3+i)'} - 1 - u_i} \cdot p_{(a_1+a_2+a_3+i)'}^{2n_{\cdot(a_1+a_2+a_3+i)'}+1-u_i} \right.\right.\right.$$

$$\left.\left.\left. \cdot \left(1 - p_{(a_1+a_2+a_3+i)'}\right)^{u_i-1} \right] \right.\right.$$

$$\left.\left. \cdot \left(\left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)^2 (diff_1) \right) \cdot \left(\binom{2(n_{\cdot 1'} - 1)}{n_{\cdot 1'} - \left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)(diff_1)} \right) \right\} \right]$$

where $diff_1 \in \{2/\sqrt{n_{\cdot 1'}}, 4/\sqrt{n_{\cdot 1'}}\}$. Now sum over $g_{a_1+a_2+1}, \dots, g_{a_1+a_2+a_3}$. Let $diff_2 = g_1 - \sum_{j'=2}^{a_1+a_2} g_{j'} - \sum_{i=1}^{a_4} \frac{2}{\sqrt{n_{\cdot(a_1+a_2+a_3+i)'}\left(\frac{u_i}{2n_{\cdot(a_1+a_2+a_3+i)'+2}}\right)\left(1-\frac{u_i}{2n_{\cdot(a_1+a_2+a_3+i)'+2}}\right)}}$

$- \sum_{d=1}^{a_3} \frac{1}{\sqrt{n_{\cdot(a_1+a_2+d)'}\left(\frac{v_d}{2n_{\cdot(a_1+a_2+d)'+2}}\right)\left(1-\frac{v_d}{2n_{\cdot(a_1+a_2+d)'+2}}\right)}}$ where $u_i \in \{1,2\}$ and

$v_d \in \left\{2, \dots, n_{\cdot(a_1+a_2+d)'}\right\}$. The joint distribution of $g_1, \dots, g_{a_1+a_2}$ is

$$f(g_1, g_2, \dots, g_{a_1+a_2})$$

$$= p_{1'}^{n_{\cdot 1'}} \cdot (1-p_{1'})^{n_{\cdot 1'}} \cdot \prod_{j'=2}^{a_1} f_5(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_1+1}^{a_1+a_2} f_4(r_{j'1} = g_{j'})$$

$$\cdot \prod_{j'=a_1+a_2+a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{v_{a_3}=2}^{n_{\cdot(a_1+a_2+a_3)'}} \dots \sum_{v_1=2}^{n_{\cdot(a_1+a_2+1)'}} \left\{ \prod_{d=1}^{a_3} \left[2 \binom{2(n_{\cdot(a_1+a_2+d)'}-1)}{v_d-2} \cdot p_{(a_1+a_2+d)'}^{v_d-1} \right.\right.\right.$$

$$\cdot \left(1-p_{(a_1+a_2+d)'}\right)^{2n_{\cdot(a_1+a_2+d)'}+1-v_d}$$

$$\left.\left.\left. +2 \binom{2(n_{\cdot(a_1+a_2+d)'}-1)}{2n_{\cdot(a_1+a_2+d)'}-v_d} p_{(a_1+a_2+d)'}^{2n_{\cdot(a_1+a_2+d)'}+1-v_d} \left(1-p_{(a_1+a_2+d)'}\right)^{v_d-1} \right] \right.$$

$$\sum_{u_{a_4}=1}^{2} \dots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2(n_{\cdot(a_1+a_2+a_3+i)'}-1)}{2n_{\cdot(a_1+a_2+a_3+i)'}-1-u_i} \cdot p_{(a_1+a_2+a_3+i)'}^{2n_{\cdot(a_1+a_2+a_3+i)'}+1-u_i} \right.\right.$$

$$\left.\left. \cdot \left(1-p_{(a_1+a_2+a_3+i)'}\right)^{u_i-1} \right] \right.$$

$$\left.\left. \cdot \left(\left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)^2 (diff_2) \right) \cdot \left(\binom{2(n_{\cdot 1'}-1)}{n_{\cdot 1'} - \left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)(diff_2)} \right) \right\} \right\} \right]$$

where $diff_2 \in \{2/\sqrt{n_{\cdot 1'}}, 4/\sqrt{n_{\cdot 1'}}\}$. Now sum over $g_{a_1+1}, \dots, g_{a_1+a_2}$. Let $diff_3 = g_1 -$

$$\sum_{j'=2}^{a_1} g_{j'} - \sum_{i=1}^{a_4} \frac{2}{\sqrt{n_{\cdot(a_1+a_2+a_3+i)'} \left(\frac{u_i}{2n_{\cdot(a_1+a_2+a_3+i)'+2}}\right)\left(1-\frac{u_i}{2n_{\cdot(a_1+a_2+a_3+i)'+2}}\right)}} -$$

$$\sum_{d=1}^{a_3} \frac{1}{\sqrt{n_{\cdot(a_1+a_2+d)'} \left(\frac{v_d}{2n_{\cdot(a_1+a_2+d)'+2}}\right)\left(1-\frac{v_d}{2n_{\cdot(a_1+a_2+d)'+2}}\right)}} - \sum_{e=1}^{a_2} \frac{2}{\sqrt{n_{\cdot(a_1+e)'} \left(\frac{x_e}{2n_{\cdot(a_1+e)'+2}}\right)\left(1-\frac{x_e}{2n_{\cdot(a_1+e)'+2}}\right)}}$$

where $u_i \in \{1,2\}$, $v_d \in \{2,\dots,n_{\cdot(a_1+a_2+d)'}\}$, and $x_e \in \{3,\dots,n_{\cdot(a_1+e)'}\}$. Then the joint distribution of $g_1,\dots,g_{a_1}$ is

$$f(g_1,g_2,\dots,g_{a_1}) = p_{1'}^{n_{\cdot 1'}} \cdot (1-p_{1'})^{n_{\cdot 1'}} \cdot \prod_{j'=2}^{a_1} f_5(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_1+a_2+a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{x_{a_2}=3}^{n_{a_1+a_2}} \cdots \sum_{x_1=3}^{n_{a_1+1}} \left\{ \prod_{e=1}^{a_2} \left[\binom{2(n_{\cdot(a_1+e)'}-1)}{x_e-3} p_{(a_1+e)'}^{x_e-1} (1-p_{(a_1+e)'})^{2n_{\cdot(a_1+e)'}+1-x_e} \right.\right.\right.$$

$$\left.\left.\left. + \binom{2(n_{\cdot(a_1+e)'}-1)}{2n_{\cdot(a_1+e)'}-x_e-1} p_{(a_1+e)'}^{2n_{\cdot(a_1+e)'}+1-x_e} (1-p_{(a_1+e)'})^{x_e-1} \right] \right.$$

$$\cdot \sum_{v_{a_3}=2}^{n_{\cdot(a_1+a_2+a_3)'}} \cdots \sum_{v_1=2}^{n_{\cdot(a_1+a_2+1)'}} \left\{ \prod_{d=1}^{a_3} \left[2 \binom{2(n_{\cdot(a_1+a_2+d)'}-1)}{v_d-2} \cdot p_{(a_1+a_2+d)'}^{v_d-1} \right.\right.$$

$$\cdot (1-p_{(a_1+a_2+d)'})^{2n_{\cdot(a_1+a_2+d)'}+1-v_d}$$

$$\left.\left. + 2 \binom{2(n_{\cdot(a_1+a_2+d)'}-1)}{2n_{\cdot(a_1+a_2+d)'}-v_d} p_{(a_1+a_2+d)'}^{2n_{\cdot(a_1+a_2+d)'}+1-v_d} (1-p_{(a_1+a_2+d)'})^{v_d-1} \right] \right.$$

$$\sum_{u_{a_4}=1}^{2} \cdots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2(n_{\cdot(a_1+a_2+a_3+i)'}-1)}{2n_{\cdot(a_1+a_2+a_3+i)'}-1-u_i} \cdot p_{(a_1+a_2+a_3+i)'}^{2n_{\cdot(a_1+a_2+a_3+i)'}+1-u_i} \right.\right.$$

$$\left.\left. \cdot (1-p_{(a_1+a_2+a_3+i)'})^{u_i-1} \right] \right.$$

$$\left.\left. \cdot \left(\left(\frac{\sqrt{n_{\cdot 1'}}}{2} \right)^2 (diff_3) \right) \cdot \left(\binom{2(n_{\cdot 1'}-1)}{n_{\cdot 1'} - \left(\frac{\sqrt{n_{\cdot 1'}}}{2} \right)(diff_3)} \right) \right\} \right\} \right\} \right]$$

where $diff_3 \in \{2/\sqrt{n_{\cdot 1'}}, 4/\sqrt{n_{\cdot 1'}}\}$. Now sum over $g_2,\dots,g_{a_1}$. Let

$diff_4 =$

$$g_1 - \sum_{i=1}^{a_4} \frac{2}{\sqrt{n_{\cdot(a_1+a_2+a_3+i)'}\left(\frac{u_i}{2n_{\cdot(a_1+a_2+a_3+i)'}+2}\right)\left(1-\frac{u_i}{2n_{\cdot(a_1+a_2+a_3+i)'}+2}\right)}} -$$

$$\sum_{d=1}^{a_3} \frac{1}{\sqrt{n_{\cdot(a_1+a_2+d)'}\left(\frac{v_d}{2n_{\cdot(a_1+a_2+d)'}+2}\right)\left(1-\frac{v_d}{2n_{\cdot(a_1+a_2+d)'}+2}\right)}} - \sum_{e=1}^{a_2} \frac{2}{\sqrt{n_{\cdot(a_1+e)'}\left(\frac{x_e}{2n_{\cdot(a_1+e)'}+2}\right)\left(1-\frac{x_e}{2n_{\cdot(a_1+e)'}+2}\right)}} -$$

$\sum_{h=2}^{a_1} \frac{2s_h}{\sqrt{n_{\cdot h'}}}$ where $u_i \in \{1,2\}$, $v_d \in \{2,\dots,n_{\cdot(a_1+a_2+d)'}\}$, $x_e \in \{3,\dots,n_{\cdot(a_1+e)'}\}$, and $s_h \in \{1,2\}$.

Then the distribution of g_1, the genetic score, for this case is

$$f(g_1) = p_{1'}^{n_{\cdot 1'}} \cdot (1-p_{1'})^{n_{\cdot 1'}} \cdot \prod_{j'=a_1+a_2+a_3+a_4+1}^{J} f_1\left(r_{j'1} = g_{j'}\right)$$

$$\cdot \left[\sum_{s_{a_1}=1}^{2} \cdots \sum_{s_2=1}^{2} \left\{ \prod_{h=2}^{a_1} \left[\binom{2}{s_h} \binom{2(n_{\cdot h'}-1)}{n_{\cdot h'}-s_h} p_{h'}^{n_{\cdot h'}} (1-p_{h'})^{n_{\cdot h'}} \right] \right.$$

$$\sum_{x_{a_2}=3}^{n_{a_1+a_2}} \cdots \sum_{x_1=3}^{n_{a_1+1}} \left\{ \prod_{e=1}^{a_2} \left[\left(\binom{2(n_{\cdot(a_1+e)'}-1)}{x_e-3} p_{(a_1+e)'}^{x_e-1} \left(1-p_{(a_1+e)'}\right)^{2n_{\cdot(a_1+e)'}+1-x_e} \right.\right.\right.$$

$$\left.\left.\left. + \binom{2(n_{\cdot(a_1+e)'}-1)}{2n_{\cdot(a_1+e)'}-x_e-1} p_{(a_1+e)'}^{2n_{\cdot(a_1+e)'}+1-x_e} \left(1-p_{(a_1+e)'}\right)^{x_e-1} \right] \right.$$

$$\cdot \sum_{v_{a_3}=2}^{n_{\cdot(a_1+a_2+a_3)'}} \cdots \sum_{v_1=2}^{n_{\cdot(a_1+a_2+1)'}} \left\{ \prod_{d=1}^{a_3} \left[2 \binom{2(n_{\cdot(a_1+a_2+d)'}-1)}{v_d-2} \cdot p_{(a_1+a_2+d)'}^{v_d-1} \right.\right.$$

$$\cdot \left(1-p_{(a_1+a_2+d)'}\right)^{2n_{\cdot(a_1+a_2+d)'}+1-v_d}$$

$$\left.\left. \left. +2 \binom{2(n_{\cdot(a_1+a_2+d)'}-1)}{2n_{\cdot(a_1+a_2+d)'}-v_d} p_{(a_1+a_2+d)'}^{2n_{\cdot(a_1+a_2+d)'}+1-v_d} \left(1-p_{(a_1+a_2+d)'}\right)^{v_d-1} \right] \right.$$

$$\cdot \sum_{u_{a_4}=1}^{2} \cdots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2\left(n_{\cdot(a_1+a_2+a_3+i)'} - 1\right)}{2n_{\cdot(a_1+a_2+a_3+i)'} - 1 - u_i} \cdot p_{(a_1+a_2+a_3+i)'}^{2n_{\cdot(a_1+a_2+a_3+i)'}+1-u_i} \right.\right.$$

$$\left.\cdot \left(1 - p_{(a_1+a_2+a_3+i)'}\right)^{u_i-1} \right]$$

$$\left.\left.\cdot \left(\binom{2}{\left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)(diff_4)} \cdot \binom{2(n_{\cdot 1'}-1)}{n_{\cdot 1'} - \left(\frac{\sqrt{n_{\cdot 1'}}}{2}\right)(diff_4)} \right) \right\} \right\} \right]$$

where $diff_4 \in \{2/\sqrt{n_{\cdot 1'}}, 4/\sqrt{n_{\cdot 1'}}\}$.

Now consider a second case where $a_1 = 0$ and $a_2 > 0$. Then

$r_{1'1} \in \left\{ 2/\sqrt{n_{\cdot 1'}\left(\frac{3}{2n_{\cdot 1'}+2}\right)\left(1 - \frac{3}{2n_{\cdot 1'}+2}\right)}, \ldots, 2/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1 - \frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)} \right\}$. Transforming to $g_1, g_2, \ldots, g_J$ gives

$$f(g_1, g_2, \ldots, g_J)$$

$$= \prod_{j'=2}^{a_2} f_4(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_2+1}^{a_2+a_3} f_3(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_2+a_3+1}^{a_2+a_3+a_4} f_2(r_{j'1} = g_{j'})$$

$$\cdot \prod_{j'=a_2+a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot\left[\left(\left(1-\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J}g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-3\right)\right.$$
$$\cdot p_{1'}^{\left(1-\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J}g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-1}$$
$$\cdot(1-p_{1'})^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J}g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)}$$
$$+\left(\left(1+\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J}g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-3\right)$$
$$\cdot p_{1'}^{\left(1+\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J}g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-1}$$
$$\left.\cdot(1-p_{1'})^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J}g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)}\right]$$

where $g_1-\sum_{j'=2}^{J}g_{j'}\in\left\{2/\sqrt{n_{\cdot 1'}\left(\frac{3}{2n_{\cdot 1'}+2}\right)\left(1-\frac{3}{2n_{\cdot 1'}+2}\right)},\ldots,2/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)}\right\}$.

Now sum over $g_{a_2+a_3+a_4+1},\ldots,g_J$. These variables are all equal to zero and $f_1(g_{j'})$ is a constant so

$$f(g_1,g_2,\ldots,g_{a_2+a_3+a_4})$$
$$=\prod_{j'=2}^{a_2}f_4(r_{j'1}=g_{j'})\cdot\prod_{j'=a_2+1}^{a_2+a_3}f_3(r_{j'1}=g_{j'})\cdot\prod_{j'=a_2+a_3+1}^{a_2+a_3+a_4}f_2(r_{j'1}=g_{j'})$$
$$\cdot\prod_{j'=a_2+a_3+a_4+1}^{J}f_1(r_{j'1}=g_{j'})$$

$$\cdot\left[\left(\left(1-\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{a_2+a_3+a_4} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-3\right)\right.$$
$$\cdot p_{1'}^{\left(1-\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{a_2+a_3+a_4} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-1}$$
$$\cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{a_2+a_3+a_4} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)}$$
$$+\left(\left(1+\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{a_2+a_3+a_4} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-3\right)$$
$$\cdot p_{1'}^{\left(1+\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{a_2+a_3+a_4} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-1}$$
$$\left.\cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{a_2+a_3+a_4} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)}\right]$$

where

$$g_1-\sum_{j'=2}^{a_2+a_3+a_4} g_{j'} \in \left\{2/\sqrt{n_{\cdot 1'}\left(\frac{3}{2n_{\cdot 1'}+2}\right)\left(1-\frac{3}{2n_{\cdot 1'}+2}\right)}, \ldots, 2/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)}\right\}.$$

Now sum over $g_{a_2+a_3+1}, \ldots, g_{a_2+a_3+a_4}$. Let

$$diff_5 = g_1-\sum_{j'=2}^{a_2+a_3} g_{j'}-\sum_{i=1}^{a_4}\frac{2}{\sqrt{n_{\cdot(a_2+a_3+i)'}\left(\frac{u_i}{2n_{\cdot(a_2+a_3+i)'}+2}\right)\left(1-\frac{u_i}{2n_{\cdot(a_2+a_3+i)'}+2}\right)}} \text{ where } u_i \in \{1,2\}.$$

Then the joint distribution of $g_1, g_2, \ldots, g_{a_2+a_3}$ is

$$f(g_1, g_2, \dots, g_{a_2+a_3})$$

$$= \prod_{j'=2}^{a_2} f_4(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_2+1}^{a_2+a_3} f_3(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_2+a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{u_{a_4}=1}^{2} \cdots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2(n_{\cdot(a_2+a_3+i)'} - 1)}{2n_{\cdot(a_2+a_3+i)'} - 1 - u_i} \cdot p_{(a_2+a_3+i)'}^{2n_{\cdot(a_2+a_3+i)'}+1-u_i} \cdot \left(1 - p_{(a_2+a_3+i)'}\right)^{u_i-1} \right] \right.\right.$$

$$\left[\left(\frac{2(n_{\cdot 1'} - 1)}{\left(1 - \sqrt{1 - 16/(n_{\cdot 1'}(diff_5)^2)}\right)(n_{\cdot 1'} + 1) - 3} \right) \cdot p_{1'}^{\left(1 - \sqrt{1-16/(n_{\cdot 1'}(diff_5)^2)}\right)(n_{\cdot 1'}+1)-1} \right.$$

$$\left. \cdot \left(1 - p_{1'}\right)^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-16/(n_{\cdot 1'}(diff_5)^2)}\right)(n_{\cdot 1'}+1)} \right.$$

$$\left. + \left(\frac{2(n_{\cdot 1'} - 1)}{\left(1 + \sqrt{1 - 16/(n_{\cdot 1'}(diff_5)^2)}\right)(n_{\cdot 1'} + 1) - 3} \right) \cdot p_{1'}^{\left(1 + \sqrt{1-16/(n_{\cdot 1'}(diff_5)^2)}\right)(n_{\cdot 1'}+1)-1} \right.$$

$$\left. \left. \left. \cdot \left(1 - p_{1'}\right)^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_5)^2)}\right)(n_{\cdot 1'}+1)} \right] \right\} \right]$$

where $diff_5 \in \left\{ 2/\sqrt{n_{\cdot 1'}\left(\frac{3}{2n_{\cdot 1'}+2}\right)\left(1 - \frac{3}{2n_{\cdot 1'}+2}\right)}, \dots, 2/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1 - \frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)} \right\}$. Now sum over $g_{a_2+1}, \dots, g_{a_2+a_3}$. Let

$diff_6 =$

$$g_1 - \sum_{j'=2}^{a_2} g_{j'} - \sum_{i=1}^{a_4} \frac{2}{\sqrt{n_{\cdot(a_2+a_3+i)'}\left(\frac{u_i}{2n_{\cdot(a_2+a_3+i)'}+2}\right)\left(1 - \frac{u_i}{2n_{\cdot(a_2+a_3+i)'}+2}\right)}} -$$

$$\sum_{d=1}^{a_3} \frac{1}{\sqrt{n_{\cdot(a_2+d)'}\left(\frac{v_d}{2n_{\cdot(a_2+d)'}+2}\right)\left(1 - \frac{v_d}{2n_{\cdot(a_2+d)'}+2}\right)}}$$ where $u_i \in \{1, 2\}$ and $v_d \in \left\{2, \dots, n_{\cdot(a_2+d)'}\right\}$. Then the joint distribution of $g_1, \dots, g_{a_2}$ is

$$f(g_1,\dots,g_{a_2}) = \prod_{j'=2}^{a_2} f_4(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_2+a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{v_{a_3}=2}^{n_{\cdot(a_1+a_2+a_3)'}} \dots \sum_{v_1=2}^{n_{\cdot(a_1+a_2+1)'}} \left\{ \prod_{d=1}^{a_3} \left[2\binom{2\left(n_{\cdot(a_2+d)'}-1\right)}{v_d-2} \cdot p_{(a_2+d)'}^{v_d-1} \cdot \left(1-p_{(a_2+d)'}\right)^{2n_{\cdot(a_2+d)'}+1-v_d} \right.\right.\right.$$

$$\left.\left.\left. +2\binom{2\left(n_{\cdot(a_2+d)'}-1\right)}{2n_{\cdot(a_2+d)'}-v_d} p_{(a_2+d)'}^{2n_{\cdot(a_2+d)'}+1-v_d} \left(1-p_{(a_2+d)'}\right)^{v_d-1} \right] \right.$$

$$\sum_{u_{a_4}=1}^{2} \dots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2\left(n_{\cdot(a_2+a_3+i)'}-1\right)}{2n_{\cdot(a_2+a_3+i)'}-1-u_i} \cdot p_{(a_2+a_3+i)'}^{2n_{\cdot(a_2+a_3+i)'}+1-u_i} \cdot \left(1-p_{(a_2+a_3+i)'}\right)^{u_i-1} \right] \right.$$

$$\left[\left(\frac{2(n_{\cdot 1'}-1)}{\left(1-\sqrt{1-16/(n_{\cdot 1'}(diff_6)^2)}\right)(n_{\cdot 1'}+1)-3} \right) \cdot p_{1'}^{\left(1-\sqrt{1-16/(n_{\cdot 1'}(diff_6)^2)}\right)(n_{\cdot 1'}+1)-1} \right.$$

$$\cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-16/(n_{\cdot 1'}(diff_6)^2)}\right)(n_{\cdot 1'}+1)}$$

$$+\left(\frac{2(n_{\cdot 1'}-1)}{\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_6)^2)}\right)(n_{\cdot 1'}+1)-3} \right) \cdot p_{1'}^{\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_6)^2)}\right)(n_{\cdot 1'}+1)-1}$$

$$\left.\left.\left. \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_6)^2)}\right)(n_{\cdot 1'}+1)} \right] \right\} \right\} \right]$$

where $diff_6 \in \left\{ 2/\sqrt{n_{\cdot 1'}\left(\frac{3}{2n_{\cdot 1'}+2}\right)\left(1-\frac{3}{2n_{\cdot 1'}+2}\right)}, \dots, 2/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)} \right\}$. Now sum over $g_2,\dots,g_{a_2}$. Let $diff_7 = g_1 - \sum_{i=1}^{a_4} \frac{2}{\sqrt{n_{\cdot(a_2+a_3+i)'}\left(\frac{u_i}{2n_{\cdot(a_2+a_3+i)'}+2}\right)\left(1-\frac{u_i}{2n_{\cdot(a_2+a_3+i)'}+2}\right)}} - \sum_{d=1}^{a_3} \frac{1}{\sqrt{n_{\cdot(a_2+d)'}\left(\frac{v_d}{2n_{\cdot(a_2+d)'}+2}\right)\left(1-\frac{v_d}{2n_{\cdot(a_2+d)'}+2}\right)}} - \sum_{e=2}^{a_2} \frac{2}{\sqrt{n_{\cdot e'}\left(\frac{x_e}{2n_{\cdot e'}+2}\right)\left(1-\frac{x_e}{2n_{\cdot e'}+2}\right)}}$ where $u_i \in \{1,2\}$,

$v_d \in \{2, \dots, n_{\cdot(a_2+d)'}\}$, and $x_e \in \{3, \dots, n_{\cdot e'}\}$. Then the distribution of the genetic score, g_1, for this case is

$$f(g_1) = \prod_{j'=a_2+a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{x_{a_2}=3}^{n_{a_2}} \cdots \sum_{x_2=3}^{n_2} \left\{ \prod_{e=2}^{a_2} \left[\binom{2(n_{\cdot e'}-1)}{x_e-3} p_{e'}^{x_e-1} (1-p_{e'})^{2n_{\cdot e'}+1-x_e} \right.\right.\right.$$

$$\left.\left.\left. + \binom{2(n_{\cdot e'}-1)}{2n_{\cdot e'}-x_e-1} p_{e'}^{2n_{\cdot e'}+1-x_e} (1-p_{e'})^{x_e-1} \right] \right.$$

$$\sum_{v_{a_3}=2}^{n_{\cdot(a_1+a_2+a_3)'}} \cdots \sum_{v_1=2}^{n_{\cdot(a_1+a_2+1)'}} \left\{ \prod_{d=1}^{a_3} \left[2 \binom{2(n_{\cdot(a_2+d)'}-1)}{v_d-2} \cdot p_{(a_2+d)'}^{v_d-1} \cdot \left(1-p_{(a_2+d)'}\right)^{2n_{\cdot(a_2+d)'}+1-v_d} \right.\right.$$

$$\left.\left. + 2 \binom{2(n_{\cdot(a_2+d)'}-1)}{2n_{\cdot(a_2+d)'}-v_d} p_{(a_2+d)'}^{2n_{\cdot(a_2+d)'}+1-v_d} \left(1-p_{(a_2+d)'}\right)^{v_d-1} \right] \right.$$

$$\cdot \sum_{u_{a_4}=1}^{2} \cdots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2(n_{\cdot(a_2+a_3+i)'}-1)}{2n_{\cdot(a_2+a_3+i)'}-1-u_i} \cdot p_{(a_2+a_3+i)'}^{2n_{\cdot(a_2+a_3+i)'}+1-u_i} \cdot \left(1-p_{(a_2+a_3+i)'}\right)^{u_i-1} \right] \right.$$

$$\left[\left(\frac{2(n_{\cdot 1'}-1)}{\left(1-\sqrt{1-16/(n_{\cdot 1'}(diff_7)^2)}\right)(n_{\cdot 1'}+1)-3} \right) \cdot p_{1'}^{\left(1-\sqrt{1-16/(n_{\cdot 1'}(diff_7)^2)}\right)(n_{\cdot 1'}+1)-1} \right.$$

$$\cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-16/(n_{\cdot 1'}(diff_7)^2)}\right)(n_{\cdot 1'}+1)}$$

$$+ \left(\frac{2(n_{\cdot 1'}-1)}{\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_7)^2)}\right)(n_{\cdot 1'}+1)-3} \right) \cdot p_{1'}^{\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_7)^2)}\right)(n_{\cdot 1'}+1)-1}$$

$$\left. \left. \left. \left. \left. \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_7)^2)}\right)(n_{\cdot 1'}+1)} \right] \right\} \right\} \right\} \right]$$

where $diff_7 \in \left\{2/\sqrt{n_{\cdot 1'}\left(\frac{3}{2n_{\cdot 1'}+2}\right)\left(1-\frac{3}{2n_{\cdot 1'}+2}\right)}, \ldots, 2/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)}\right\}$.

Next consider case 3 where $a_1 = a_2 = 0$ and $a_3 > 0$. In this case $r_{1'1} \in \left\{1/\sqrt{n_{\cdot 1'}\left(\frac{2}{2n_{\cdot 1'}+2}\right)\left(1-\frac{2}{2n_{\cdot 1'}+2}\right)}, \ldots, 1/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)}\right\}$. Thus the joint distribution of $g_1, g_2, \ldots, g_J$ is

$$f(g_1, g_2, \ldots, g_J) = \prod_{j'=3}^{a_3} f_3(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_3+1}^{a_3+a_4} f_2(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[2 \cdot \left(\left(1 - \sqrt{1 - 4/\left(n_{\cdot 1'} \left(g_1 - \sum_{j'=2}^{J} g_{j'} \right)^2 \right)} \right) (n_{\cdot 1'} + 1) - 2 \right) \right.$$

$$\cdot p_{1'}^{\left(1-\sqrt{1-4/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-1}$$

$$\cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-4/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)}$$

$$+ \left(\left(1 + \sqrt{1 - 4/\left(n_{\cdot 1'} \left(g_1 - \sum_{j'=2}^{J} g_{j'} \right)^2 \right)} \right) (n_{\cdot 1'} + 1) - 2 \right)$$

$$\cdot p_{1'}^{\left(1+\sqrt{1-4/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)-1}$$

$$\left. \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-4/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J} g_{j'}\right)^2\right)}\right)(n_{\cdot 1'}+1)} \right]$$

where $g_1 - \sum_{j'=2}^{J} g_{j'} \in \left\{1/\sqrt{n_{\cdot 1'}\left(\frac{2}{2n_{\cdot 1'}+2}\right)\left(1-\frac{2}{2n_{\cdot 1'}+2}\right)}, \ldots, 1/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)}\right\}$.

Now sum out $g_2, \ldots, g_{a_3}$. Let $diff_8 = g_1 - \sum_{j'=a_3+1}^{J} g_{j'} - \sum_{d=2}^{a_3} \frac{1}{\sqrt{n_{\cdot d'}\left(\frac{v_d}{2n_{\cdot d'}+2}\right)\left(1-\frac{v_d}{2n_{\cdot d'}+2}\right)}}$ where $v_d \in \{2, \ldots, n_{\cdot d'}\}$. Then the joint distribution of $g_1, g_{a_3+1}, \ldots, g_J$ is

$$f(g_1, g_{a_3+1}, \cdots, g_J) = \prod_{j'=a_3+1}^{a_3+a_4} f_2(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{v_{a_3}=2}^{n_{a_3}} \cdots \sum_{v_2=2}^{n_2} \left\{\prod_{d=2}^{a_3} \left[2\binom{2(n_{\cdot d'}-1)}{v_d-2} p_{d'}^{v_d-1}(1-p_{d'})^{2n_{\cdot d'}+1-v_d} + 2\binom{2(n_{\cdot d'}-1)}{2n_{\cdot d'}-v_d} p_{d'}^{2n_{\cdot d'}+1-v_d}(1-p_{d'})^{v_d-1}\right]\right.\right.$$

$$\cdot \left[2 \cdot \left(\frac{2(n_{\cdot 1'}-1)}{\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_8)^2)}\right)(n_{\cdot 1'}+1)-2}\right) \cdot p_{1'}^{\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_8)^2)}\right)(n_{\cdot 1'}+1)-1} \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_8)^2)}\right)(n_{\cdot 1'}+1)}\right.$$

$$\left.\left. + \left(\frac{2(n_{\cdot 1'}-1)}{\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_8)^2)}\right)(n_{\cdot 1'}+1)-2}\right) \cdot p_{1'}^{\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_8)^2)}\right)(n_{\cdot 1'}+1)-1} \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_8)^2)}\right)(n_{\cdot 1'}+1)}\right]\right\}\right]$$

where $diff_8 \in \left\{1/\sqrt{n_{\cdot 1'}\left(\frac{2}{2n_{\cdot 1'}+2}\right)\left(1-\frac{2}{2n_{\cdot 1'}+2}\right)}, \ldots, 1/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)}\right\}$. Now sum over $g_{a_3+1}, \ldots, g_{a_3+a_4}$. Let $diff_9 = g_1 - \sum_{j'=a_3+a_4+1}^{J} g_{j'} - \sum_{d=2}^{a_3} \frac{1}{\sqrt{n_{\cdot d'}\left(\frac{v_d}{2n_{\cdot d'}+2}\right)\left(1-\frac{v_d}{2n_{\cdot d'}+2}\right)}} -$

$\sum_{i=1}^{a_4} \frac{2}{\sqrt{n_{\cdot(a_3+i)'}\left(\frac{u_i}{2n_{\cdot(a_3+i)'}+2}\right)\left(1-\frac{u_i}{2n_{\cdot(a_3+i)'}+2}\right)}}$ where $u_i \in \{1,2\}$ and $v_d \in \{2,\dots,n_{\cdot d'}\}$. The joint distribution of $g_1, g_{a_3+a_4+1}, \dots, g_J$ is

$$f(g_1, g_{a_3+1}, \dots, g_J) = \prod_{j'=a_3+a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{u_{a_4}=1}^{2} \cdots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2(n_{\cdot(a_3+i)'}-1)}{2n_{\cdot(a_3+i)'}-1-u_i} \cdot p_{(a_3+i)'}^{2n_{\cdot(a_3+i)'}+1-u_i} \cdot \left(1-p_{(a_3+i)'}\right)^{u_i-1} \right] \right.$$

$$\sum_{v_{a_3}=2}^{n_{a_3}} \cdots \sum_{v_2=2}^{n_2} \left\{ \prod_{d=2}^{a_3} \left[2\binom{2(n_{\cdot d'}-1)}{v_d-2} p_{d'}^{v_d-1} (1-p_{d'})^{2n_{\cdot d'}+1-v_d} \right.\right.$$

$$\left.\left. + 2\binom{2(n_{\cdot d'}-1)}{2n_{\cdot d'}-v_d} p_{d'}^{2n_{\cdot d'}+1-v_d} (1-p_{d'})^{v_d-1} \right] \right.$$

$$\cdot \left[2 \cdot \left(\frac{2(n_{\cdot 1'}-1)}{\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_9)^2)}\right)(n_{\cdot 1'}+1)-2} \right) \cdot p_{1'}^{\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_9)^2)}\right)(n_{\cdot 1'}+1)-1} \right.$$

$$\left. \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_9)^2)}\right)(n_{\cdot 1'}+1)} \right.$$

$$\left. + \left(\frac{2(n_{\cdot 1'}-1)}{\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_9)^2)}\right)(n_{\cdot 1'}+1)-2} \right) \cdot p_{1'}^{\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_9)^2)}\right)(n_{\cdot 1'}+1)-1} \right.$$

$$\left. \left. \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_9)^2)}\right)(n_{\cdot 1'}+1)} \right] \right\} \right]$$

where $diff_9 \in \left\{ 1/\sqrt{n_{\cdot 1'}\left(\frac{2}{2n_{\cdot 1'}+2}\right)\left(1-\frac{2}{2n_{\cdot 1'}+2}\right)}, \dots, 1/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)} \right\}$. Now sum over $g_{a_3+a_4+1}, \dots, g_J$. These variables are all equal to zero and the $f_1(g_{j'})$'s are all constants.

Let $diff_{10} = g_1 - \sum_{d=2}^{a_3} \frac{1}{\sqrt{n_{\cdot d'}\left(\frac{v_d}{2n_{\cdot d'}+2}\right)\left(1-\frac{v_d}{2n_{\cdot d'}+2}\right)}} - \sum_{i=1}^{a_4} \frac{2}{\sqrt{n_{\cdot (a_3+i)'}\left(\frac{u_i}{2n_{\cdot (a_3+i)'}+2}\right)\left(1-\frac{u_i}{2n_{\cdot (a_3+i)'}+2}\right)}}$

where $u_i \in \{1,2\}$ and $v_d \in \{2,\ldots,n_{\cdot d'}\}$. Thus the distribution of the genetic score, g_1, for this case is

$$f(g_1) = \prod_{j'=a_3+a_4+1}^{J} f_1\left(r_{j'1} = g_{j'}\right)$$

$$\cdot \left[\sum_{u_{a_4}=1}^{2} \cdots \sum_{u_1=1}^{2} \left\{ \prod_{i=1}^{a_4} \left[\binom{2\left(n_{\cdot (a_3+i)'}-1\right)}{2n_{\cdot (a_3+i)'}-1-u_i} \cdot p_{(a_3+i)'}^{2n_{\cdot (a_3+i)'}+1-u_i} \cdot \left(1-p_{(a_3+i)'}\right)^{u_i-1} \right] \right.$$

$$\cdot \sum_{v_{a_3}=2}^{n_{a_3}} \cdots \sum_{v_2=2}^{n_2} \left\{ \prod_{d=2}^{a_3} \left[2\binom{2(n_{\cdot d'}-1)}{v_d-2} p_{d'}^{v_d-1} (1-p_{d'})^{2n_{\cdot d'}+1-v_d} \right.\right.$$

$$\left.\left. + 2\binom{2(n_{\cdot d'}-1)}{2n_{\cdot d'}-v_d} p_{d'}^{2n_{\cdot d'}+1-v_d} (1-p_{d'})^{v_d-1} \right] \right.$$

$$\cdot \left[2 \cdot \left(\frac{2(n_{\cdot 1'}-1)}{\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_{10})^2)}\right)(n_{\cdot 1'}+1)-2} \right) \cdot p_{1'}^{\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_{10})^2)}\right)(n_{\cdot 1'}+1)-1} \right.$$

$$\left. \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1-\sqrt{1-4/(n_{\cdot 1'}(diff_{10})^2)}\right)(n_{\cdot 1'}+1)} \right.$$

$$\left. + \left(\frac{2(n_{\cdot 1'}-1)}{\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_{10})^2)}\right)(n_{\cdot 1'}+1)-2} \right) \cdot p_{1'}^{\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_{10})^2)}\right)(n_{\cdot 1'}+1)-1} \right.$$

$$\left. \left. \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-\left(1+\sqrt{1-4/(n_{\cdot 1'}(diff_{10})^2)}\right)(n_{\cdot 1'}+1)} \right] \right\} \right]$$

where $diff_{10} \in \left\{ 1/\sqrt{n_{\cdot 1'}\left(\frac{2}{2n_{\cdot 1'}+2}\right)\left(1-\frac{2}{2n_{\cdot 1'}+2}\right)}, \ldots, 1/\sqrt{n_{\cdot 1'}\left(\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)\left(1-\frac{n_{\cdot 1'}}{2n_{\cdot 1'}+2}\right)} \right\}$.

Next consider case 4 where $a_1 = a_2 = a_3 = 0$ and $a_4 > 0$. Then the joint distribution of $g_1, g_2, \dots, g_J$ is

$$f(g_1, g_2, \dots, g_J)$$

$$= \prod_{j'=2}^{a_4} f_2(r_{j'1} = g_{j'}) \cdot \prod_{j'=a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left(\frac{2(n_{\cdot 1'} - 1)}{(n_{\cdot 1'} + 1)\left(1 + \sqrt{1 - 16/\left(n_{\cdot 1'}\left(g_1 - \sum_{j'=2}^{J} g_{j'}\right)^2\right)}\right)} - 3 \right)$$

$$\cdot p_{1'}^{(n_{\cdot 1'}+1)\left(1+\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J} g_{j'}\right)^2\right)}\right)-1}$$

$$\cdot (1 - p_{1'})^{2n_{\cdot 1'}+1-(n_{\cdot 1'}+1)\left(1+\sqrt{1-16/\left(n_{\cdot 1'}\left(g_1-\sum_{j'=2}^{J} g_{j'}\right)^2\right)}\right)}$$

where $g_1 - \sum_{j'=2}^{J} g_{j'} \in \left\{ 2/\sqrt{n_{\cdot 1'}\left(\frac{1}{2n_{\cdot 1'}+2}\right)\left(1 - \frac{1}{2n_{\cdot 1'}+2}\right)}, 2/\sqrt{n_{\cdot 1'}\left(\frac{2}{2n_{\cdot 1'}+2}\right)\left(1 - \frac{2}{2n_{\cdot 1'}+2}\right)} \right\}$.

Now sum out $g_2, \dots, g_{a_4}$. Let $diff_{11} = g_1 - \sum_{j'=a_4+1}^{J} g_{j'} - \sum_{i=2}^{a_4} \frac{2}{\sqrt{n_{\cdot i'}\left(\frac{u_i}{2n_{\cdot i'}+2}\right)\left(1-\frac{u_i}{2n_{\cdot i'}+2}\right)}}$ where $u_i \in \{1, 2\}$. Then the joint distribution of $g_1, g_{a_4+1}, \dots, g_J$ is

$$f(g_1, g_{a_4+1}, \dots, g_J)$$

$$= \prod_{j'=a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{u_{a_4}=1}^{2} \cdots \sum_{u_2=1}^{2} \left\{ \prod_{i=2}^{a_4} \left[\binom{2(n_{\cdot i'} - 1)}{2n_{\cdot i'} - u_i - 1} p_{i'}^{2n_{\cdot i'}+1-u_i} (1 - p_{i'})^{u_i - 1} \right] \right.\right.$$

$$\cdot \left(\frac{2(n_{\cdot 1'} - 1)}{(n_{\cdot 1'} + 1)\left(1 + \sqrt{1 - 16/(n_{\cdot 1'}(diff_{11})^2)}\right) - 3} \right)$$

$$\cdot p_{1'}^{(n_{\cdot 1'}+1)\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_{11})^2)}\right)-1}$$

$$\left.\left. \cdot (1 - p_{1'})^{2n_{\cdot 1'}+1-(n_{\cdot 1'}+1)\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_{11})^2)}\right)} \right\} \right]$$

where $diff_{11} \in \left\{ 2/\sqrt{n_{\cdot 1'}\left(\frac{1}{2n_{\cdot 1'}+2}\right)\left(1 - \frac{1}{2n_{\cdot 1'}+2}\right)}, 2/\sqrt{n_{\cdot 1'}\left(\frac{2}{2n_{\cdot 1'}+2}\right)\left(1 - \frac{2}{2n_{\cdot 1'}+2}\right)} \right\}$. Now sum over $g_{a_4+1}, \dots, g_J$. Let $diff_{12} = g_1 - \sum_{i=2}^{a_4} \frac{2}{\sqrt{n_{\cdot i'}\left(\frac{u_i}{2n_{\cdot i'}+2}\right)\left(1-\frac{u_i}{2n_{\cdot i'}+2}\right)}}$ where $u_i \in \{1, 2\}$. Then the distribution of the genetic score, g_1, for this case is

$$f(g_1) = \prod_{j'=a_4+1}^{J} f_1(r_{j'1} = g_{j'})$$

$$\cdot \left[\sum_{u_{a_4}=1}^{2} \cdots \sum_{u_2=1}^{2} \left\{ \prod_{i=2}^{a_4} \left[\binom{2(n_{\cdot i'}-1)}{2n_{\cdot i'}-u_i-1} p_{i'}^{2n_{\cdot i'}+1-u_i} (1-p_{i'})^{u_i-1} \right] \right.\right.$$

$$\cdot \left(\frac{2(n_{\cdot 1'}-1)}{(n_{\cdot 1'}+1)\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_{12})^2)}\right)-3} \right)$$

$$\cdot p_{1'}^{(n_{\cdot 1'}+1)\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_{12})^2)}\right)-1}$$

$$\left.\left. \cdot (1-p_{1'})^{2n_{\cdot 1'}+1-(n_{\cdot 1'}+1)\left(1+\sqrt{1-16/(n_{\cdot 1'}(diff_{12})^2)}\right)} \right\} \right]$$

Where $diff_{12} \in \left\{ 2/\sqrt{n_{\cdot 1'}\left(\frac{1}{2n_{\cdot 1'}+2}\right)\left(1-\frac{1}{2n_{\cdot 1'}+2}\right)}, 2/\sqrt{n_{\cdot 1'}\left(\frac{2}{2n_{\cdot 1'}+2}\right)\left(1-\frac{2}{2n_{\cdot 1'}+2}\right)} \right\}$.

Now consider case 5 where $r_{11} = \cdots = r_{J1} = 0$, then $g_1 = 0$ and the joint distribution of $g_1, g_2, \ldots, g_J$ is

$$f(g_1, g_2, \ldots, g_J)$$

$$= \prod_{j'=1}^{J} \left\{ \sum_{t=3}^{n_{\cdot j'}} \left[\binom{2(n_{\cdot j'}-1)}{t-1} \cdot p_{j'}^{t-1} \cdot (1-p_{j'})^{2n_{\cdot j'}+1-t} + \binom{2(n_{\cdot j'}-1)}{2n_{\cdot j'}+1-t} \right.\right.$$

$$\left. \cdot p_{j'}^{2n_{\cdot j'}+1-t} \cdot (1-p_{j'})^{t-1} \right] + \sum_{t=1}^{2} \left[\binom{2(n_{\cdot j'}-1)}{t-1} \cdot p_{j'}^{t-1} \cdot (1-p_{j'})^{2n_{\cdot j'}+1-t} \right]$$

$$\left. + \binom{2(n_{\cdot j'}-1)}{n_{\cdot j'}} \cdot p_{j'}^{n_{\cdot j'}} \cdot (1-p_{j'})^{n_{\cdot j'}} \right\}.$$

Notice that the above distribution does not depend on the individual $g_{j'}$'s and for each $g_{j'}$ there is only one possible value of zero. Hence when the ancillary variables are summed out the marginal distribution of the genetic score when $g_1 = 0$ is

$$f(g_1) = \prod_{j'=1}^{J} \left\{ \sum_{t=3}^{n_{\cdot j'}} \left[\binom{2(n_{\cdot j'}-1)}{t-1} \cdot p_{j'}^{t-1} \cdot (1-p_{j'})^{2n_{\cdot j'}+1-t} + \binom{2(n_{\cdot j'}-1)}{2n_{\cdot j'}+1-t} \cdot p_{j'}^{2n_{\cdot j'}+1-t} \right. \right.$$

$$\left. \cdot (1-p_{j'})^{t-1} \right] + \sum_{t=1}^{2} \left[\binom{2(n_{\cdot j'}-1)}{t-1} \cdot p_{j'}^{t-1} \cdot (1-p_{j'})^{2n_{\cdot j'}+1-t} \right]$$

$$\left. + \binom{2(n_{\cdot j'}-1)}{n_{\cdot j'}} \cdot p_{j'}^{n_{\cdot j'}} \cdot (1-p_{j'})^{n_{\cdot j'}} \right\}.$$

As demonstrated in the previous derivations the distribution of the genetic score can be tedious to calculate. It requires the values of $r_{11}, r_{21}, \ldots, r_{J1}$ in order to know which case the resulting distribution the function lies in. Additionally there exist nuisance parameters, $p_{j'}$, $j' = 1, .., J$. Since these are unknown the exact probabilities in the distribution cannot be calculated. Additionally the functions are sensitive to the $p_{j'}$'s. For example if a $p_{j'}$ is increased then the probability the genetic score equals zero is decreased. For these reasons the distribution of the genetic scores is not used to calculate a test statistic rather a Krusal-Wallis test is used since the distribution of the genetic scores is skewed and possibly contains outliers.

The multinomial weighted sum statistic has the advantage that it simultaneously considers all variants at a locus. This is not the only approach to take. A marker by marker approach such as the one detailed in the next section can pinpoint associations at variants.

2.2 Single Marker Analysis for Multinomial Data

A Single Marker Analysis (SMA) is also proposed to perform association analysis. The hypotheses tested are:

H_0: The phenotype and marker are statistically independent.

H_a: There is an association between the phenotype and marker.

In this method each marker is tested individually then a multiple testing correction is used to determine if the result is significant enough to warrant rejecting the null hypothesis.

First a phenotype by genotype contingency table is constructed. Observations missing the phenotype or genotype are excluded from the test. As in the typical contingency analysis, for each cell in the table the estimated expected cell count is calculated as the row total times the column total divided by the number of observations in the table. If the expected cell count for any cell in the table is less than five then an exact test is used. This routine returns the exact p-value for an observed contingency table using the hypergeometric distribution with fixed row and column totals. If all of the cell counts were five or greater then a Chi-square test of independence is run using Pearson's statistic.

2.2.1 Multiple Testing Correction

The above described SMA only provides results for a single test at a single marker. In practice these tests are used multiple times at different markers to search for an association across the locus. Therefore it is important to use a multiple testing correction when making conclusions. The False Discovery Rate (FDR) controlling procedure proposed by Benjamini and Hochberg (1995) is used to adjust for the large number of tests being simultaneously run. This method differs from a Family-Wise Error Rate (FWER) control in that it controls the number of false positives rather than the probability of making a single Type I error. This method has been shown to have higher power than a traditional FWER control (Benjamini & Hochberg, 1995). In general this method assumes the test statistics are independent. There are special cases of dependency of the test statistics where the results still hold (Benjamini & Yekutieli, 2001). One of these cases is for positively correlated tests. When linkage disequilibrium exists the tests are positively correlated (Verhoeven, Simonsen, & McIntyre, 2005). An adjustment can be made to the procedure if dependency outside of the special cases exists (Benjamini & Yekutieli, 2001).

Define Q as the proportion of false rejections among all rejections, then the Benjamini and Hochberg FDR is the $E(Q)$. The Benjamini and Hochberg procedure considers testing *D* null hypotheses, $H_{01}, \ldots, H_{0D}$, using observed significance levels, $P_1, \ldots, P_D$. The first step in the procedure is to order the observed significance levels so that $P_{(1)} \leq \cdots \leq P_{(D)}$. Denote the null hypothesis corresponding to the ordered observed significance level $P_{(l)}$ as $H_{0(l)}$. Find

$$d = max\left\{l: P_{(l)} \leq \frac{l}{D}\alpha\right\}$$

and reject $H_{0(1)}, \ldots, H_{0(d)}$. If the above *d* does not exist then none of the null hypotheses are rejected. Let D_0 be the number of true null hypotheses. For this procedure it is the case that (Benjamini & Hochberg, 1995)

$$E(Q) \leq \frac{D_0}{D}\alpha \leq \alpha.$$

Hence the procedure controls the FDR at the $\alpha \cdot D_0/D$ level.

2.3. Multinomial Logistic Regression

Multinomial logistic regression is also proposed to test for an association between a nominal phenotype with more than two categories and rare variants. This method is the multivariate generalized linear model approach. Multinomial logistic regression is called many different names. It can also be called polychotomous, polytomous, or baseline-category logistic regression. The model begins by choosing a baseline category. Continuing the notation above let the categories in the phenotype be indexed with $k = 1, \ldots, K$. Any category can be chosen as the baseline. For clarity let the baseline category by the K^{th} category. The multinomial logistic model simultaneously compares category *K* with the other $K - 1$ categories. Let *Y* be the phenotype. Dummy variables were created for the genotypes at the variants. Normally two dummy variables would be needed for each marker since there are three genotypic categories.

However since these markers are rare variants it is possible only one dummy variable is necessary for a given marker. Let $j^* = 1, \ldots, J^*$ index the dummy variables of genotypes. Let X_{iJ^*} be the $j^{*\text{th}}$ dummy variable for the i^{th} individual. Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ^*})'$ be the column vector for the i^{th} individual. Extending the notation from a logistic model, let for category *k*

$$\pi_k(\boldsymbol{X}_i) = P(Y = k|\boldsymbol{X}_i)$$

with $\sum_k \pi_k(\boldsymbol{X}_i) = 1$. Under this set up for an arbitrary $\boldsymbol{X}$, the counts in the *K* categories have a multinomial distribution with probabilities $\pi_1(\boldsymbol{X}), \ldots, \pi_{K-1}(\boldsymbol{X})$ such that $\pi_K(\boldsymbol{X}) = 1 - \sum_k \pi_k(\boldsymbol{X})$. The baseline category is paired with each other category in a logit model. The model is then *K*– 1 simultaneous models

$$ln\frac{\pi_k(\boldsymbol{X})}{\pi_K(\boldsymbol{X})} = a_k + \boldsymbol{\beta}_k'\boldsymbol{X}$$

for *k* = 1, …, *K* – 1 with a_k an intercept term and $\boldsymbol{\beta}_k$ a column vector of coefficients for the k^{th} model (Agresti, 2002). Rewriting the above equations, the probability of category *k* is

$$\pi_k(\boldsymbol{X}) = \frac{exp(a_k + \boldsymbol{\beta}_k'\boldsymbol{X})}{1 + \sum_{h=1}^{K-1} exp(a_h + \boldsymbol{\beta}_h'\boldsymbol{X})}$$

where $\alpha_K = 0$ and $\boldsymbol{\beta}_K = 0$. For the phenotype of the i^{th} individual let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK})$ where y_{ik} is one if the phenotype is in category *k* and zero otherwise. Notice for each individual $\sum_k y_{ik} = 1$. Formally define the parameters for the k^{th} logit as $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kJ^*})'$. Also note that $\pi_K(\boldsymbol{X}_i) = 1 - \sum_{k=1}^{K-1} \pi_k(\boldsymbol{X}_i)$, $\pi_K(\boldsymbol{X}_i) = 1/\left[1 + \sum_{k=1}^{K-1} exp(a_k + \boldsymbol{\beta}_k'\boldsymbol{X})\right]$ and $y_{iK} = 1 - \sum_{k=1}^{K-1} y_{ik}$. To find the log likelihood first consider the contribution of individual *i*

$$\ln\left[\prod_{k=1}^{K} \pi_k(\boldsymbol{X}_i)^{y_{ik}}\right] = \sum_{k=1}^{K-1} y_{ik} \ln \pi_k(\boldsymbol{X}_i) + \left(1 - \sum_{k=1}^{K-1} y_{ik}\right) \ln \pi_K(\boldsymbol{X}_i)$$

$$= \sum_{k=1}^{K-1} y_{ik} \ln \pi_k(\boldsymbol{X}_i) - \sum_{k=1}^{K-1} y_{ik} \ln \pi_K(\boldsymbol{X}_i) + \ln \pi_K(\boldsymbol{X}_i)$$

$$= \sum_{k=1}^{K-1} y_{ik} [\ln \pi_k(\boldsymbol{X}_i) - \ln \pi_K(\boldsymbol{X}_i)] + \ln \frac{1}{1 + \sum_{k=1}^{K-1} \exp(a_k + \boldsymbol{\beta}_k' \boldsymbol{X})}$$

$$= \sum_{k=1}^{K-1} y_{ik} \ln \frac{\pi_k(\boldsymbol{X}_i)}{\pi_K(\boldsymbol{X}_i)} - \ln \left[1 + \sum_{k=1}^{K-1} \exp(a_k + \boldsymbol{\beta}_k' \boldsymbol{X}) \right]$$

$$= \sum_{k=1}^{K-1} y_{ik} \exp(a_k + \boldsymbol{\beta}_k' \boldsymbol{X}) - \ln \left[1 + \sum_{k=1}^{K-1} \exp(a_k + \boldsymbol{\beta}_k' \boldsymbol{X}) \right]$$

Note that the constant term $n!/[y_{i1}! \ldots y_{iK}!]$ in the multinomial distribution was excluded above since it does not contribute to the information about the parameters. Now consider the full log likelihood based on *n* independent observations.

$$\ln \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \pi_k(\boldsymbol{X}_i)^{y_{ik}} \right] = \sum_{i=1}^{n} \ln \left[\prod_{k=1}^{K} \pi_k(\boldsymbol{X}_i)^{y_{ik}} \right]$$

$$= \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K-1} y_{ik} \exp(a_k + \boldsymbol{\beta}_k' \boldsymbol{X}) - \ln \left[1 + \sum_{k=1}^{K-1} \exp(a_k + \boldsymbol{\beta}_k' \boldsymbol{X}) \right] \right\}$$

The model is fit and maximum likelihood estimates are found by maximizing the log likelihood function using the Newton-Raphson or similar algorithm (Croissant). The null hypothesis of no association is $H_o: \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K = 0$. A likelihood ratio test is used to test this hypothesis. The log likelihood of the full model, $\ln L$, is compared to the log likelihood under the null hypothesis, $\ln L_0$. The test statistic is (Hosmer & Lemeshow, 2000, p. 270)

$$-2(\ln L_0 - \ln L) \dot{\sim} \chi^2_{J^*(K-1)}$$

and $H_0$is rejected for large values of the test statistic.

2.4 Multinomial Logistic Regression with Collapsing

The genotypes of the rare variants are converted to dummy variables for use in the multinomial logistic regression. Since the rare allele made some genotypes occur only a few times in the data set it is possible that all individuals in a phenotypic category had the same value of the dummy variable. This phenomenon is called quasi-complete separation. Due to the nature of the data, quasi-complete separation occurs frequently. Quasi-complete separation is unavoidable in the data sets since dummy variables for the rare variants are necessary. When quasi-complete separation occurs between an independent variable and the dependent variable in multinomial logistic regression, the maximum likelihood estimate for the coefficient to the independent variable does not exist (Albert & Anderson, 1984). Since the MLE does not exist a maximum point of the log likelihood function does not exist. Rather the log likelihood is maximized in the limit. This leads to problems when the maximization routine tries to find a maximum that does not exist. Combining variables together is recommended by Allison (2008) as a possible way to allow the routine to find a solution. Using this suggestion any dummy variables with quasi-complete separation are collapsed together to form one dummy variable. This dummy variable takes the value of one if at least one variant from the collapsing group contains the rare allele and is zero otherwise. Multinomial logistic regression is then used to test for an association as described in the previous section.

CHAPTER III

SIMULATIONS

To assess the performance of the previously proposed tests, simulations were run under different scenarios that might affect the type I error and power. Section 3.1 gives the factors considered. Section 3.2 details the steps in the simulation process.

3.1 Scenarios Considered

The factors considered in the simulations were sample size, number of phenotypic categories, and heritability under the alternative hypothesis. Initially sample sizes of 500, 1000, and 2000 were utilized in the simulations. Three, five, and seven phenotypic categories were considered. For simulations under the alternative hypothesis the heritability in the broad sense was varied with lambda set as 0.01, 0.05, 0.1, 0.2, 0.3, 0.5, or 0.8.

In order to make recommendations on sample sizes additional simulations were run for the heritability in the broad sense of 5%, 10% and 20%. These heritability levels were chosen based on the fact that most genome wide association studies using common variants can account for 5% to 10% of the heritability (Asimit & Zeggini, 2010). Sample sizes were chosen so that the proposed weighted sum test and SMA could achieve at least 80% power with three, five, and seven phenotypic categories.

3.2 Steps in the Simulations

Generation of genetic markers and phenotypes was accomplished by modifying a procedure described by Morris and Zeggini (2010). The goal in generating the data was to simulate sampling diploid individuals with rare variants and a multinomial phenotype from a population of 20,000. One thousand iterations were run for each combination of the parameters given in the previous section. The following steps describe how a single iteration of the algorithm was run. The steps include the details used for simulating under the null or alternative hypothesis.

1. A population of 40,000 haplotypes in a 50,000 base pair region is created using the ms program (UNIX platform) by Hudson (2002). Recombination is assumed and a crossover rate of 1 cM per million base pairs is used as in Morris and Zeggini. Additionally a per base mutation rate of 10^{-8} and an effective population size of 10,000 is also taken from Morris and Zeggini. Justification for these choices of parameters is not provided by Morris and Zeggini. It is assumed these parameters are reasonable for simulating a population of haplotypes. Two parameters, θ and ρ, are required in the call to the ms program. The first is calculated as $\theta = 4N_0\mu$ where $N_0$is the diploid effective population size and μ is the locus neutral mutation rate. The second parameter is the cross over rate parameter, ρ. It is calculated as $\rho = 4N_0r$ where r is the probability of crossover between ends of the locus. The correctly scaled parameters as used by Morris and Zeggini aee calculated as:

$$\theta = 4(10{,}000)(50{,}000)(10^{-8}) = 20$$

$$\rho = 4(10{,}000)(50{,}000 - 1)\left(\frac{1/100}{1{,}000{,}000}\right) = 19.9996 \approx 20$$

The call to the ms program is "./ms nsam nrep –t θ –r ρ nsites > output/temp1" where nsam = 40000, nrep = 1, $\theta = 20$, $\rho = 20$, and nsites = 50000. The "> ouput/temp1" section of the call saves the ms data to an external file for later.

2. The data set produced by ms is read into R (R Development Core Team, 2011). The minor allele frequency (MAF) is calculated for each marker. Only rare variants, markers with a MAF of 5% or less, are kept for analysis. It is important to note that Morris and Zeggini (2010) as well as Li and Leal (2008) define a rare variant as a marker with a MAF of 1% or less. For simulations under the alternative hypothesis of association, markers are randomly selected to be causal so that the total MAF of the markers is approximately 10%.
3. At this point it is necessary to capture a population parameter for simulations under the alternative before the sample is collected. The use of this parameter is explained in a later step. It is desired to calculate the probability that all causal alleles are the wild type in the individual. Let b_i be the number of rare causal alleles across the whole locus for diploid individual i. Hence it is desired to calculate $P(b_i = 0)$ in the population. Since the data set contains haplotypes not individuals the probability can not be directly calculated. Originally, pairing of all 40,000 haplotypes was considered to create individuals so this quantity could be calculated. However the procedure proved to be too computationally intensive to run in a reasonable amount of time. A theoretical equivalent is produced using the haplotypes. Let B_{1i} be the number of causal rare alleles in haplotype 1 and B_{2i} be the number of causal rare alleles for haplotype 2 paired with haplotype 1. Then $b_i = B_{1i} + B_{2i}$ since the sum in the individual can be broken into the haplotypes. Due to the fact that haplotypes are randomly paired and assumed independent it can be shown:

   $P(b_i = 0) = P(B_{1i} = 0 \cap B_{2i} = 0)$

$$= P(B_{1i} = 0) \cdot P(B_{2i} = 0)$$

$$= [P(B_{1i} = 0)]^2$$

Hence the only differences between $P(b_i = 0)$ and $[P(B_{1i} = 0)]^2$ are due to the exact random pairing performed. The final equation is simply the square of the probability that all causal alleles are wild type on one haplotype. The quantity $[P(B_{1i} = 0)]^2$ is calculated from the generated population and saved for later in the simulation process when the parameter $P(b_i = 0)$ is needed. Under the null hypothesis this quantity is not calculated since there are no causal alleles.

4. A sample of 2N haplotypes is randomly selected. These haplotypes are randomly paired together to create diploid organisms. Since a sample is taken from the population it is possible that a rare variant had a MAF of greater than 5%. It is also possible that for some markers no rare alleles made it into the sample.
5. For generating phenotypic data under the alternative hypothesis of association, b_i, the number of rare causal alleles across the whole locus for each individual is calculated. The quantity λ is defined as the heritability in the broad sense. Morris and Zeggini (2010) simulate the phenotype from a $N(I(b_i > 0), \sigma^2)$ but do not give a formula for the variance or standard deviation. They simply state that "the standard deviation, σ, is determined by the spectrum of causal variants and their joint contribution, λ, to the phenotypic variance" (Morris & Zeggini, 2010). To derive the formula for the variance it may be helpful to recall from Bain and Englehardt (Introduction to Probability and Mathematical Statistics, 1992) Theorem 5.4.3:

$$Var(Y) = E_X[Var(Y|X)] + Var_X[E(Y|X)].$$

The conditional distribution $y_i|b_i \sim N(I(b_i > 0), \sigma^2)$ is the normal variable that Morris and Zeggini generated in their simulations. Note that for the mean in the normal variable it is the case that $I(b_i > 0) \sim Bernoulli\big(P(b_i > 0)\big)$. It is desired to avoid making any

assumptions about the individual markers. Linkage disequilibrium is possible under the following results. Given the above conditional distribution it is necessary to know the variance of the resulting, y_i's. The variance is:

$$Var(y_i) = E_{b_i}[Var(y_i|b_i)] + Var_{b_i}[E(y_i|b_i)]$$

$$= E_{b_i}[\sigma^2] + Var_{b_i}[I(b_i > 0)]$$

$$= \sigma^2 + P(b_i > 0)\big(1 - P(b_i > 0)\big)$$

$$= \sigma^2 + [1 - P(b_i = 0)][P(b_i = 0)]$$

The first term in the last equality is the variance of the environmental effects, $Var(E)$, in the usual equation $Var(P) = Var(G) + Var(E)$. The second term is the variance of the genetic effects. The heritability in the broad sense is written as:

$$\lambda = \frac{Var(G)}{Var(G) + Var(E)}.$$

Solving for the $Var(E)$ obtains

$$Var(E) = \frac{1 - \lambda}{\lambda} Var(G).$$

Thus the variance in the conditional distribution needed to be

$$\sigma^2 = \frac{1 - \lambda}{\lambda} [1 - P(b_i = 0)][P(b_i = 0)]$$

in order to have the correct heritability in y_i. Recall it was not computationally efficient to calculate $P(b_i = 0)$ in the population. Thus the theoretical equivalent $[P(B_{1i} = 0)]^2$ saved in step 3 is used in its place. Under the null hypothesis of no association a random variable from a standard normal distribution is generated in place of y_i. To obtain the categorical phenotypes, the empirical percentiles of the y_i's are used to divide the data into the desired number of categories. For example when five categories are needed the quintiles are used to divide the y_i's into five categories. The generated data set thus consists of a categorical phenotype and all rare variants. For simulations under the null

hypothesis these rare variants are all non-causal. For simulations under the alternative hypothesis these rare variants are a mixture causal and non-causal variants.

6. The proposed methods from Chapter 2 are applied to the generated data set.
    a. The proposed weighted sum statistic test deriving a p-value from the appropriate chi-square distribution is the test of interest. The p-value is compared to a 0.05 level to determine a decision. In addition to this version of the test, a permutation test is run to find an empirical p-value for the weighted sum test statistic. This is done to show that the permutation test is not necessary when the observations are independent.
    b. The previously described single marker analysis (SMA) is also performed. If the expected cell count for any cell in the contingency table is less than five then an exact test through the fisher.test() function in R was used. This function uses the FEXACT routine created by Mehta and Patel (1986) and modified by Clarkson, Fan, and Joe (1993) to run the Fisher's exact test when the table is larger than 2×2. If all of the expected cell counts are greater than five then the usual Chi-square test is run in R. A false discovery rate (FDR) of $\alpha = 0.05$ is used in the algorithm. The procedure results in a decision for each test. If any decision is to reject then the whole SMA is counted as significant. Thus a significant result does not mean that all variants were rejected. Rather it means that at least one test on the locus produced a decision to reject.
    c. Two versions of multinomial logistic regression (with and without collapsing) are also considered to test for association as described in Chapter 2. Linear dependencies and dummy variables with only one value are eliminated by using a Gaussian elimination function provided by John Fox (2007) on the R help forum. This is necessary because the multinomial logistic regression function in R does not tolerate linear dependencies or redundant variables. To run the multinomial

logistic regression the R package mlogit (Croissant) is utilized. The package can only handle eighty-five variables at a time so if necessary the independent variables are split into two groups and two models are run. In this case eighty-five variables are included in the first model and the remaining in the second model. If splitting the variables is necessary then a Bonferonni correction to a 0.05 level is used to adjust the type I error for the two tests. In the case of two models if either of the decisions are to reject then the result is counted as significant. If only one test is needed then a 0.05 level is used for making a decision.

7. A tally of the number of significant tests is kept. If a test is significant then one is added to the counting variable for that test. During initial testing it became clear that there is a huge problem with using the mlogit package for the multinomial logistic regression. The model fitting failed numerous times when the hessian became computationally singular. For both versions of the multinomial logistic regression routines the number of times the routine succeeded are counted.

As noted above these steps were run 1000 times for each combination of the parameters. The proportion of significant tests was used to estimate the type I error under the null hypothesis and power under the alternative. Estimates for the multinomial logistic regression methods used the number of times the routines succeeded not the number of tests attempted to calculate these proportions. The results of the simulations are presented in the following chapter.

CHAPTER IV

RESULTS

The results of the simulations described in chapter III are detailed here. Section 4.1 discusses of the number of rare variants generated in the simulations as well as the number of causal variants simulated. Section 4.2 presents the Type I error estimates of the methods considered. Power estimates are provided and discussed for the viable methods in section 4.3. Section 4.4 gives sample size recommendations for the recommended methods.

4.1 Number of Markers and Causal Markers

In order for this work to be comparable to future work, the number of variants considered in each test is needed. Many previous researchers fixed the number of variants generated in each data set and the number of causal variants (Basu & Pan, 2011; Li & Leal, 2008; Madsen & Browning, 2009). The data generation procedure used in these simulations allows the number of variants and causal variants to fluctuate. Table 4.1 provides the five number summaries plus the standard deviation of the number of rare variants and causal variants generated in the simulations and testing. The number of rare variants includes 101,900 iterations of the data generation procedure (due to a failure in the file system on Pistol Pete at Oklahoma State University 100 iterations of the data was lost). The number of rare variants was generated under both the null

Number of Rare Variants						
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
218	310	328	329.3	348	476	29.1

Number of Causal Rare Variants						
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	Standard Deviation
3	13	17	18.4	23	69	7.7

Table 4.1: Statistics for the Number of Rare Variants and Causal Variants

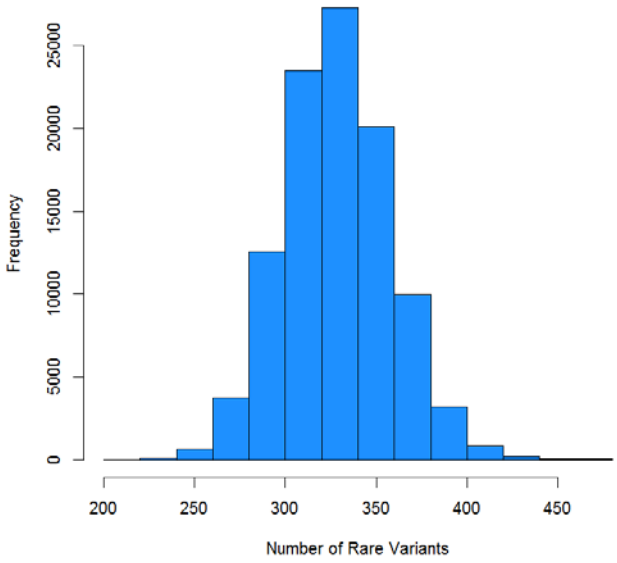

Figure 4.1: Histogram of the Number of Rare Variants

and alternative hypothesis. The number of causal variants came exclusively from simulations under the alternative hypothesis. The histogram of the number of rare variants generated in each simulation is provided in Figure 4.1. Both Table 4.1 and Figure 4.1 show the empirical distribution is bell shaped. The mean is approximately 329.3 rare variants generated. This is many more than the 100 variants that Madsen and Browning used in their simulations (2009). It is also many more than the 5 to 20 variants Li and Leal (2008) simulated and the 8 to 72 variants Basu and Pan (2011) used. Figure 4.2 gives the histogram of the number of causal variants generated under the alternative hypothesis. The number of causal rare variants is slightly right

skewed. Recall from the description of the data generation process that the total MAF of causal variants is set to 10%. Hence a large number of causal variants means that many of them are extremely rare.

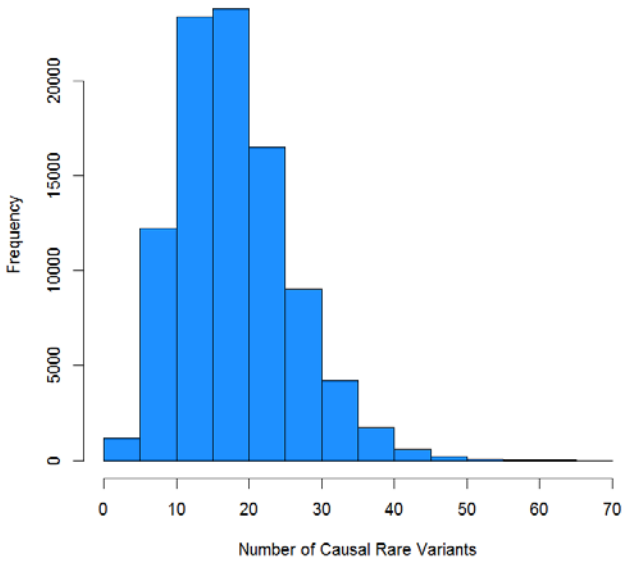

Figure 4.2: Histogram of the Number of Causal Variants

4.2 Comparison of Type I Error

The Type I error estimates for all of the methods applied to the simulated data are presented in Table 4.2. Each estimate is based on 1000 iterations. The data for these simulations are generated under the null hypothesis of no association. Table 4.2 gives the proportion of false rejections for each method considered for all tests at a 0.05 level. The multinomial logistic regression methods encountered failures in the routine for some iterations. The symbol † is used to mark scenarios where failures occurred. The estimates given in these cells are based on the successful iterations not the number of iterations tried. Additionally estimates that are significantly different from 0.05 at a 0.05 level are marked with the symbol *. Estimates different

from 0.05 are determined using the appropriate rejection region for a z-test of $H_0: p = 0.05$ versus $H_a: p \neq 0.05$.

Sample Size	Method	Number of Phenotypic Categories 3	5	7
500	MNWSS	0.045	0.055	0.040
	NMWSSP	0.045	0.058	0.043
	SMA	0.01*	0.012*	0.009*
	MLOGIT	0.3567†*	0.0951†*	0.0089†*
	MLOGITC	0.019*	0.019*	0.0130†*
1000	MNWSS	0.042	0.040	0.055
	NMWSSP	0.042	0.042	0.056
	SMA	0.013*	0.013*	0.018*
	MLOGIT	0.232*	0.0911†*	0.0067†*
	MLOGITC	0.018*	0.0285†*	0.0311†*
2000	MNWSS	0.055	0.050	0.060
	NMWSSP	0.053	0.052	0.053
	SMA	0.018*	0.013*	0.012*
	MLOGIT	0.2142†*	0.0612†*	0.0115†*
	MLOGITC	0.006*	0.019*	0.0262†*

†A portion of these tests failed and the results are most likely biased
*The Type I Error Rate is significantly different from 0.05 at a 0.05 level.
Estimates based on 1000 iterations.

Table 4.2: Type I Error Estimates for All Methods at a 0.05 Level

First note that the multinomial weighted sum statistic (MNWSS) gives a Type I error rate at the desired level of 0.05. Also note that the results are very similar to the results for the multinomial weighted sum statistic with a permutation test (MNWSSP). This shows that there is not a bias in using the distribution based test over the permutation test for these sample sizes. Since the distribution based test is computationally simpler, it is recommended over the permutation test. Madsen and Browning used a permutation test in their method and did not consider using a distributional quantity. Given the results of this study it may be possible to simplify their procedure by using a distribution based test.

The single marker analysis (SMA) for a multinomial phenotype is very conservative for all sample sizes and number of categories included in the simulation study. This is expected since an adjustment was made for multiple tests. Basu and Pan found that their SMA on case control data was conservative (2011).

The multinomial logistic regression (MLOGIT) had an inflated Type I Error rate when the phenotype had three and five categories. This is consistent with results for logistic regression on case control status (Li & Leal, 2008). For this reason Li and Leal excluded logistic regression from power simulations. Since multinomial logistic regression has an extremely inflated type I error rate it is not a viable method for determining an association between genetic markers and a multinomial phenotype. The method appears to be conservative when the phenotype has seven categories. This might be due to the large number of failures in the routine for seven phenotypic categories versus three or five phenotypic categories. For a sample size of 2000 the multinomial logistic regression routine failed in 0.1% of the simulations for 3 phenotypic categories, 0.4% for five phenotypic categories, and 30.6% for seven phenotypic categories. Failures in the multinomial logistic regression routine also make it unreliable as a method of association. For these reasons multinomial logistic regression is excluded from consideration in the power simulations.

		Number of Phenotypic Categories		
Sample Size	Method	3	5	7
500	MLOGIT	0.002	0.001	0.217
	MLOGITC	0	0	0.001
1000	MLOGIT	0	0.001	0.258
	MLOGITC	0	0.002	0.002
2000	MLOGIT	0.001	0.004	0.306
	MLOGITC	0	0	0.007
Estimates Based on 1000 iterations.				

Table 4.3: Proportion of Failures in the Multinomial Logistic Regression Routine

The multinomial logistic regression with collapsing (MLOGITC) was considered as an alternative to the multinomial logistic regression. Recall that variants with quasi-complete separation were collapsed together into a single variable in this method. The Type I Error estimates for MLOGITC are conservative instead of inflated as in the multinomial logistic regression. Unfortunately the method still experiences failures in the multinomial logistic regression routine. These failures make the MLOGITC unreliable. Hence it is not considered in the subsequent power simulations.

4.3 Comparison of Power

This section examines the power estimates of the multinomial weighted sum statistic (MNWSS), the multinomial weighted sum statistic with a permutation test (MNWSSP), and a single marker analysis for multinomial phenotypes (SMA). The multinomial logistic regression methods are excluded for the previously mentioned reasons. Results will be grouped together by sample size since this factor influenced the power the most.

Figure 4.3 illustrates the power estimates for a sample size of 500 with three, five, and seven phenotypic categories as heritability is increased. The multinomial weighted sum statistic and multinomial weighted sum statistic with a permutation test have very similar results for most of the simulations. Hence the two lines are overlaid in these figures and many subsequent ones. This further illustrates the fact that a permutation test is not necessary. In each of these figures the multinomial weighted sum statistic starts out with a higher power than the SMA. Recall that the multinomial weighted sum statistic had the correct Type I Error rate while the SMA was conservative. The single marker analysis overtakes the multinomial weighted sum statistic between a heritability of 1% and 10%. The single marker analysis quickly reaches a power of 1 while the multinomial weighed sum statistic increases but does not reach a power of 1. The

similarity of these three figures shows that the methods are only slightly influenced by the number of phenotypic categories for a sample size of 500.

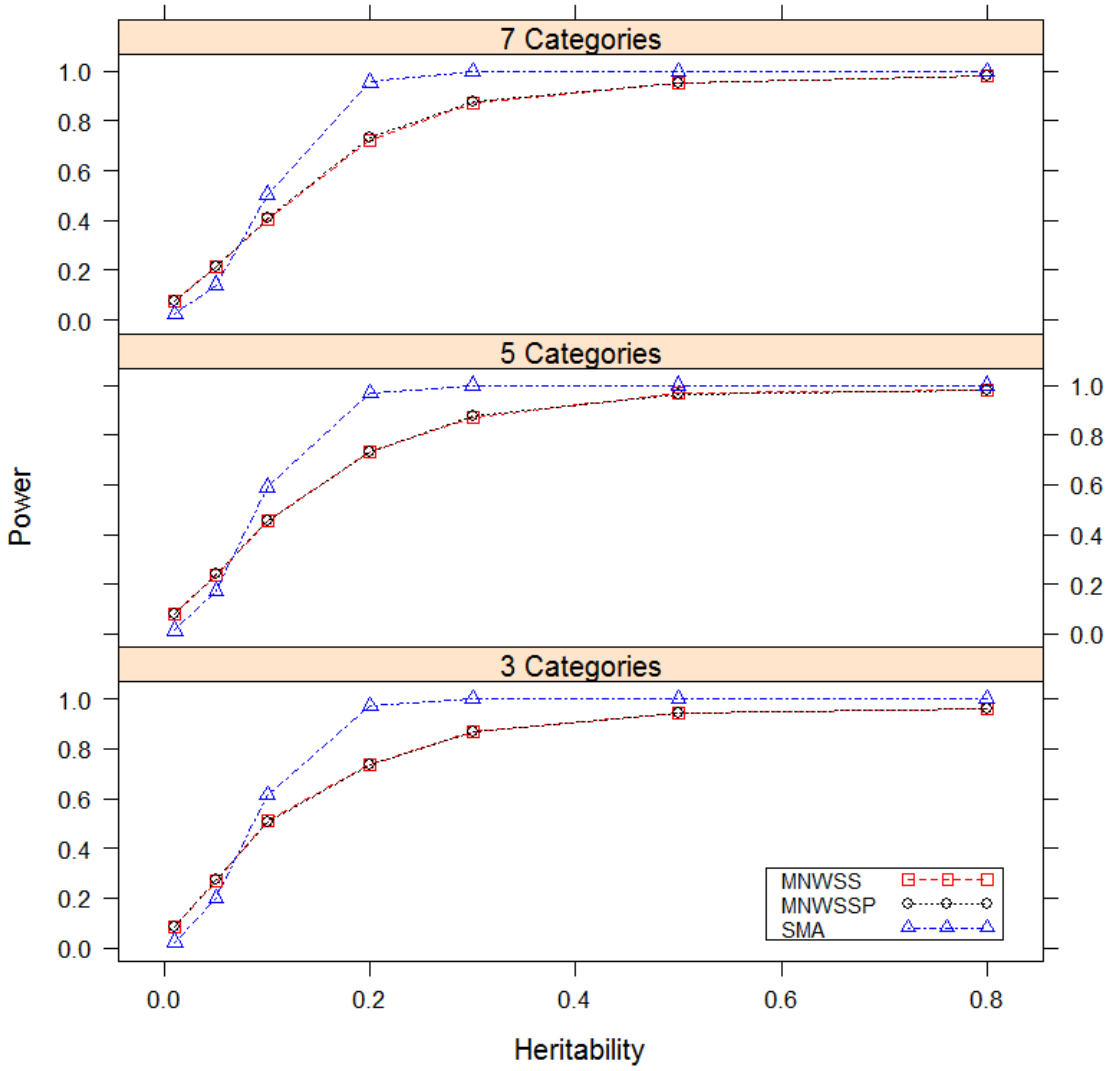

Figure 4.3: Power Comparison for a Sample Size of 500 with Three, Five, and Seven Phenotypic Categories

Figure 4.4 gives the power estimates of the methods for a sample size of 1000 with three, five, and seven phenotypic categories. Again the multinomial weighted sum statistic starts out with a higher power than the single marker analysis at a heritability of 1%. Here the SMA

overtakes the MNWSS for the heritability between 1% and 5%. Again the SMA quickly reaches a power of 1 while the MNWSS increases but never reaches a power of 1.

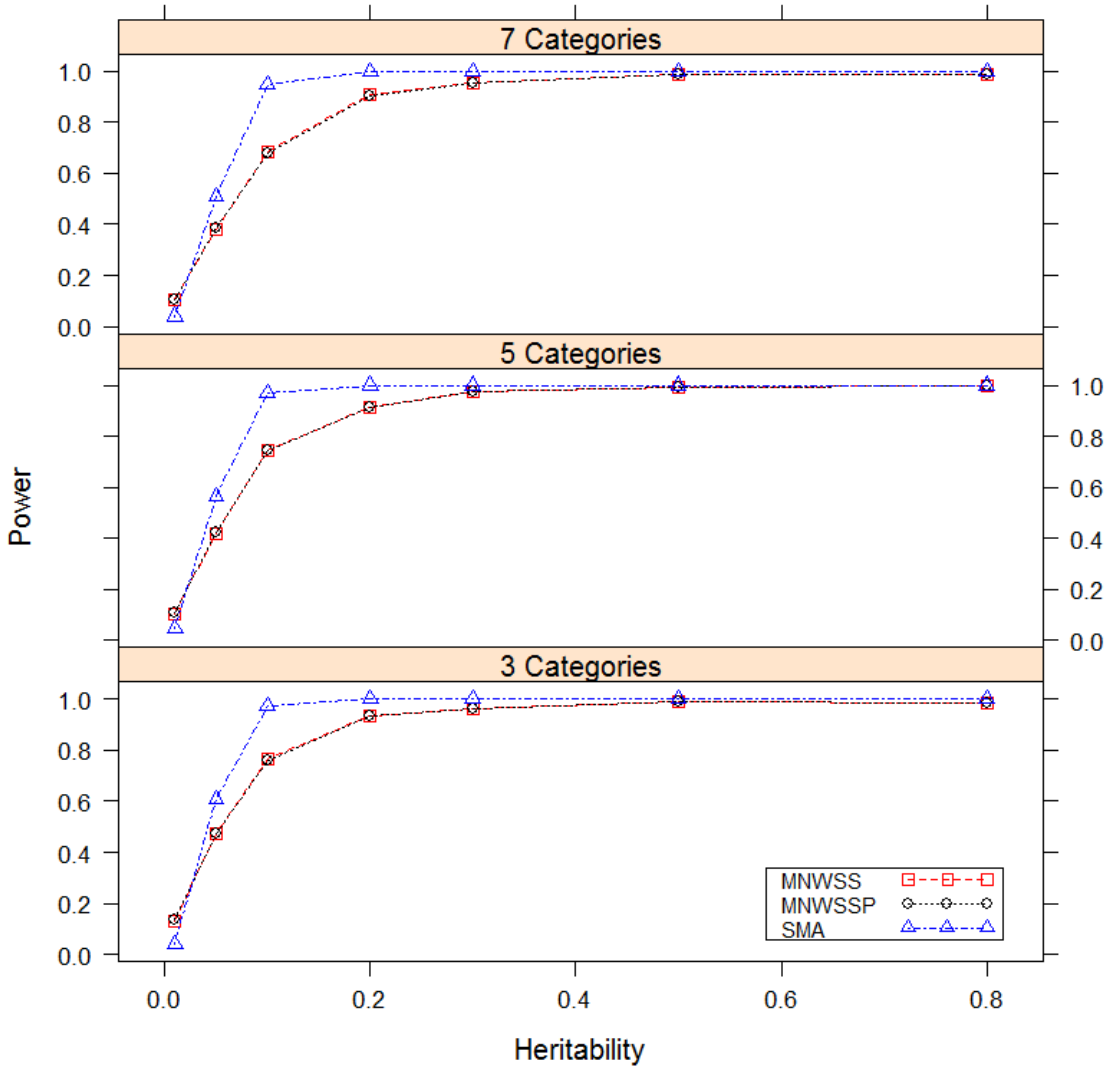

Figure 4.4: Power Comparison for a Sample Size of 1000 with Three, Five, and Seven Phenotypic Categories

Figure 4.5 gives the power comparisons for a sample size of 2000 with three, five, and seven phenotypic categories. Again for a heritability of 1% the MNWSS has higher power than the SMA. The SMA overtakes the WSS for the heritability between 1% and 5%. The single marker analysis reaches a power of 1 quickly while the MNWSS increases but does not reach 1.

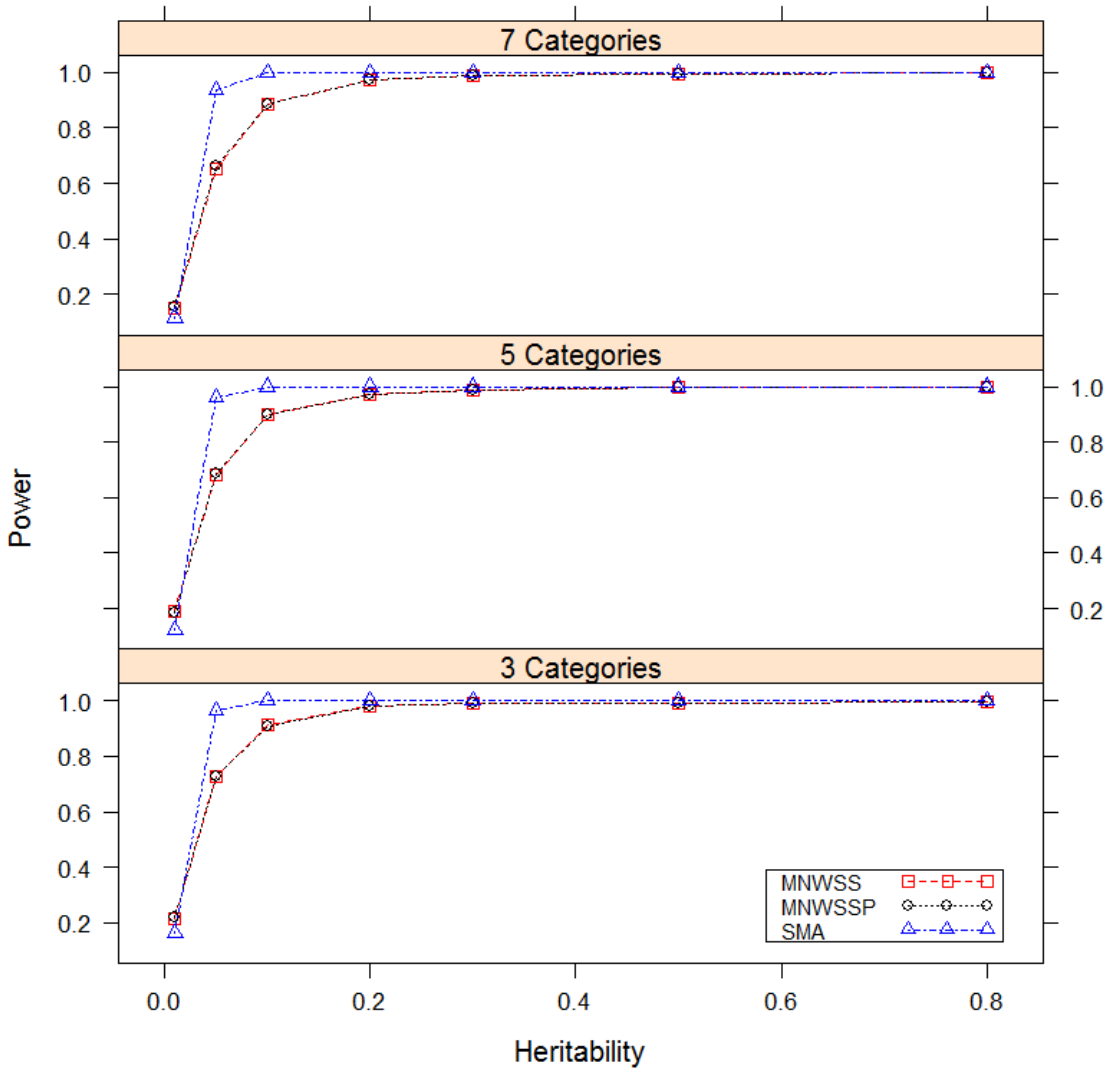

Figure 4.5: Power Comparison for a Sample Size of 2000 with Three, Five, and Seven Phenotypic Categories

Figure 4.6 gives a side by side comparison of power for all of the sample sizes and phenotypic categories. This figure illustrates that as sample size increases, the power of the methods increase and stabilize at or near one faster. This figure also shows that there are only small differences between the plots for the different number of phenotypic categories. These differences were detailed earlier.

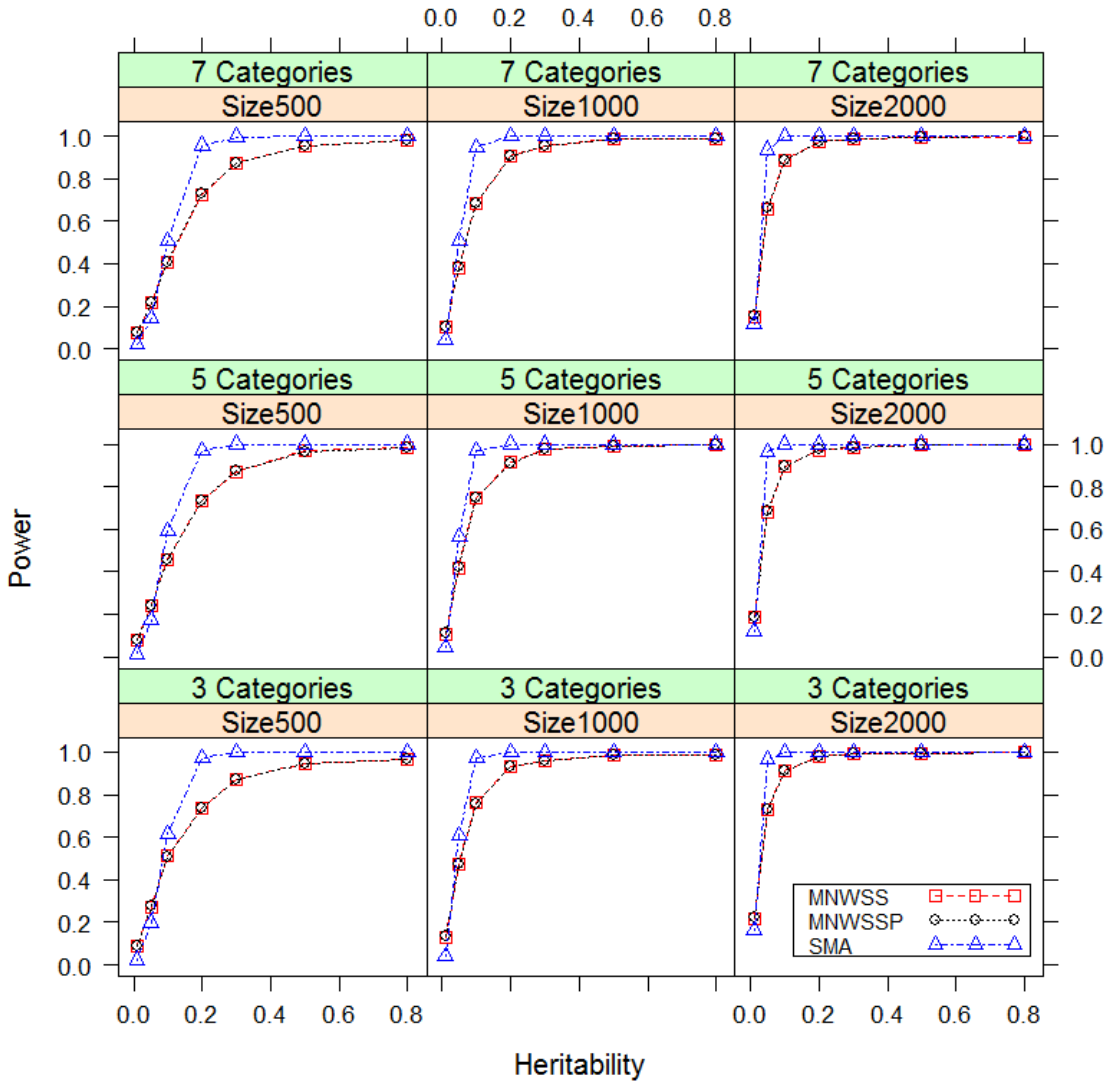

Figure 4.6: Side by Side Power Comparison for Sample Size and Phenotypic Categories

For each scenario considered above the single marker analysis has a lower power than the multinomial weighted sum statistic when the heritability is at 1%. Recall that the multinomial weighted sum statistic had the correct type I error rate while the single marker analysis was conservative due to the adjustment for multiple tests. The single marker analysis quickly gains power and over takes the multinomial weighted sum statistic as the heritability increases. The single marker analysis reaches and maintains a power of 1 while the multinomial weighted sum statistic increases in power more slowly and never reaches one. These results are very contrary to

the findings reported by Madsen and Browning for their study of the dichotomous case (2009). Madsen and Browning reported that the power of their weighted sum statistic for case control data increased quickly to 1 and remained there as the population attributable risk increased. They also showed that their single marker analysis always had a power less than 20% for any level of population attributable risk. These results may be specific to the way the data was generated. Madsen and Browning started with the case control status then generated the genotypes at each variant independently. The data simulation process for this dissertation study generated a population of haplotypes using a coalescent process then allowed the alleles to help determine the phenotype. This process should result in a population that is more realistic of the genetic structure of a real population. Madsen and Browning only used a Fisher's exact test on each variant in their single marker analysis. This dissertation proposes using the chi-square test of independence when all cell counts are five or greater and an exact test when a cell count fall below this threshold. Also Madsen and Browning used the Dunn-Sidak correction (Abdi, 2007) on the smallest p-value to determine the significance of the single marker analysis results. This work proposes using the False Discovery Rate controlling procedure of Benjamini and Hochberg (1995). The Benjamini and Hochberg procedure has been shown to be more powerful than the Bonferonni correction.

4.4 Sample Size Recommendations

It is intended that these methods will be used by future researchers. Therefore sample size recommendations are necessary for each of the methods. Since the power heavily depends on the heritability of the phenotype it is extremely important to have a good estimate of the heritability before proceeding. The heritability of the trait will vary from phenotype to phenotype. Asimit and Zeggini report that current GWAS on common variants can "explain at most 5% - 10% of the heritable component of disease" (2010). In the absence of an estimated heritability due to rare variants, heritability levels of 5%, 10%, and 20% will be investigated. The

highest heritability level is added since it is believed that rare variants can collectively contribute more than common variants to the variation of the phenotype. For heritability levels of 30% and above a sample size of 500 will be enough to give a power greater than 80% for both the multinomial weighted sum statistic and the single marker analysis. Results will be discussed by the heritability of the phenotype below. For all of the results below a size 0.05 test is considered.

Figure 4.7 visualizes the power versus sample size for a heritability level of 5% with three, five, and seven phenotypic categories. It is desired to determine approximately how many observations are necessary to achieve 80% power for both the multinomial weighted sum statistic method and the single marker analysis. The horizontal dashed gray line marks 80% power.

First consider three phenotypic categories. For the multinomial weighted sum statistic just a little over 2500 observations are needed to reach 80% power. The single marker analysis needs between 1000 and 1500 observations to reach this power. Interpolation between these two points gives approximately 1294 observations to reach 80% power. Slightly increasing this estimate and the estimates based on interpolation below would be prudent since this is a straight line interpolation of a convex line. Also the points used in the interpolation are estimates of the true power. The large sample sizes required by both methods are mostly due to the low heritability of the phenotype. The large number of non-causal variants in the simulated data might also be affecting the power of the multinomial weighted sum statistic.

Next consider the sample sizes for a heritability of 5% and five phenotypic categories. The multinomial weighted sum statistic needs significantly more than 2500 observations to achieve an 80% power. Interpolation between 2500 and 3000 yields a sample size of approximately 2654. Similar to the results for three categories the single marker analysis requires between 1000 and 1500 observations to achieve 80% power. Again using interpolation approximately 1393 observations are needed.

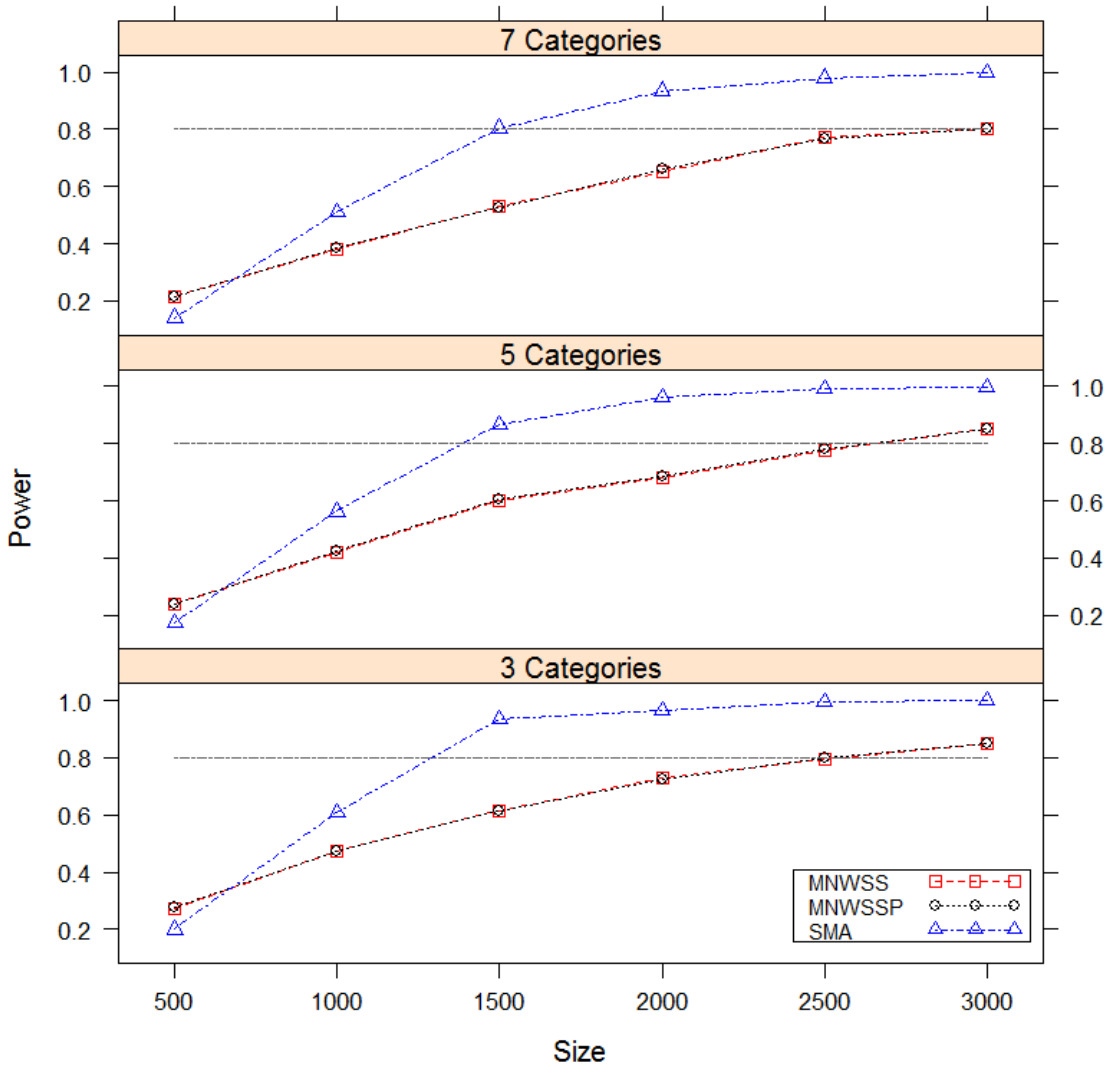

Figure 4.7: Power versus Sample Size for a Heritability of 5% with Three, Five, and Seven Phenotypic Categories

Finally consider the power curves for a heritability of 5% and seven phenotypic categories. No interpolation is needed for this scenario. The multinomial weighted sum statistic requires approximately 3000 observations to reach a power of 80%. The single marker analysis only needs 1500 observations to reach this power. Hence the multinomial weighted sum statistic requires double the observations needed by the single marker analysis to reach 80% power when the heritability is 5% and there are seven categories in the phenotype.

Overall for a heritability of 5% as the number of phenotypic categories increased the sample size requirement to reach 80% power also increased. The single marker analysis uses either a chi-square test or an exact test at each variant. These tests are dependent on the number of observations in each cell. Including more phenotypic categories in the test increases the number of cells. Hence the data gets spread over more cells as the phenotypic categories increase. The multinomial weighted sum statistic uses the Kruskal-Wallis test in the procedure. Increasing the number of phenotypic categories increases the number of populations the test. So once again the data is spread out as the number of phenotypic categories increases. This spreading out of the data could account for the larger sample size requirements in both methods as the number of phenotypic categories increases.

Next study a heritability level of 10%. Figure 4.8 gives the power curves for a heritability of 10% with three, five, and seven phenotypic categories. For three phenotypic categories the multinomial weighted sum statistic requires between 1000 and 1500 to reach the threshold while the single marker analysis needs between 500 and 750 observations. Again using interpolation an approximation can be found for a more exact sample size required. For the multinomial weighted sum statistic approximately 1183 observations are needed. For the single marker analysis approximately 675 observations are needed.

Next review the power curves for a heritability of 10% and five phenotypic categories displayed in Figure 4.8. To reach 80% power the multinomial weighted sum statistic needs between 1000 and 1500 observations. Interpolating as before, approximately 1255 individuals are needed. For the single marker analysis between 500 and 750 observations are required to reach an 80% power. Straight line interpolation yields approximately 695 observations.

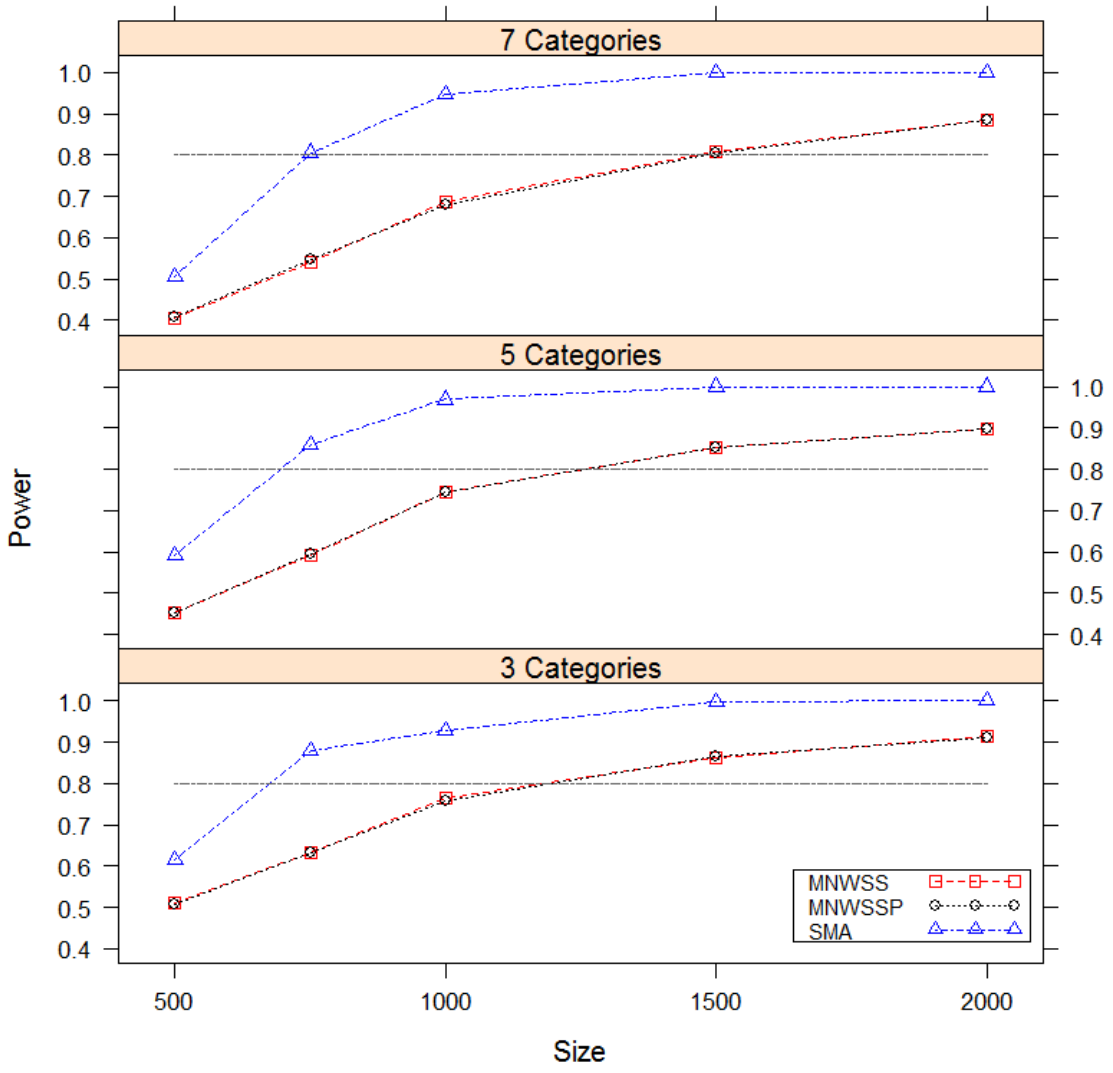

Figure 4.8: Power versus Sample Size for a Heritability of 10% with Three, Five, and Seven Phenotypic Categories

The final plot in Figure 4.8 maps the power versus sample size for seven phenotypic categories when the heritability is 10%. No interpolation is necessary for this set of parameters. The multinomial weighted sum statistic requires approximately 1500 observations to achieve 80% power. The single marker analysis only requires 750 observations to achieve this same level. Hence with seven phenotypic categories and a heritability of 10%, the single marker

analysis requires only half as many observations to reach 80% power as the multinomial weighted sum statistic.

Overall the sample size requirements for a heritability of 10% were less than for a heritability of 5%. This is expected since the heritability is a measure of the strength of the association between the phenotype and variants. Similar to the results for the heritability at 5%, the sample size requirements increased as the number phenotypic categories increased.

Finally examine the power versus sample size for the heritability at 20%. Figure 4.9 plots the power curves for three, five, and seven categories in the phenotype. Once again the gray dashed line represents the desired 80% power. These curves show that both methods reach 80% power with smaller sample sizes than considered above. Also the trend of increasing sample size with increasing phenotypic categories continues.

Start with the results for three phenotypic categories. The multinomial weighted sum statistic needs between 500 and 750 observations to reach 80% power. Interpolation gives approximately 623 observations. A sample size of 500 is more than adequate to achieve 80% power for the single marker method.

The middle plot in Figure 4.9 displays the power versus sample size for a heritability level of 20% and five phenotypic categories. Once again the multinomial weighted sum statistic requires between 500 and 750 individuals to achieve 80% power. Approximating using straight line interpolation yields 643 observations. Similar to before a sample size of 500 is more than enough for the single marker analysis to achieve 80% power.

The top plot in Figure 4.9 displays the power curves for seven phenotypic categories with a heritability level of 20%. The sample size to reach 80% power for the weighted sum statistic lies between 500 and 750 observations. Interpolating once again finds an approximate sample

size of 685. A sample size of 500 is more than enough to achieve 80% for the single marker analysis.

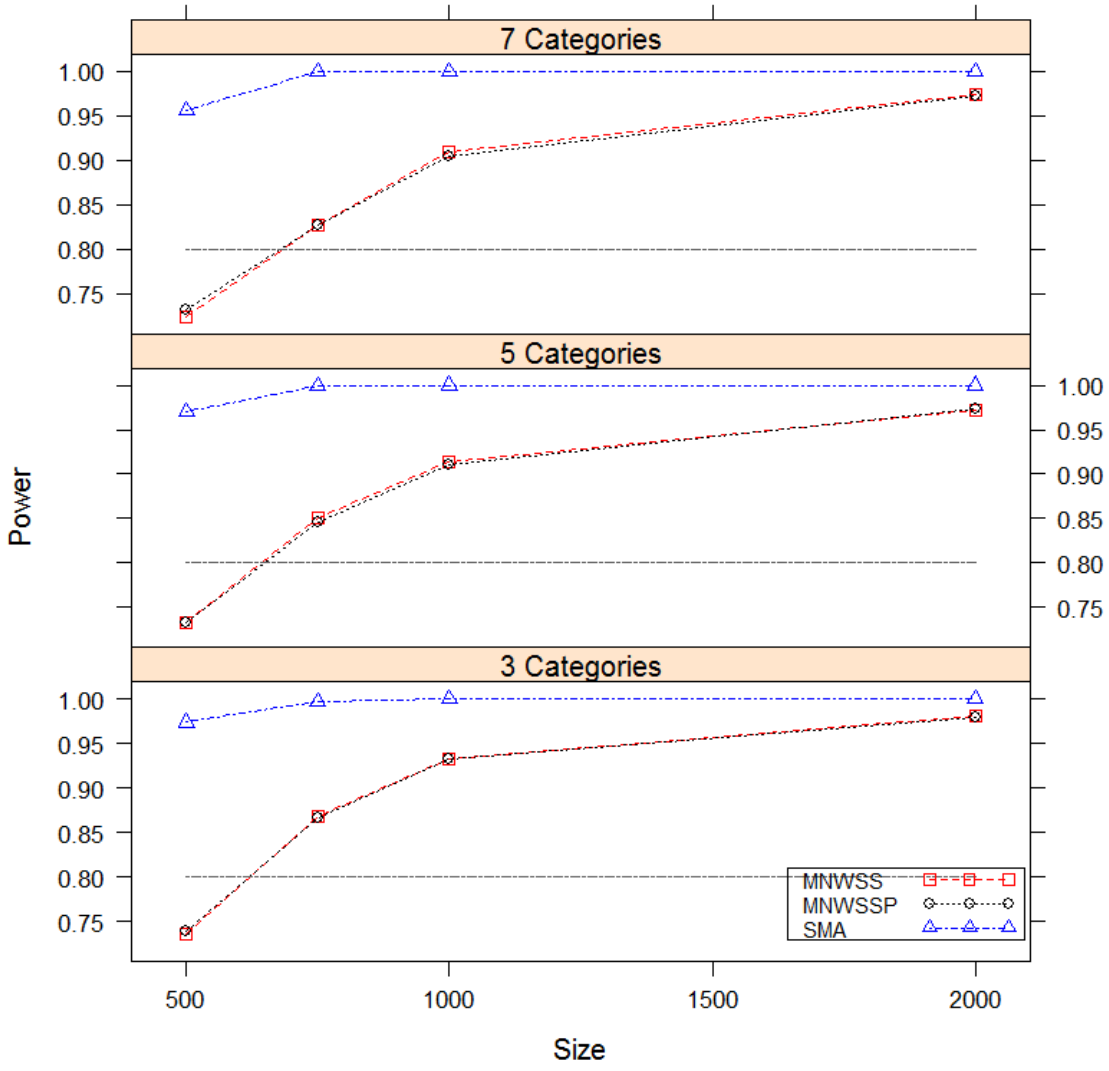

Figure 4.9: Power versus Sample Size for a Heritability of 20% with Three, Five, and Seven Phenotypic Categories

4.5: Summary

The multinomial weighted sum statistic and the single marker analysis have a reasonable or conservative Type I Error rate. However the power for these methods is low when the heritability is less than 10%. The results presented here are dependent on the structure of the

simulated data. The large number of non-causal variants in the simulated data may have reduced the power for the MNWSS. It has been demonstrated that pooling methods lose power when a large number of non-causal variants are included in the test (Basu & Pan, 2011). For a heritability of 10% to 20% the single marker analysis is recommended even though it is very conservative. For heritability greater than 20% both the multinomial logistic regression and the single marker analysis reach reasonably high power. This does not mean that the multinomial weighted sum statistic has a higher power than the single marker analysis for this heritability. Rather both have approximately 80% power or greater when the heritability is greater than 20%.

The multinomial logistic regression is not recommended. The inflated Type I Error rate of the method makes it unsuitable as a method of association. Additionally the unpredictability of the multinomial logistic regression routine makes it unreliable as a method. Collapsing of variants that had quasi-complete separation helped in fixing the inflated Type I Error rate but it did not completely resolve the issues with failures in the routine. For this reason it is also not recommended as a method for association. Further research is needed to devise modifications to the multinomial logistic regression methods to make them viable.

Sample size recommendations were made for heritability levels of 5%, 10%, and 20%. Estimates based on straight line interpolation should be slightly increased as noted above. The heritability levels used in the sample size study were arbitrarily chosen. Other levels may better suit the expectations of researchers. Since the power greatly depends on the heritability, additional simulations may be necessary for a given heritability level. The results of the simulation study showed that the proposed single marker analysis required many less observations than the multinomial weighted sum statistic to reach 80% power. The simulation study also revealed that as the number of phenotypic categories increases the required sample size to reach 80% power increases for both methods.

CHAPTER V

APPLICATION

In the previous chapter it was shown that the proposed multinomial weighted sum statistic (MNWSS) and the single marker analysis (SMA) performed well in simulations for reasonable sample sizes. Ultimately the methods should be applied by analyzing real data. The purpose of this chapter is to show that the methods are viable for analysis on an empirical data set. The data analyzed here is a resequencing study of participants in the Dallas Heart Study at the University of Texas Southwestern. Section 5.1 will begin by describing the data sets (including phenotypes in Section 5.1.1 and genotypes in Section 5.1.2) and findings other researchers have published on the same data sets. Section 5.2 will describe the analysis of the data using the proposed methods. Section 5.3 will present the results. Lastly section 5.4 compares the results of the proposed methods with other researchers' findings and provides a discussion of the conclusions.

5.1 Dallas Heart Study Data

An association between multiple rare variants in the ANGPTL4 gene and plasma triglyceride was reported in a resequencing study of Dallas Heart Study participants (Romeo, et al., 2007). Further research using the same cohort found an association between rare variants in the ANGPTL3 and ANGPTL5 genes with plasma triglyceride level (Romeo, et al., 2009). These

rare variant associations were further confirmed in an investigation of several case control methods (Price, et al., 2010). In each of these studies the quantitative phenotype, plasma triglyceride level, was adjusted for race and gender. Then a categorical phenotype was created using the quartile membership. Only individuals in the top and bottom quartiles of the adjusted plasma triglyceride distribution were included in the tests for association. This was necessary since the tests (see below) could only accommodate two categories. Excluding the middle fifty percent of the distribution decreased the significance of the tests compared to a quantitative test on all of the observations (Price, et al., 2010). The original studies performed at the University of Texas Southwestern used a Fisher's exact test on the number of individuals with nonsynonymous variants (nucleotide mutations that change the amino acid sequence) in the categories. Several proposed case control methods were run on the data by Price et al. (2010).

5.1.1 Phenotype Data Set

The data consists of two separate data sets. The first is a set of phenotypic variables and several covariates. The second is a data set of genotypes. The phenotypic data set contains 3557 observations, including 1986 females and 1571 males. For each research participant a subject ID, gender, ethnicity, age, body mass index (BMI), statin, and plasma triglyceride were given. Some of the observations included missing values for one or more of the variables. Treatment of missing values is discussed in Section 5.2. The ethnic groups break down as follows: 603 Hispanic, 1832 Non-Hispanic Black, 1047 Non-Hispanic White, and 75 other. The ages of the participants ranged from 18 to 65 years old. Body mass index was also considered as a phenotype by the original researchers. Their studies found no significant associations between BMI and any of the genes (Romeo, et al., 2007; Romeo, et al., 2009). The BMI ranges between 14.45 and 65.23. A histogram of the variable of interest, plasma triglyceride level, is presented in figure 5.1. This histogram was generated using the provided data. The plasma triglyceride level ranges from 21 to 1669.

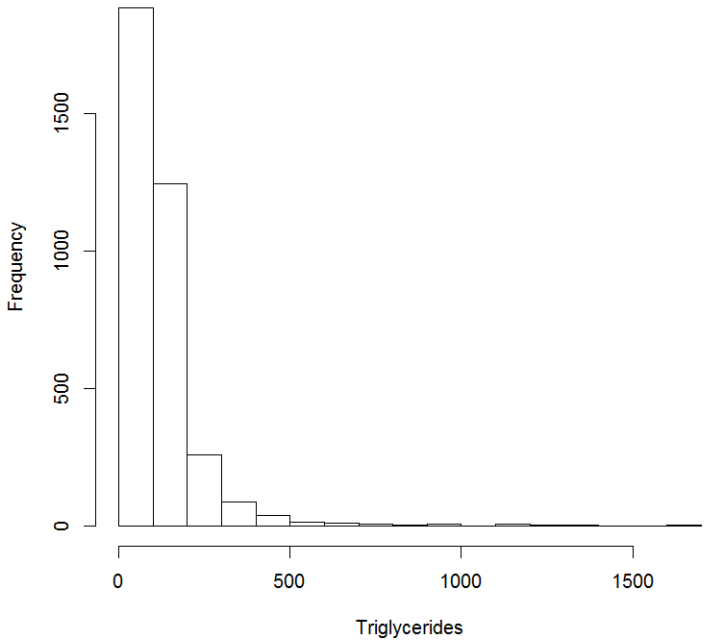

Figure 5.1: Histogram of Triglycerides

5.1.2 Genotype Data Set

The second data set contains genotypes. The data set is a list of mutant genotypes. There are 15819 records in the data set. For each record the gene, subject ID, mutation ID, mutation type, and genotype are given. The genotypes are coded as 1 for heterozygous and 2 for homozygous for the mutant allele. None of the wild type homozygous genotypes are listed in the data set. Rather they are implied by their absence. There are three separate genes with genotype data available. They are ANGPTL3, ANGPTL4, and ANGPTL5. The subject ID is not unique in this data set since some subjects have multiple mutations. Subjects with only wild type alleles are not listed in the genotype data set. The mutation ID gives the name of the mutation. There are 282 unique mutations in the data set. They are typed as frame shift, intronic, missense, noncoding, nonsense, or synonymous. The minor allele frequencies of the mutations are

presented in figure 5.2. There are eleven mutations that are not rare variants. The remaining 271 mutations are rare variants. The inclusion/exclusion of these variants in the analysis is discussed in section 5.3.

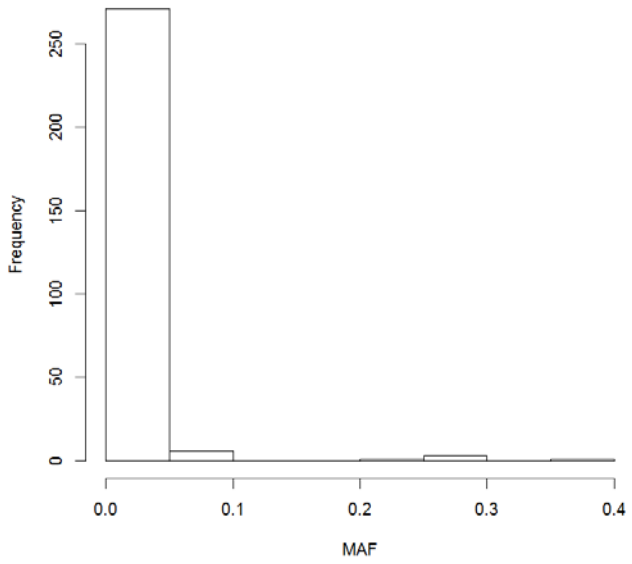

Figure 5.2: Histogram of Minor Allele Frequencies

5.2 Analysis Methods

The phenotype, plasma triglyceride level, was adjusted and categorized by previous researchers prior to analysis. Before adjusting the phenotype individuals taking statins, lipid lowering medicines, were excluded from the analysis. This was also done by the original studies (Romeo, et al., 2007; Romeo, et al., 2009). The original studies also excluded individuals with diabetes, men consuming more than 30g of alcohol a day, and women consuming more than 20g of alcohol a day. Since diabetes information and alcohol consumption were not made available it is unknown whether these individuals remain in the data set for this analysis. The adjustment and categorization described below was adapted from the procedure used by Price et al. (2010). First eight groups were formed by the combinations of gender and ethnicity. Within each group the

plasma triglyceride was standardized. This adjusted triglyceride level was then categorized using the quartiles. The first category is the lowest quartile. The second category is the middle 50% of the distribution. The last category is the upper quartile. Previous researchers discarded the second category (Romeo, et al., 2007; Romeo, et al., 2009; Price, et al., 2010). For this research it is kept and included in the analysis since the methods previously proposed can accommodate more than two categories. Individuals missing plasma triglyceride were given a missing value for the phenotype.

The genotype data was not in a commonly used format and required some restructuring. As mentioned before, the genotype data is a list of mutant genotypes and does not include wild type genotypes. An individual will have multiple observations in the genotype file if the individual has a mutant genotype at multiple SNPs. Likewise an individual will have no observations in the genotype file if all genotypes are the wild type homozygous. This necessitated some additional processing so that the proposed methods could be applied. Prior to running the analysis a data set containing one observation for each participant and one variable for each variant was created from the genotype data file. For each variant the genotype is coded as "0" for the homozygous wild type, "1" for heterozygous, or "2" for mutant homozygous.

The original researchers analyzed each gene separately (Romeo, et al., 2007; Romeo, et al., 2009). They also did not use all 282 mutations found in the genotype file. Only nonsynonymous sequence variants were considered for analysis. In addition to excluding individuals in the middle 50% of the distribution, variants that contained individuals in both the top and bottom quartiles of adjusted plasma triglycerides were also excluded. A list of the variants in the reduced data set is provided in Romeo et. al. 2009. Table 5.1 below gives the number of variants for the full and reduced data sets. Of the 88 variants in the ANGPTL3 gene only 17 were included in the original analysis. For the ANGPTL4 gene only 14 of the 94 variants were used. Only nine of the 100 variants in the ANGPTL5 gene were utilized in the original test.

For comparison purposes the proposed methods were applied to both the full set and reduced set of variants for each gene. It should be noted that for the reduced set of variants the MAF's are all less than 0.1%.

	Number of Variants	
Gene	Full	Reduced
ANGPTL3	88	17
ANGPTL4	94	14
ANGPTL5	100	9

Table 5.1: Number of Variants in the Full and Reduced Data Sets

Price and authors (2010) took a different approach when analyzing the data set. Rather than analyzing each gene individually they applied their methods to the whole data set. They did not specify whether they used the full set of variants or the reduced set of variants. Additionally they excluded individuals in the ethnic group "Other". Although analyzing each gene individually is more informative, for comparison purposes the proposed tests were run using all three genes together with the ethnic group "Other" excluded. Both the full set and reduced set were considered in this analysis.

5.3 Results

For information purposes the MNWSS and SMA were first run on the full variant sets for each gene. This approach represents a naïve use of the data and does not require any information on the type or functionality of the mutations. Results for the full and reduced data sets by gene are provided in table 5.2. For the MNWSS none of the tests are significant after correcting for multiple tests. The observed significance levels for ANGPTL3, ANGPTL4, and ANGPTL5 genes were 0.2646, 0.4454, and 0.0237 respectively. For the SMA none of the tests produced a decision to reject the null hypothesis of no association after adjusting for the multiple tests. Since the original studies only considered the reduced set of variants, a comparison cannot be made with these observed significance levels.

		Data Set	
Gene	Method	Full	Reduced
ANGPTL3	MNWSS	0.2624	0.0165
	SMA	All "Do Not Reject"	All "Do Not Reject"
	Romeo, et. al. 2009		0.064
ANGPTL4	MNWSS	0.4454	0.0044
	SMA	All "Do Not Reject"	All "Do Not Reject"
	Romeo, et. al. 2007		0.016
ANGPTL5	MNWSS	0.0237	0.119
	SMA	All "Do Not Reject"	All "Do Not Reject"
	Romeo, et. al. 2009		0.022

Table 5.2: Results for the Full and Reduced Data Sets by Gene

The original study only provides results for the reduced set of variants. They also did not account for the multiple tests being simultaneously performed. For the ANGPTL3 gene the researchers report an observed significance level of 0.064 for the test of association (Romeo, et al., 2009). Although this is not significant at the researchers' chosen 0.05 level it is close to significance. The MNWSS run on the reduced set produces a p-value of 0.0165 for the test of association between the ANGPTL3 gene and triglycerides. The SMA did not produce any decisions to reject. The original study reports a p-value of 0.016 for the test of association between the ANGPTL4 gene and triglycerides (Romeo, et al., 2007). The MNWSS run on the reduced set of variants in the ANGPTL4 gene produces a p-value of 0.0044. The SMA did not produce any decisions to reject. For the ANGPTL5 gene the original study reports a p-value of 0.022 for their test of association (Romeo, et al., 2009). The MNWSS gains an observed significance level of 0.1190 when run on this reduced set. The SMA again did not produce any decisions to reject the null hypothesis of no association. The original researchers did not correct for the multiple tests being performed. If they had used the Benjamini and Hochberg (1995) FDR controlling method on the three tests considered here they would have found the results of the tests for the ANGPTL4 and ANGPTL5 genes significant. The observed significance levels for the MNWSS for the ANGPTL3 and ANGPTL4 genes are significant after using the Benjamini

and Hochberg (1995) FDR controlling method to account for the multiple tests. Also the MNWSS yields p-values below the reported p-values in the original studies for these two genes. The SMA did not find any significant associations for any of the genes. Further inspection shows that the individual tests in the SMA tend to yield high p-values. It should be noted here that the data set analyzed by the proposed methods may be slightly different from the one used in the original study. Since information on diabetes and alcohol consumption is not contained in the data set analyzed by the proposed methods, individuals excluded in the original analysis may be included for the new results. This could result in some differences between the results of the original studies and the results of the proposed methods.

As mentioned before Price et al. did not consider the genes separately but rather ran one test combining all three genes (2010). They also did not relay whether they used the full or the reduced set of genes or whether they excluded individuals taking statins. Their work proposes five different tests of association for case control data. Their results (assuming the reduced data set) and the results of the proposed methods are presented in table 5.3. Their proposed fixed threshold tests produced p-values of 0.013 for a one percent threshold and 0.00007 for a five percent threshold. Price et al.'s weighted approach yields a p-value of 0.0020. Their variable threshold test outputs a p-value of 0.00038. Their recommended variable threshold test plus Polyphen weights yields a p-value of 0.00002. Now consider the tests proposed in this work. The SMA did not produce any rejections of the null hypotheses of no association for either the full or reduced set of variants. Using the full set of variants the MNWSS finds a p-value of 0.1092 when all three genes are combined. Running the MNWSS test on only the reduced set of variants with all three genes together produces a p-value of 0.00000393. These results suggest that Price et al. (2010) uses only the reduced set of variants when running their analyses since their weighed approach is very similar to Madsen and Browning's (2009). Comparing the results

for the reduced set of variants, the MNWSS has a p-value lower than any of the tests proposed by Price et al.

	Data Set	
Method	Full	Reduced
MNWSS	0.1092	0.00000393
SMA	All "Do Not Reject"	All "Do Not Reject"
Fixed Threshold 1%†		0.013
Fixed Threshold 5%†		0.00007
Weigthed†		0.002
Variable Threshold†		0.00038
Variable Threshold + Polyphen†		0.00002

†Price, et. al. 2010

Table 5.3: Results of Combining All Genes in the Full and Reduced Data Sets

5.4 Discussion

This chapter demonstrates that the proposed methods are viable for data analysis. The MNWSS was able to detect associations in two of the three genes at a 0.05 level after adjusting for the multiple tests. The original studies report associations in two of the three genes at a 0.05 level without adjusting for multiple tests. Additionally for the genes where the MNWSS produces a significant result, the p-values are smaller than the p-values from the original studies. When the MNWSS is applied to all three genes together on the reduced set of variants the result is more significant than any of the results presented by Price and authors (2010) in their analysis of the data.

The SMA did not produce any decisions to reject the null hypothesis of no association after correcting for multiple tests. Further analysis revealed that the individual p-values at each variant were high. Thus no one variant is strongly associated with plasma triglyceride levels. Rather collectively the rare variants are associated with the phenotype.

Comparing the results for the full and reduced set of variants highlights the importance of choosing which variants to include in the test. The reduced set excludes synonymous variants

which do not alter the resulting protein structure and hence are not likely to have an effect. The reduced set also excludes variants for which there are individuals in both the top and bottom quartiles that had the mutation. This in effect excluded all of the common variants. This cherry picking of variants reversed the decisions on all of the tests for the MNWSS. For the ANGPTL3 and ANGPTL4 genes the p-values were reduced from 0.2646 to 0.0165 and from 0.4454 to 0.0044 by selecting variants for inclusion in the test. On the other hand for the ANGPTL5 gene the p-value was increased from 0.0237 to 0.1190. Recall for this gene that 100 different mutations were collected. However, only nine of them made it into the reduced set of variants. In this case the researchers may have thrown out some important mutations.

CHAPTER VI

CONCLUSIONS

This dissertation study investigated three new methods to test for an association between a nominal phenotype and multiple rare variants. Since methods in this area were lacking the proposed methods came from extending methods currently used to test for an association between a dichotomous phenotype and multiple rare variants. The methods proposed and evaluated here provide a starting point into association analysis for data with a multinomial phenotype and multiple rare variants.

There is still room for a great deal of exploration in this area. Since the inception of this project there has been an explosion of methods for association between a dichotomous phenotype and multiple rare variants. Many of these methods could be extended to the case of a multinomial phenotype. Also this project encountered difficulties in using multinomial logistic regression as a method of association. Collapsing of variants with quasi-complete separation was tested as a quick fix to the problems. However this fix did not solve all of the issues. Additional modifications to multinomial logistic regression are necessary for it to be a viable method of association. There are many variations on logistic regression to test for association between a dichotomous phenotype and multiple rare variants (Basu & Pan, 2011). Some of these methods might be modified for the multinomial case.

The methods proposed here ignored possible epistasis, or interactions between variants. This assumption is common in rare variant association methods. Research into how rare variant methods behave in the presence of epistasis is needed.

Another area that needs to be addressed in methods of rare variant analysis is the inclusion of covariates. Most data sets contain covariates such as age, gender, and ethnicity that affect the phenotype. Neither of the proposed methods can accommodate covariates. There are methods of rare variant analysis for dichotomous and quantitative phenotypes that can include covariates. For dichotomous phenotypes the weighted SSU test with permutations (Basu & Pan, 2011) and kernel-machine test (Wu, et al., 2010) specifically allow for including multiple covariates. For quantitative phenotypes Morris and Zeggini's (2010) tests can include covariates. However these methods may not be suited to all data sets. For example the SSU test with permutations and the kernel-machine test are both very computationally intensive. These methods would not be suited for analyzing a large data set. Rare variant analysis methods for both dichotomous and multinomial phenotypes that can include covariates without computational complexity would be extremely useful.

REFERENCES

Abdi, H. (2007). The Bonferonni and Sidak Corrections for Multpile Comparisons. In N. Salkind, *Encyclopedia of Measurement and Statistics.* Thousand Oaks, CA: Sage.

Agresti, A. (2002). *Categorical Data Analysis.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*(1), 1-10.

Allison, P. D. (2008). Convergence Failures in Logistic Regression. *SAS Global Forum 2008: Statistics and Data Analysis*, (pp. 1-11).

Asimit, J., & Zeggini, E. (2010). Rare Variant Association Analysis Methods for Complex Traits. *Annual Review of Genetics, 44*, 293-308.

Bain, L. J., & Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics* (2nd ed.). Belmont, California: Duxbury Press.

Bansal, V., Ondrej, L., Torkamani, A., & Schork, N. J. (2010). Statistical Analysis Strategies for Association Studies Involving Rare Variants. *Nature Reviews Genetics, 11*, 773-785.

Basu, S., & Pan, W. (2011). Comparison of Statistical Tests for Disease Association with Rare Variants. *Genetic Epidemiology, 35*, 606-619.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, 75*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annuals of Statistics, 29*(4), 1165-1188.

Clarkson, D. B., Fan, Y.-a., & Joe, H. (1993). A Remark on Algorithm 643: FEXACT: An Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables. *ACM Transactions on Mathematical Software, 19*(4), 484-488.

Conover, W. J. (1999). *Practical Nonparametric Statistics.* New York: John Wiley & Sons, Inc.

Croissant, Y. (n.d.). *Estimation of Multinomial Logit Models in R: The mlogit packages*. Retrieved December 1, 2011, from http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf

Feng, T., Elson, R. C., & Zhu, X. (2011). Detecting Rare and Common Variants for Complex Traits: Sibpair and Odds Ratio Weighted Sum Statistics (SPWSS, ORWSS). *Genetic Epidemiology, 35*, 398-409.

Fox, J. (2007, July 9). *Re: [R] Row-Echelon Form*. Retrieved December 2011, from tolstoy.newcastle.edu.au/R/ez/help/07/09/24777.html

Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R., & Amos, C. I. (2008). Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. *The American Journal of Human Genetics, 82*, 100-112.

Han, F., & Pan, W. (2010). A Data-Adaptive Sum Test for Diease Association with Multiple Common or Rare Varaints. *Human Heredity, 70*, 42-54.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide Association Studies for Common Diseases and Complex Traits. *Nature Reviews Genetics, 6*, 95-108.

Hoffmann, T. J., Marini, N. J., & Witte, J. S. (2010). Comprehensive Approach to Analyzing Rare Genetic Variants. *PLoS ONE, 5*(11).

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* New York City: John Wiley & Sons, Inc.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher nuetral model of genetic variation. *Bioinformatics, 18,* 337-338.

Ionita-Laza, L., Buxbaum, J. D., Laird, N. M., & Lange, C. (2011). A New Testing Strategy to Identify Rare Variants with Either Risk or Protective Effect on Disease. *PLoS Genetics, 7*.

Lawrence, R., Day-Williams, A. G., Elliott, K. S., Morris, A. P., & Zeggini, E. (2010). CCRaVAT and QuTie - Enabling Analysis of Rare Variants in Large-scale Case Control and Quantitative Trait Association Studies. *BMC Bioinformatices, 11*, 527.

Li, B., & Leal, S. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases; Application to Analysis of Sequence Data. *The American Journal of Human Genetics, 83*, 311-321.

Liu, D. J., & Leal, S. M. (2010). A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genetics, 6*(10).

Mackay, T. F. (2009). Q&A: Genetic Analysis of Quantitative Traits. *Journal of Biology, 8*(23), 23.

Madsen, B. E., & Browning, S. R. (2009). A Groupwise Associaton Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics, 5*(2).

Mehta, C. R., & Patel, N. R. (1986). Algorithm 643 FEXACT: A FORTRAN Subroutine for Fisher's Exact Test on Unordered rXc Contingency Tables. *AMC Transactions on Mathematical Software, 12*(2), 154-161.

Morris, A. P., & Zeggini, E. (2010). An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies. *Genetic Epidemiology, 34*, 188-193.

Mukhopadhyay, I., Feingold, E., Weeks, D. E., & Thalamuthu, A. (2010). Association Tests Using Kernel-Based Measures of Multi-Locus Genotype Similarity Between Individuals. *Genetic Epidemiology, 34*, 213-221.

Neale, B. J., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., et al. (2011). Testing for an Unusual Distribution of Rare Variants. *PLos Genetics, 7*(3).

Pan, W. (2009). Asymptotic Tests of Association with Multiple SNPs in Linkage Disequilibrium. *Genetic Epidemiology, 33*, 497-507.

Panagiotou, O. A., Evangelou, E., & Ioannidis, J. P. (2010). Genome-wide Significant Associations for Variants with Minor Allele Frequency of 5% or Less - An Overview. *American Journal of Epidemiology, 172*, 869-889.

Price, A. L., Kryukov, G. V., DeBakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J., et al. (2010). Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *The American Journal of Human Genetics, 86*, 832-838.

Pritchard, J. K. (2001). Are Rare Variants Responsible for Susceptibility to Complex Diseases? *American Journal of Human Genetics, 69*, 124-137.

Pritchard, J. K., & Cox, N. J. (2002). The Allelic Architecture of Human Disease Genes: Common Disease - Common Variant ... or Not? *Human Molecular Genetics, 11*, 2417-2423.

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Reich, D. E., & Lander, E. S. (2001). On the Allelic Spectrum of Human Disease. *Trends in Genetics, 17*(9), 502-510.

Risch, N., & Merikangas, K. (1996). The Future of Genetic Studies of Complex Diseases. *Science, 273*, 1516-1517.

Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H., et al. (2007). Population-based Resequencing of ANGPTL4 Uncovers Variations That Reduce Triglycerides and Increase HDL. *Nature Genetics, 39*(4), 513-516.

Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L. A., Boerwinkle, E., Hobbs, H. H., et al. (2009). Rare Loss-of-function Mutations in ANGPTL Family Members Contribute to Plasma Triglyceride Levels in Humans. *The Journal of Clinical Investigation, 119*(1), 70-79.

Schaid, D. J., McDonnell, S. K., Hebbring, S. J., Cunningham, J. M., & Thibodeau, S. N. (2005). Nonparmetric Tests of Association of Multiple Genes with Human Diease. *American Journal of Human Genetics, 76*, 780-793.

Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. Rare Allele Hypotheses for Complex Disease. *Current Opinion in Genetics & Development, 19*, 212-219.

Smith, D. J., & Lusis, A. J. (2002). The Allelic Structure of Common Disease. *Human Molecular Genetics, 11*, 2455-2461.

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Magnusson, K. P., et al. (2007). Genetic Determinants of Hair, Eye, and Skin Pigmentation in Europeans. *Nature Genetics, 39*, 1443-1452.

Thalamuthu, A., Zhao, J., Keong, G. T., Kondragunta, V., & Mukhopadhyay, I. (2011). Association Tests for Rare and Common Variants Based on Genotypic and Phenotypic Measures of Similarity Between Individuals. *BMC Proceedings, 5*(Suppl 9).

VanLiere, J. M., & Rosenberg, N. A. (2008). Mathematical Properties of the r2 Measure of Linkage Disequilibrium. *Theoretical Population Biology, 74*(1), 130-137.

Verhoeven, K. J., Simonsen, K. L., & McIntyre, L. M. (2005). Implementing False Discovery Rate Control: Increasing Your Power. *OIKOS, 108*, 643-647.

Wessel, J., & Schork, N. J. (2006). Generalized Genomic Distance-Based Regression Methodology for Multilocus Association Analysis. *The American Journal of Human Genetics, 79*, 792-806.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., et al. (2010). Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics, 86*, 929-942.

Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., & Zollner, a. S. (2010). Extending Rare-Variant Testing Strategies: Analysis of Noncoding Sequence and Imputed Genotypes. *The American Journal of Human Genetics, 87*, 604-617.

Zeggini, E., & Asimit, J. L. (2010). An Evaluation of Power to Detect Low-Frequency Variant Associations using Allele-Matching Tests that Account for Uncertainty. *WSPC - Proceedings* (pp. 100-105). eproceedings.worldscinet.com.

APPPENDICES

Table A.1: Power Comparison for a Sample Size of 500

		Lambda						
Categories	Method	0.01	0.05	0.1	0.2	0.3	0.5	0.8
3	MNWSS	0.086	0.271	0.51	0.735	0.869	0.944	0.963
	MNWSSP	0.086	0.277	0.506	0.739	0.869	0.944	0.963
	SMA	0.023	0.2	0.616	0.974	1	1	1
	MLOGIT	0.454	0.781	0.9599†	1.0†	1	1.0†	1.0†
	MLOGITC	0.057	0.257	0.661	0.9590†	0.997	0.999	0.998
5	MNWSS	0.078	0.238	0.454	0.732	0.871	0.967	0.98
	MNWSSP	0.078	0.239	0.453	0.732	0.874	0.965	0.98
	SMA	0.014	0.172	0.591	0.97	1	1	1
	MLOGITC	0.0250†	0.188	0.505	0.833	0.963	0.9940†	0.9990†
7	MNWSS	0.073	0.214	0.406	0.724	0.87	0.95	0.98
	MNWSSP	0.073	0.216	0.408	0.731	0.875	0.951	0.98
	SMA	0.022	0.141	0.505	0.956	0.997	1	1
	MLOGITC	0.019	0.114	0.292	0.6167†	0.855	0.9920†	0.9970†

†A portion of these tests failed and the results are most likely biased
MNWSS = Multinomial Weighted Sum Statistic, MNWSSP = Multinomial Weighted Sum Statistic with Permutation Test, SMA = Single Marker Analysis, MLOGIT = Multinomial Logistic Regression, MLOGITC = Multinomial Logistic Regression with Collapsing of Variants with Quasi-Complete Separation

Table A.2: Power Comparison for a Sample Size of 1000

		Lambda						
Categories	Method	0.01	0.05	0.1	0.2	0.3	0.5	0.8
3	MNWSS	0.131	0.471	0.765	0.932	0.961	0.989	0.987
	MNWSSP	0.134	0.473	0.758	0.933	0.961	0.989	0.985
	SMA	0.039	0.609	0.972	1	1	1	1
	MLOGIT	0.355	0.84	0.997	1.0†	1.0†	1.0†	1.0†
	MLOGITC	0.056	0.577	0.974	1	1	1	1
5	MNWSS	0.105	0.417	0.746	0.913	0.976	0.991	0.995
	MNWSSP	0.109	0.423	0.744	0.911	0.974	0.991	0.995
	SMA	0.046	0.563	0.969	1	1	1	1
	MLOGITC	0.053	0.557	0.951	0.999	1	1	1.0†
7	MNWSS	0.102	0.379	0.685	0.909	0.956	0.985	0.989
	MNWSSP	0.104	0.385	0.681	0.904	0.956	0.986	0.989
	SMA	0.039	0.509	0.948	1	1	1	1
	MLOGITC	0.0662†	0.4769†	0.8739†	0.9990†	1.0†	1.0†	1.0†

†A portion of these tests failed and the results are most likely biased
MNWSS = Multinomial Weighted Sum Statistic, MNWSSP = Multinomial Weighted Sum Statistic with Permutation Test, SMA = Single Marker Analysis, MLOGIT = Multinomial Logistic Regression, MLOGITC = Multinomial Logistic Regression with Collapsing of Variants with Quasi-Complete Separation

Table A.3: Power Comparison for a Sample Size of 2000

		Lambda						
Categories	Method	0.01	0.05	0.1	0.2	0.3	0.5	0.8
3	MNWSS	0.214	0.728	0.913	0.981	0.994	0.992	0.999
	MNWSSP	0.22	0.727	0.911	0.98	0.993	0.992	0.999
	SMA	0.161	0.965	1	1	1	1	1
	MLOGIT	0.415	0.969†	0.9990†	1	1.0†	1.0†	1.0*†
	MLOGITC	0.1	0.941	1	1	1	1.0†	1
5	MNWSS	0.189	0.68	0.898	0.972	0.985	0.997	0.998
	MNWSSP	0.187	0.685	0.897	0.973	0.985	0.997	0.998
	SMA	0.121	0.962	1	1	1	1	1
	MLOGITC	0.114	0.938	1	1	1	1	1.0†
7	MNWSS	0.149	0.653	0.886	0.974	0.988	0.994	0.997
	MNWSSP	0.153	0.662	0.886	0.972	0.989	0.994	0.998
	SMA	0.115	0.935	1	1	1	1	1
	MLOGITC	0.1247†	0.9057†	0.9990†	1.0†	1.0†	1.0†	1.0†

†A portion of these tests failed and the results are most likely biased
MNWSS = Multinomial Weighted Sum Statistic, MNWSSP = Multinomial Weighted Sum Statistic with Permutation Test, SMA = Single Marker Analysis, MLOGIT = Multinomial Logistic Regression, MLOGITC = Multinomial Logistic Regression with Collapsing of Variants with Quasi-Complete Separation

Table A.4: Power versus Sample Size for a Heritability of 5%

		Sample Size					
Categories	Method	500	1000	1500	2000	2500	3000
3	MNWSS	0.271	0.471	0.613	0.728	0.796	0.849
	MNWSSP	0.277	0.473	0.612	0.727	0.799	0.848
	SMA	0.200	0.609	0.934	0.965	0.996	0.999
5	MNWSS	0.238	0.417	0.599	0.680	0.777	0.852
	MNWSSP	0.239	0.423	0.604	0.685	0.780	0.850
	SMA	0.172	0.563	0.865	0.962	0.990	0.998
7	MNWSS	0.214	0.379	0.532	0.653	0.770	0.800
	MNWSSP	0.216	0.385	0.526	0.662	0.768	0.801
	SMA	0.141	0.509	0.804	0.935	0.980	0.998

Table A.5: Power versus Sample Size for a Heritability of 10%

		Sample Size				
Categories	Method	500	750	1000	1500	2000
3	MNWSS	0.510	0.632	0.765	0.861	0.913
	MNWSSP	0.506	0.633	0.758	0.867	0.911
	SMA	0.616	0.879	0.927	0.998	1.000
5	MNWSS	0.454	0.593	0.746	0.852	0.898
	MNWSSP	0.453	0.594	0.744	0.851	0.897
	SMA	0.591	0.859	0.969	1.000	1.000
7	MNWSS	0.406	0.540	0.685	0.807	0.886
	MNWSSP	0.408	0.548	0.681	0.804	0.886
	SMA	0.505	0.806	0.948	0.999	1.000

Table A.6: Power versus Sample Size for a Heritability of 20%

		Sample Size			
Categories	Method	500	750	1000	2000
3	MNWSS	0.735	0.868	0.932	0.981
	MNWSSP	0.739	0.866	0.933	0.980
	SMA	0.974	0.997	1.000	1.000
5	MNWSS	0.732	0.851	0.913	0.972
	MNWSSP	0.732	0.846	0.911	0.973
	SMA	0.970	0.999	1.000	1.000
7	MNWSS	0.724	0.827	0.909	0.974
	MNWSSP	0.731	0.827	0.904	0.972
	SMA	0.956	1.000	1.000	1.000

VITA

Janae Elizabeth Nicholson

Candidate for the Degree of

Doctor of Philosophy

Thesis: METHODS OF ASSOCIATION FOR GENOME DATA WITH RARE VARIANTS AND A MULTINOMIAL RESPONSE

Major Field: Statistics

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Statistics at Oklahoma State University, Stillwater, Oklahoma in May, 2013.

Completed the requirements for the Master of Science in Statistics at Oklahoma State University, Stillwater, Oklahoma in 2005.

Completed the requirements for the Bachelor of Science in Mathematics at University of Kansas, Lawrence, Kansas in 2000.

Experience:

Research Statistician at MPSI Systems, Tulsa, Oklahoma from 2000 to 2003.

Biostatistics Intern at Quintiles, Kansas City, Missouri during the summer of 2004.

Research and Teaching Assistant for Oklahoma State University, Stillwater, Oklahoma from 2003 to 2011.

Professional Memberships: American Statistical Association

Name: Janae Nicholson Date of Degree: May, 2013

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: METHODS OF ASSOCIATION FOR GENOME DATA WITH RARE VARIANTS AND A MULTINOMIAL RESPONSE

Pages in Study: 97 Candidate for the Degree of Doctor of Philosophy

Major Field: Statistics

Scope and Method of Study:

A rare variant is a Single Nucleotide Polymorphism (SNP) with a minor allele frequency (MAF) of 5% or less. Approximately 60% of human SNPs are rare variants**.** New rapid genotyping technologies now make it possible to efficiently survey these rare variants. Many new statistical methods are being developed to analyze the associations between rare variants and phenotypes. Current methods have focused on dichotomous phenotypes such as case/control status or quantitative phenotypes such as weight or cholesterol level. Rare variant association methods for multinomial phenotypes, or categorical outcomes with more than two possibilities, have not been adequately addressed. The purpose of this study is to develop new methods of rare variant association analysis for a multinomial phenotype. Several new methods are proposed and evaluated using simulations.

Findings and Conclusions:

Simulations showed that two of the proposed methods are viable for rare variant association analysis with multinomial phenotypes. These methods have the correct or conservative Type I error rate and reasonable power for large samples with a moderate heritability. The viable methods are applied to resequencing data from the Dallas Heart Study. One of the methods detected an association between a categorized plasma triglyceride level and the ANGPTL3 and ANGPTL4 genes.

ADVISER'S APPROVAL: Dr. Lan Zhu