7-17-2009

An Application of Armitage Trend Test to Genome-wide Association Studies

Nigel A. Scott

AN APPLICATION OF ARMITAGE TREND TEST TO GENOME-WIDE ASSOCIATION STUDIES

by

NIGEL A. SCOTT

Under the Direction of Dr. Yixin Fang

ABSTRACT

Genome-wide Association (GWA) studies have become a widely used method for analyzing genetic data. It is useful in detecting associations that may exist between particular alleles and diseases of interest. This thesis investigates the dataset provided from problem 1 of the Genetic Analysis Workshop 16 (GAW 16). The dataset consists of GWA data from the North American Rheumatoid Arthritis Consortium (NARAC). The thesis attempts to determine a set of single nucleotide polymorphisms (SNP) that are associated significantly with rheumatoid arthritis. Moreover, this thesis also attempts to address the question of whether the one-sided alternative hypothesis that the minor allele is positively associated with the disease or the two-sided alternative hypothesis that the genotypes at a locus are associated with the disease is appropriate, or put another way, the question of whether examining both alternative hypotheses yield more information.

INDEX WORDS: False discovery rate, Genetic analysis workshop, Rheumatoid arthritis, Sequentially rejective bonferroni, Single nucleotide polymorphisms

AN APPLICATION OF ARMITAGE TREND TEST TO GENOME-WIDE ASSOCIATION STUDIES

by

NIGEL A. SCOTT

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2009

AN APPLICATION OF ARMITAGE TREND TEST TO GENOME-WIDE ASSOCIATION STUDIES

by

NIGEL A. SCOTT

Committee Chair: Dr. Yixin Fang

Committee: Dr. Yu-sheng Hsu
Dr. Yuanhui Xiao

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
August 2009

DEDICATION

This thesis is dedicated to the woman I love, my wife Carmen, and my wonderful son Nicholas.

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Dr. Yixin Fang, for all of his guidance and support. I would like to thank all my professors in the Department for all their encouragement and guidance over the years. Especially, I would like to thank the committee members, Dr. Hsu and Dr. Xiao, for their advice and input in completing this thesis. Special thanks to Xin Huang for his help with programming in R and Latex. Finally, I would like to thank my family and friends for all of their support and understanding.

The data set, from North American Rheumatoid Arthritis Consortium (NARAC), is proved by Genetic Analysis Workshop (GAW) 16. And it was originally analyzed by Plenge *et al.* (2007). I would like to thank both GAW 16 who provided the dataset and the research group who generated the dataset. Following the rule, I will not publish any result related to GAW 16 data in any journal until we get the permission.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Genome-wide Association (GWA) Studies

There has been a long history, dating back to 1985, behind decoding the human genome and its potential uses. The Human Genome Project which spanned over 20 years, essentially sequenced the human genome and allowed researchers to study what all humans have in common (Roberts, 2001). In addition to the information gained from the Human Genome Project, the results released by the International HapMap Consortium (2003) further advanced our understanding of the human DNA by studying the variability of the DNA in several world populations.

There are many exciting successes in this area. For example, Risch and Meringkan (1996) and WTCCC (2007) list several diseases such as Huntington's disease and Alzheimer's disease in which researchers have found genetic basis. However, Risch and Meringkan (1996) also indicated that there had not been many successes in more complex diseases, due to the modest association of some genes to these diseases. They also suggested that the method of linkage analysis used for detection has low power in finding linkages between diseases and genes. Instead, they showed that GWA studies have much higher power and can detect associations for the more complex diseases.

The detection of strategically selected markers called Single Nucleotide Polymorphisms (SNPs), play a vital role in GWA studies. SNPs are essentially variations in the DNA sequence of chromosomes. According to International HapMap Consortium (2003), a section of DNA in a chromosome region will have a sequence of bases consisting of A, T, C, or G. Whenever these sequences vary across the regions, they are referred to as SNPs. On average, these variations in the chromosome occur at a rate of one variant per 1,000 bases as reported by Wang *et al.* (1998). It was also estimated that about 10 million SNPs account for about 90% of the variation in the human population, and the other 10% results from rare variants in the population (Kruglyak and Nickerson, 2001). However, according to Gabriel *et al.*

(2002) and Carlson (2003) most of these variations can be studied by genotyping 200,000 to 1,000,000 tag SNPs across the human genome.

In order to conduct a GWA study, researchers take samples of DNA from a group of people with the disease of interest, and samples of DNA from a control group without the disease. These samples are then tested for the presence of SNPs that can highlight genetic abnormalities. If it is found that these abnormalities are significantly present in individuals with the disease relative to those without the disease, then those mutations can be considered as being associated with an increased risk of the disease.

GWA studies have already proven to have some successes. As noted in WTCCC (2007), these studies have been able to find significant associations between specific genetic mutations and certain diseases such as type-II diabetes, Parkinson's disease and Crohn's disease. However, other more complex diseases, such as rheumatoid arthritis, still hold a challenge for researchers.

1.2 NARAC Dataset

The dataset in problem 1 of the Genetic Analysis Workshop 16 (GAW 16) is part of the dataset used by Plenge *et al.* (2007), in which a GWA study was performed on cases, from North America and Sweden, with anti-CCP positive rheumatoid arthritis. The GAW 16 dataset is based on the data from North America. After removing duplicated and contaminated samples, this dataset consists of 868 cases and 1,194 controls and contains 545,080 SNPs. Actually, the initial North American cases were taken from several rheumatology clinics that make up the NARAC. The cases were randomly drawn from these clinics and the patients were self-identified as having white ancestry and they are matched with control subjects according to similar self-identified ethnic background.

1.3 Challenges for Statistical Methods

GWA studies have been made possible due to the improvement of technology. For example, the WTCCC used the Affymetrix GeneChip 500K Mapping Array Set that allowed it the ability to study 7 diseases in approximately 2,000 cases and 3,000 controls. As pointed

out by Gabriel *et al.* (2002), the variations in the human population can be studied by genotyping 200,000 to 1,000,000 tag SNPs across the genome. While the technology has made it possible to study associations between diseases and SNPs, it has highlighted some statistical challenges in terms of analyzing such large datasets that has plagued the statistics for years.

One issue is the curse of high dimensionality (Hastie, Tibshirani and Friedman, 2009). This problem occurs when the number of variables are much larger than the sample size. Liang and Kelemen (2008) classified the statistical methods used to address high dimensionality problems into three groups: filtering methods, wrapping methods and embedded methods. The filtering methods are the most popular, because they are convenient and fast.

When filtering methods are used to address high dimensionality problems, the next big problem created is the problem of multiple testing. Hypothesis testing is used in GWA studies to determine which SNPs are most significantly associated with the disease of interest. Each SNP that is analyzed constitutes one hypothesis test. In traditional hypothesis testing, the significant level is often set at 5%. However, as shown in the table below as the number of SNPs tested increases, the number of SNPs falsely claimed to be significant increases, provided that all the SNPs are non-significant.

Table 1.1: Increase in False Positive ($\alpha = 5\%$)

Number of SNPs Tested	False Positive Incidence
100	5
10,000	500
500,000	25,000

Note: Assume that all the SNPs are non-significant

Tons of methods have been proposed to combat the problem of multiple testing, as discussed by Hastie *et al.* (2009). According to Lee (2004), the Sequentially Rejective Bonferroni (SRB) method introduced by Holm (1979), uses an application of Bonferroni correction to ordered p-values. The False Discovery Rate (FDR) method proposed by Benjamini and Hochberg (1995) is similar to the SRB method in that it uses the ordering of p-values. The

FDR method adjusts these p-values as a proportion of the rank of the p-value to the total number of hypothesis tests. These two methods are discussed further in this thesis.

1.4 Armitage Trend Test

One of the disadvantages of the case-control GWA studies is that they are prone to a number of biases including population stratification, as pointed out by Pearson and Manolio (2008). Despite the debate on the importance of considering confounding due to population stratification in GWA studies using case-control designs, Thomas and Witte (2002) and Wacholder *et al.* (2002), the Armitage's trend tests can correct for population stratification to some extent, as suggested by Armitage (1955), Sasieni (1997) and Schaid and Jacobsen (1999). Some other methods were also developed for the same purpose, based on the Armitage's trend test, such as the genomic control approach discussed by Devlin and Roeder (1999) and Reich and Goldstein (2001).

However, there is still a question as to whether the one-sided or the two-sided alternative hypothesis is appropriate, or put another way, whether or not examining both the one-sided and the two-sided alternative hypotheses can give more information. The dataset for problem 1 of GAW 16 provides us with a chance to address this question, because it is a part of a combined sample from the NARAC and the Swedish Epidemiological Investigation of Rheumatoid Arthritis (EIRA). The results from the combined sample is used as a reference.

1.5 Organization of Thesis

The remainder of this thesis is organized as follows. Chapter two provides an overview of the three classical methods, difference of proportions, relative risk and odds ratio, used to analyze case-control data. Chapter three presents the theory behind the Armitage's trend test for both the one-sided and the two-sided alternative hypotheses as well as the motivation for its use in this thesis. Chapter four discusses the SRB method and the FDR method, and their use in finding a threshold to determine the significant SNPs. Chapter five compares the results of the one-sided and the two-sided alternatives when using the Armitage's trend test and the three classical methods. Chapter six presents some conclusions and discussion.

Chapter 2

THREE CLASSICAL METHODS

2.1 Notation

The response variable in a GWA study is the status of the disease of interest, either diseased or non-diseased. The explanatory variable at a SNP is the genotype: AA, Aa or aa. The data at a SNP are shown in a 2×3 contingency table, as displayed in Table 2.1. Throughout the thesis, denote the major allele as "A", and the minor allele as "a". Here the minor allele is the one with smaller frequency.

Table 2.1: Genotype Distribution at a SNP

	AA	Aa	aa	Total
Case	n_{10}	n_{11}	n_{12}	N_1
Control	n_{00}	n_{01}	n_{02}	N_0
Total	N_{+0}	N_{+1}	N_{+2}	N

In Table 2.1, N_1 and N_0 denote the numbers of subjects in the case group and the control group respectively, and N_{+0}, N_{+1} and N_{+2} denote the numbers of the subjects with genotypes AA, Aa and aa respectively. Let N denote the total numbers of subjects in the study. Let n_{10}, n_{11} and n_{12} denote the numbers of subjects with genotypes AA, Aa and aa in the case group and n_{00}, n_{01} and n_{02} the corresponding numbers in the control group.

At any SNP, each subject has two alleles, a major allele and a minor allele. By counting the frequency of the minor allele "a" and the major allele "A", we can convert Table 2.1 to a 2×2 contingency table, as displayed in Table 2.2.

Table 2.2: Allele Distribution at a SNP

	"a"	"A"	Total
Case	$\tilde{n}_{11} = 2n_{12} + n_{11}$	$\tilde{n}_{10} = 2n_{10} + n_{11}$	$\widetilde{N}_1 = 2N_1$
Control	$\tilde{n}_{01} = 2n_{02} + n_{01}$	$\tilde{n}_{00} = 2n_{00} + n_{01}$	$\widetilde{N}_0 = 2N_0$
Total	$\widetilde{N}_{+1} = 2N_{+2} + N_{+1}$	$\widetilde{N}_{+0} = 2N_{+0} + N_{+1}$	$\widetilde{N} = 2N$

In Table 2.2, $\widetilde{N}_1$ and $\widetilde{N}_0$ denote the numbers of alleles in the case group and the control group respectively, and $\widetilde{N}_{+1}$ and $\widetilde{N}_{+0}$ denote the numbers of the minor and major alleles in the dataset. Let $\widetilde{N}$ denote the total number of alleles in the study. Let $\tilde{n}_{11}$ and $\tilde{n}_{10}$ denote the numbers of minor and major alleles in the case group and $\tilde{n}_{01}$ and $\tilde{n}_{00}$ the numbers of minor and major alleles in the control group. Let P_1 denote the population proportion of the minor allele "a" in the case group, and P_0 denote the population proportion of the minor allele "a" in the control group. Define the sample proportion of the minor allele in the case group as $\hat{p}_1 = \widetilde{n}_{11}/\widetilde{N}_1$ and the sample proportion of the minor allele in the control group as $\hat{p}_0 = \widetilde{n}_{01}/\widetilde{N}_0$. Agresti, (2007) describes three classical methods, difference of proportions, relative risk and odds ratio to analyze 2×2 tables, such as Table 2.2.

2.2 Difference of Proportions

Because $D = P_1 - P_0$ compares the population proportion of the minor alleles in the case group with that in the control group, if the difference of proportions is zero, then there is no association between the SNP and the risk of disease. If $P_1 > P_0$, then there exists a positive association between the minor allele and the risk of disease.

The difference of sample proportions $\hat{d} = \hat{p}_1 - \hat{p}_0$ estimates D. Let $\hat{p} = \widetilde{N}_{+1}/\widetilde{N}$ be the pooled sample proportion of the minor allele. Under the null hypothesis that there is no association between the SNP and the risk of disease, the estimated standard error of $\hat{d}$ is,

$$SE(\hat{d}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{\widetilde{N}}}. \tag{2.1}$$

To test the H_0: $P_1 = P_0$, the test statistic is

$$Z_1 = \frac{\hat{p}_1 - \hat{p}_0}{SE(\hat{d})}, \tag{2.2}$$

which asymptotically follows the standard normal distribution under the H_0.

For the one-sided alternative hypotheses that the minor allele is positively associated with the disease of interest, that is $P_1 > P_0$, the p-value is $\text{Prob}(N(0,1) \geq Z_1^{obs})$. For the two-sided alternative hypotheses that the genotypes are associated with the disease of interest, that is $P_1 \neq P_0$, the test based on Z_1 is equivalent to the test based on Z_1^2, which asymptotically follows a chi-squared distribution with one degree of freedom, χ_1^2. The p-value is $\text{Prob}(\chi_1^2 \geq Z_1^{2obs})$.

2.3 Relative Risk

The method based on difference of proportions is not effective when the proportions are near 0 (Agresti, 2007). The method of relative risk can solve this problem. The relative risk is defined as $\rho = P_1/P_0$, which can be estimated by the sample relative risk of $\hat{\rho} = \hat{p_1}/\hat{p_0}$. If $\rho = 1$, then there is no association between the SNP and the risk of disease. If $\rho > 1$, then there exists a positive association between the minor allele and the risk of disease.

To avoid skewness in the asymptotic distribution of $\hat{\rho}$, the $\log \hat{\rho}$ is considered. The estimated standard error of $\log \hat{\rho}$ is (Agresti, 2007)

$$SE(\log \hat{\rho}) = \sqrt{\frac{(1-\hat{p_1})}{\widetilde{N}_1 \hat{p_1}} + \frac{(1-\hat{p_0})}{\widetilde{N}_0 \hat{p_0}}}. \tag{2.3}$$

To test the H_0: $P_1 = P_0$ the test statistic is

$$Z_2 = \frac{\log \frac{\hat{p_1}}{\hat{p_0}}}{SE(\log \hat{\rho})}, \tag{2.4}$$

which asymptotically follows the standard normal distribution under the H_0.

For the one-sided alternative hypotheses that the minor allele is positively associated with the disease of interest, that is $P_1 > P_0$, the p-value is $\text{Prob}(N(0,1) \geq Z_2^{obs})$. For the two-sided alternative that the genotypes are associated with the disease of interest, that is $P_1 \neq P_0$, the test based on Z_2 is equivalent to the test based on Z_2^2, which asymptotically follows a chi-squared distribution with one degree of freedom, χ_1^2. The p-value is $\text{Prob}(\chi_1^2 \geq Z_2^{2obs})$.

2.4 Odds Ratio

Another method used is the odds ratio test statistic which is based on the ratio of the odds of the minor allele in the case group with that in the control group. The odds of the minor allele in the case group is defined as $odds_1 = P_1/(1 - P_1)$ and the odds of the minor allele in the control group is defined as $odds_2 = P_0/(1 - P_0)$. The odds ratio can be defined as $\theta = \frac{P_1}{1-P_1}/\frac{P_0}{1-P_0}$, which can be estimated by the sample odds ratio $\hat{\theta} = \frac{\hat{p}_1}{1-\hat{p}_1}/\frac{\hat{p}_0}{1-\hat{p}_0}$. If $\theta = 1$, then there is no association between the SNP and the risk of the disease. If $\theta > 1$, then there exists a positive association between the minor allele and risk of the disease.

To avoid skewness in asymptotic distribution of $\hat{\theta}$, the $\log\hat{\theta}$ is considered. The estimated standard error of the $\log\hat{\theta}$ is (Agresti, 2007)

$$SE(\log\hat{\theta}) = \sqrt{\frac{1}{\widetilde{n}_{11}} + \frac{1}{\widetilde{n}_{10}} + \frac{1}{\widetilde{n}_{01}} + \frac{1}{\widetilde{n}_{00}}}. \tag{2.5}$$

To test the H_0: $P_1 = P_0$ the test statistic is,

$$Z_3 = \frac{\log\hat{\theta}}{SE(\log\hat{\theta})}, \tag{2.6}$$

which asymptotically follows the standard normal distribution under the H_0.

For the one-sided alternative hypotheses that the minor allele is positively associated with the disease of interest, that is $P_1 > P_0$, the p-value is $\text{Prob}(N(0,1) \geq Z_3^{obs})$. For the two-sided alternative hypotheses that the genotypes are associated with the disease of interest, that is $P_1 \neq P_0$, the test based on Z_3 is equivalent to the test based on Z_3^2, which asymptotically follows a chi-squared distribution with one degree of freedom, χ_1^2. The p-value is $\text{Prob}(\chi_1^2 \geq Z_3^{2obs})$.

Chapter 3

ARMITAGE TREND TEST

3.1 Population Stratification

Case-control designs are useful in analyzing GWA studies to answer the question of which SNPs are most significantly associated with an increase risk of disease. However, the main drawback of case-control studies is that they are prone to population stratification.

Population stratification is a form of confounding that can occur specifically in genetics studies, such as GWA studies. This type of confounding occurs when two or more subgroups of the population under study display a large variation in the allele frequencies of the gene being investigated. These subgroups also differ from the rest of the population in the risk of disease. There are many causes of population stratification. One possible cause is migration. The migrated group of people maybe susceptible to a particular disease and has become part of a larger population (Thomas and Witte, 2002). In this situation, a case-control study will detect a false association between the population and the disease of interest, that is really being caused by the association between the migrated subgroup and the disease of interest.

There are several examples of population stratification that highlight the seriousness of the problem, and the adverse effects it can have on the results of a study, Thomas and Witte (2002). One of those examples involves a genetic association study that suggested an inverse association between variants in the immunoglobulin haplotype $Gm^{3,5,13,14}$ and non-insulin-dependent diabetes mellitus in members of the Gila River Indian Community. After further investigation of this association, it was determined that the inverse association was due to the Caucasian heritage among the community. In fact, there was an association between heritage and $Gm^{3,5,13,14}$ and between Caucasian heritage and risk of the specific diabetes mellitus. Once the data was corrected for heritage, the results no longer reflected the inverse association.

Figure 3.1 was presented in Wacholder *et al.* (2002) to explain population stratification. Case-control studies can consist of participants with several unknown backgrounds, such as

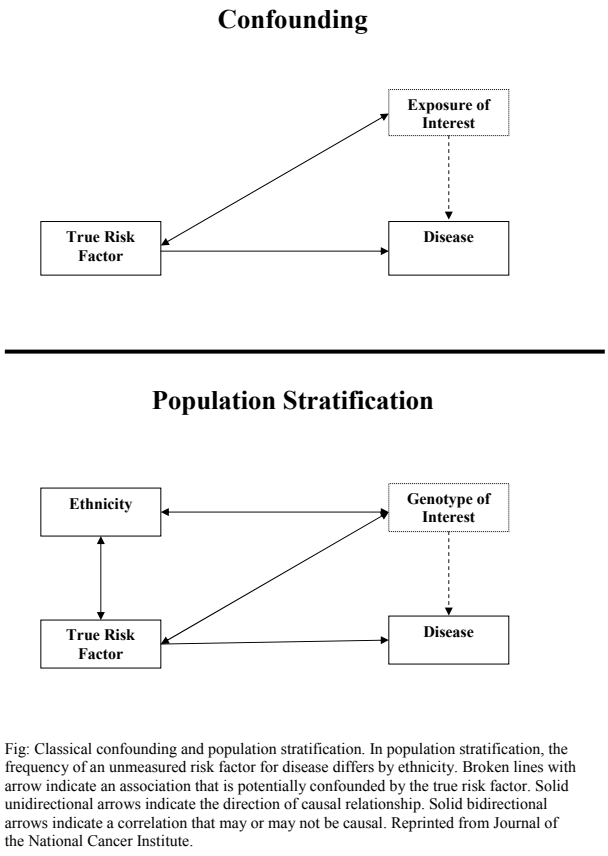

Fig: Classical confounding and population stratification. In population stratification, the frequency of an unmeasured risk factor for disease differs by ethnicity. Broken lines with arrow indicate an association that is potentially confounded by the true risk factor. Solid unidirectional arrows indicate the direction of causal relationship. Solid bidirectional arrows indicate a correlation that may or may not be causal. Reprinted from Journal of the National Cancer Institute.

Figure 3.1: Confounding and Population Stratification

ethnicity. When these participants are pooled together in the study, an association may be observed between a genetic variant and the disease of interest. However, if the participants are broken up based on their background, such as ethnicity, then the observed association between a genetic variant and the disease of interest may no longer exist.

3.2 Armitage Trend Test

This disadvantage is overcomed, to some extent, by applying the Armitage's trend test, as suggested by Armitage (1955), Sasieni (1997), and Schaid and Jacobsen (1999). Sasieni (1997) discusses the use of three different approaches for the analysis of genetic case-control

data. The first approach differentiates the subjects into three categories, those in which the allele of interest is recessive, those in which the allele of interest is co-dominant, an heterozyous allele, and those in which the allele of interest is dominant, an homozygous allele. This approach uses a standard 2×3 contingency table, as in Table 2.1. The second approach is to combine the heterozygous and homozygous separate from the subjects without the allele of interest. The third approach is to consider the allele frequency for the cases and controls, which in effect doubles the sample size. These last two approaches use a 2×2 contingency table. Sasieni (1997) concluded that an analysis based on treating alleles as individual entities is valid only when the Hardy-Weinberg equilibrium holds. As a result, he warns against this allelic based analysis and recommends that genetic case-control data be analyzed using genotype approach. Below is the generic contingency table from Sasieni (1997), denoting the major allele as "A", and the minor allele as "a". Table 3.1 is identical to Table 2.1, except for the addition of the three scoring systems in Table 3.1.

Table 3.1: Scores for Armitage's Trend Test

	AA	Aa	aa	Total
Case	n_{10}	n_{11}	n_{12}	N_1
Control	n_{00}	n_{01}	n_{02}	N_0
Total	N_{+0}	N_{+1}	N_{+2}	N
Score	x_0	x_1	x_2	

In order to analyze such 2×3 contingency tables, a test statistic based on scores was developed to determine whether there is a linear trend in proportions (Armitage, 1955; Sasieni, 1997; and Schaid and Jacobsen, 1999). Let x_j denote the score associated with the j^{th} column of the contingency table, j=0, 1, 2. The Armitage's trend test statistic is

$$X_A^2 = \frac{N(N\Sigma n_{1j}x_j - N_1\Sigma N_{+j}x_j)^2}{N_1N_0[N\Sigma N_{+j}x_j^2 - (\Sigma N_{+j}x_j)^2]}. \tag{3.1}$$

This statistics has an approximate chi-square distribution with one degree of freedom under the null hypothesis. When considering the two-sided alternative hypothesis that the

genotypes at a SNP are associated with the disease of interest, this statistic can be employed. An important part of this statistic is the choice of the scoring system. According Armitage (1955), while the score does not affect the validity of the test, it does affect the power of the test. If there is no prior information or known relationship between the columns, it can be difficult to determine a scoring system. However, in analyzing the contingency table above, there are three common scoring systems used. The scoring system can be chosen as one of the following: (1) co-dominant score: $x_0 = 0$, $x_1 = 1$, and $x_2 = 2$; (2) dominant score: $x_0 = 0$, $x_1 = 1$ and $x_2 = 1$; (3) recessive score: $x_0 = 0$, $x_1 = 0$, and $x_2 = 1$. Since the disease can be consider rare, the minor allele can be assumed to be of interest in terms of increased risk of the occurrence of the disease. So the system is in favor of this minor allele.

3.3 One-Sided and Two-Sided Alternative Hypotheses

The Armitage's trend test statistic can also be adjusted for the one-sided alternative hypotheses. There are two one-sided alternatives, (i) the alternative that the minor allele is positively associated with the disease of interest; (ii) the alternative that the major allele is positively associated with the disease of interest. As mentioned above, the disease can be considered to be rare, so that the first alternative hypothesis is more logical. However, if no prior information is known, both alternative hypotheses would have to be considered. So Armitage's trend test for the one sided alternative hypotheses is,

$$Z_A = \frac{\sqrt{N}(N\Sigma n_{1j}x_j - N_1\Sigma N_{+j}x_j)}{\sqrt{N_1N_0[N\Sigma N_{+j}x_j^2 - (\Sigma N_{+j}x_j)^2]}}. \tag{3.2}$$

Under the null hypothesis, this statistic is approximately distributed with a $N(0,1)$. The same scoring systems described above can be used.

If the one-sided alternative hypotheses (i) is considered, the p-value = Prob$(N(0,1) \geq Z_A^{obs})$; if the one-sided alternative hypotheses (ii) is considered, the p-value = Prob$(N(0,1) \leq Z_A^{obs})$; if two-sided alternative hypotheses is considered, the p-value = Prob$(\chi_1^2 \geq X_A^{obs})$.

Chapter 4

FINDING THRESHOLDS

4.1 Notation

A statistical challenge in dealing with the data from a GWA study is that the number of variables is much larger than the sample size. The problem is so-called multiple comparison. In the GWA study, each SNP to be analyzed constitutes one hypothesis test. When testing hundred thousands of hypotheses simultaneously, it is important to determine a threshold for selecting the most significant SNPs. While tons of methods have been proposed to deal with this problem, this thesis reviews and compares two approaches: the SRB method and the FDR method.

Some notations for these two methods are defined as follows. Assume that there are n null hypotheses $H_1, H_2, H_3, \cdots,$ and H_n. Their corresponding alternative hypotheses are denoted by $K_1, K_2, K_3, \cdots,$ and K_n. To test for these null hypotheses, test statistics are $Y_1, Y_2, Y_3, \cdots,$ and Y_n. Let the n critical regions be $C_1, C_2, C_3, \cdots,$ and C_n. Particularly, in this thesis, the test for each null hypotheses could be one of the three classical methods and the Armitage trend tests with the three difference scores.

4.2 Sequentially Rejective Bonferroni Method

One approach in dealing with the multiple testing problem is the Sequentially Rejective Bonferroni (SRB) method. This was proposed by Holm (1979). The method is based on the Bonferroni test and requires the type-I error to be as small as possible. Philosophically, for each of these n tests, the probability of committing a type-I error is less than or equal to a small predetermined value α. Rejecting a null hypothesis constitutes making a discovery, that the alternative hypotheses is statistically significant.

Now let the corresponding p-values generated from the test statistics, $Y_1, Y_2, Y_3, \cdots,$ and Y_n, be $P_1, P_2, P_3, \cdots,$ and P_n, where $P_k = \alpha_k(Y_k)$, where k=1, 2, $\cdots$, n. When these p-values are ordered $P^{(1)} \leq P^{(2)} \leq P^{(3)} \leq \cdots \leq P^{(n)}$, along with their corresponding hypotheses, $H^{(1)} \leq H^{(2)} \leq H^{(3)} \leq \cdots \leq H^{(n)}$, the most significant ones would have the smallest p-values.

As shown in Table 1.1, the number of false positives increases as the number of null hypotheses tested increases. The SRB method attempts to solve this multiple testing problem by adjusting the significant level α, for each hypotheses tested, before comparing it with the p-values. Specifically, these p-values are compared to corresponding levels denoted by

$$\frac{\alpha}{n}, \frac{\alpha}{n-1}, \cdots, \frac{\alpha}{1}. \tag{4.1}$$

The hypotheses are rejected until no other rejections are possible. Since the most important hypotheses would have the smallest p-values, they are compared with the smallest level of $\alpha/(n-i+1)$, where $i = 1, 2, 3, \cdots, n$. The least important hypotheses are compared with increasingly larger levels.

4.3 False Discovery Rate Controlling Method

Another approach to combating the multiple testing problem that arises in GWA studies comes from an idea proposed by Benjamini and Hochberg (1994), referred to as False Discovery Rate (FDR). They noted that the classical approaches despite their uses in industries are less likely used in genetic.

Many multiple testing procedures, such as the SRB method, are based on controlling the type-I error. The FDR method takes a philosophically different approach, in that it takes into account the number of hypotheses that are falsely rejected, that is the number of false discoveries. As a result, the FDR can be defined as the expected proportion of errors among the reject hypotheses.

Table 4.1: Number of Errors Committed When Testing n Null Hypotheses

	Declared non-significant	Declared significant	Total
True Null Hypotheses	U	V	n_0
Non-true Null Hypotheses	T	S	$n - n_0$
	$n - R$	R	n

The FDR method considers testing simultaneously n null hypotheses of which n_0 are true. The situation is summarized in Table 4.1. Benjamini and Hochberg (1994) suggested that

the unknown random variable $Q = V/R$, can be used to represent the proportion of errors committed by falsely rejecting the null hypotheses. This is the proportion of the rejected null hypotheses which are erroneously rejected. Let the FDR be represented by Q_c which is the expectation of Q. Then we have

$$Q_c = E(Q) = E(\frac{V}{R}). \tag{4.2}$$

The FDR method is conducted as follows. Consider testing $H_1, H_2, \cdots,$ and H_n based on the corresponding p-values $P_1, P_2, \cdots,$ and P_n. When these p-values are ordered $P^{(1)} \leq P^{(2)} \leq P^{(3)} \leq \cdots \leq P^{(n)}$, along with their corresponding hypotheses, $H^{(1)} \leq H^{(2)} \leq H^{(3)} \leq \cdots \leq H^{(n)}$, the most significant ones would have the smallest p-values. Let $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$. Define the following Bonferroni-type multiple testing procedure: Let k be the largest i for which

$$P_{(i)} \leq \frac{i}{n} q^*; \tag{4.3}$$

then reject all $H_{(i)}$, $i = 1, 2, \cdots, k$. As stated by Theorem I in Benjamini and Hochberg (1994), for independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at q^*.

4.4 Comparison of SRB method and FDR method

These methods take two philosophically different approaches to the multiple comparison problem. With these two approaches, there is some measure of tradeoff that exists between the type-I and type-II errors.

The SRB method focuses on keeping the type-I error small. This is accomplished by taking the predetermined value of α, and reducing it by a factor of $1/n, 1/(n-1), 1/(n-2), \cdots, 1/1$. These adjusted levels are then compared to the p-values to determine significance. The FDR method instead focuses on controlling the expected number of falsely rejected hypotheses, q^*. This is accomplished by taking the q^* and reducing it by a factor

of $1/n, 2/n, 3/n, \cdots, n/n$. These adjusted levels are then compared with the p-values to determine significance.

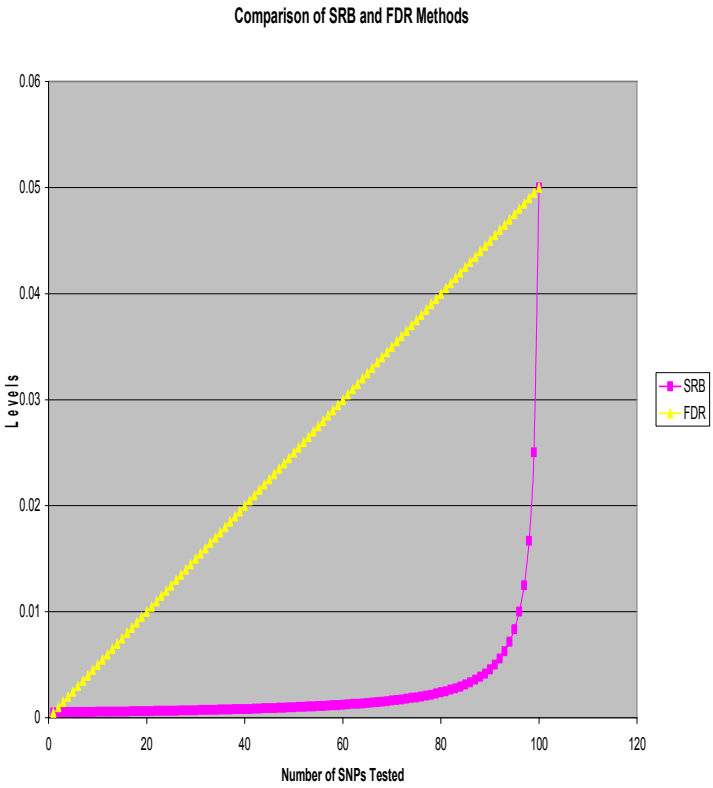

Figure 4.1: Comparison of Levels from SRB and FDR Methods
Note: Assume there are 100 SNPs to be tested. For SRB method α=0.05. For FDR method q^*=0.05.

Figure 4.1 compares these two approaches by looking at the factors that adjust α for the SRB method and q^* for the FDR method. It shows that $1/n, 1/(n-1), 1/(n-2), \cdots, 1/1 \leq 1/n, 2/n, 3/n, \cdots, n/n$ at all corresponding levels. This means that the FDR method has a larger type-I error, and hence more false positives than the SRB method.

Chapter 5

RESULTS

5.1 Thresholds for NARAC Dataset

The Bonferroni correction is the most popular method for adjusting the type-I error, α. It accomplishes this by using α/n, where n is the number of hypotheses to be tested. This adjusted value is compared with the p-values from the statistical tests. For the NARAC dataset, n = 545,080. The LOD threshold from the Bonferroni correction is $-\log(\alpha/n)$. If α=0.05, this gives a threshold of 9.17, which will be considered a strict threshold for the NARAC dataset.

The SRB and FDR methods can also be used to test hundreds of thousands of hypotheses simultaneously and indicate which SNPs are significantly associated with the disease of interest. To illustrate this, the test based on the difference of proportions and the Armitage's Trend test with the three types of scores are conducted for each SNP. The tests based on the relative risk and the odds ratio are not shown here since they yield similar results to the test based on the difference of proportions. For each test, all chromosomes are pooled together and their p-values are arranged in ascending order. The LOD values are calculated using $-\log P_{(i)}$, where $i = 1, 2, \cdots, n$ and the log is in base 10.

By using the SRB method, if $\alpha = 0.05$ then the SNPs are considered to be significant once $-\log P_{(i)} \geq -\log \alpha/(n-i+1)$. Table 5.1 illustrates this procedure for both one-sided alternative that the minor allele is positively associated with the disease of interest and the two-sided alternative that the genotypes are associated with the disease of interest using the test based on the difference of proportions. For the one-sided alternative, the first 110 SNPs in Table 5.1 are considered to be significant, since at the 111^{th} SNP, we observe for the first time that $-\log P_{(i)} < -\log \alpha/(n-i+1)$. For the two-sided alternative, the first 267 SNPs in Table 5.1 are considered to be significant, since at the 268^{th} SNP, we observe for the first time that $-\log P_{(i)} < -\log \alpha/(n-i+1)$.

Similarly, by the FDR method, if the FDR is controlled at q^*=0.05 then the SNPs are considered to be significant once $-\log P_{(i)} \geq -\log(i/n)q^*$. Table 5.2 illustrates this procedure for both one-sided alternative that the minor allele is positively associated with the disease of interest and the two-sided alternative that the genotypes are associated with the disease of interest using the test based on the difference of proportions. For the one-sided alternative, the first 305 SNPs in Table 5.2 are considered to be significant, since at the 306^{th} SNP, we observe for the first time that $-\log P_{(i)} < -\log(i/n)q^*$. For the two-sided alternative, the first 1,571 SNPs in Table 5.2 are considered to be significant, since at the $1,572^{\text{th}}$ SNP, we observe for the first time that $-\log P_{(i)} < -\log \alpha/(n-i+1)$. Other tables using the Armitage's trend tests with the three types of scores are provided in Appendix A.

Table 5.1: SRB Method Applied to All Chromosomes Using Difference of Proportions

One-Sided Alternative			Two-Sided Alternative		
i	$-\log P_{(i)}$	$-\log \frac{\alpha}{n-i+1}$	i	$-\log P_{(i)}$	$-\log \frac{\alpha}{n-i+1}$
1	15.95458977	7.037490243	1	15.95458977	7.037490243
2	15.95458977	7.037489446	2	15.95458977	7.037489446
3	15.65355977	7.037488649	3	15.95458977	7.037488649
4	15.65355977	7.037487853	4	15.95458977	7.037487853
5	15.65355977	7.037487056	5	15.95458977	7.037487056
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
106	7.371783044	7.037406576	263	7.117567861	7.037281443
107	7.280601029	7.037405779	264	7.093959244	7.037280646
108	7.251640762	7.037404982	265	7.08718124	7.037279849
109	7.102001717	7.037404185	266	7.056862889	7.037279052
110	7.073791831	7.037403388	267	7.045368703	7.037278255
111	6.925704639	7.037402591	268	7.022062247	7.037277458

5.2 Two-Sided Alternative with SRB and FDR

This thesis attempts to conduct the tests based on the NARAC dataset, and hopes to produce results similar to those in Plenge *et al.* (2007) based on the NARAC and EIRA datasets. The results from that paper identify SNPs on chromosome 9 which contain the common genetic variant at the TRAF1-C5, as being associated with the disease of interest.

Table 5.2: FDR Method Applied to All Chromosomes Using Difference of Proportions

One-Sided Alternative			Two-Sided Alternative		
i	$-\log P_{(i)}$	$-\log \frac{i}{n}q^*$	i	$-\log P_{(i)}$	$-\log \frac{i}{n}q^*$
1	15.95458977	7.041392685	1	15.95458977	7.037490243
2	15.95458977	6.740362689	2	15.95458977	6.736460247
3	15.65355977	6.56427143	3	15.95458977	6.560368988
4	15.65355977	6.439332694	4	15.95458977	6.435430252
5	15.65355977	6.342422681	5	15.95458977	6.338520239
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
301	4.574769683	4.56282619	1567	3.846768666	3.842421246
302	4.568560359	4.561385742	1568	3.846747367	3.842144185
303	4.5668634	4.559950057	1569	3.844239762	3.841867299
304	4.564884748	4.558519102	1570	3.843674598	3.841590591
305	4.560209023	4.557092846	1571	3.842915061	3.841314058
306	4.546526523	4.555671259	1572	3.839681445	3.841037701

To this end, this thesis focuses on chromosome 9 in its analysis of results. The results from the other chromosomes can be analyzed similarly, but are not reported in this thesis. Based on Tables 5.1 and 5.2 the FDR gives a treshold of about 4, while the SRB gives a threshold of about 7. The Bonferroni correction yields a strict threshold of 9.17 and the SRB is closer than the FDR to the Bonferroni correction. So a threshold of 7 is used.

Tables 5.3 and 5.4 show the SRB and FDR methods respectively. These tables are based on the two-sided alternative hypotheses that the genotypes are associated with the disease of interest. The threshold from the FDR method suggest that 100 SNPs from chromosome 9 are significant. The first 15 of these SNPs are illustrated in Table 5.4. The SRB method suggest that 12 SNPs are significant. The SNPs reported in Plenge *et al.* (2007) are marked by asterisks in both tables.

In Table 5.3, results from Armitage's trend tests, which corrects for population stratification to some extent, are shown in columns X^2_{A1}, X^2_{A2} and X^2_{A3}. These three columns correspond to the three type of scoring systems. For the SNPs with asterisks, X^2_{A1} seems to be more significant than X^2_{A2} and X^2_{A3}. This suggests that these SNPs are more likely

to be co-dominant. For the remaining SNPs, X^2_{A3} is slightly more significant than X^2_{A1}, but X^2_{A2} is not significant. This shows that these SNPs are very likely to be recessive. Table 5.4 is similarly analyzed, in that the SNPs with asterisks are co-dominant, while those without the asterisks are recessive.

Another interesting observation from Table 5.3, is that the SNP “rs10985073" does not appear in this table, even though Plenge *et al.* (2007) has it as being significant. This can be explained by the fact that the SRB method, while it has a smaller probability of committing type-I errors compared to the FDR method, it has a larger probability of committing type-II errors compared to the FDR. The SRB has a larger probability of having false negatives. This is the case with SNP “rs10985073". It does not show up as being significant, but if the results of Plenge *et al.* (2007) can be trusted, it is significant.

Most important to the goal of this thesis is that the result from the difference of proportions shown in column Z^2_1 show that more SNPs are reported to be significant than in Plenge *et al.* (2007). The other two classical methods report similar outcomes. The difference of proportions was used in Plenge *et al.* (2007) to show significance. The question is whether these additional SNPs are truly significant, or are they false positives. This question is answered partially by looking at the results from the one-sided alternative hypotheses.

5.3 One-Sided Alternative with SRB and FDR

This thesis attempts to determine whether the one-sided alternative hypothesis that the minor allele is positively associated with the disease of interest or the two-sided alternative hypothesis that the genotypes at a locus are associated with the disease interest is appropriate. Tables 5.5 and 5.6 illustrates the results of the one-sided alternative hypotheses using the SBR and FDR methods respectively. Both Tables show that fewer SNPs are significant compared to the two-sided alternative.

In Table 5.6 suggest that only 25 SNPs are significant using the FDR threshold compared to the 100 SNPs under the two-sided alternative. The six SNPs reported in Plenge *et al.* (2007) are marked with asterisks. The FDR threshold shows that 19 other SNPs are

also significant. As discussed in Chapter 4, the FDR method has larger type-I error when compared to the SBR method. The FDR has a larger probability of selecting false positives. These 19 SNPs are very likely false positives.

The results in Table 5.5 using the SRB threshold are completely consistent with the results reported in Plenge *et al.* (2007). The SRB method has smaller type-I error compared to the FDR method. It has a smaller probability of selecting SNPs that are false positives. Moreover, by considering the one-sided alternative, the number of false positive SNPs reported under the two-sided alternative is significantly reduced. For the NARAC dataset, it seems more reasonable to consider the one-sided alternative that the minor allele is positively associated with the disease of interest.

These results are made even more clear by observing Figure 5.1 that shows the graphs the LOD values of the test based on difference of proportions and the Armitage's trend test using the three types of scores. The graphs compare the two-sided and one-sided alternatives. It is observed that under the two-sided alternative more SNPs appear to be above the threshold and hence more are significant, than compared to the one-sided alternative. This further supports the result that the one-sided alternative is more appropriate. The graphs for the other chromosomes are provided in Appendix B.

Table 5.3: Chromosome 9 - Two Sided Alternative Using SRB

SNP	Z_1^2	X_{A1}^2	X_{A2}^2	X_{A3}^2
rs872863	15.65355977	14.77849851	1.514451286	15.10949173
rs12380341	11.61068906	10.01190276	0.357222945	13.45216265
rs7854383	8.848399893	8.349044334	1.426720752	8.425534747
rs11792145	8.583459392	6.530999627	0.035692369	12.24026001
*rs2900180	8.205139225	8.188440621	5.195706388	6.093046561
*rs3761847	7.905145845	7.745195252	5.923341449	5.027178064
*rs881375	7.644376052	7.630402983	4.814960586	5.711575544
*rs1953126	7.558423781	7.532747742	5.048066573	5.443400865
rs11185665	7.541426497	6.479544704	0.325826012	9.681503269
*rs10760130	7.422629646	7.296316207	6.032196468	4.398693388
rs16929545	7.222324725	6.706251792	1.010274200	7.615575785
rs7021867	7.169175803	7.145823736	3.429323902	6.060016234

Table 5.4: Chromosome 9 - Two-Sided Alternative Using FDR

SNP	Z_1^2	X_{A1}^2	X_{A2}^2	X_{A3}^2
rs872863	15.65355977	14.77849851	1.514451286	15.10949173
rs12380341	11.61068906	10.01190276	0.357222945	13.45216265
rs7854383	8.848399893	8.349044334	1.426720752	8.425534747
rs11792145	8.583459392	6.530999627	0.035692369	12.24026001
*rs2900180	8.205139225	8.188440621	5.195706388	6.093046561
*rs3761847	7.905145845	7.745195252	5.923341449	5.027178064
*rs881375	7.644376052	7.630402983	4.814960586	5.711575544
*rs1953126	7.558423781	7.532747742	5.048066573	5.443400865
rs11185665	7.541426497	6.479544704	0.325826012	9.681503269
*rs10760130	7.422629646	7.296316207	6.032196468	4.398693388
rs16929545	7.222324725	6.706251792	1.0102742	7.615575785
rs7021867	7.169175803	7.145823736	3.429323902	6.060016234
*rs10985073	6.979571033	6.871828281	5.630131282	4.190117723
rs10815605	6.911648334	5.377070055	0.108460003	9.081435534
rs2087358	6.702586771	5.676697464	0.048732253	9.267784016

Table 5.5: Chromosome 9 - One-Sided Alternative Using SRB

SNP	Z	Z_{A1}	Z_{A2}	Z_{A3}
*rs2900180	8.50616922	8.489470624	5.496736384	6.394076556
*rs3761847	8.206175845	8.046225245	6.224371444	5.32820806
*rs881375	7.945406048	7.931432977	5.115990582	6.01260554
*rs1953126	7.859453775	7.833777737	5.349096569	5.744430861
*rs10760130	7.723659641	7.597346203	6.333226464	4.699723384
*rs10985073	7.280601029	7.172858277	5.931161278	4.491147719

Table 5.6: Chromosome 9 - One-Sided Alternative Using FDR

SNP	Z	Z_{A1}	Z_{A2}	Z_{A3}
*rs2900180	8.50616922	8.489470624	5.496736384	6.394076556
*rs3761847	8.206175845	8.046225245	6.224371444	5.32820806
*rs881375	7.945406048	7.931432977	5.115990582	6.01260554
*rs1953126	7.859453775	7.833777737	5.349096569	5.744430861
*rs10760130	7.723659641	7.597346203	6.333226464	4.699723384
*rs10985073	7.280601029	7.172858277	5.931161278	4.491147719
rs10821376	6.311121125	6.226259449	4.977581247	4.187365559
rs10122120	6.092889305	5.965322782	3.553945037	4.974853751
rs1412224	5.918572721	5.667379507	3.562094393	4.548336867
rs7037866	5.849851994	5.770187608	2.642452562	5.628537838
rs3802400	5.812590713	5.734669784	2.840231366	5.354178503
rs942152	5.705178895	5.680209038	4.033059385	4.121280492
rs540124	5.559100556	5.189371179	3.943581378	3.788547028
rs9409575	5.525573516	5.458565469	2.762855576	4.880818507
rs306772	5.441733507	5.373340922	2.815296589	4.589856973
rs2578240	5.282875011	5.305390553	2.56494468	5.050914524
rs10758875	5.164580164	5.240910111	2.579385278	4.988672234
rs913588	5.062264344	5.064095565	2.628907803	4.821522642
rs2025324	4.92685557	4.767655552	2.977829096	3.936825417
rs3802401	4.779773872	4.783660502	1.846746222	4.449794008
rs965474	4.766941026	4.698495465	4.27441813	2.872064904
rs1468673	4.758838365	4.700002288	2.608394768	4.136229612
rs3897745	4.689205229	4.727279688	1.580076088	4.564260452
rs7022212	4.601264948	4.360589882	2.363891926	3.8173894
rs548348	4.564884748	4.508921901	3.704820517	3.084481137

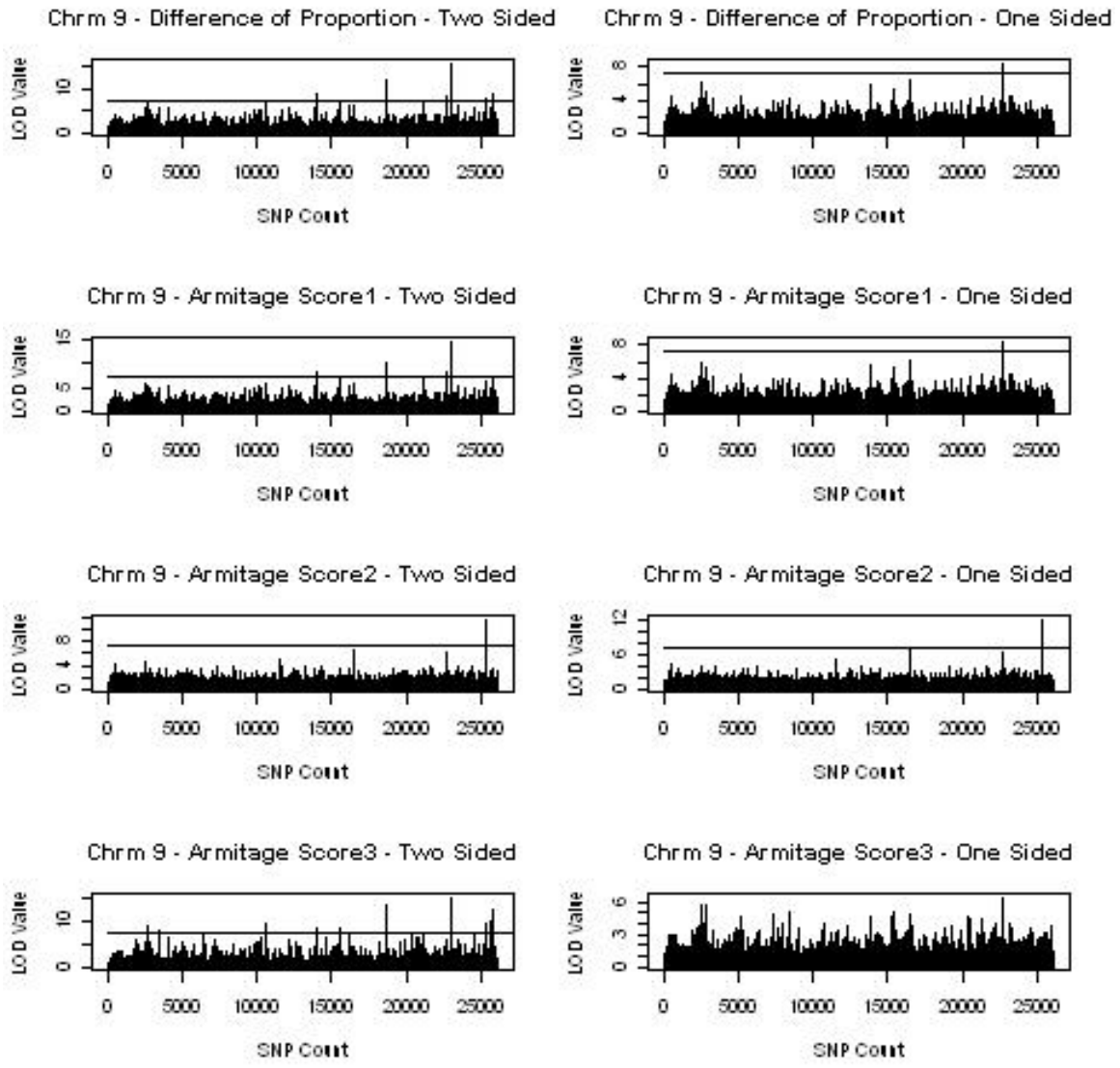

Figure 5.1: Chromosome 9 Two Sided and One Sided

Chapter 6

CONCLUSIONS

This thesis attempts to determine a set of SNPs that are significantly associated with rheumatoid arthritis using the NARAC data. In addition, this thesis also attempts to address the question of whether the one-sided alternative hypotheses that the minor allele is positively associated with the disease of interest or the two-sided alternative hypotheses that the genotypes are associated with the disease of interest is appropriate. The results from this thesis are compared with the work from Plenge *et al.* (2007), under the assumption that if results were trustful based in the combined sample (NARAC and EIRA), then similar results can be obtained based on a part of the sample (NARAC).

From the analysis in chapter 5, concentrating on the one-sided alternative tends to remove much of the noise that is present when considering the two-sided alternative. This yields the similar results as Plenge *et al.* (2007). The Armitage Trend test, which controls for population stratification to some extent, can also be used to determine which SNPs are likely co-dominant, recessive and dominant.

The SRB and FDR methods are used to deal with the problem of multiple comparisons in GWA studies. The SRB method, because of a smaller type-I error, seems to work better than the FDR method. However, a smaller type-I error indicates a larger type-II error and hence smaller power. The FDR method would have larger power compared with the SRB method. So care must be taken when deciding which method to use, since there is often a tradeoff between type-I error and power.

BIBLIOGRAPHY

[1] A. Agresti. An Introduction to Categorical Data Analysis. 2nd ed. *Hoboken, NJ: Wiley, 2007.*

[2] R. Armitage. Tests for Linear Trends in Proportions and Frequencies. *Biometrics* **11** (1955), 375-386.

[3] Y. Benjamini, and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. it J Royal Statistical Society **57** (1995), 289-300.

[4] C. Carlson *et al.* Additional SNPs and Linkage-Disequilibrium Analysis are Necessary for Whole-Genome Association Studies in Humans. *Nature Genet* **33** (2003), 518-521.

[5] F. Collins *et al.* New Goals for the U.S. Human Genome Project: 1998-2003. *Science* **282** (1998), 682-689.

[6] B. Devlin, and K. Roeder. Genomic Control for Association Studies. *Biometrics* **55** (1999), 997-1004.

[7] S. Gabriel *et al.* The Structure of Haplotype Blocks in the Human Genome. *Science* **296** (2002), 2225-2229.

[8] T. Hastie, R. Tibshirani, and J. Friedman. The Elements of Statistical Learning. 2nd ed. *New York: Springer-Verlag, 2009.*

[9] S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Statist* **6** (1979), 65-70.

[10] L. Kruglyak and D. Nickerson. Variation is the Spice of Life. *Nature Genet* **27** (2001), 234-236.

[11] M. Lee. Analysis of Microarray Gene Expression Data. *The Netherlands: Kluwer, 2004.*

[12] Y. Liang and A. Kelemen. Statistical Advances and Challenges for Analyzing Correlated High Dimensional SNP Data in Genomic Study for Complex Diseases. it Statistics Surveys **2** (2008), 43-60.

[13] J. Ott. Analysis of Human Genetic Linkage. 3rd ed. *Baltimore, MD: University Press, 1999.*

[14] T. Pearson, T. Manolio. How to Interpret a Genome-wide Association Study. *JAMA* **299** (2008), 1335-1345.

[15] R. Plenge, M. Seielstad, L. Padyukov *et al.* TRAF1-C5 as a Risk Locus for Rheumatoid Arthritis-a Genomewide Study. *N Engl J Med* **357** (2007), 1199-1209.

[16] D. Reich and D. Goldstein. Detecting Association in a Case-Control Study while Correcting for Population Stratification. *Genet Epidemiol* **20** (2001), 4-16.

[17] N. Risch and K. Meringkangas. The Future If Genetic Studies of Complex Human Disease. *Science* **273** (1996), 1516-1517.

[18] L. Roberts. Controversial From the Start. *Science* **291** (2001), 1182-1188.

[19] P. Sasieni. From Genotypes to Genes: Doubling the Sample Size. *Biometrics* **53** (1997), 1253-1261.

[20] D. Schaid and S. Jacobsen. Biased Tests of Association: Comparisons of Allele Frequencies When Departing from Hardy-Weinberg Proportions. *Am J Epidemiol* **149** (1999), 706-711.

[21] The International HapMap Consortium. The International HapMap Project. *Nature*, **426** (2003), 789-796.

[22] The Wellcome Trust Case Control Consortium. Genome-wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls. *Nature* **447** (2003), 661-683.

[23] D. Thomas, and J. Witte. Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? *Cancer Epidemiol Biomarkers Prev* **11** (2002), 505-512.

[24] S. Wacholder, N. Rothman, and N. Caporaso. Counterpoint: Bias from Population Stratification Is Not a Major Treat to the Validity of Conclusions from Epidemiological Studies of Common Polymorphisms and Cancer. *Cancer Epidemiol Biomarkers Prev* **11** (2002), 512-520.

[25] D. Wang *et al.* Large Scale Identification, Mapping, and Genotyping of Single Nucleotide Polymorphisms in the Human Genome. *Science* **280** (1998), 1077-1082.

APPENDIX A

METHODS APPLIED TO ALL CHROMOSOMES

Table A.1: SRB Method Applied to All Chromosomes Using Armitage Trend Test: Score 1

One-Sided Alternative			Two-Sided Alternative		
i	$-\log P_{(i)}$	$-\log \frac{\alpha}{n-i+1}$	i	$-\log P_{(i)}$	$-\log \frac{\alpha}{n-i+1}$
1	15.95458977	7.037490243	1	15.95458977	7.037490243
2	15.95458977	7.037489446	2	15.95458977	7.037489446
3	15.95458977	7.037488649	3	15.95458977	7.037488649
4	15.95458977	7.037487853	4	15.95458977	7.037487853
5	15.95458977	7.037487056	5	15.95458977	7.037487056
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
105	7.29933708	7.037407373	204	7.094042645	7.037328472
106	7.249840288	7.037406576	205	7.081299247	7.037327675
107	7.172858277	7.037405779	206	7.070780584	7.037326878
108	7.168687338	7.037404982	207	7.065220047	7.037326081
109	7.11051051	7.037404185	208	7.037803289	7.037325284
110	6.977747803	7.037403388	209	7.02652442	7.037324487

Table A.2: FDR Method Applied to All Chromosomes Using Armitage Trend Test: Score 1

One-Sided Alternative			Two-Sided Alternative		
i	$-\log P_{(i)}$	$-\log \frac{i}{n}q^*$	i	$-\log P_{(i)}$	$-\log \frac{i}{n}q^*$
1	15.95458977	7.037490243	1	15.95458977	7.037490243
2	15.95458977	6.736460247	2	15.95458977	6.736460247
3	15.95458977	6.560368988	3	15.95458977	6.560368988
4	15.95458977	6.435430252	4	15.95458977	6.435430252
5	15.95458977	6.338520239	5	15.95458977	6.338520239
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
281	4.594796099	4.588783923	1339	3.916876178	3.910709666
282	4.590472185	4.587241135	1340	3.914022299	3.910385445
283	4.58912837	4.585703807	1341	3.913043637	3.910061465
284	4.586435957	4.584171903	1342	3.90987487	3.909737727
285	4.58600428	4.582645383	1343	3.909431274	3.90941423
286	4.581096046	4.58112421	1344	3.908970886	3.909090974

Table A.3: SRB Method Applied to All Chromosomes Using Armitage Trend Test: Score 2

One-Sided Alternative			Two-Sided Alternative		
i	$-\log P_{(i)}$	$-\log \frac{\alpha}{n-i+1}$	i	$-\log P_{(i)}$	$-\log \frac{\alpha}{n-i+1}$
1	15.65355977	7.037490243	1	15.95458977	7.037490243
2	15.65355977	7.037489446	2	15.65355977	7.037489446
3	15.25561977	7.037488649	3	15.47746852	7.037488649
4	15.17643852	7.037487853	4	15.47746852	7.037487853
5	15.10949173	7.037487056	5	15.47746852	7.037487056
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
45	7.363966079	7.037455184	103	7.288854606	7.037408966
46	7.358192526	7.037454388	104	7.222977769	7.03740817
47	7.218073116	7.037453591	105	7.0692883	7.037407373
48	7.15520363	7.037452794	106	7.062936083	7.037406576
49	7.134668063	7.037451997	107	7.05716253	7.037405779
50	6.998258493	7.0374512	108	6.984901697	7.037404982

Table A.4: FDR Method Applied to All Chromosomes Using Armitage Trend Test: Score 2

One-Sided Alternative			Two-Sided Alternative		
i	$-\log P_{(i)}$	$-\log \frac{i}{n}q^*$	i	$-\log P_{(i)}$	$-\log \frac{i}{n}q^*$
1	15.65355977	7.037490243	1	15.95458977	7.037490243
2	15.65355977	6.736460247	2	15.65355977	6.736460247
3	15.25561977	6.560368988	3	15.47746852	6.560368988
4	15.17643852	6.435430252	4	15.47746852	6.435430252
5	15.10949173	6.338520239	5	15.47746852	6.338520239
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
193	4.762563575	4.751932934	205	4.779919099	4.725736382
194	4.762431027	4.749688513	206	4.744296879	4.723623023
195	4.761208396	4.747455632	207	4.728191006	4.721519897
196	4.752455696	4.745234172	208	4.727199986	4.719426908
197	4.748502961	4.743024017	209	4.727158648	4.717343957
198	4.740603973	4.740825053	210	4.705387958	4.715270948

Table A.5: SRB Method Applied to All Chromosomes Using Armitage Trend Test: Score 3

One-Sided Alternative			Two-Sided Alternative		
i	$-\log P_{(i)}$	$-\log \frac{\alpha}{n-i+1}$	i	$-\log P_{(i)}$	$-\log \frac{\alpha}{n-i+1}$
1	15.95458977	7.037490243	1	15.95458977	7.037490243
2	15.95458977	7.037489446	2	15.95458977	7.037489446
3	15.95458977	7.037488649	3	15.95458977	7.037488649
4	15.95458977	7.037487853	4	15.95458977	7.037487853
5	15.65355977	7.037487056	5	15.95458977	7.037487056
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
82	7.172270273	7.037425701	381	7.060629554	7.037187371
83	7.16350943	7.037424904	382	7.059987948	7.037186574
84	7.158937026	7.037424107	383	7.057190494	7.037185776
85	7.086642167	7.03742331	384	7.053331656	7.037184979
86	7.039949146	7.037422514	385	7.048459333	7.037184182
87	7.035507398	7.037421717	386	7.026866128	7.037183384

Table A.6: FDR Method Applied to All Chromosomes Using Armitage Trend Test: Score 3

One-Sided Alternative			Two-Sided Alternative		
i	$-\log P_{(i)}$	$-\log \frac{i}{n}q^*$	i	$-\log P_{(i)}$	$-\log \frac{i}{n}q^*$
1	15.95458977	7.037490243	1	15.95458977	7.037490243
2	15.95458977	6.736460247	2	15.95458977	6.736460247
3	15.95458977	6.560368988	3	15.95458977	6.560368988
4	15.95458977	6.435430252	4	15.95458977	6.435430252
5	15.65355977	6.338520239	5	15.95458977	6.338520239
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
207	4.766319169	4.721519897	1811	3.780647366	3.779571793
208	4.746166609	4.719426908	1812	3.780436936	3.77933205
209	4.73330389	4.717343957	1813	3.780397869	3.779092439
210	4.727998954	4.715270948	1814	3.780043274	3.77885296
211	4.721933118	4.713207788	1815	3.778860012	3.778613614
212	4.710493483	4.711154382	1816	3.776151949	3.778374399

APPENDIX B

GRAPHS OF LOD SCORES FOR SNPs

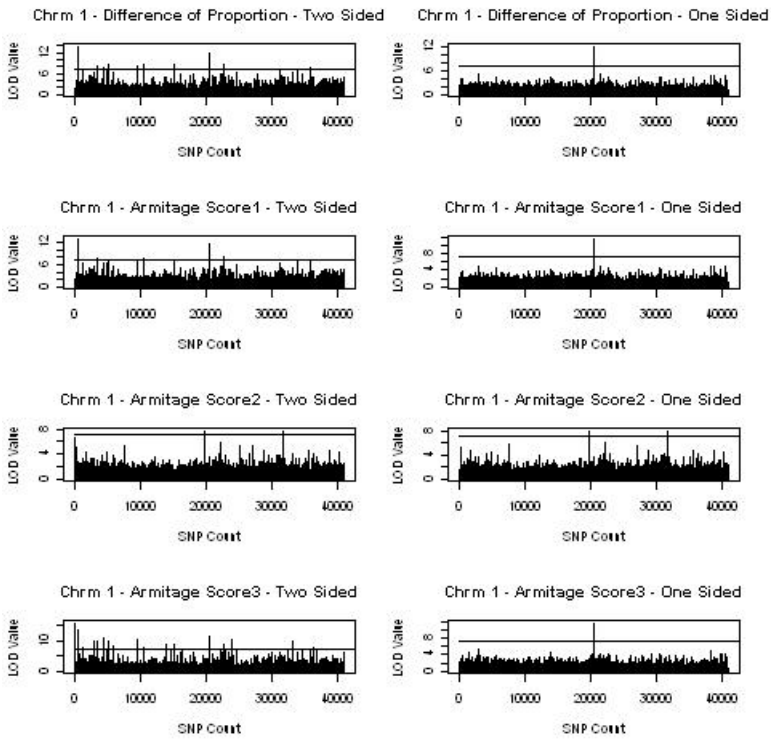

Figure B.1: Chromosome 1: One and Two Sided

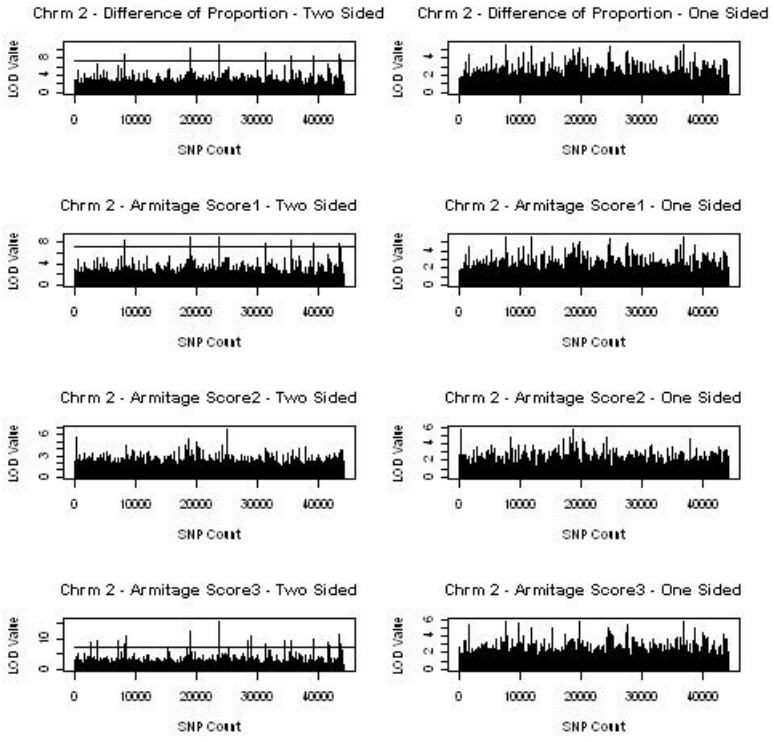

Figure B.2: Chromosome 2: One and Two Sided

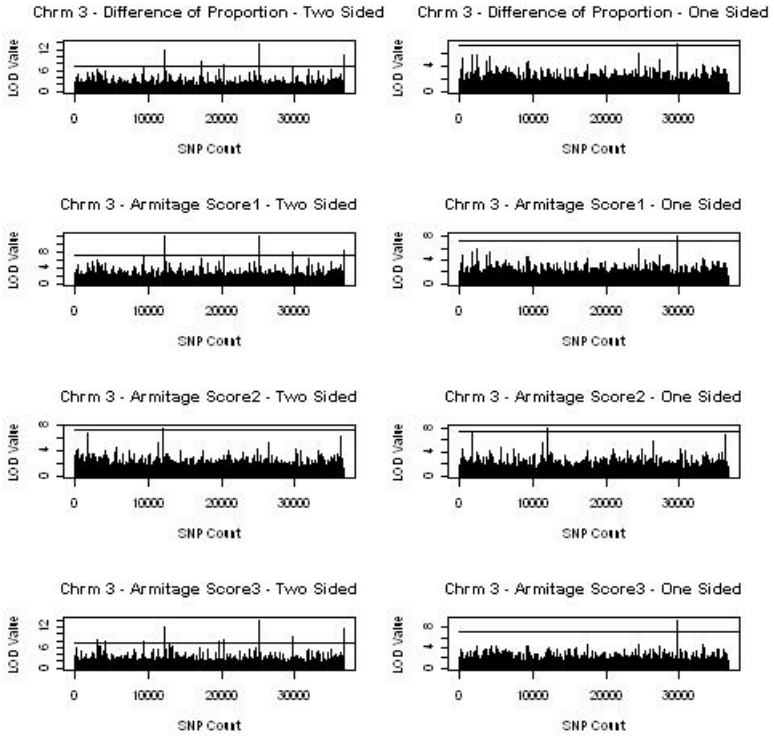

Figure B.3: Chromosome 3: One and Two Sided

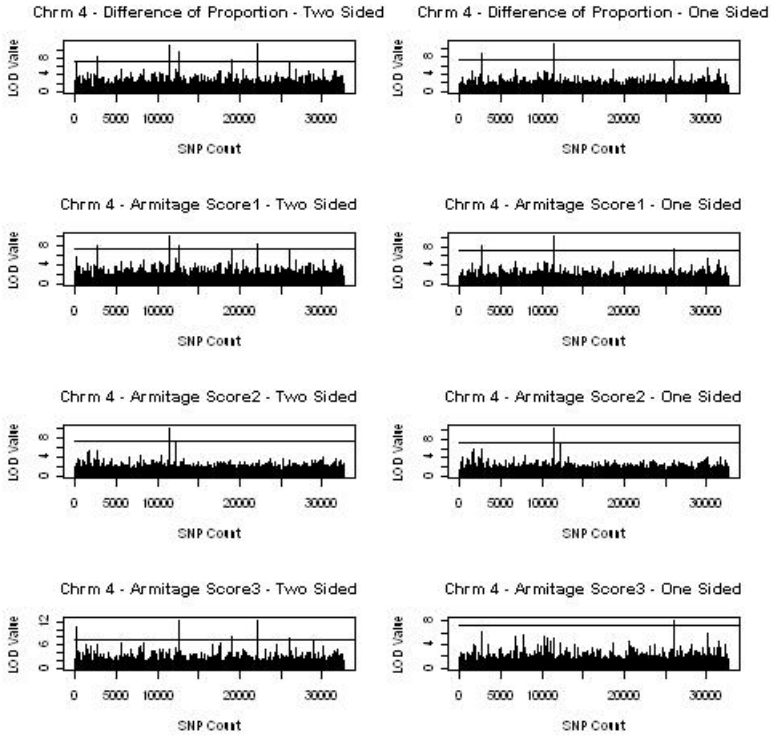

Figure B.4: Chromosome 4: One and Two Sided

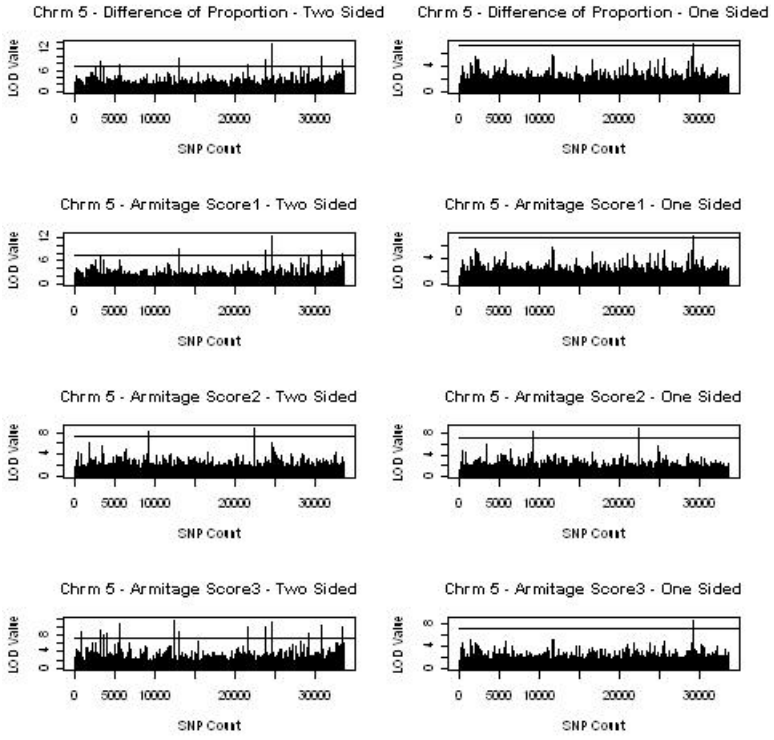

Figure B.5: Chromosome 5: One and Two Sided

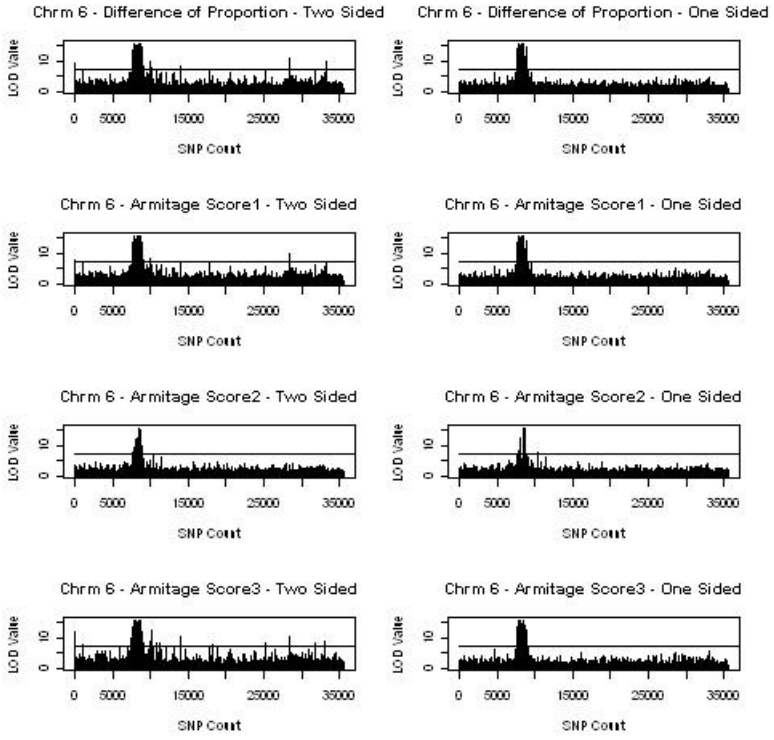

Figure B.6: Chromosome 6: One and Two Sided

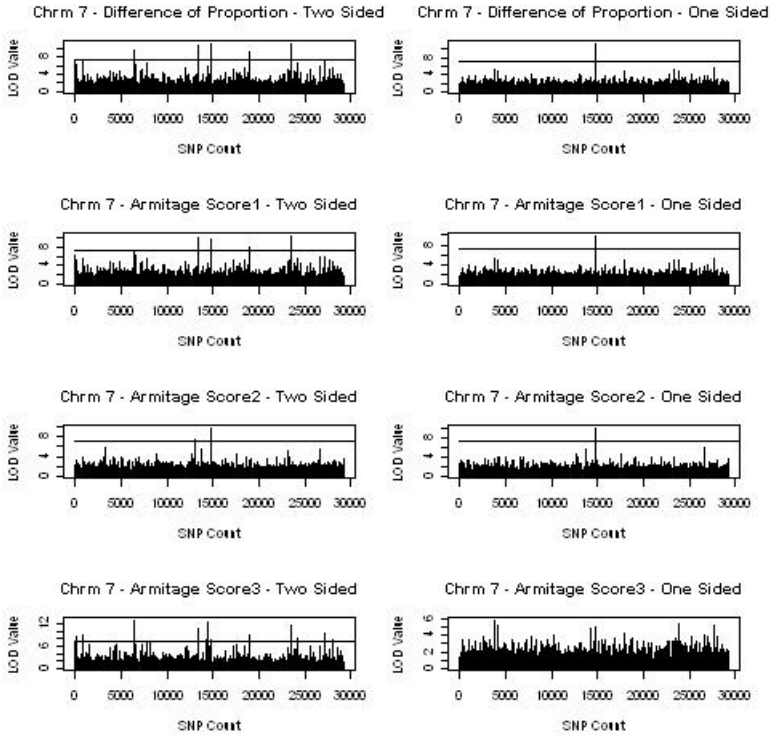

Figure B.7: Chromosome 7: One and Two Sided

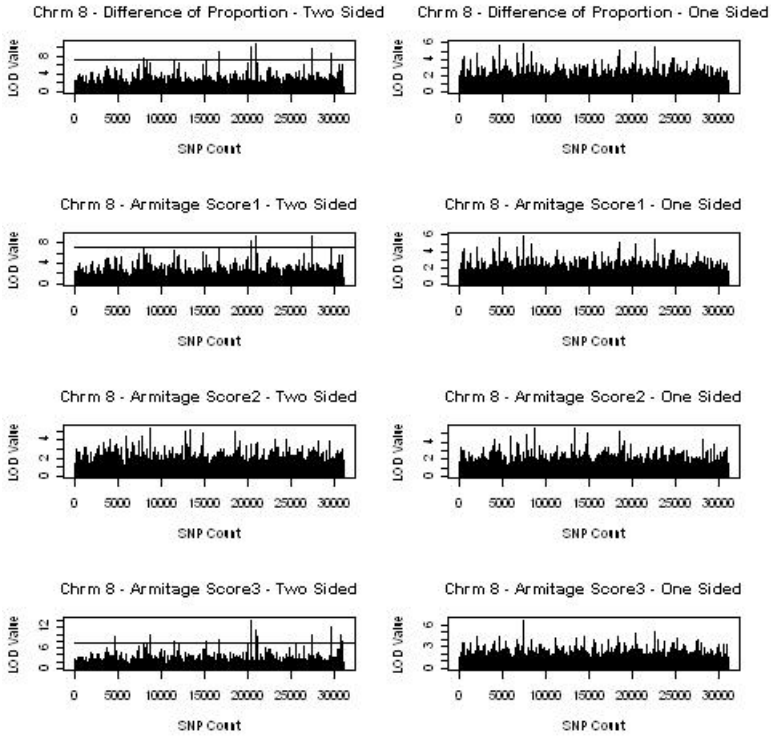

Figure B.8: Chromosome 8: One and Two Sided

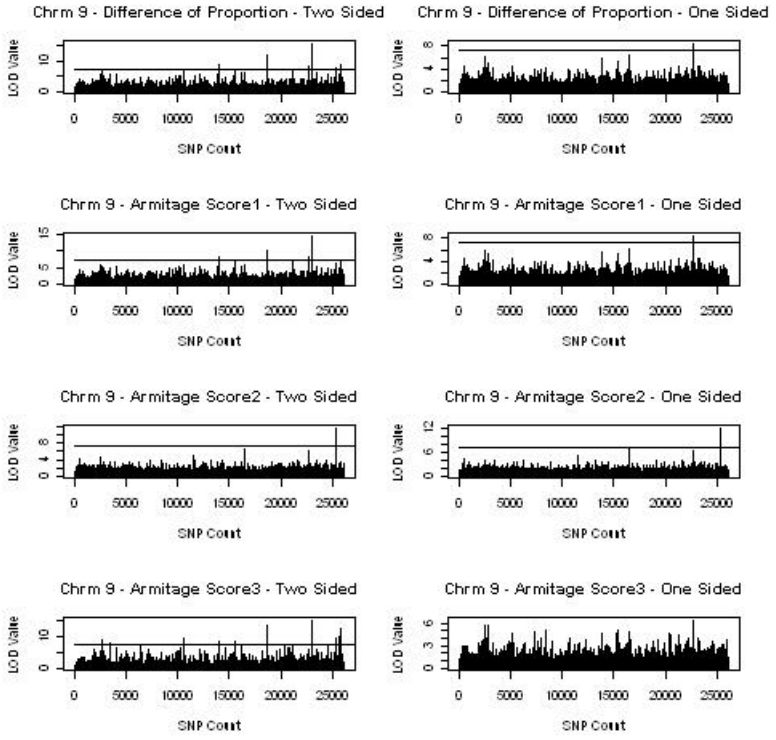

Figure B.9: Chromosome 9: One and Two Sided

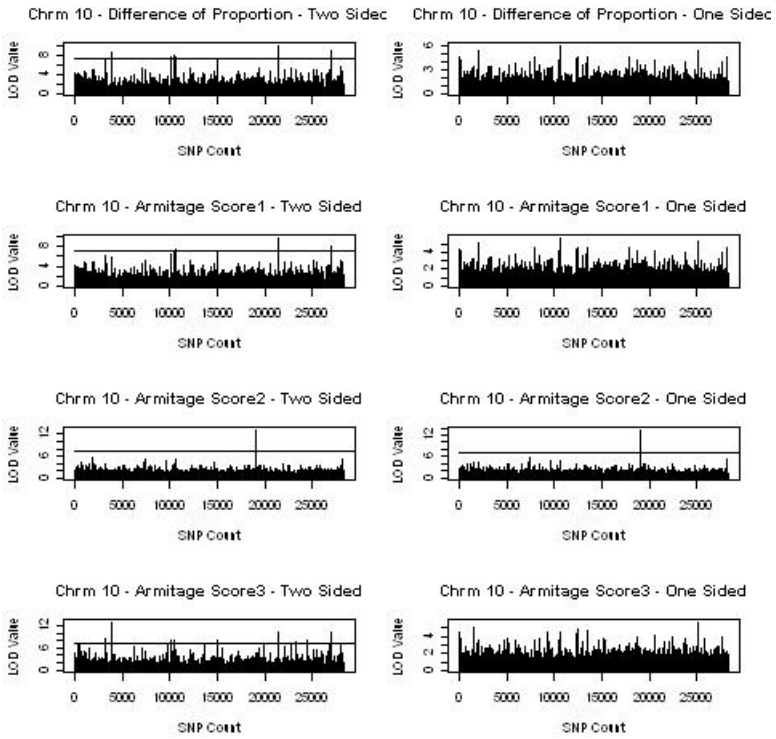

Figure B.10: Chromosome 10: One and Two Sided

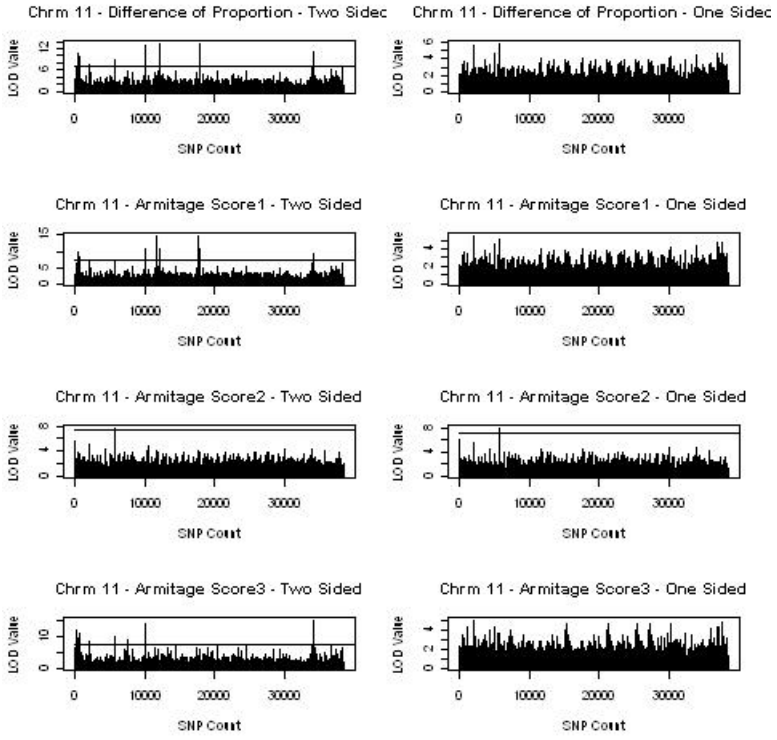

Figure B.11: Chromosome 11: One and Two Sided

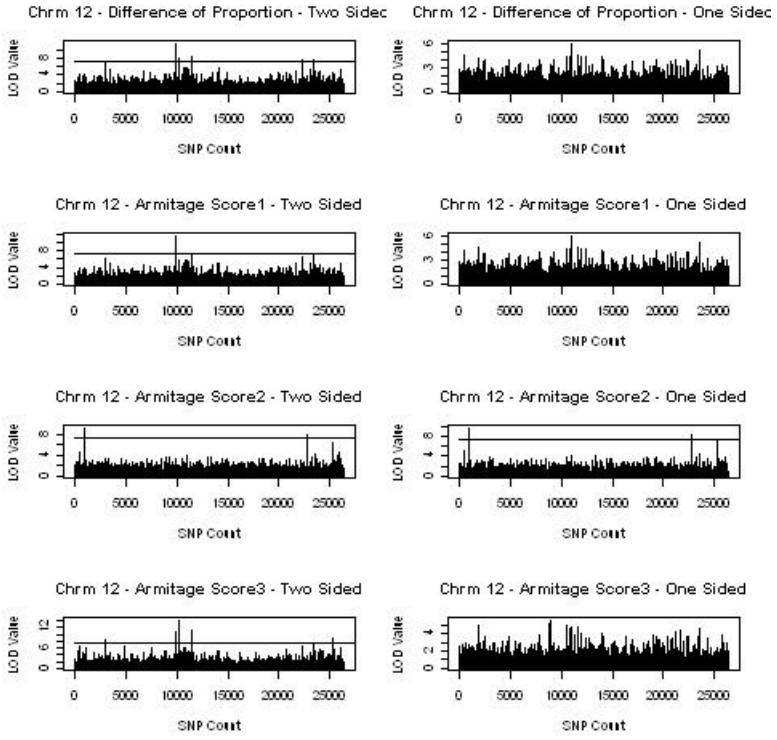

Figure B.12: Chromosome 12: One and Two Sided

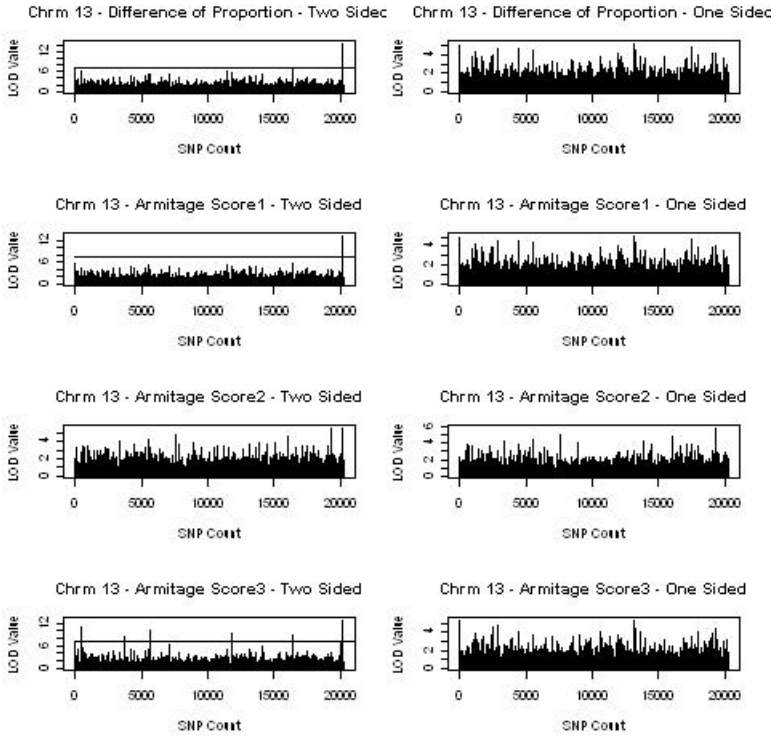

Figure B.13: Chromosome 13: One and Two Sided

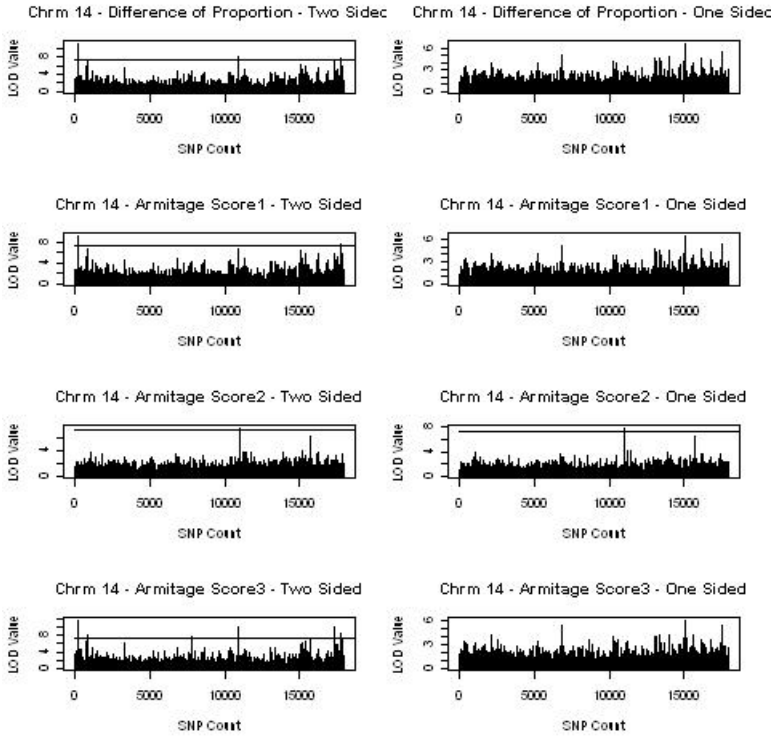

Figure B.14: Chromosome 14: One and Two Sided

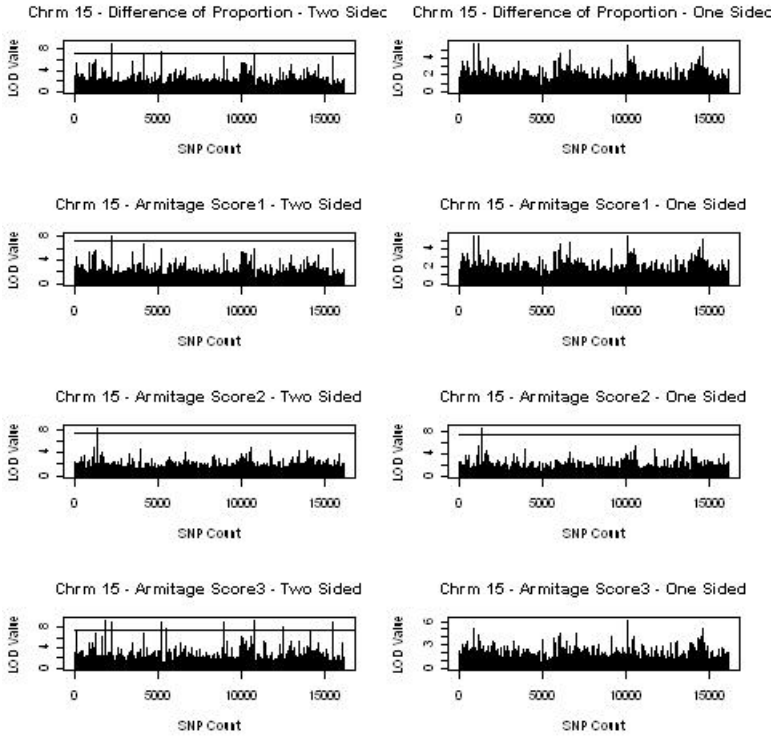

Figure B.15: Chromosome 15: One and Two Sided

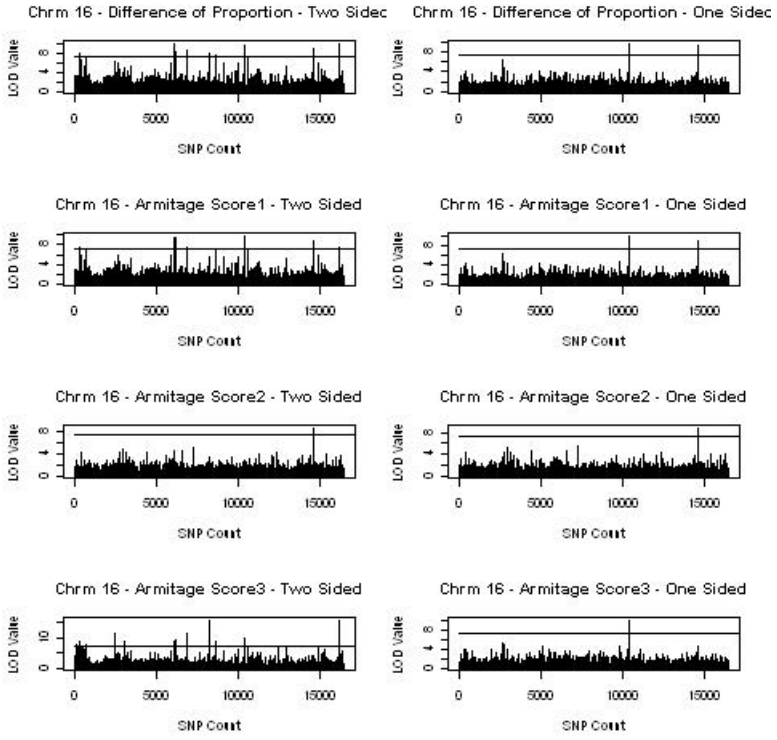

Figure B.16: Chromosome 16: One and Two Sided

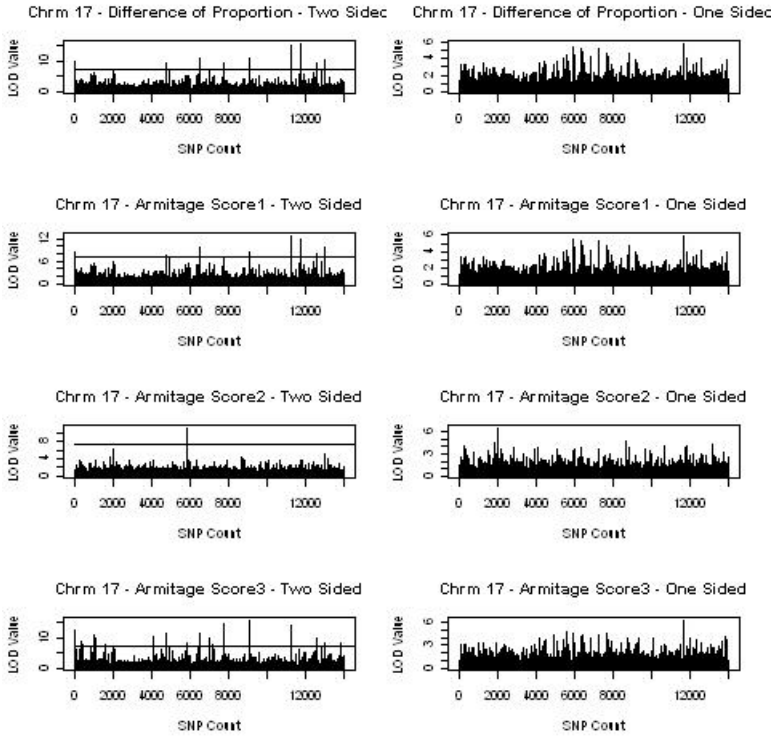

Figure B.17: Chromosome 17: One and Two Sided

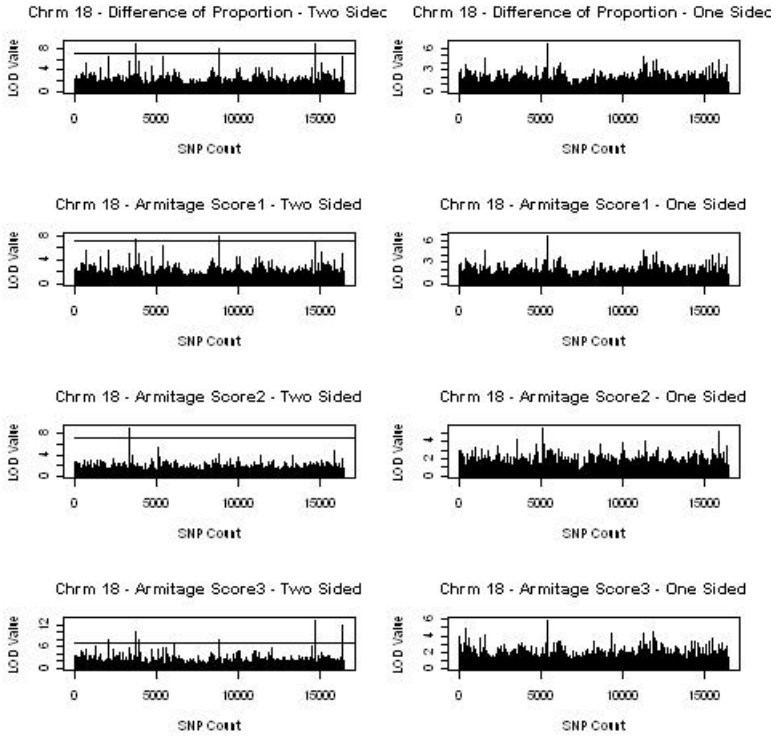

Figure B.18: Chromosome 18: One and Two Sided

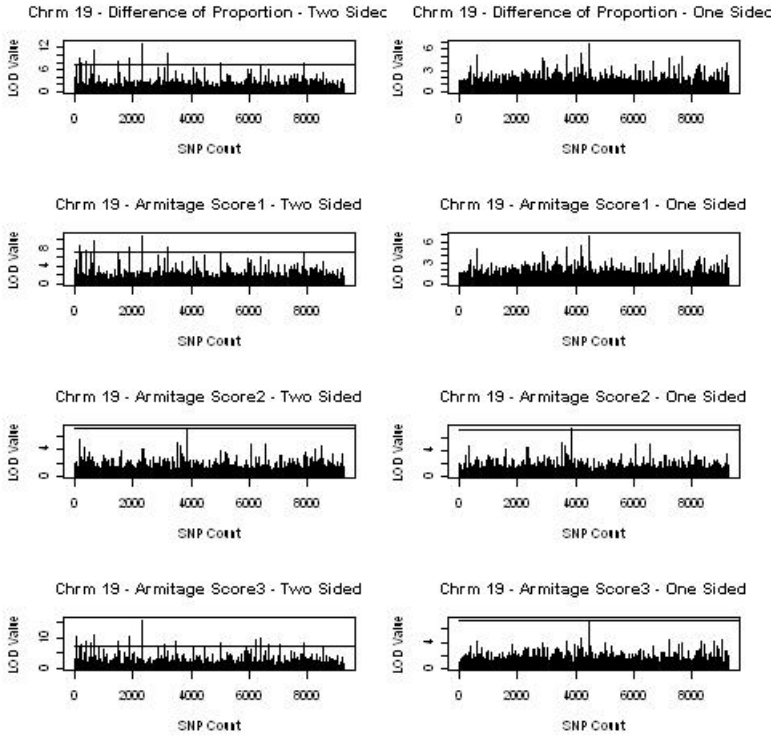

Figure B.19: Chromosome 19: One and Two Sided

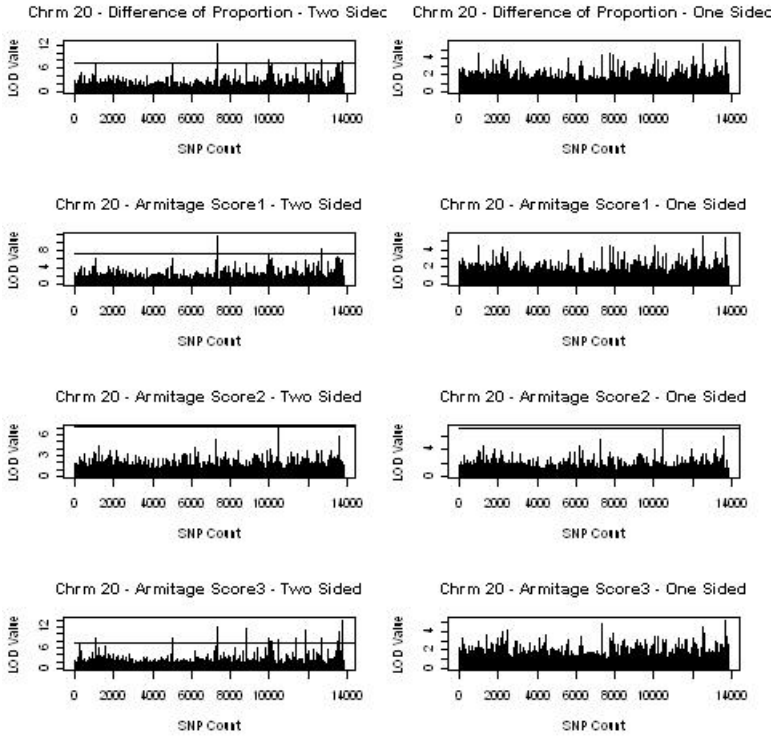

Figure B.20: Chromosome 20: One and Two Sided

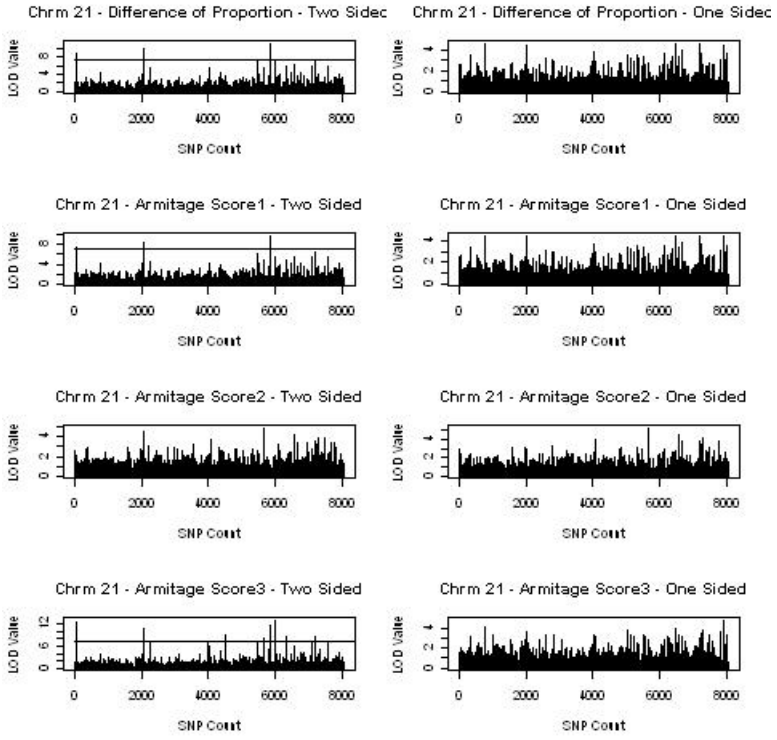

Figure B.21: Chromosome 21: One and Two Sided

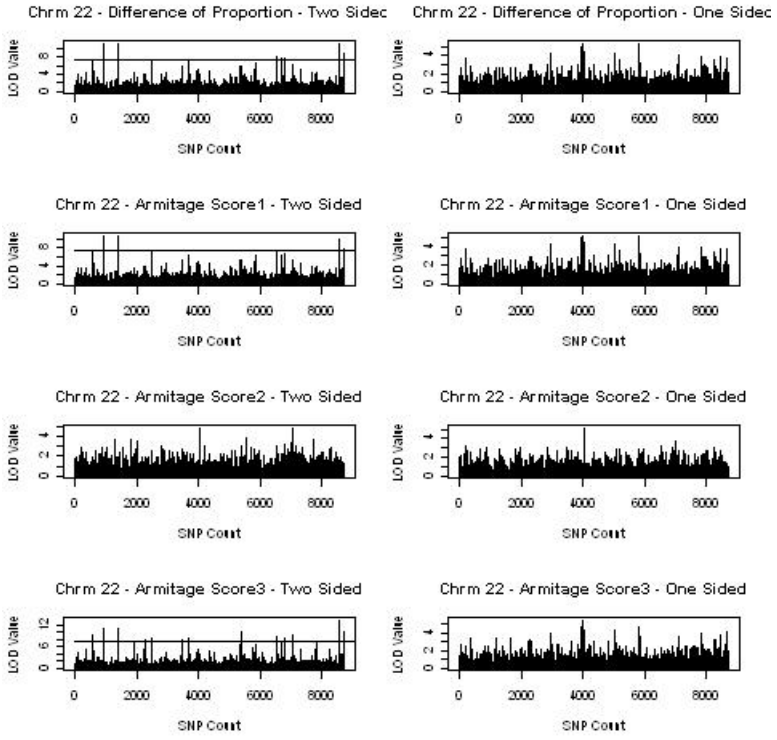

Figure B.22: Chromosome 22: One and Two Sided