Statistical Methods for Rare Variant Association

Mohammed Nasser D Alshahrani

Department of Mathematics and Statistics

University of Leeds

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

January 2018

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

This thesis is dedicated to those who turned their eyes away from other people fail, and applauded their attempts and encouraged them to stand for nothing. For those who care about others and light a candle when its dark.

For my parents, wife, and my son and daughter. For Dr. Arief Gusnanto, Professor Charles Taylor and Professor Jenny Barrett.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor, Dr Arief Gusnanto, for the continuous assistance and support of my PhD study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me throughout my research and writing of this dissertation. I could not have thought to have a better advisor and mentor for my study.

In addition to my advisor, I would like to thank those others who have given their invaluable support in my research and writing toward this goal: professors Charles Taylor and Jenny Barrett – for their encouragement, insightful comments, time and guidance. I have been extremely lucky to have these supervisors, who cared so much about my work, and who responded to my questions and queries so promptly. Each of the members of my Dissertation Committee has provided me extensive personal and professional guidance; each has taught me a big deal about both scientific research and life in general.

To all members of my family: I am extremely grateful to my parents Nasser Alshahrani and Gashma Alqahtani for their love, prayers, caring, and the sacrifices they made in educating and preparing me for my future. The quality of their support has allowed me to feel them near to me despite their being so far away. I am very much thankful to my wife Munirah Alqahtani for her patience and for being by my side for all the years of my study, through every journey. She has provided the home that I worked out of and the motivation to continue when I faltered. Thanks to my son Nasser and daughter Diala for their love, understanding, prayers and continuing support. Also, I

express my thanks to my brothers and sisters for their assistance and valuable prayers.

Lastly, but by no means in the least, I wish to express my gratitude to Dr Abdulrahman Al-khedhairi and Dr Farhan H. Al-Juaidi for their support in my first steps toward my higher education studies. And, most surely, I would be remiss indeed were I not to express my deep gratitude to the Prince Sattam bin Abdulaziz University for their unfailing support and assistance through funding my PhD study and research.

Abstract

Deoxyribonucleic acid (DNA) sequencing allows researchers to conduct more complete assessments of low-frequency and rare genetic variants. In anticipation of the availability of next-generation sequencing data, there is increasing interest in investigating associations between complex traits and rare variants (RVs). In contrast to association studies of common variants (CVs), due to the low frequencies of RVs, common wisdom suggests that existing statistical tests for CVs might not work, motivating the recent development of several new tests that analyze RVs, most of which are based on the idea of pooling/collapsing RVs.

Genome-wide association studies (GWAS) based on common SNPs gained more attention in the last few years and have been regularly used to examine complex genetic compositions of diseases and quantitative traits. GWASs have not discovered everything associated with diseases and genetic variations. However, recent empirical evidence has demonstrated that low-frequency and rare variants are, in fact, connected to complex diseases.

This thesis will focus on the study of rare variant association. Aggregation tests, where multiple rare variants are analyzed jointly, have incorporated weighting schemes on variants. However their power is very much dependent on the weighting scheme. I will address three topics in this thesis: the definition of rare variants and their call file (VCF) and a description of the methods that have been used in rare variant analysis. Finally, I will illustrate challenges involved in the analysis of rare variants and propose different weighting schemes for them. Therefore, since the efficiency of rare variant studies might be

considerably improved by the application of an appropriate weighting scheme, choosing the proper weighting scheme is the topic of the thesis. In the following chapters, I will propose different weighting schemes, where weights are applied at the level of the variant, the individual or the cell (i.e. the individual genotype call), as well as a weighting scheme that can incorporate quality measures for variants (i.e., a quality score for variant calls) and cells (i.e., genotype quality).

Abbreviations

GWAS	Genome-Wide Association Study.
VCF	Variants Call Format.
GQ	Genotype Quality.
G	Genotype Matrix with 0/0,1/0 and 1/1 elements.
X	Genotype Matrix $n \times p$.
Z	Covariate Matrix $q \times p$.
$\mathcal{F}$	Minor allele frequency (MAF).
$S(\gamma)$	Score test incorporating variants weight.
$S(\psi)$	Score test incorporating individuals weight.
$S(\gamma\psi)$	Score test incorporating variants and individuals weight.
Ψ	$n \times n$ matrix with individual weights at diagonal.
Γ	$p \times p$ matrix with variant weights at diagonal.
$\boldsymbol{\gamma}$	Vector $p \times 1$ for variant weights.
$\boldsymbol{\psi}$	Vector $n \times 1$ for individual weights.
$\mathbf{U}$	Vector $p \times 1$ for score function.
$\mathfrak{q}$	The extra information based on variants level.
ω	The variant weight incorporating extra information.
Ω	$n \times p$ cell weights.
$w_j(\mathcal{F})$	Variant weights which are function of MAF.

Contents

Chapter 1

Introduction

1.1 Introduction

1.1.1 Background

The human genome consists of building blocks called base pairs, which are part of the deoxyribonucleic acid (DNA) in the chromosomes of every person. There are more than three billion base pairs in human DNA, and over 99% are identical between individuals. A base pair that varies in a population and has two variants is called a single-nucleotide polymorphism (SNP). Nucleotides are the building blocks of nucleic acids and comprise three subunit molecules: a five-carbon sugar, a nitrogenous base, and a phosphate.

In each person's DNA, SNPs appear, on average, once every 300 nucleotides, and there are nearly 10 million SNPs in the human genome. Different variations of DNA, such as structural variants, including insertions and deletions (INDELs), in which one or several nucleotides are added or missing in part of a DNA sequence. Block substitutions involve several adjacent nucleotides, while inversions involve changes in the order of nucleotides in a genomic region, and copy-number variants (CNVs) involve DNA sequences of > 1 kb that appear in various numbers and are similar to a reference genome. In most cases, single-nucleotide variants are created by replacements during the replication of a nucleotide that carries a given base to another nucleotide, which, in turn, carries a different base.

SNPs can act as biological markers, which are known locations on chromosomes, and as such, SNPs help scientists determine which genes are associated with particular diseases. However, when SNPs appear within a gene or near a gene, they can also have a more direct function in a disease process by affecting the gene's function. SNPs have two variants, called alleles. They are often referred to as A and a, or major and minor alleles, respectively. a minor allele generally has a population frequency of less than 0.5. This frequency is called the minor allele frequency (MAF). Based on MAFs, SNPs can be divided into three groups: common SNPs with MAFs greater than 0.05, low-frequency SNPs with MAFs between 0.01 and 0.05, and rare SNPs, or rare variants, with MAFs less than 0.01. These thresholds differ across studies and are discussed in more detail in section 1.1.2.

Genotype refers to the genetic composition of cells. For every trait of an individual, such as hair colour, eye colour, height and weight, a cell has instructions for two alleles, or alternative forms of a gene. The genotype of an individual leads to this pair of alleles at an SNP (i.e., AA, Aa or aa). One of the alleles comes from the person's mother, and the other comes from the person's father. An individual's genotype is his or her unique combination of these alleles, which can be heterozygous (i.e., different)or homozygous (i.e., the same). A person's phenotype refers to a trait that can be physically observed, such as a morphology or behaviour.

Many genetic studies have established that thousands of loci, or positions on chromosomes, contribute to common polygenic human diseases, which are diseases caused by multiple genes, as well as traits. This finding was noted in a genome-wide association study (GWAS) catalogue (Burdett, 2017). However, recent studies have focused mainly on common variants because they provide a fantastic opportunity to investigate the impact of common variants on complex diseases.

GWASs have noted that common SNPs have gained prominence in recent years and are used routinely to examine complex genetic compositions of diseases and quantitative traits. More than $52,491$ disease-associated common variants have been catalogued using GWAS techniques (Duncan & Brown, 2013). Such studies rely on systematic evaluations of common genetic variants, which usually

involve high MAFs Visscher *et al.* (2012). The studies have been vital to advancing our knowledge concerning disease pathologies, for example, central nervous system functioning in individuals predisposed to obesity and the pathways of macular degeneration in cases of age-related vision loss (Klein *et al.*, 2005).

Nonetheless, while there has been a significant amount of work completed and progress made in this area, many genetic contributions to complex traits remain unexplained (Teslovich *et al.*, 2010). GWASs have not discovered all the associations between diseases and genetic variations. In fact, genetic variants discovered to date explain only a small proportion of estimated heritability, and thus, only a small proportion of variations in phenotypes can be currently explained based on genetic differences. However, some studies have used indirect statistical methods to suggest that common variants explain at least 30%, and potentially more, of many diseases' and traits' heritabilityLee *et al.* (2012b), Yang *et al.* (2011).

A key example of such a study is a GWAS that explored type 2 diabetes, or T2D (Mendelian Inheritance in Man, or MIM, number 125853). The study examined more than $150,000$ patients and identified > 70 loci at appropriate levels of genome-wide significance; however, its examination was only helpful in explaining 11% of T2D's heritability. A similar GWAS found the same trend in Crohn's disease patients. It examined $210,000$ subjects but was able to account for only 23% of the estimated variation (h^2) in inherited cases (Franke *et al.*, 2010). Loci identified in GWASs often have only modest effects on disease risk and quantitative trait variations, and hence, a movement in this area from theoretical knowledge to usable knowledge and clinical applications is likely to be slow (Figure 1.1).

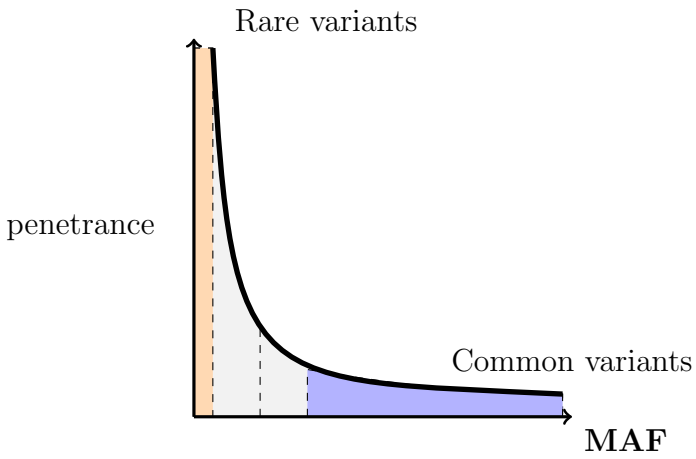


Figure 1.1: Frequencies of published association variants. Variants can be classified by the frequency (on the horizontal axis) and its effect (i.e., penetrance, which is the proportion of individuals carrying a variant who also manifest a specific phenotype) on the vertical axis. Note that rare variants (typically $< 0.1\%$ to $< 5\%$) are highly penetrant and often associated with severe developmental disorders, while common variants have modest effects.

This gap regarding heritability, which is often called 'the problem of missing heritability', has some potential explanations. While some sources of missing heritability remain unclear, one hypothesis is that most of the missing heritability is found in rare genetic variants Manolio *et al.* (2009), Zuk *et al.* (2014). The case for the important role of rare variants relies on the idea that alleles which predispose someone to disease are likely to be deleterious and, thus, are kept at low frequencies via purifying selections Kryukov *et al.* (2007).

Indeed, rare variants play important roles in most human diseases. For example, many Mendelian disorders and rarer types of more common diseases are caused by highly penetrant but rare genetic variants Gibson (2012). Losses of functional variants, which prevent the generation of functional proteins, are known to be particularly rare.

Recent empirical evidence has demonstrated that low-frequency and rare variants are, in fact, connected to complex diseases. However, commercial genotyping arrays have, until recently, largely ignored this part of the allele frequency spec-

trum. As there is no reliable catalogue of rare variants to support array designs, and genome-wide surveys of rare variations generally need more assays than existing arrays support, the focus of most studies remains on common variants.

Recent advances in DNA sequencing technologies have led to significant changes in population and medical genetics. DNA sequencing allows researchers to conduct more complete assessments of low-frequency and rare genetic variants and investigate how such variants influence complex traits MacArthur *et al.* (2012). Next-generation sequencing (NGS) technologies offer high-throughput parallel-sequencing approaches that can generate billions of short-sequence analyses more cost effectively than older technologies. NGSs make short reads aligned to reference genomes, do researchers can successfully identify any genotype site at which there are differences between individuals. As the cost of sequencing continues to fall, exome sequencing, which transcribes portions of genomes, and whole-genome sequencing (WGS) of complex diseases may become increasingly achievable Gudmundsson *et al.* (2012).

However, while genome sequencing provides a means of investigating low-frequency and rare variants in complex diseases, it is not without challenges. First, conducting deep WGS of large numbers of patients is, and will likely remain, expensive. This type of research sequences large numbers of replicate reads for specific regions, so various alternative strategies, including targeted sequencing, exome sequencing, low-depth WGS, and extreme phenotype sampling, have been proposed as more cost-effective options. Second, achieving statistically significant results using classical single variant-based association tests for low-frequency and rare variants is difficult without significantly increasing sample sizes. Hence, costs increase unless sizes of effects are very large. Third, requisite multiple-test corrections are poorly understood and difficult to use effectively. Rare variants are diverse and occur infrequently, so rare-variant association studies (RVAS), which are similar to common-variant association studies (CVAS) Zuk *et al.* (2014), require extensive sample collection and detailed statistical analyses to detect genetic associations with diseases. To address these limitations, RVAS need further development.

Some early RVAS efforts were based on the notion that rare variants related to common diseases can be identified using a small number of samples, and some

discoveries were made Cohen *et al.* (2004), Bonnefond *et al.* (2012). Nonetheless, analytical methodologies for RVAS are unfixed, although many scholars have proposed a wealth of potential methods. In light of this situation, researchers are turning to new statistical methods designed for RVASs with the aim of increasing the validity and power of each finding. These methods evaluate associations of multiple variants in a biologically relevant region, such as a gene, rather than testing the effects of single variants, which was commonplace in some GWASs Lee *et al.* (2014).

1.1.2 The Definition of Rare Variants

There is little agreement among scholars as to what constitutes a genuinely rare variant. A variant is defined as either a locus or an allele at a locus, which was described in detail by Saint Pierre & Génin (2014). Frazer *et al.* (2009) posited that rarity referred to an MAF at a locus when viewing the locus as a variant, so the researcher described a rare variant as a genetic alternative that has an MAF of $< 1\%$. However, this definition varies between scholars, and acceptability thresholds also vary. In Frazer *et al.* (2009), a cut-off of 1% is recommended with rare variants that define alleles with frequencies of $< 1\%$, while in Gorlov *et al.* (2011), the threshold is 5%. Bodmer & Bonilla (2008) relied on an upper limit of 1%, but conversely, the author suggested a lower limit of 0.1% to distinguish genuinely rare variants from a third category of variants, which included 'clearly deleterious mutations'. For many scientists, recognizing that these frequency boundaries are not absolute and that there can be overlaps between low-frequency common variants and high-frequency rare variants can be helpful in understanding the larger debate.

Cirulli & Goldstein (2010) established four categories based on variant frequencies. The first is very common variants, which have frequencies between 5% and 50%. The second is less common variants, which have frequencies of between 1% and 5%. The third is rare but not private variants, which have frequencies of $< 1\%$ (these are polymorphic in one or more major human populations). Finally, the fourth category is private variants, which only appear in the proband's immediate relatives.

Alternatively, Zuk *et al.* (2014) defined common variants as those that occur regularly enough to allow individual tests with control groups. Given actual sample sizes, the researchers noted that variants with frequencies as low as 0.5%, which appear at least once in every 100 human subjects, can be classified as common variants. The researchers defined rare variants in contrast to common variants, as less frequently than once in every 100 human subjects.

As a definition, this thesis uses rare variants of less than 0.05. It classifies MAFs between 0.01 and 0.05 as rare variants and can be classified as high-frequency rare variants. Common variants appear in between 0.05% and 0.5% of subjects. Finally, (see Figure 1.2 for the definition) it classifies rare variants as extremely rare, moderately rare, and common, which are explained in later sections. Note that the number of individuals is $n = 2000$, so MAF $= 0.005$ is the threshold for extremely rare variants. This classification of variants by MAF will be used throughout the thesis.

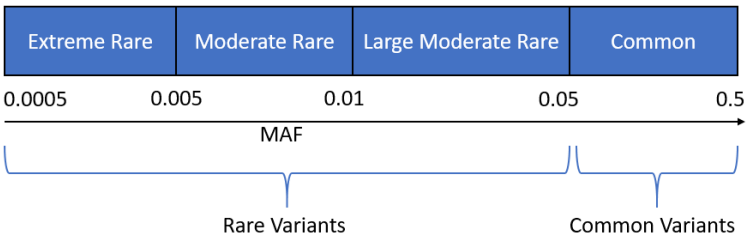


Figure 1.2: Definition used for the levels of MAF in this thesis.

1.1.3 Types of variation

Genetic variants are categorized into two classes based on the composition of their nucleotides: single nucleotide variants or structural variants. Single nucleotide variants are created by changes at a single nucleotide position on the DNA sequence. In most cases, single nucleotide variants are generated by a replacement during the replication of a nucleotide that carries a given base to a nucleotide that carries a different base. For example, an SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a stretch of DNA. Such a replacement generates change in neither the DNA sequence length nor the order of

the nucleotide. This replacement is called a point mutation and is often unique so that only two variants can be observed at any position. SNPs are the most obvious common type of genetic variation between people. Each SNP denotes a difference in a single strand of DNA–the nucleotide. Structural variants represent all other kinds of DNA variations. This is, therefore, a rather widely varied class of variants that includes INDELs, where one or a few nucleotides are missing in some DNA sequences, block substitutions, where a change occurs in the string of adjacent nucleotides, CNVs, which are stretches of DNA sequences larger than 1 kb present in variable numbers that are different in the normal population, and inversions, where the order of the nucleotides in a given genomic region is changed (Saint Pierre & Génin, 2014).

1.2 Overview of the Thesis

1.2.1 Objective

Due to the low power, it is impossible to analyse each variant individually. Indeed, a high proportion of rare variants appear in only one or two people in a large sample. For this reason, standard GWAS analyses cannot produce stable estimates of rare variants. Thus, methods were developed to combine variants within regions. The objectives of this thesis are as follows:

- Develop statistical and inference methods to analyse rare variant associations, as association analyses of rare variants require statistical methods that can effectively combine information across variants and estimate the overall effect of data;
- Investigate schemes for weighting variants in analyses; and
- Consider how genotyping errors commonly occur and remain a challenge in sequencing studies; because quality scores can be good measurements of genotyping accuracy, and genotype quality scores are offered by automated biotechnologies, the aim is to develop an association test that incorporates genotyping quality scores to improve statistical power and inference.

1.2.2 Outline of the Thesis

The thesis is divided into ten chapters. Chapter 2 introduces variant-call format structures and explores data analysis. Chapter 3 reviews the literature on RVASs with the aim of introducing RVA tests and examining the differences between such tests. Chapter 4 discusses score tests and their distributions by incorporating weighting schemes. Chapters 5 and 6 introduce different weighting schemes based on variant levels. In Chapter 5, rare variant regions are discussed, while Chapter 6 extends that discussion to whole regions (rare and common). Chapter 6 introduces combined weighting schemes. Finally, in Chapters 7, 8 and 9, new weighting schemes that incorporate external information regarding score tests are examined. For example, variant measures, such as variant-based quality (quality calls) and genotype quality, are examined. Chapter 7 develops a weighting scheme based on variants while incorporating information based on quality calls. Chapter 8 develops an individual weighting scheme and combines it with a variant weighting scheme (marginal weight). Chapter 9 develops a weighting scheme based on variants and individual levels (cell weights). Finally, Chapter 10 summarises this thesis's major findings and provides suggestions for future research.

Chapter 2

Variant Call Format and Exploratory Data Analysis

2.1 Variant Call-Format (VCF) Files

A VCF file is a standardised file with text that represents INDELs, SNPs, and structural variation calls. VCF files contain meta-information lines, header lines, and data lines. The files also contain genotype information about samples for all positions in genomes. In this chapter, VCF files are explained in detail, and some examples are given to illustrate concepts using real VCF data. The VCF version discussed is VCF 4.2.

2.1.1 Fixed Fields

The header line describes the eight fixed mandatory columns. The columns are named:

1. chromosome (CHROM),
2. position (POS),
3. identification (ID),
4. reference (REF),
5. alternative (ALT),

6. quality (QUAL),
7. filter (FILTER), and
8. information (INFO).

Descriptions of the column names are provided below.

1. CHROM: An identifier from a reference genome. All entries for a specific CHROM should form a contiguous block within a VCF file. For CHROMs, an alphanumeric string is required.
2. POS: A reference position; the first base is 'position 1'. Positions are sorted numerically, in increasing order, within reference CHROMs. Integers are required.
3. ID: A semicolon-separated list of unique variant descriptions. If an ID is a dbSNP variant, its (rs) numbers can be used. IDs should not exist in multiple data records, and if an ID is unavailable, a missing value should be used. An alphanumeric string is required.
4. REF: A base that contains only the letters A, C, G, T, and N. These letters should always be uppercase, and multiple bases can be authorised. Values in POS fields relate to the positions of first bases in sequences. For INDELs, a reference string, or sequence, must carry a base before an event, which must be exhibited in a POS field. REFs require strings.
5. ALT: A comma-separated list of named non-reference alleles that are in at least one sample. Options are base strings built from the bases A, C, G, T, and N, and options fit the angle-bracketed ID string ("ID"). If there are no non-reference alleles, then a missing value should be employed. Bases should always be uppercase. Alphanumeric strings without commas, white spaces, and angle brackets are permitted in ID strings.
6. QUAL: A numeric Phred-scaled quality score for an assertion made in an ALT, for example, $-10\log_{10}$ prob (call in ALT is wrong). If an ALT is "." (no variant), then it should appear as $-10\log_{10}$ p (variant). However, if an

ALT is not ".", then it should appear as $-10\log_{10}$ p (no variant). A high QUAL score means that it is a high assurance call. It should be noted that although integer Phred scores have traditionally been used, the QUAL field can be a floating point to permit higher resolutions for low confidence calls if desired.

7. FILTER: 'PASS' indicates that a site, or POS, has passed all filters. If a POS has not passed all filters, a list of codes for filters that fail will be presented, and semicolons will separate items in the list; for example, "$q10; s50$" symbolises that the QUAL at a POS is under 10, and the number of samples with INFO is below 50% of the total number of samples. "0" is reserved, and it should not be used as a filter string. If filters have not been applied, then this field should be set to a missing value. FILTERs are alphanumeric strings.

It should be noted that following header blocks and field names, lines or blocks describe single variants; several properties of such variants are described in columns.

Type of Variation	Alignment	VCF represent
SNP	1 2 3 4 A C G T A T G T	POS REF ALT 2 C T
Insertion	1 2 3 4 5 A C - G T A C T G T	POS REF ALT 2 C CT
Deletion	1 2 3 4 A C G T A - - T	POS REF ALT 1 ACG A

Table 2.1: Type of SN variation.

Examples of different genotypes represented in VCF data are given in Table (2.2).

CHROM	POS	REF	ALT	ALT1	Type
1	150737433	C	G		SNP
1	150737444	T	C		SNP
1	150737565	TTCTCTCTCTCTC	TTCTCTCTCTCTCTCTC		INSERTION
1	150781527	CAA	CA		DELETION
1	150782111	A	ACACAC	ACAC	INSERTION
1	150782176	CAGA	CAGAGA		INSERTION

Table 2.2: Examples of variation types taken from real VCF data (call outputs). Only references, alternative alleles, and types are shown.

2.1.2 Genotype Field

If genotype data are included in a VCF, a FORMAT column header is needed. The acronym 'GT' is used to represent genotypes and encrypt them as allele values separated by either \ or | in such headers. / means a genotype is unphased, and | means it is phased. The expression 0 is used if they are reference alleles (alleles in REF fields), while 1 is used if they are first alleles in ALTs, and 2 is used if they are second alleles in ALTs.

Polyploid cells and organisms are living beings that have more than two paired (homologous) sets of chromosomes, while triploids have three sets, and tetraploids have four. There are many types of these organisms, and each is labelled according to the number of chromosome sets in its nuclei. For diploid cells, example labels include 0/1, 1 | 0, and 1/2. Additionally, for triploid calls, example labels include 0/0/1. However, for haploid calls, for instance, for Y male non-pseudoautosomal X cells, or mitochondria, only one allele value should be given. If a call variant cannot be made for a sample at a given locus, "." should be defined for a specific missing allele in a GT field (e.g., "./." would be used for a diploid genotype, and "." would be used for a haploid genotype). A GT field is encoded as a typed integer vector. It should be noted that the data used in this chapter is diploid.

1/1 and 0/0	Three samples encoded sequentially
./.	Two missing alleles
0/1/2	A tetraploid with alleles

To provide an example, if a REF is A and an ALT is T, 0/0 means AA, 1/0 means AT, and 1/1 means TT. However, sometimes one will see INDELs in data. For instance, if one lets a REF be A in the ALT ACACAC,ACAC, the meaning of the variation 0/0 is AA. Similarly, 2/0 means $ACAC/A$, 2/1 means $ACAC/ACACAC$, and 2/2 means $ACAC/ACAC$.

```
[HEADER LINES]
#CHROM  POS ID      REF ALT QUAL    FILTER  INFO          FORMAT         NA12878
chr1    873762  .       T   G   5231.78 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047   A   G   3931.66 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431  C   T   71.77   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:1,3:4:25.92:103,0,26
chr1    974165  rs9442391   T   C   29.84   LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:14,4:14:60.91:61,0,255
```

Figure 2.1: An example of genotype data in the VCF format.

Using the last column in Table 2.1, the tags in the FORMAT column can be described.

- GT: The genotype of this sample. For a diploid person, the GT field indicates the two alleles given by the sample, encrypted by a 0 for the reference allele (REF), 1 for the first alternative allele (ALT), 2 for the second ALT allele, and so on. In the case of a single ALT allele (by far the more common case), GT will be either:
  - 0/0, meaning that the sample is homozygous;
  - 0/1, meaning that the sample is heterozygous and carries 1 copy of each REF and ALT allele; or,
  - 1/1, meaning that the sample is a homozygous ALT.
- GQ: The quality of the genotype, or Phred-scaled probability that the true genotype is the one provided in the GT field. In a diploid situation, if GT is 0/1, then GQ is actually $L(0/1)/(L(0/0)+L(0/1)+L(1/1))$; note that L is the likelihood that the representation is 0/0, 0/1, or 1/1 under the model built for the next generation sequence dataset.
- AD and DP: These are equivalent fields that describe two primary ways of recording the data depth for this sample at this position or site. Data depth is explained in the next section.

- PL: The likelihoods of the given genotypes (here, 0/0, 0/1, and 1/1) are presented in this field. These are the normalised, Phred-scaled likelihoods for all of the genotypes 0/0, 0/1, and 1/1, with no priors. To illustrate, this would be L data given that the true genotype is 0/1 for the heterozygous case. The very likely genotype (in the GT field) is scaled; its $P = 1.0$ (0 when Phred-scaled), and the other likelihoods indicate their Phred-scaled likelihoods relative to this most likely genotype. It is rounded to the nearest integer.

  A PL has three numbers. The first, second, and third numbers relate to the probability that the site is homozygous, heterozygous, or homozygous, respectively, for the alternate allele. As the number increases, the probability that the sample is that genotype decreases. Hence, if a PL is $485, 0, 535$, the software is relatively certain that the sample is heterozygous rather than a homozygous reference or homozygous alternate. Moreover, the GT confirms this by being 1/0. If the first and last numbers had been lower, then the quality of the SNP would be reduced, and the genotype would be less reliable. An example of a PL is given in Figure (2.2) At position 150737433, which is the first line of the figure, the PL is $0, 205, 255$ (i.e., the likelihoods of the three given genotypes) at this position in individual number 1, which is named $X06\_0006e\_1$ (replicate 1 for individual $X06\_0006e$). The replicate number is the last number of the individual's name (the individuals represented in the columns in the table here). The genotype is homozygous in reference 0/0 based on the PL result.

CHROM	POS	X06_0006e_1	X06_0006e_2	X06_0042e_1	X06_0042e_2
1	150737433	0,205,255	0,202,255	0,181,255	0,247,255
1	150737444	0,247,255	0,250,255	0,202,255	0,255,255
1	150737518	0,255,255	0,255,255	0,255,255	0,255,255
1	150737565	255,81,0	255,96,0	0,117,255	0,157,255
1	150761952	0,244,255	0,244,255	0,144,255	0,187,255
1	150761982	0,255,255	0,255,255	0,208,255	0,238,255
1	150762071	0,255,255	0,255,255	0,255,255	0,255,255
1	150762097	0,255,255	0,255,255	0,255,255	0,255,255
1	150762119	0,255,255	0,255,255	0,255,255	0,255,255
1	150762171	255,255,0	255,255,0	0,255,255	0,255,255
1	150779931	0,255,255	0,255,255	0,255,255	0,255,255
1	150780019	0,255,255	0,255,255	0,255,255	0,255,255
1	150780760	0,229,255	0,238,255	0,129,255	0,196,255

Figure 2.2: Illustration of a PL as shown in VCF output. PL (the likelihoods of the given genotypes) is the likelihood of a GT (genotype).

- DP: this field describes the total depth of reads that passed the caller's internal quality control metrics (i.e., the depth of a high-quality read).

CHROM	POS	X06_0006e_1	X06_0006e_2	X06_0042e_1	X06_0042e_2	X06_0071e_1	X06_0071e_2	X06_0076e_1	X06_0076e_2	X06_0115e_1	X06_0115e_2
1	150737433	68	67	60	82	57	49	68	40	53	80
1	150737444	82	83	67	92	68	54	72	50	54	89
1	150737518	104	121	86	121	99	73	100	95	93	125
1	150737565	27	34	39	53	42	29	51	41	39	56
1	150761952	81	81	48	62	65	63	80	65	58	96
1	150761982	93	113	69	79	84	93	102	87	74	138
1	150762071	144	187	114	146	128	142	160	150	119	202
1	150762097	136	205	117	149	131	160	175	156	142	234
1	150762119	132	199	114	146	126	152	152	150	151	232
1	150762171	131	176	99	145	132	135	153	182	137	230
1	150779931	130	181	112	142	125	133	146	144	149	196
1	150780019	122	171	118	154	121	131	132	145	172	191
1	150780760	76	79	43	65	67	48	65	45	56	89

Figure 2.3: Output of DP (read depth) as shown in VCF output or a VCF file. The rows represent the genomic position, while the columns represent individuals with replicates.

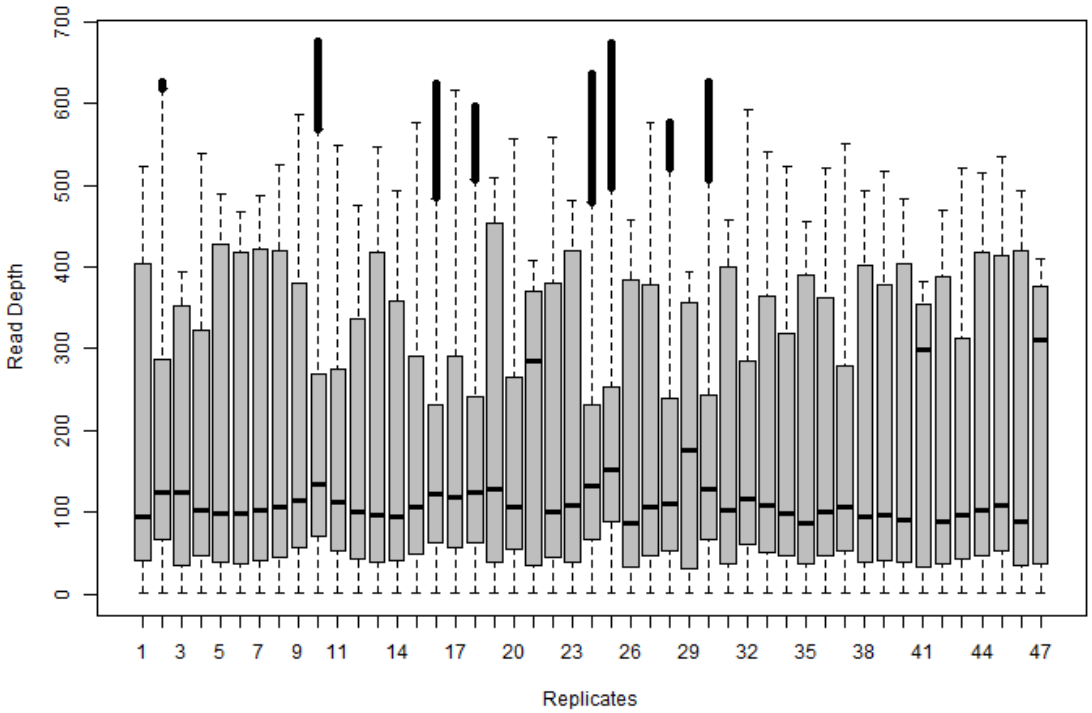


Figure 2.4: Read depth at each replicate in an individual based on the real replicate data that we have, which include rare variants in many genes, such as *PARP1* and *TYROSINASE*.

Examples of how PLs and GQs appear in VCFs are presented in Table 2.1. To understand the genotypes for $NA12878$ at $chr1 : 899282$ in Table 2.1, refer to $chr1$ 899282 $rs28548431$ C T [CLIPPED] $GT : PL : DP : GQ$ $0/1 : 1,3 : 4 : 25.92 : 103, 0, 26$. At this position (site), the genotype is called $GT = 0/1$, which presents REF and ALT alleles as C/T. The confidence indicated by $GQ = 25.92$ is low, largely because there was only a total of 4 reads at this position ($DP = 4$). The reason for uncertainty is noticeable in the PL field, where $PL(0/1) = 0$ (the normalised value that agrees with a likelihood of 1.0), but there is a possibility that the subject is homozygous with the variant allele since $PL(1/1) = 26$, which corresponds to the probability $10^{-2.6}$, or 0.0026. It is clear that the subject is not homozygous with the REF allele because $PL(0/0) = 103$; this corresponds to a

probability of $10^{-10.3}$, which is a very low number.

2.1.3 Phred Quality Score

A Phred quality score is an integer that represents the estimated probability of an error (i.e., the estimated probability that a base is incorrect). Phred quality scores are attached to every nucleotide base call in automated sequencer traces. Moreover, they are widely accepted as they characterise the quality of DNA sequences. Phred quality scores are represented by Q and are logarithmically linked to the base-calling error probability P;

$$Q = -10 \log_{10} P$$

where P is the error probability for the base. The Phred value is rounded to the nearest integer and enables a higher resolution for a low confidence call.

$$P = 10^{-Q/10}$$

A low Q score can increase false-positive variant calls. A high-quality score indicates high confidence calls. Applying the Q formula, $Q = 10$ means a 1 in 10 chance the base is wrong, and 90% accuracy of the base call and P will be 0.1. A threshold of $Q = 10$ is usually applied.

2.2 Example Dataset

Sequencing data were provided from a case-control study of melanoma, including sequence data from 61 gene subsets of 1317 cases and 697 controls. Three datasets were used. The first set is replicate data, which were sequenced twice (technological replication). The replicate data contains 47 samples; 23 individuals (22 pairs and 1 trio) were sequenced at $44,382$ sites. The data was filtered, and 2785 sites were removed due to poor quality, leaving $41,597$ sites. Most of these sites were considered rare variants due to the low minor allele frequencies; however, some are common variants. A total of $41,597$ biallelic variants, including nucleotide polymorphisms (SNPs) and insertion/deletions (INDELs), have been identified. The second set is for the region covering the gene *PARP1* on chromosome 1,

which has 547 sites. The third set is for the region covering gene *TYROSINASE* (*TYR1*) on chromosome 11; this gene has 318 sites.

2.2.1 General Notation

There are three main concepts in this chapter: genotype data, the case-control phenotype, and weight. Assume n subjects are sequenced in a region with p variant sites observed, where $n = 2014$ and p is based on the gene that we are analysing, such as the *PARP1* gene $p = 571$ or *TYR* gene $p = 318$. For the i^{th}, subject y_i denotes the phenotype variable, and $G_i = (g_{i1}, g_{i1}, \ldots, g_{ip})$ is the genotype design matrix. The elements of the genotype matrix are $g_{ij} = 0, 1, 2$ for common homozygous, heterozygous, and rare homozygous variants, respectively.

2.3 Pilot Study to Investigate Experimental Error

2.3.1 Introduction

In the example data, there are 22 individuals; each individual has two replicates, with the exception of one individual who has three. It is important to investigate the differences between replicates in each individual. Across all 23 individuals, how many loci are there where the replicates differ? We will call each individual's data the genotype of that individual.

In this section, studying and investigating the differences between the replicates will be illustrated in the first section, followed by an investigation of the missing values in the data. Table (2.3.1) shows the count of genotypes in the replicate dataset.

Type	./.	0/0	1/0	1/1	2/0	2/1	2/2	3/0	3/1	3/2	3/3
Count	69404	1743946	82315	53828	836	3419	374	21	568	311	37

Table 2.3: Counts of genotypes in the replicate dataset.

2.3.2 Notations

For $i = 1 \ldots, n$ and $j = 1, \ldots, p$, let g_{ijk} be the genotype of individual i at locus j for replicate k. There is one individual with 3 replicates, so k will be 1, 2, or 3. Note that in the replicates data, $p = 41597$ and $n = 23$. Also, for $i = 1 \ldots, n$ and $j = 1, \ldots, p$, let D_{ij} denote the matrix of the difference between replicates and M_{ij} denote the missing genotype at individual i and position j or M_{ijk} denote the missing genotype at individual i, position j, and replicate k.

2.3.3 Differences Between Replicates

In this section, we will identify differences between replicates for each individual. Any difference is a result of some experimental error. Consider a new matrix: its element is 1 when the genotype of replicate one at a position j and individual i is different from the genotype of replicate two at the same position and individual, and 0 otherwise. We denote the matrix of differences D as follows:

$$D_{ij} = \begin{cases} 0 & \text{if} \quad g_{ij1} = g_{ij2} \\ 1 & \text{if} \quad g_{ij1} \neq g_{ij2} \end{cases}$$

For the triple one, if all the replicates are the same, they will be denoted by 0, otherwise 1. The total number of differences in individual i across all positions is $d_{i.} = \sum_{j=1}^{p} D_{ij}$ for $i = 1, \ldots, n$. We use two methods to sum up the differences between the positions: either ignore or consider the missing values. Let $d_{j.}$ be the summation of differences between individuals at position j; $d_{.j} = \sum_{i=1}^{n} D_{ij}$.

Individual i	$d_{i.}$ with missing	$d_{i.}$ without missing values
1	777	283
2	863	327
3	869	306
4	827	283
5	759	287
6	809	323
7	823	281
8	731	277
9	767	282
10	849	311
11	831	302
12	782	252
13	1116	369
14	839	369
15	838	293
16	851	359
17	836	282
18	733	282
19	806	257
20	900	325
21	800	277
22	849	307
23	851	290

Table 2.4: The number of differences between the replicates with and without missing values per individual.

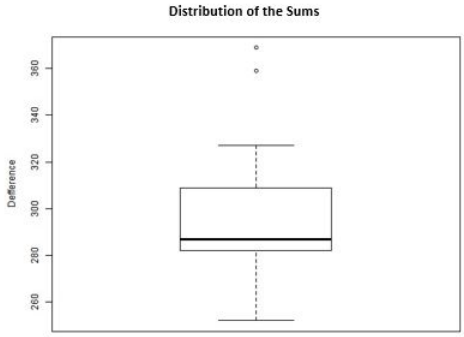


Figure 2.5: Distribution of differences for individual i. The mean number of differences per i individual is 298 (without missing values).

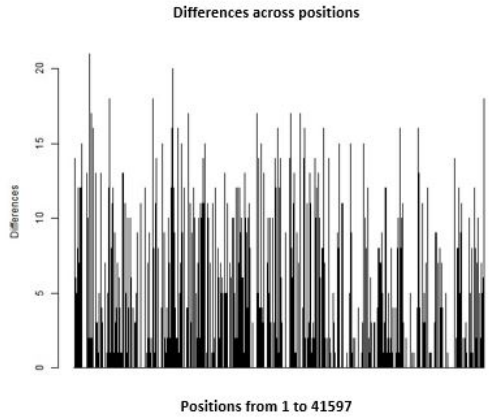


Figure 2.6: Summation of differences at j positions. For example, at position number 1, 14 out of 23 individuals have different replicates at this specific position.

Thus, the percentage of positions at which replicates differ is 0.7%. Out of a total of $956,731$ pairs (i.e., the elements D_{ij}), $6,860$ are different. The number of differences varies considerably at each position. For example, one position has different replicates for just one individual, while other positions have different replicates for more than 12 individuals. Moreover, one position differs for 21

individuals, and 438 positions differ for just one individual. Table 2.5 illustrates the frequencies of possible numbers of differences. For example, there are 40,129 positions with no differences in any individuals.

Summation of different $d_{\cdot j}$	Total Number of positions
0	40129
1	438
2	196
3	143
4	126
5	94
6	64
7	76
8	56
9	49
10	57
11	39
12	41
13	23
14	21
15	15
16	13
17	8
18	5
19	2
20	1
21	1

Table 2.5: Frequencies of differences among 23 individuals. There are 40,129 positions (96%) where all the replicates in all the individuals are the same. No replicates differ in these positions.

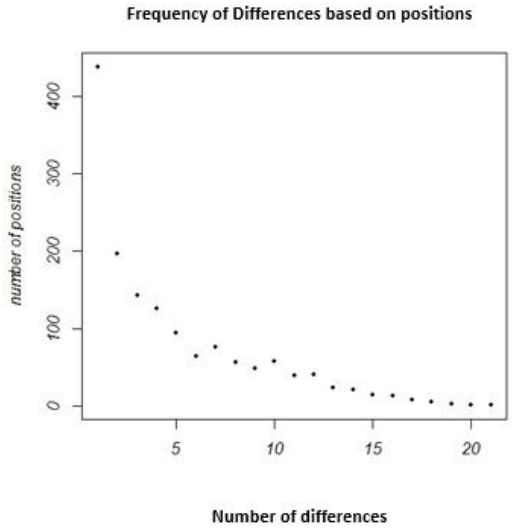


Figure 2.7: An illustration of numbers of positions versus the number of differences among individuals. For example, there are 438 positions with only one observed individual; their replications differ from one another. There is also one position with 21 individuals; the replications of the individuals differ from one another.

2.4 Missing Values

In VCFs, there are many types of genotypes; one type includes symbols, such as "./.", which indicate missing values. The data's genotypes, including the missing values, are shown in Table 2.6.

Type	./.	0/0	1/0	1/1	2/0	2/1	2/2	3/0	3/1	3/2	3/3
Count	69344	1743841	82132	54143	834	3426	383	24	574	322	36

Table 2.6: A summary of genotypes in the replicate data.

The missing values can be considered two ways, or with two matrices. In the first, which is based on 47 replicates, every individual has two copies, so the dimensions of the matrix are $41,597$ rows (positions of SNPs) and 47 columns corresponding to copies for individuals. The second matrix is based on 23 individuals; its dimensions are 41,597 rows and 23 columns (individuals). In Table 2.6, there are $69,344$ genotypes with the value "./.", which means the genotype values are missing, based on the first matrix, which is a replicate one. At a position with this value, one cannot say whether the genotype is homozygous or heterozygous. There are $2,741$ positions out of $41,597$ with at least one missing value, which is 6% of the positions. Based on the individual matrix (the missing values for both replicates), the number of cells with missing values is $40,054$ out of $956,731$, or 4%. There are $2,444$ positions out of $41,597$ with at least one missing value (6% of the positions).

Let M_{ij} be a missing matrix with a binary value. Let1 indicate a missing value, and let 0 indicate a value that is not missing. For $i = 1, 2, \ldots, n$ individuals, there are $j = 1, 2, \ldots, p$ variants (positions) and $k = 1, 2$ replicates:

$$M_{ijk} = \begin{cases} 1 & \text{if} \quad g_{ijk} = \text{"./."} \\ 0 & \text{if} \quad g_{ijk} \neq \text{"./."} \end{cases}$$

For replicate k of individual i, let $m_{i.k}$ be the total number of missing values in each replicate, so $m_{i.k} = \sum_{j=1}^{p} M_{ijk}$.
The total number of missing values in each replicate is shown in table 2.7.

Replicate 1	Total of Missing $m_{i.k}$	Replicate 2	Total of Missing $m_{i.k}$
1.1	1357	1.2	1353
2.1	1518	2.2	1388
3.1	1567	3.2	1648
4.1	1563	4.2	1541
5.1	1491	5.2	1465
6.1	1401	6.2	1469
7.1	1534	7.2	1494
8.1	1411	8.2	1341
9.1	1426	9.2	1445
10.1	1418	10.2	1500
11.1	1506	11.2	1453
12.1	1507	12.2	1359
13.1	1413	13.2	1530
14.1	1324	14.2	1546
15.1	1360	15.2	1561
16.1	1188	16.2	1396
17.1	1435	17.2	1587
18.1	1489	18.2	1452
19.1	1407	19.2	1486
20.1	1612	20.2	1669
21.1	1502	21.2	1479
22.1	1569	22.2	1479
23.1	1542	23.2	1675
Total $m_{..k}$	69344		

Table 2.7: Total of missing values in each replicate

Table 2.8 shows the number of missing values for each individual. Note that a position is counted as having a missing value if one or both replicates are missing.

Individuals i	The total of missing values in i individuals
Ind 1	1602
ind 2	1721
ind 3	1889
ind 4	1824
ind 5	1714
ind 6	1678
ind 7	1785
ind 8	1603
ind 9	1678
ind 10	1728
ind 11	1744
ind 12	1698
ind 13	1725
ind 14	1711
ind 15	1733
ind 16	1538
ind 17	1788
ind 18	1696
ind 19	1712
ind 20	1928
ind 21	1752
ind 22	1795
ind 23	1889

Table 2.8: Total of missing values for each individual

2.4.1 Relationship Between Read Depth and Missing Values

Read depth, or the depth in a genomic position, is equal to the number of reads aligned to a position. It describes the total depth of reads that the pass the internal caller quality check (i.e., the depth of coverage of each position for each sample). When there is no read at position j for individual i, ($DP = 0$). It is

considered a missing value, which is denoted in VCF as "./.".

2.4.2 Modelling Differences using Logistic Regression

This section presents model differences and read depth. As the chapter's focus is on read depth, the data is fit with a summation of replicate read depths, and square roots. The predictors for the model are the differences that occur between replicates at given positions.

Let r_{ijk} be the read depth for individual i, position j, and replicate k.

$$\text{Summation} = \sum_{j=1}^{p=41597} \sum_{i=1}^{n=23} (r_{ij1} + r_{ij2})$$

$$\text{Square Root} = \sum_{j=1}^{p=41597} \sum_{i=1}^{n=23} \sqrt{r_{ij1} \times r_{ij2}}$$

The Response is the Difference Between Replicates

Let the indicator for the presence of a difference between replicates at a given position for a given individual be the response variable y. It will be assumed that each position is independent of the others. This assumption may not be accurate, but it will be used for simplicity. Let D_{ij} denote the value 0 or the value 1 for individual i at position j, as defined in 2.3.3. The matrix is

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1p} \\ d_{21} & d_{22} & \cdots & d_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{np} \end{pmatrix}$$

where $p = 41597$ is the total number of positions, and $n = 23$ is the total number of individuals. Since the response is binary, to use a generalised linear model to fit the model, this matrix is converted to a vector y.

$$y = \begin{pmatrix} d_{11} \\ d_{12} \\ d_{13} \\ \vdots \\ d_{1p} \\ \vdots \\ d_{2p} \\ \vdots \\ d_{np} \end{pmatrix}$$

Significant results appear when one fits y with the summation and the square root of the product as explanatory variables. The final model is y with the summation and square root of the product.

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-3.2552	0.0157	-207.13	2e-16
summation	-0.0143	0.0002	-81.87	2e-16
square root	0.0014	0.0000	38.59	2e-16

Table 2.9: The output of fitting a logistic regression model, where the response is disagreement between replicates, and the covariates are the summation and the square root of read depth.

This model suggests that the summation and square root of the read depth are predictive of difference (disagreements) between replicates. It can be concluded that as the differences between replicates increase, the read depth decreases.

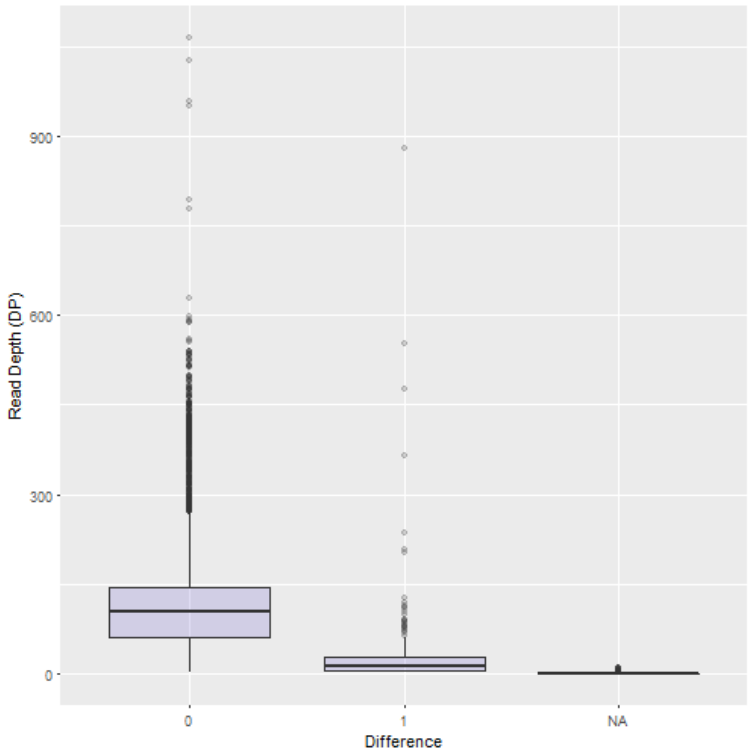


Figure 2.8: The summation of read depths between replicates at given position j and individual 2 versus disagreement between replicates. In this example, we choose individual 2 with its replicates.

2.5 Conclusion

In this chapter, VCF is introduced, and examples are given using real data. An exploratory analysis shows it can be concluded that missing values from the VCF examples are associated with read depth. If there is no read at a specific genomic position, then there is a missing value. In addition, the analysis shows that there are relationships between replicate differences (disagreements) and read depths; when read depths decrease, differences (disagreement) increase.

Chapter 3

Association Testing for Rare Genetic Variants

3.1 Methods for Testing Rare Variants

Advances in NGS technologies have provided unprecedented opportunities to discover rare variants and evaluate their effects on disease risks and trait variations, as rare variants are important when studying complex human diseases and traits. However, when allele frequencies are very low, the likelihood of observing rare variants in study samples is small. The resulting lack of variation in data usually causes statistical tests of association to be significantly underpowered.

Recently, studies of rare variants have proposed different testing strategies for genotype and phenotype associations that are based on aggregating association information across multiple SNPs into single tests. It is generally recognized that a good strategy for analysing rare variants is combining them into units of association. The purpose of aggregation is to enrich association signals and reduce penalties that result from conducting multiple tests.

There are two categories of the most common gene-level association tests. The first is burden tests, which are linear tests that detect specific associations if all variants are in one direction. They were designed for detecting associations of genotypic burden scores summarized from sets of rare variants. Burden tests are used by Morgenthaler & Thilly (2007), Li & Leal (2008), Morris & Zeggini (2010), Madsen & Browning (2009), and Price *et al.* (2010). The second category

is quadratic tests, which can be used across a wide range of genetic models. The tests keep individual rare variants as individual variables and evaluate whether at least one multiple rare variant is associated with a trait. Quadratic tests are used by Neale *et al.* (2011), Wu *et al.* (2011), and Lee *et al.* (2012a). This paper theoretically and empirically examines both classes of tests and provides several new insights concerning them. Additionally, novel simulation studies are conducted that complement the empirical investigation and illuminate comparisons of the methods.

The key feature shared by the two types of tests is that they test collective rather than individual effects of multiple rare variants as entire groups. Therefore, once associations between groups of rare variants are identified, further analyses are sometimes required to determine which variants cause the associations. Indeed, there are many methods for detecting rare variant associations, categorize in three main categories, shown in Table 3.1.

Association methods		
Category	Description	Tests
Burden Test	Collapses rare variants into genetic scores	CAST, CMC, WST
Variance Component Test	Tests variance of genetic tests	SKAT, SSU, C-alpha
Combined Test	Combines Burden and Variance Component Tests	SKAT-O, Fisher Methods

Table 3.1: A summary of the methods used for detecting rare variant associations.

Burden tests, which are also called linear tests, combine rare variant information into single scores or variables. Quadratic tests use score functions and evaluate distributions of the genetic effects instead of combining groups of variants (Lee *et al.* (2014)). In this chapter, popular tests that fall under the category of a burden or quadratic test are discussed:

- Weighted sum statistic tests (WST)

- Replication-based strategies Test (RBT)
- C-alpha test
- Sequence kernel association tests (SKAT)

Sections 3.3 and 3.4 discuss the mathematical details, as well as insights into the advantages and limitations, of each method.

3.2 Data Description and Model

We consider a case-control study with a total sample size of n. Assuming we test the association between genetic variants and phenotype in a candidate region or gene which includes p SNPs in n individuals. For individual i, let $y_i = 1$ or $y_i = 0$ denote a case or control, respectively, with $i = 1 \ldots, n$, let μ_i be the mean of y_i, $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ represent allele counts $x_{ij} \in \{0, 1, 2\}$ (assuming additive trait model) for p variants of interest, where $j = 1, \ldots, p$, and $\mathbf{z}_i^T = (z_{i1}, \ldots, z_{ik})$ represent covariates, and where $k = 1, \ldots, q$. We assume that y_i follows a distribution in the exponential-likelihood family and consider the following generalised linear model:

$$h(\mu_i) = \beta_0 + \mathbf{z}_i^T \boldsymbol{\alpha} + \mathbf{x}_i^T \boldsymbol{\beta}, \tag{3.1}$$

where $h(\mu_i) = logit(\mu_i)$ for a binary phenotype, β_0 is an intercept, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are regression coefficients for the covariates $\mathbf{z}_i^T$ and allele counts $\mathbf{x}_i^T$, respectively.

We test the association between genotype and phenotype where the null hypothesis is:

$$H_0 : \beta = \beta_1 = \cdots = \beta_p = 0 \tag{3.2}$$

In the simulated study, we generate a number of variants. We consider a sample of 1000 cases and 1000 controls. For each model, we simulate $1,000$ datasets. We assume a multiplicative trait model with the odds ratio (OR) and frequency of the minor allele at the causative SNP equal to m. We consider m in

the 1–3 range and a multiplicative model with an OR ranging from 1 (for the type I error rate) to 3. When the OR is greater than 1, we consider the SNP to have a direct effect on the disease, so we call it a causal variant or SNP. Simulation of the SNPs assumed Hardy–Weinberg equilibrium (HWE).

The simulated LD patterns are defined by $\rho_{\grave{i}j}$, where $\grave{i}$ and j are the location index of markers and the trait locus on the gene, respectively. The patterns of LD in the cases and controls are the same. We consider tow scenarios, namely $\rho_{\grave{i}j} = 0.5^{|\grave{i}-j|}$, or ρ, is randomly sampled from a uniform distribution between 0.3 and 0.9. There are 100 SNPs per gene in the simulation.

3.3 Burden Test

The burden, or linear, test creates a burden score for each subject by collapsing the SNPs with minor allele frequencies (MAFs) below a particular threshold and relates the score to the trait of interest. For example, the combined multivariate and collapsing (CMC) method splits the variants into subgroups based on their MAFs and collapses them within each subgroup. The burden test collapses information for multiple genetic variants into a single genetic score Asimit *et al.* (2012), Morgenthaler & Thilly (2007), Morris & Zeggini (2010).

3.3.1 Cohort allelic sums test (CAST)

The Cohort Allelic Sums Test (CAST (Morgenthaler & Thilly (2007)) is one of the first tests based on a collapsing technique for rare variants, which involves collapsing genotypes across rare variants to create a super variant (random vector):

For simplicity, we will present score values or the genetic score summary as

$$C_i = \sum_{j=1}^{p} t_j x_{ij}, \tag{3.3}$$

where t_j is a threshold indicator which will be equal 1 in CAST, C_i is the score value, and x_{ij} is the genotype matrix; its elements belong to $(0, 1, 2)$.

Suppose all rare variants have the same effect, i.e.

$$\beta_1 = \beta_2 = \ldots = \beta_p = \beta$$

Then:

$$\mathbf{x}_i^T \boldsymbol{\beta} = (\underbrace{x_{i1} + x_{i2}, \ldots + x_{ip}}_{C_i})\boldsymbol{\beta}$$

So, our new model is:

$$h(\mu_i) = \beta_0 + \mathbf{z}_i^T \boldsymbol{\alpha} + \mathbf{C}_i^T \boldsymbol{\beta}, \tag{3.4}$$

$$\begin{cases} C_i = 1 & \text{if any } x_{ij} > 0 \quad \text{(any rare variant present)} \\ C_i = 0 & \text{otherwise} \end{cases}$$

In CAST, it is assumed that the appearance of any rare variant increases disease risk, and the genetic score is set to $C_i = 0$ given no minor alleles in a region, and $C_i = 1$ otherwise (see Table 3.2). Then, the test is performed to detect the association between a phenotype and new C_i. A genetic score summary (3.3) tests the association between a phenotype and C_i, which will be used in upcoming sections.

y	x_1	x_2	x_3		C
1	1	0	0		1
1	1	0	1		1
.	.	.	.		.
.	.	.	.		.
0	0	0	0		0
0	0	0	0		0

Table 3.2: Illustration of the pooling the SNPs in CAST.

3.3.2 Combined multivariate and collapsing (CMC) tests

This test is a modification of CAST with an extension to improve its power when both rare and common variants are present. Therefore, if the data only has rare variants, then CMC will be the same as CAST. Like CAST, CMC collapses

rare variants but in different categories of MAF and evaluates the joint effect of both common and rare variants by using Hoteling's t-test, which is based on a multivariate distribution comparable to the F-distribution. As a generalization, the combined multivariate and collapsing (CMC) method splits the variants into subgroups based on their MAFs and collapses the variants within each subgroup Li & Leal (2008). We can express this step of the test as follows:

- Group variants based on their MAFs $\mathcal{F}$.
- Collapse each group using CAST approach.
- Perform Hoteling's t-test.

3.3.3 Weighted sum test (WST)

The Weighted sum statistic test (WST) is a burden test proposed by Madsen & Browning (2009) that compares the number of mutations in a group of RVs between a sample of affected individuals (case group) and an unrelated group of unaffected individuals (the control group). For each variant j, the rarer allele is considered the mutation, and the weight is calculated based on the MAF. Since causal variants often have a higher MAF in cases and a lower MAF in controls, and thus, a higher combined MAF across cases and controls, this method may down-weight the causal variants, which will reduce the test's effectiveness. Hence, Madsen & Browning (2009) suggests that only the MAFs of controls should be used to calculate weights. The weight used in this test is the inverse of the variance of (MAF) in controls, and then summed the weighted rare variants. The weighted-sum test (WST) of Madsen & Browning (2009) uses the Wilcoxon rank-sum test and obtains p values by permutation.

Let n_0 be the number of individuals in the control group, n_1 the number of individuals in the case group, and n the total number of individuals. Then, define the weight as

$$w_j = \sqrt{n_{.j} q_j (1 - q_j)}, \tag{3.5}$$

where

$$q_j = \frac{m_{0j} + 1}{2n_{0j} + 2}, \tag{3.6}$$

and m_{0j} is the number of mutant alleles observed for variant j in the unaffected individuals (controls); n_{0j} is the number of unaffected individuals genotyped for variant j, and $n_{.j}$ is the total number of individuals (affected and unaffected) genotyped for variant j Madsen & Browning (2009). Clearly, q_j is the estimated MAF from the control group only.

The genetic score of each individual i is calculated as

$$L_i = \sum_{j=1}^{p} \frac{x_{ij}}{\hat{w}_j} \tag{3.7}$$

All individuals are ranked according to their genetics score L_i. The sum of the ranks of affected individuals is then calculated:

$$T_{wst} = \sum_{i \in A} rank(L_i),$$

where A is the population of cases (affected individuals). However, T_{wst} can be re-written to use the genetic score summary 3.3 and be consistent with other methods in notation:

$$T_{wst} = \sum_{i \in A} rank(C_i),$$

where C_i is the genetic score summary in (equation 3.3), and t_j is equivalent to $1/w_j$ in the WST.

Under the null-hypothesis and the assumption that the genotypes of the affected individuals are independent of each other, T_{wst} is the sum of n_1 independently and identically distributed (iid) random variables. Based on the central limit theorem, it is approximately normally distributed since n_1 is typically large. The affected and unaffected status is permuted among the total population k times to obtain a sample $(T_1, T_2, \ldots, T_k)$; here, T is the T_{wst}. The sample's mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ are then calculated, and the standardized score-sum is found as

$$z = \frac{T_{wst} - \hat{\mu}}{\hat{\sigma}}.$$

z follows an approximately standard normal distribution under the null hypothesis. A p-value for the association can be obtained by comparing z to the quantiles of the standard normal distribution.

3.3.4 Replication-based strategy (RBT)

Ionita-Laza *et al.* (2011)proposes a replication-based strategy (RBT) based on a weighted sum of statistics and partitioning variants observed among cases and controls into distinct groups according to the respective observed frequencies of the minor allele; (m_0, m_1) denotes the group containing all variants that have exactly m_0 copies of the minor allele in the controls and exactly m_1 copies of the minor allele in affected cases. Let $m_{m_0}^{m_1}$ be the size of the group (i.e, number of variants) (m_0, m_1). Since we are interested in identifying risk variants, we will consider only groups with $m_1 > m_0$. Table 3.3 shows the different groups (Ionita-Laza *et al.*, 2011).

m_0/m_1	1	2	3	...
0	m_0^1	m_0^2	m_0^3	...
1		m_1^2	m_1^3	...
2			m_2^3	...

Table 3.3: Variants are classified according to the number of times they appear in controls (m_0) and in cases (m_1). Only variants that appear more frequently in cases than in controls or those more likely to be risk variants are shown.

We define the following weighted-sum statistic S where each variant in group (m_0, m_1) is assigned a weight $w_{m_0}^{m_1}$:

$$S = \sum_{m_0=0}^{N_r} \sum_{m_1 > m_0} m_{m_0}^{m_1} w_{m_0}^{m_1} \tag{3.8}$$

where N_r is the upper threshold of the number of occurrences of a variant amongst the controls.

Madsen & Browning (2009) use data-dependent weights, with

$$w_{m_0}^{m_1} = \frac{m_1}{\sqrt{q(1-q)}}$$

where $q = \frac{m_0+1}{2n_0+1}$ is the estimated frequency based on controls only, and n_0 is the number of controls.

Since the number of changes (mutations) at a rare variant position follows an approximate Poisson distribution, the probability $P(m_0, m_1)$ of observing a variant position with at most m_0 mutations in controls and at least m_1 mutations in cases under the null hypothesis can be calculated by a probability mass function of the Poisson distribution as

$$P(m_0, m_1) = \frac{e^{\hat{f}}\hat{f}^{m_0}}{m_0!} \times \frac{e^{\hat{f}}\hat{f}^{m_1-1}}{(m_1-1)!}.$$

where $\hat{f} = \frac{m_0+m_1}{2}$ is the estimated SNP frequency based on the observed number of appearances in both cases and controls, and $P(m_0, m_1)$ is the Poisson distribution function with parameters m_0 and m_1.

Note that, as the observed frequency in cases compared to controls increases (i.e., as $m_1 - m_0$ increases), the weight increases, and hence, S also increases. We can evaluate the significance of S by applying a standard permutation procedure to permute the control/cases case labels randomly, thereby quantifying the significance of which S is higher than expected under the null. The power of RBT is less sensitive to the direction of variant effects in a genetic region of interest (Ionita-Laza *et al.*, 2011).

3.3.5 The limitations of linear tests

The main limitation of linear tests is they assume that all tested variants influence the phenotype in the same direction. Also, to achieve reasonable power, the tests require large proportions of causal variants. Burden methods rest on the assumption that all rare variants in a set of a group are causal variants associated with a phenotype with the same effect direction. When this assumption is violated, it results in a substantial loss of power (Rivas *et al.* (2011), Basu & Pan (2011), Lee *et al.* (2014)). Moreover, some burden tests only use qualitative information such

as CAST, CMC, WST and C-alpha, and cannot use quantitative traits; however, most of these tests can be easily extended to quantitative traits such as WST. Therefore, there is no adjustment for covariates for most of them such as CAST, CMC, WST and C-alpha since they use pooling strategy, however, they can be extended to incorporate covariates such as C-alpha test (Wu *et al.* (2011)).

3.4 Quadratic Tests

Based on the marginal model (3.1) for variant j, a 'score statistic' can be defined as

$$U_j = \sum_{i=1}^{n} x_{ij}(y_i - \mu_i). \tag{3.9}$$

where μ_i is the estimated mean of $\mathbf{y}_i$ under the null hypothesis (3.2) and obtained by the null model $h(\mu_i) = \beta_0 + \mathbf{z}_i^T \boldsymbol{\alpha}$. When disease risk is increased, the value of U_j will be positive, and when variant j is associated with decreased risk, the value of U_j is negative. The derivation of these score statistics is presented in a later chapter. Quadratic tests can be used to overcome directional issues, which are challenging in linear statistics. Therefore, covariates must be included in some quadratic tests since the covariates can be controlling population stratification, which is important in genetic association studies.

3.4.1 C-alpha tests

C-alpha is a well-established and powerful test for the presence of a mixture of biased and unbiased coins (Neyman & Scott (1965), Zelterman & Chen (1988)). Neale *et al.* (2011) proposes the C-alpha score test and applies it to testing the degree of association in a group of rare variants. Under the assumption that the rare variants are randomly distributed across the subjects, the probability of observing a specific variant m_1 times in the cases out of m total can be evaluated by the binomial (m, p) distribution. This distribution (m, p) evaluates the probability of observing a particular variant in m_1 affected cases amongst a total population of m. Under the balanced sample of cases and controls, it means that $p = 0.5$ and m_1 is $0, 1$, and 2 for $m = 2$ are expected with probabilities of

0.25, 0.5, and 0.25, respectively. We typically will observe a higher proportion of doubletons with $m_1 = 2$ and/or $m_1 = 0$ than expected, if some variants are detrimental or protective. Because each variant cannot provide sufficient information to draw a firm conclusion about the association, the C-alpha test is applied to detect a pattern across the full collection of rare variants in the target region.

If the region being investigated has no alleles associated with a specific phenotype, then the counts should follow a binomial distribution with m being equal to the number of copies of the observed variant. Assuming that rare variants are distributed randomly across all individuals, a C-alpha test can be used to identify the presence of a pattern across the full set of rare variants in the target region. For a j^{th} variant observed m_j times, we assume that m_{1j} follows the binomial distribution (m_j, p_j) under the null hypothesis $H_0 : p_j = p_0$ where $p_0 = \frac{m_1}{m}$. The C-alpha test statistic T compares the variance of each observed count with its expected variance under the assumption of a binomial distribution.

$$T = \sum_{j=1}^{m} [(m_{1j} - m_j p_o) - m_j p_o (1 - p_o)]$$

where m_{1j} is the number observations of a j^{th} variant in affected cases out of a total of n individuals (i.e., the number of copies of the j^{th} variant type in the affected cases, and m_j is the number of copies of the j^{th} variant type).

To standardize the test statistic, we require c, the variance of T:

$$c = \sum_{n=2}^{\max n} m(n) \sum_{u=0}^{n} [(u - np_o)^2 - np_o(1 - p_o)]^2 f(u|n, p_0)$$

where $m(n)$ is the number of variants with n copies, and $f(u|n, p_0)$ is the probability, assuming a binomial model, of observing u copies of a j^{th} variant.

The resulting standardized test statistic is $Z = T/\sqrt{c}$. We reject the null hypothesis if Z is larger than expected, using a standard one-tailed normal distribution as a reference (Neale *et al.*, 2011).

The C-alpha method is only appropriate with qualitative traits since it depends on testing homogeneity for a set of binomial proportions rather than logistic regression. It cannot be adjusted to covariate, and no common variants are included in the test since there is no weight in the test. Therefore, if we included common variants, they would dominate the signal from rare variants.

3.4.2 Sequence kernel association test (SKAT)

Sequence kernel association tests (SKAT) are a flexible approach to identifying associations between variants in a region (both rare and common) and a dichotomous (i.e., case-control) phenotype that also allows for covariate adjustment, for example, to account for population stratification. SKATs analytically calculate a p-value for each region while accommodating for covariates. Adjustments for multiple comparisons are required for analyzing multiple regions, for example, with the Bonferroni correction or FDR control. SKATs use a multiple-regression model to directly regress the phenotype on both genetic variants in a given region and covariates and allows different variants to have different magnitudes: positive, negative, and zero. SKAT does not require the selection of thresholds (Wu *et al.*, 2011).

SKAT was developed using a variance-component score test in a mixed-model framework by considering rare variants. It was developed to test the regression coefficients of variants.

Recall model (3.1); SKATs test H_0 by assuming that each $\boldsymbol{\beta}_j$ follows a random distribution with a mean of 0 and variance of ;$w_j\tau$, where τ is a variance component and w_j is a pre-specified weight for the variant j. The null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ is equivalent to testing $\tau = 0$.

Assume that n individuals are sequenced in a region where p variant sites have been observed. For the i^{th} subject, y_i denotes the phenotype variable and $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip})$ are the genotypes for the p variant sites. We assume an additive allelic model, letting $x_{ij} = 0$, 1, or 2 denote the number of copies of the minor allele, although dominant and recessive models can also be considered. Then, we can introduce the test, which is a variance component test.

The score test's advantage is that the null model $P(y_i = 1) = \beta_0 + \mathbf{z}_i^T\boldsymbol{\alpha}$ only needs to be fitted for dichotomous traits. Specifically, the variance-component score statistic is

$$Q = (\mathbf{y} - \boldsymbol{\mu})^T K (\mathbf{y} - \boldsymbol{\mu})$$

where the kernel $K = XWX^T$, $\boldsymbol{\mu}$ is the predicted mean of y under H_0; that is, $\hat{\boldsymbol{\mu}} = logit^{-1}(\alpha_0 + \mathbf{z}_i^T\alpha)$ for dichotomous traits. We estimate α_0 and α under the null hypothesis by regressing y only on the covariates Z. Here, X is an $n \times p$

matrix with a $(i,j)^{\text{th}}$ entry equal to the genotype of variant j of subject i, with $W = diag(w_1, \ldots, w_p)$ containing the weights of the p variants.

$K = \sum_{j=1}^{p} w_j x_{ij} x_{i'j}$. $K(.,.)$ is called the *weighted linear kernel function*, which is a measure of the similarity between i^{th} and i'^{th} subjects in the region using the p markers. (There are also two choices of kernels to model epistatic effects.) Q follows a mixture of chi-square distributions which can be approximated with the method of Davies or by Satterthwaite method as a scaled chi-square distribution, $k\chi^2_v$, in which the scale parameter,k, and the degrees of freedom, v, are calculated via moment matching. A good choice of weights can improve power. Each weight w_j is pre-specified, using only genotype information, and no information about the outcome.

Wu *et al.* (2011) set $w_j \sim Beta(MAF_j, a_1, a_2)$. The beta distribution density function with pre-defined parameters a_1 and a_2 is evaluated at the sample MAF (across both cases and controls) for the j^{th} variant in the data. If rarer variants are expected to have larger effects, then setting $0 < a_1 < 1$ and $a_2 > 1$ will allow for up-weighting rarer variants and down-weighting more common weights. (Wu *et al.*, 2011) recommends setting $a_1 = 1$ and $a_2 = 25$ as this up-weights rare variants while still placing substantial non-zero weights for variants with MAF $1 - -5\%$ (Wu *et al.*, 2011).

3.5 Numerical Power Comparisons

For this chapter, simulation studies were conducted to examine the performances of burden and quadratic tests. Comparisons were made between the CMC, $CAST$, WST, and $C - alpha$ tests in terms of proportions of causal variants with different MAFs, as well as the inclusion and exclusion of rare variants (large frequencies) in the simulated data, the inclusion and exclusion of common variants in the simulated data, and, finally, different directions. These comparisons illuminate differences between the two types of tests and rare variant associations. We will not cover all the explained test above, we only cover some from each category (burden and non-burden), since our focus is just to illustrate the difference between categories not all methods.

3.5.1 Simulation

Genotype data for $2,000$ individuals were generated for each individual either 200 or 100 SNPs were simulated. These variants varied in the number of associations and direction of effects. The numbers of extremely rare and moderately rare variants, which have MAFs of $0.0005 - 0.005$ and $0.005 - 0.01$, respectively, also varied; the default percentages are 40%, 40%, and 20% for ERV, MRV, and CV, respectively. The percentages may have changed, and sometimes, we excluded the common variants. This will be specified in the captions of the figures. The effect size of causal variants varied from strong to low and was represented by an odds ratio, so $OR = 3$ represented a large effect, and $OR = 1.5$ represented a low effect. Also, effects were considered with different directions, namely $OR = 0.3$ and $OR = 0.2$ represented large and small effects with opposite directions. Additionally, relationships between MAFs and effects were considered as in Figure 3.4. Each rare variant had a mutation rate or MAF uniformly distributed between 0.0005 and 0.01, while each common variant had a mutation rate or MAF uniformly distributed between 0.05 and 0.5.

To obtain the genotype matrix X, $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_p)$ was generated using a multivariate normal distribution with a variance of 1 and a pairwise correlation of $\mathbf{z}_{\grave{i}}$ and $\mathbf{z}_j$ at $0.5^{|\grave{i}-j|}$; $1 \leq \grave{i}, j \leq p$ between any two latent components. The simulated LD patterns are defined by ρ_{ij}, where $\grave{i}$ and j are the location index of markers in the gene, respectively. We consider two scenarios namely $\rho_{\grave{i}j} = 0.5^{|\grave{i}-j|}$ or ρ is randomly sampled from a uniform distribution between 0.3 and 0.9. A threshold was noted for each latent vector component to obtain a vector of a binary variable, for example, $(\mathbf{d})$, which represented haplotypes. Two vectors of haplotypes $\mathbf{d}_1$ and $\mathbf{d}_2$ were generated for each individual. Then, the two independently generated haplotypes (d) were combined by taking the sum $\mathbf{x} = \mathbf{d}_1 + \mathbf{d}_2$ with the vector X $(0/1/2)$, which represents a genotype. The threshold for component j, for example, c_j, was obtained so that $P(d = 1)$ was controlled to mimic rare or common variants. Finally, given the odds vector β, the disease status Y $(0/1)$ was generated for each X, such that

$$\text{Logit}Pr(Y = 1) = \frac{e^{Pr(Y=1)}}{1 + e^{Pr(Y=1)}} = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta},$$

where $\beta_0 = log(0.05/(1-0.05))$. For the null case, $OR = 1$ was set; for non-null cases (causal cases), $OR = 3$ was set for extremely rare variants, $OR = 2$ was set for moderately rare variants, and $OR = 1.5$ was set for common variants.

Each allele of a haplotype is generated by dichotomizing the marginal normal distribution, and the cut-off was determined by an allele frequency that randomly sampled from a uniform distribution between 0.0005 and 0.005 as extremely rare variants, 0.005 and 0.01 as moderately rare variants, 0.01 and 0.05 as large moderately rare variants, and 0.05 and 0.5 as common variants.

Throughout the simulations, we fixed the test significance level at $\alpha = 0.05$. The results were based on 1000 independent replicates for each set-up. Note that this simulation's settings are explained in detail in Chapter 5.

3.5.2 Results

Burden (WST, CMC, and CAST) and non-burden (C-alpha) tests were compared; the types of burden tests were not. Based on numerical calculations, it was found that burden tests outperformed quadratic tests if all or almost all the SNPs were causal variants. Therefore, burden tests outperformed the quadratic tests when causal variants were in the same direction, regardless of whether they are protective or deleterious SNPs.

However, the burden test performed poorly when there were few causal variants under consideration and when causal variants had in different directions. Nonetheless, burden tests performed well when percentages of the causal variants were large compared to non-causal variants.

Therefore, CMC and $CAST$ performed poorly when causality was found in extremely rare variants, such as in the extreme MAF range $[1/n - 5/n]$), even though the OR was relatively large. The $C - alpha$ test performed better when there were only rare variants in the study's data. Further, the WST was better regarding the inclusion of common variants, which was due to the weighting scheme (see Figure 3.1).

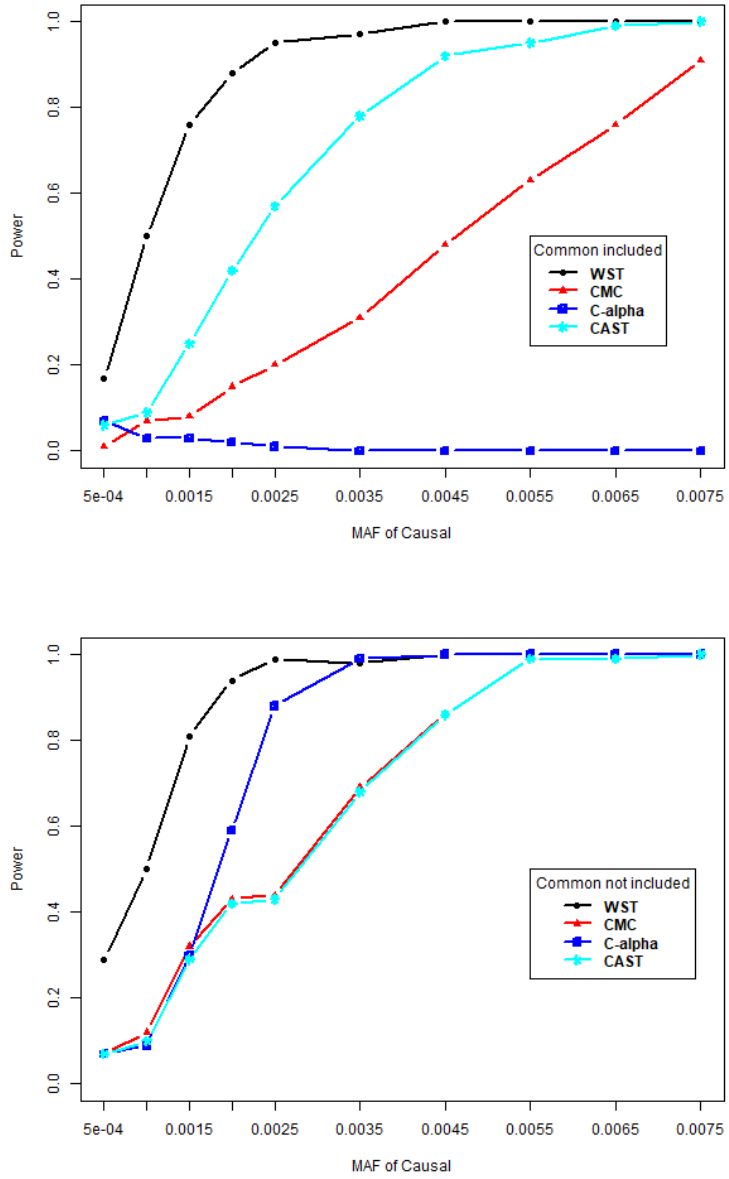


Figure 3.1: The impact of MAFs of causal on statistical power. A total of 100 variants were generated with MAFs between $0.0005 - 0.01$, including 15% causal variants in the simulated data with fixed effect size (OR=3) . Then we increased MAFs of causal variants which appear on the X-axis. The top figure includes common and rare variants with large MAFs $(0.01 - 0.5)$, but the bottom figure does not include them.

The $C - alpha$ test performed very poorly when the data contained common variants or rare variants with large frequencies between $(0.01 - 0.05)$ because no weight was included in the test. Therefore, the values of common variants dominated signals of association in rare regions. The CMC method also had a reduction of the power when the data included common variants, but the reduction was not as large as that of the $C - alpha$ test (see Figure 3.1). $CAST$ performed well when there were common variants; however, it performed poorly when there were large proportions of moderately rare variants, regardless of the effects and MAFs of causal variants (see Figure 3.2).

The CMC and $CAST$ methods performed poorly when causal variants were in the extreme MAF range. However, an improvement was seen when the number of causal variants increased significantly (see Figure 3.2). Nonetheless, the WST performed better, especially when the number of extreme causals was large, non-causal variants randomly had rare MAFs $(0.0005, 0.01)$, and no common variants were included (see Figure 3.2). Increasing the number of causal variants increased the power of the WST, CMC, and $C - alpha$ tests.

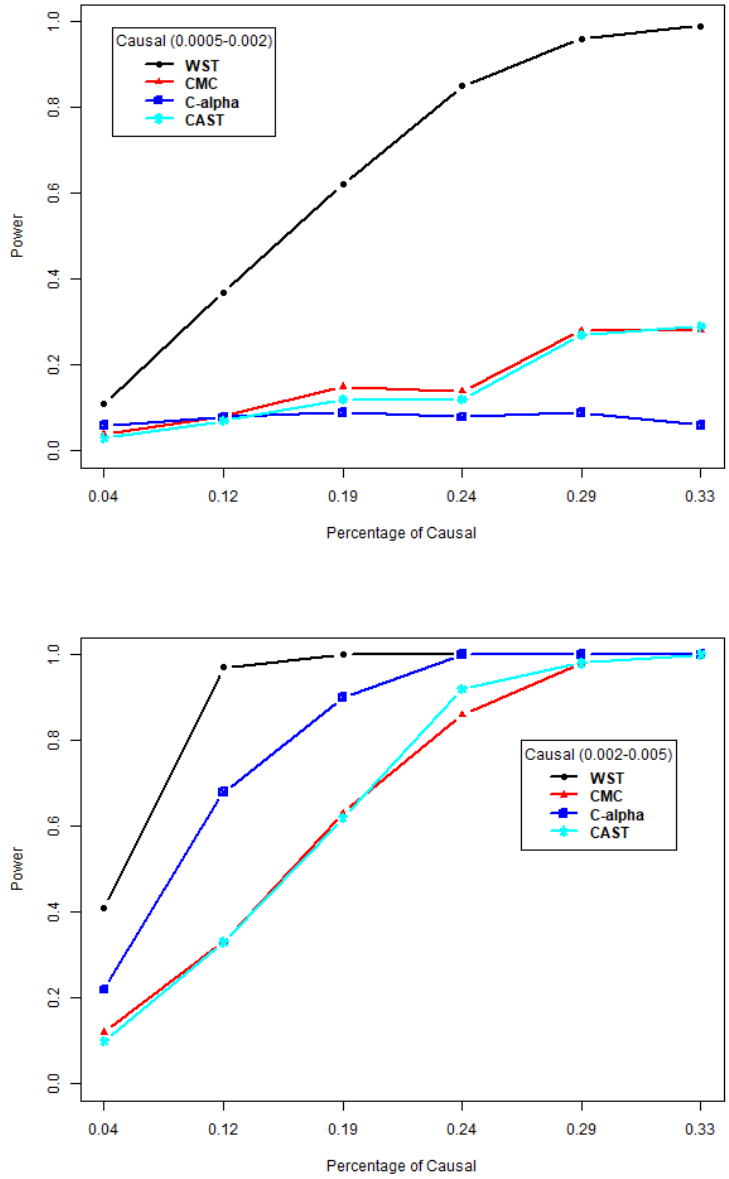


Figure 3.2: The impact of the increasing the percentage of causal variants on the power. A total of 100 variants were generated with MAFs from $0.0005-0.01$, and causal variants increased on the X-axis. Fixed MAFs between $0.0005-0.002$, which are very low (i.e., extremely rare variants), are shown in the top figure, and MAFs of $0.002-0.005$ are shown in the bottom figure. The size effect is fixed (OR=3).

When the data contained large moderately rare variants $(0.01, 0.05)$, the CMC test performed poorly, but WST performed well due the inclusion of a weighting scheme, as shown in Figure 3.3. Therefore, CMC performed poorly even when the effect of the causal variants was increased, and the MAF of the causal variants was increased. The $C - alpha$ test performed poorly when the percentage of moderately rare variants were large; the same data sample was used for the CMC and $CAST$. However, when the effect increased, and the causal variants had large MAFs, the $C - alpha$ performed better, even though there was a large proportion of moderately rare variants (see Figure 3.4).

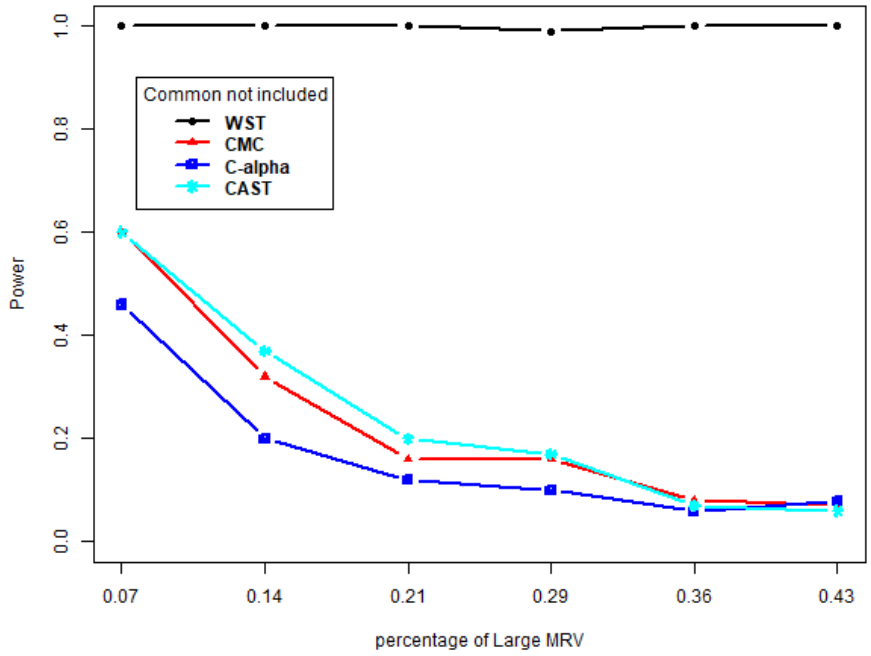


Figure 3.3: The impact of including large moderately rare variants with MAFs from $0.01 - 0.05$ on the power of tests. A total of 40% of the data included causal variants, when $OR = 3$, that were in the same direction and had MAFs between 0.0005 and 0.002. Initially, data were simulated with 90% rare variants with MAF less than 0.01. Then, the amount of large moderately rare variants in the data $(0.01 - 0.05)$ was increased, and the amount of data less than 0.01 was decreased.

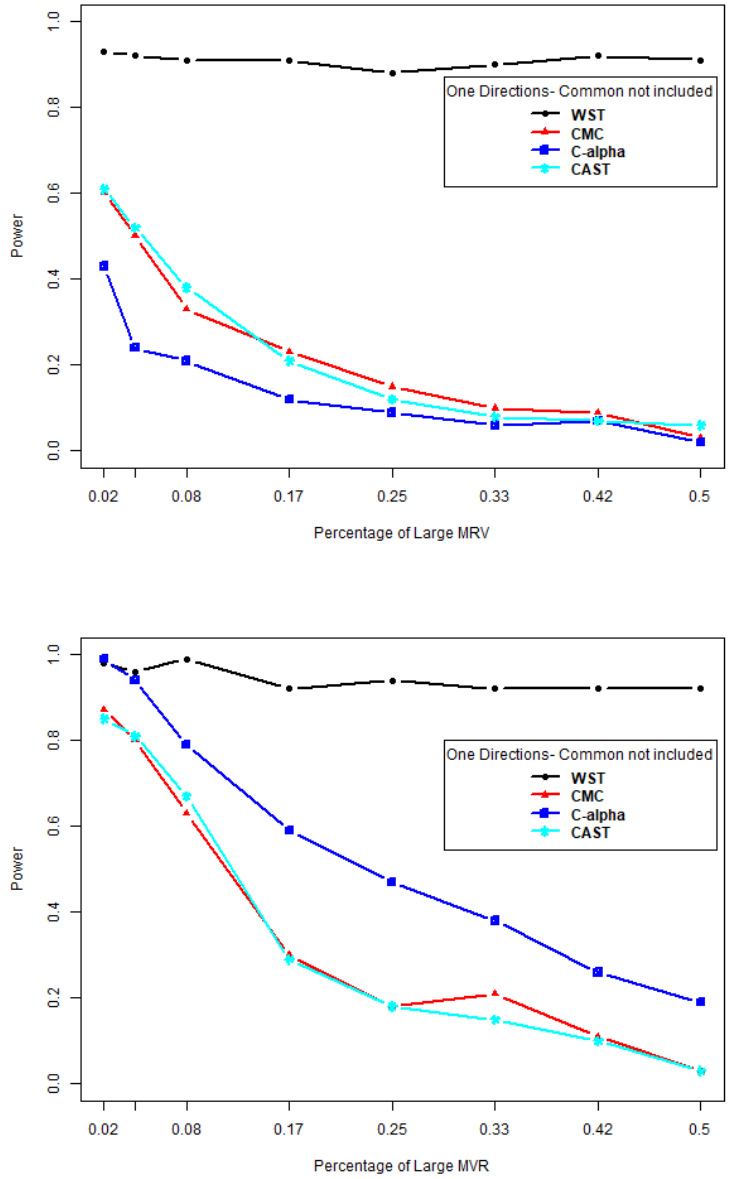


Figure 3.4: The impact of including moderately rare variants on the power of tests. The causal variants fixed in this analysis were extremely rare, and 15% of the data comprised causal variants ($OR = 3$). A total of 100 variants were generated with MAFs from $0.0005 - 0.005$. Then, the amount of extremely rare variants was reduced, and the number of moderately rare variants with MAFs ranging from $0.01 - 0.05$ was increased. The bottom figure is the same as the top figure, except causal variants are moderately rare variants with MAFs ranging from $0.005 - 0.008$ and effect size fixed at $OR = 2$.

The CMC and $CAST$ performed better when no common variants were included in the data (see Figure 3.5). In the figures, it can be seen that the $C-alpha$ test performed poorly as there was a large proportion of moderately rare variants in the analysis. This result confirms the argument that the $C - alpha$ test performs better when the data includes extremely rare variants. It should be noted that WST was not affected by the inclusion of common variants since it has a weighting scheme.

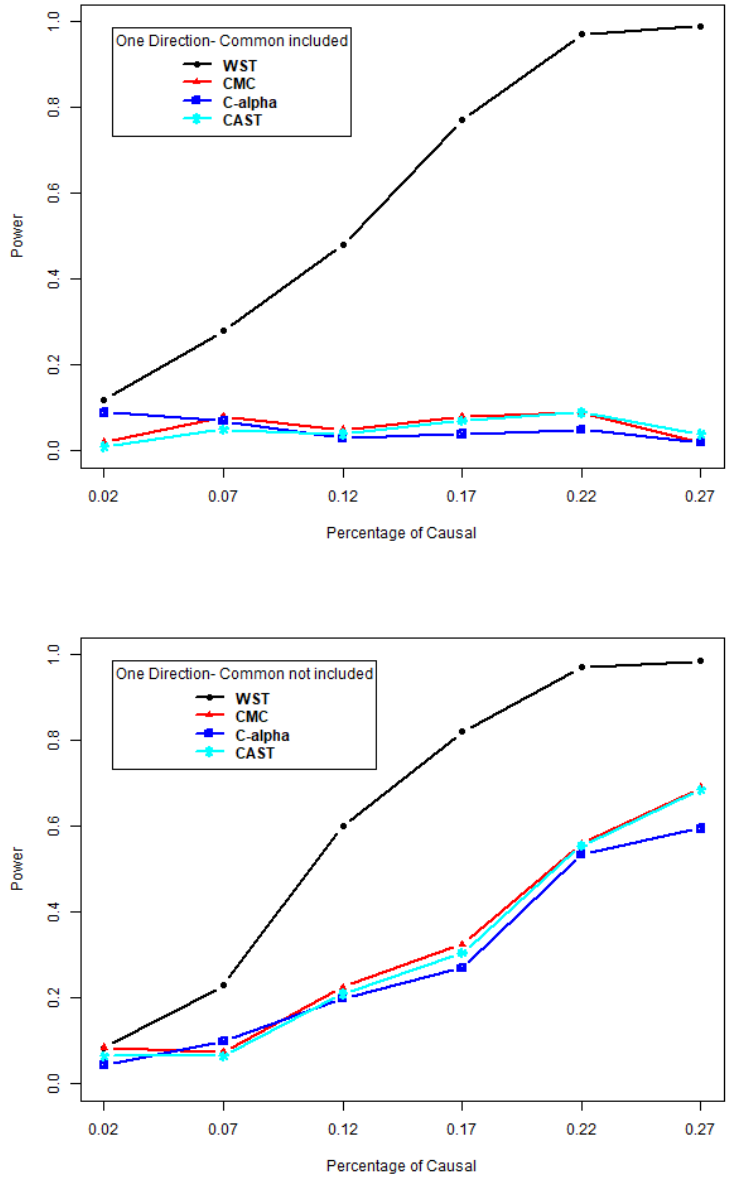


Figure 3.5: The impact of including common variants while we increase the percentage of causal variants. Causal variants were fixed with MAFs of $0.0005, 0.002$ and $OR = 3$, and the effects were in one direction. Data were generated between MAFs $(0.0005, 0.01)$. In the top figure, data were generated with common variants, and in the bottom figure, data were generated without common variants.

WST and CMC were under-powered when causal variants of different directions (i.e., neutral and protective variants) were included (see Figure 3.6) and when causal variants in one direction were included.

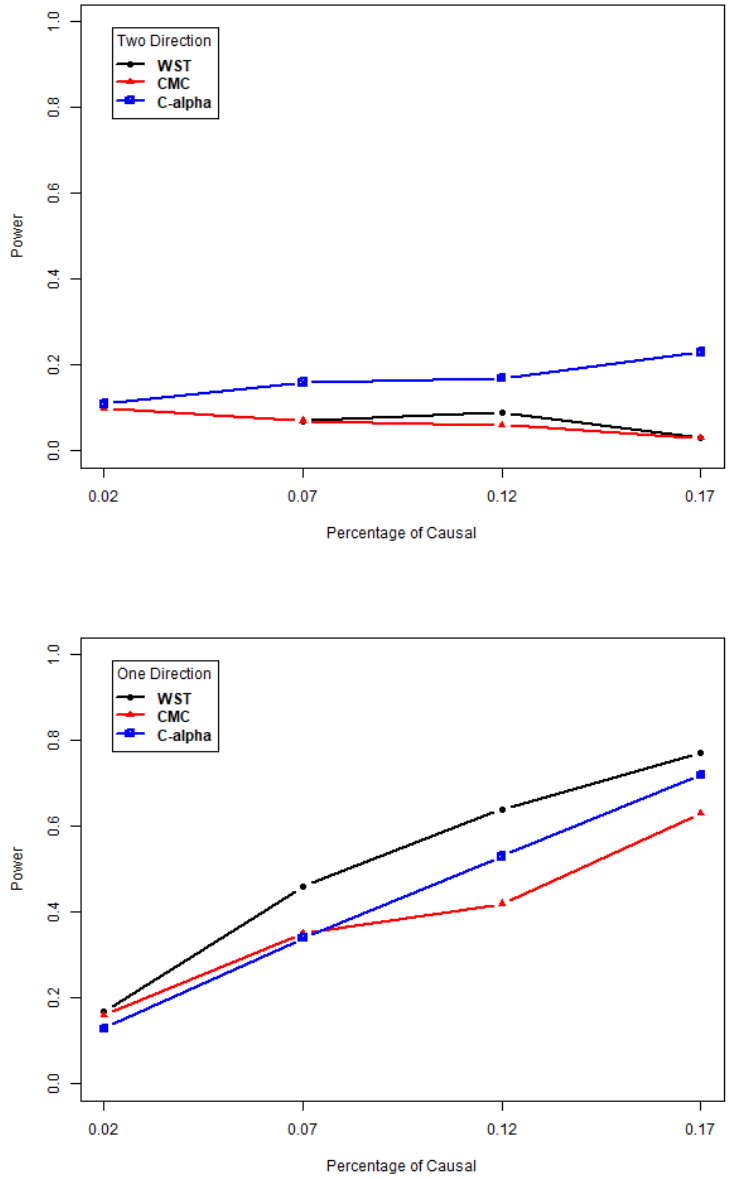


Figure 3.6: The impact of increasing the percentage of causal variants with different directions(effects) on the tests, we show that when the percentage of the causal variants increased from 2% to 17%. The causal variants were fixed with MAFs of $0.0005, 0.002$; $OR = 3$ and $OR = 0.3$, equally. Data were generated between MAFs of $0.0005, 0.01$. The top figure includes data with two directions, and the bottom figure includes data generated with one direction.

The $C - alpha$ test performed better when there were no common variants in the data sample. Additionally, when the causal variants were very rare, the test performed better, especially if the effect sizes were in a different direction. The test was robust with the inclusion of risks and neutral variants, which is an important advantage of quadratic tests. The simulations showed that the $C - alpha$ test could perform better under three conditions: if the data did not include common variants, if the number of causal variants was large, and if the effect was large. In the simulations, we did not accommodate for covariations. The CMC test does not accommodate for covariations, as Basu & Pan (2011) mentions, while the SKAT and WST can accommodate covariations (see Table 3.4 for a summary of the properties of all the tests).

3.6 Conclusion

The likelihood of observing rare variants in study samples is small when allele frequencies are very low. Consequently, statistical tests of the association become underpowered because of variations in the data. This is a major issue with RVAS. As previously mentioned, various statistical methods were proposed to increase the power of such tests, and most were based on testing multiple rare variants within genetic units, such as in (Zhang, 2015). These methods were categorised as burden and non-burden tests. This chapter explored the limitations of these tests and confirmed using weights is critical in RVAS, which is the main focus of this thesis. Thus, different weight schemes were incorporated into quadratic tests for this thesis. It should be noted that burden tests are not robust, but quadratic methods are, in regard to the directions of effects of causal variants. In addition to these topics, collapsing a group of rare variants in a gene or a region was also discussed in this chapter. As previously mentioned, $CAST$, developed by Morgenthaler & Thilly (2007), collapses rare variants and then compares collapsed allele frequencies of case and control groups. The development of CAST was a milestone in RVASs and was built upon for many other tests that use collapsing methods. The CMC method, introduced by Li & Leal (2008), is one such extension in which rare variants are collapsed in various subgroups. Collapsing rare and common variants is practised in tests of association. Note that $CAST$

and CMC methods require a fixed MAF threshold to define common and rare variants.

Madsen & Browning (2009) proposed the WST method, which includes both common and rare variants, but in the test, the variants are weighted according to their MAFs. Thus, common variants are given small weights, while rare variants are given large weights. In addition to this test, other robust methods were introduced in this chapter; they relate to the directions of causal-variant effects. Neale *et al.* (2011) developed the $C - alpha$ test, which compares expected variances to actual variances of rare variant distributions in case and control groups. Moreover, Wu *et al.* (2011) introduced $SKAT$, a variance-component score test that tests for associations in given regions between variants (common and rare) while adjusting for covariates. Both $C - alpha$ and SKAT test the variances of effects rather than the means of effects.

Test	Two Direction	AC	Sen. to CV	Target
CAST	No	No	No	RV
CMC	No	No	No	RV
WST	No	No	No	RV
RPT	No	No	No	RV
C alpha	Yes	No	Yes	RV
SKAT	Yes	Yes	No	RV- not all CV

Table 3.4: A summary of the properties of the tests discussed in this chapter. They are considered in relation to rare variants. This summary includes the tests' sensitivity to association directions, ability to adjust for covariates (AC), and sensitivity to common variants(CV) and target variants. This conclusion are from the literature and it was confirmed from the simulation.

Chapter 4

Score Test

4.1 Introduction

In this chapter, the score test, which is based on a logistic model, is used as a global (non-SNP-based) test to determine the associations among a set of rare variants. Existing collapsing methods, such as the burden test, have less power than a non-burden test because they ignore heterogeneity and only evaluate the marginal effects of SNPs. We also consider various score tests that incorporate weighting schemes and investigate the distribution of the tests. We use a score (quadratic) test for two reasons: it can overcome the issue of direction effects, and it only requires fitting the model under a null hypothesis. Therefore, the issue of estimation in rare variants can be avoided. The proposed test is built based on a logistic model in which no covariates are included.

The purpose of this chapter is to introduce the logistic model and derive a standard score test from its distribution, incorporate the variant weights, introduce theories of the quadratic form in normal distribution, and derive the distribution of the weighted score test on the variant level. We will evaluate the type I error rates and power at different settings for the simulation data.

4.2 Logistic Model

The logistic model is an example of a generalized linear model (GLM), which is an extension of the traditional linear model. The traditional linear model assumes that errors are normally distributed. However, since the data that we consider in this thesis are based on a phenotype (trait) or response, as in the form of a case-control study, a logistic model is used.

A group of p SNPs and a trait, y, are under consideration. The objective is to test whether there is an association between y and one or more of the SNPs. For a random sample of unrelated individuals, n, let y_i be a measured trait value for each individual, i, with $\mathbf{y} = (y_1, \ldots, y_n)^T$. Let x_{ij} denote the SNP genotype for individual i $(i = 1, \ldots, n)$ and $j = 1, \ldots, p$ with $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$. We also assume that the coding of SNPs is based on an additive model, where x_{ij} denotes that the variant is present in one allele $(x_{ij} = 1)$, present in both alleles $(x_{ij} = 2)$, or is absent $(x_{ij} = 0)$. We assume that there is no adjustment for covariates. Consider the logistic model with fixed effects as

$$\text{logit}P(y_i = 1) = \log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}, \tag{4.1}$$

where β is the $p \times 1$ vector of the parameters (i.e. a fixed effect).

We want to test the null hypothesis

$$H_0 : y \quad \text{and} \quad x \quad \text{are independent,}$$

which is equivalent to

$$H_0 : \beta_j = 0 \quad j = 1, \ldots, p.$$

We propose a method for testing H_0 based on score statistics. We assume that $\mathbf{y}$ is defined so that an SNP with $\beta > 0$ is termed deleterious, while $\beta < 0$ is protective, and $\beta = 0$ is neutral; both deleterious and protective SNPs are 'causal' variants.

4.3 Standard Score Test

Rao's score test is a statistical test of a simple null hypothesis; it determines whether a parameter of interest β is equal to some particular value β_0. It is the

most powerful test when the true value of β is close to β_0. The score test does not require an estimate of the information under the alternative hypothesis. This provides a potential advantage over other tests, such as the Wald test and the generalized likelihood ratio test (GLRT). This makes testing practical when the unconstrained maximum likelihood estimate is a boundary point in parameter space.

The model is based on the logit link function 4.1, in which the likelihood is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i},$$

where $p(x_i)$ is defined as $\frac{e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}}}{1 - e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}}}$. The score $U(\boldsymbol{\beta})$ is defined as

$U(\boldsymbol{\beta}) = \dfrac{\partial \log L(\boldsymbol{\beta} \mid x)}{\partial \boldsymbol{\beta}}$.

The Fisher information is

$$I(\boldsymbol{\beta}) = -\operatorname{E}\left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log L(X; \boldsymbol{\beta}) \middle| \boldsymbol{\beta}\right].$$

The statistic to test $H_0 : \boldsymbol{\beta} = \beta_0$ is

$$S(\beta_0) = \frac{U(\beta_0)^2}{I(\beta_0)}.$$

Based on that, the score test is derived from the likelihood of the binomial. The likelihood of a phenotype, y, given data X can be derived from a GLM for exponential family data (McCullagh & Nelder, 1989) according to the following:

$$L(y \mid X) = \exp \frac{y\eta - b(\eta)}{a(\phi)} - c(y, \phi),$$

where a, b, and c are known functions. The expression $\eta = h(\mathbf{x}^T \boldsymbol{\beta})$ for any function, h, which is in a logistic case, (logit)function, and ϕ is a dispersion parameter.

Under the assumption that $Pr(y = 1|X = x) = p(x; \boldsymbol{\beta})$ for function p, which is parametrized by $\boldsymbol{\beta}$, parameterize function $\boldsymbol{\beta}$ and further assume that observations are independent of each other.

Since the model is logistic, recall that in a sequence of Bernoulli trials, $y_1, \ldots, y_n$, where there is a constant probability of success, $p(x_i)$, the likelihood (i.e. the conditional likelihood function) is calculated as follows:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

To derive the score function, take the first derivative of the likelihood. We consider the log likelihood given by the following:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \{y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i))\} \\
&= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1 - p(x_i)} \\
&= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}) \\
&= \sum_{i=1}^{n} -\log\left(1 + e^{(\beta_0 + x_i^T \boldsymbol{\beta})}\right) + \sum_{i=1}^{n} y_i(\mathbf{x}_i^T \boldsymbol{\beta})
\end{aligned}$$

To locate the maximum likelihood estimate, we first take the derivative of log likelihood with respect to $\boldsymbol{\beta}$ and then solve the first derivative after setting it to zero. Thus, we differentiate with respect to β_j

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} -\frac{1}{1 + e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}}} e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}} \mathbf{x}_i^T + \sum_{i=1}^{n} y_i \mathbf{x}_i^T,$$

which leads to

$$\begin{aligned}
\frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} (y_i - p(\mathbf{x}_i^T; \boldsymbol{\beta})) x_{ij}; \\
&= \sum_{i=1}^{n} (y_i - \mu_i) x_{ij},
\end{aligned}$$

where $j = 1, \ldots, p$. We can re-write it in matrix form:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = X^T(\mathbf{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu}$ is a vector, and its elements are $\mu_i = p(x_i) = \frac{e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}}}$, which are the estimation under null.

4.3.1 Distribution of the Standard Score Test

The central limit theorem (CLT) is a key concept in probability theory because it implies the probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions. We assume that $x_1, \ldots, x_n$ are independent samples from $p_{\boldsymbol{\beta}}(x)$ and that $\boldsymbol{\beta}$ is not a boundary parameter (from the Lyapunov CLT in (Pawitan, 2001)). Thus,

$$\log L(\boldsymbol{\beta}) = \sum_i \log p_\beta(x_i)$$

$$U(\boldsymbol{\beta}) = \sum_i \frac{\partial}{\partial \boldsymbol{\beta}} \log p_\beta(x_i)$$

$$I(\boldsymbol{\beta}) = -\sum_i \frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log p_\beta(x_i)$$

$$\mathcal{I}(\boldsymbol{\beta}) = E(I(\boldsymbol{\beta})), \tag{4.2}$$

where E represents the expectation symbol. Note that $\mathcal{I}(\boldsymbol{\beta})$ represents the expected Fisher information, while $I(\boldsymbol{\beta})$ is the observed Fisher information. From each x_i, the individual score statistic is

$$y_i = \frac{\partial}{\partial \boldsymbol{\beta}} \log p_\beta(x_i).$$

Then, $y_1, \ldots, y_n$ are identical independent samples with a zero mean and a variance-covariance equivalent to the expected Fisher information.

Recall that the mean of $U(\boldsymbol{\beta}) = 0$ and that it is based on a theorem which is stated as follows: under the assumption of a regularity condition so that we can take the derivative under the integral sign, we have

$$E(U(\boldsymbol{\beta})) = 0. \tag{4.3}$$

To prove that we are following Pawitan (2001), we consider the continuous

case without a loss of generality:

$$\begin{aligned} EU(\boldsymbol{\beta}) &= \int S(\boldsymbol{\beta}) p_\beta(x) dx \\ &= \int \left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}) \right) p_\beta(x) dx \\ &= \int \frac{\frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta})}{L(\boldsymbol{\beta})} p_\beta(x) dx \\ &= \int \frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}) dx \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} \int p_\beta(x) dx = 0. \end{aligned} \tag{4.4}$$

According to the theory given above, $y_1, \ldots, y_n$ are an *iid* sample with a mean equal zero. We will consider one variable, y_1, for simplicity.

$$Ey_1 = 0$$

and variance

$$var(y_1) \equiv \mathcal{I}_1(\beta).$$

Based on the CLT, we can get

$$\sqrt{n}(\bar{y} - 0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1(\beta)),$$

or we can re-write it as

$$\frac{U(\beta)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_1(\beta)).$$

When the sample size becomes large, we have approximately

$$(\mathcal{I}_1(\beta))^{-1/2} U(\beta) \sim \mathcal{N}(0, I).$$

It can be written as a vector parameter:

$$U(\boldsymbol{\beta}) \sim \mathcal{N}(0, \mathcal{I}(\boldsymbol{\beta})),$$

where $U(\boldsymbol{\beta})$ is a vector and $\mathcal{I}(\boldsymbol{\beta})$ is a matrix.

Let $g = (\mathcal{I}_1(\boldsymbol{\beta}))^{-1/2}U(\boldsymbol{\beta})$. Since g_j is an independent standard normal variable, we can write $g^T g$ as a summation notation:

$$g^T g = \sum_{j=1}^{p} g_j^2, \tag{4.5}$$

which presents the sum of the squares of p standard normal variables. Thus,

$$g^T g \sim \chi_p^2.$$

4.4 Weighted Score Test

Rare variants may have greater effects on disease risk than common variants; thus, using a weight that allows the rare variants to contribute more to the test than the common ones can increase its power. Also, a suitable weight can help reduce the influence of common variants in the set-based association. In the burden test, which is based on the collapse technique, the underlying assumption is that all variants have the same effect on the trait, and there is no heterogeneity. To allow for individual variant effects, the current study uses a weighting scheme. There are many ways to choose the weighting functions. One can use an external weight, or another weight can be used that is estimated from the data or chosen from estimated regression coefficients Lin & Tang (2011).

In this thesis, as previously mentioned, different weight schemes are used based on the variant level, the individual level, both the variant and individual levels and, finally, the cell level. These schemes may help detect an association between rare variants and traits and may help increase the power of the test. In section (4.7), a score test is derived, incorporating a weight based on the variant level. The distribution and power of the test and the type I error are investigated. Before we derive the score test with the variant weights, we will review some theories based on the quadratic form in normal distributions.

4.5 Statistical Theories of Quadratic Form

The distributions of our method can be based on quadratic theories. We can use two theorems:

Theorem 4.5.1 *Let $\mathcal{W}$ denote a matrix of rank p, and let $\mathcal{A} \sim \mathcal{N}(0, I)$. Then, $\mathcal{A}^T\mathcal{W}\mathcal{A} \sim \chi^2_p$ only if $\mathcal{W}$ is an orthogonal projection on a space of rank p. (Box et al., 1954)*

Theorem 4.5.2 *According to Box et al. (1954), if $\mathcal{A}$ denotes a column vector of p random variables $\mathcal{A}_1, \ldots, \mathcal{A}_p$ has an expectation of zero and is distributed in a multi-normal distribution with $p \times p$ variance covariance matrix Σ,*

$$\mathcal{A} \sim \mathcal{N}(0, \Sigma).$$

If $\mathcal{Q} = \mathcal{A}^T\mathcal{W}\mathcal{A}$ is any real quadratic form of rank $r \leq p$, then $\mathcal{Q}$ is distributed as

$$X = \sum_{j=1}^{r} \lambda_j \chi^2(1),$$

where the χ^2 variate is distributed independently of every other, and the λs are the r real non-zero latent roots of matrix $\mathcal{W}\Sigma$.

Theorem 4.5.3 *Suppose that $\mathcal{A} \sim \mathcal{N}(\mu, \Sigma)$, where rank $(\Sigma) = p$. The random variable $q = \mathcal{A}^T\mathcal{W}\mathcal{A}$ has the same distribution as the random variable $X = \sum_{i=1}^{n} d_i s_i$, where d_i are the latent roots of the matrix $\mathcal{W}\Sigma$, and s_i are independent non-central χ^2 random variables, each with one degree of freedom.*

Cochran (1934) stated that Theorems 4.5.1 and 4.5.2 are special cases of Theorem 4.5.3.

Thus, when we incorporate the weight in the score test, as we will discuss later, the score test does not follow the chi-square distribution with degree of freedom p; rather, it is based on a quadratic form approximation, as stated above.

4.6 Approximation for the Distribution of the Quadratic Form

Under the assumption that the score function follows a normal distribution with a mean of zero and variance/covariance approximated with Fisher information $I(\boldsymbol{\beta})$, we recall section 4.3.1,

$$U(\boldsymbol{\beta}) \sim \mathcal{N}(0, \mathcal{I}(\boldsymbol{\beta})), \tag{4.6}$$

and

$$\mathcal{I}(\boldsymbol{\beta})^{-1/2}U(\boldsymbol{\beta}) \sim \mathcal{N}(0, I). \tag{4.7}$$

The score test, including weights, also has a quadratic form of normal distribution. The distribution of linear combinations of chi-square variables has been studied by several authors over the last four decades (Liu *et al.*, 2008). Many methods are used to compute the p-value; for instance, 'Davies' is an exact method that computes the p-value by inverting the characteristic function of the mixture chi-square (Davies, 1980). Liu *et al.* (2008) gives an approximate method which matches the first three moments.

In this section, we use the method described by Liu *et al.* (2009) to approximate the distribution of linear combinations of chi-square variables. Let W denote a $p \times p$ symmetric and non-negative definite matrix of rank p, and let $A \sim \mathcal{N}(\mu_A, V)$. Then, $Q(A) = A^T W A$. The goal is determining how the tail probability of $Q(A)$ can be estimated.

$$P(Q(A) > t) \tag{4.8}$$

Let R be an orthogonal $p \times p$ matrix which converts $T = V^{1/2}WV^{1/2}$ into the diagonal form $\Lambda = (\lambda_1, \ldots, \lambda_p) = RTR^T$, where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Since $y = RV^{-1/2}W$ is normally distributed with mean $\mu_y = RV^{-1/2}\mu_A$, variance I_p, $Q(A)$ can be written as a weighted sum of independent chi-square variables, as shown in 4.5.3.

$$Q(A) = A^T W A = y^T \Lambda y = \sum_{j=1}^{p} \lambda_j \chi^2_{h_j}(\delta_j),$$

where $h_j = 1$ and $\delta_j = \mu^2_{yi}$. Thus, Imhof (1961) gives the cumulant generating function of $Q(A)$ as

$$K(t) = \frac{1}{2}\sum_{j=1}^{p} h_j \log(1 - 2t\lambda_j) + \sum_{j=1}^{p} \frac{\delta_j \lambda_j t}{1 - 2t\lambda_j}.$$

The formula for the k^{th} cumulant of $Q(A)$ is

$$\kappa_k = 2^{k-1}(k-1)!\left(\sum_{j=1}^{p} \lambda_j^k h_j + k\sum_{j=1}^{p} \lambda_j^k \delta_j\right).$$

Then,

$$\sum_{j=1}^{p} {\lambda_j}^k h_j = \text{trace}(\Lambda^k) = \text{trace}((RTR^T)^k) = \text{trace}(T^k) = \text{trace}((WV)^k)$$

and

$$\sum_{j=1}^{p} {\lambda_j}^k \delta_j = \mu_y^T \Lambda^k \mu_y = \mu_A^T (WV)^{k-1} W \mu_A.$$

The mean and standard deviation of $Q(A)$ are

$$\mu_Q = \kappa_1 = c_1$$

and

$$\sigma_Q = \sqrt{\kappa_2} = \sqrt{2c_2},$$

where $c_k = \sum_{j=1}^{p} {\lambda_j}^k h_j + k \sum_{j=1}^{p} {\lambda_j}^k \delta_j$.

The skewness and kurtosis of $Q(A)$ are

$$\beta_1 = \sqrt{8s_1}$$

and

$$\beta_2 = 12s_2,$$

where s_1 and s_2 are $\frac{c_3}{c_2^{3/2}}$ and $\frac{c_4}{c_2^2}$, respectively.
The non-central $\chi_l^2(\delta)$ given in Liu *et al.* (2008) is used to approximate the distribution of

$$Q(A) = \sum_{j=1}^{p} \lambda_j \chi_{h_j}^2(\delta_j).$$

The tail probability is then approximated by

$$P(Q(A) > t) = P\left(\frac{Q(A) - \mu_Q}{\sigma_Q} > t^*\right)$$

, which is approximated to

$$P(Q(A) > t) = P\left(\frac{\chi_l^2(\delta) - \mu_A}{\sigma_A} > t^*\right) = P(\chi_l^2(\delta) > t^*\sigma_A + \mu_A)$$

, where $t^* = (t - \mu_Q)/\sigma_Q$, $\mu_A = E(\chi_l^2(\delta)) = l + \delta$, $\sigma_A = \sqrt{var(\chi_l^2(\delta))} = \sqrt{2a}$, and $a = \sqrt{l + 2\delta}$.

The parameters l and δ are determined so that the skewnesses are equal for both $\chi_l^2(\delta)$ and $Q(A)$, and the difference between their kurtoses is minimized. The skewness of $\chi_l^2(\delta)$ is $\sqrt{8(a^2 + \delta)/a^3}$. Given that the skewnesses of $\chi_l^2(\delta)$ and $Q(A)$ are equal, then $\delta = s_1 a^3 - a^2$ (Liu *et al.*, 2009).

4.7 Construction of Weighted Score Test Based on Variant Weights

Here, we propose score test statistics which incorporate variant weights. We assume that rare variants are causing the disease, so we are looking for weights incorporated based on variants that can allow rare variants to have a greater influence on the test statistics. To construct the weighted score test, we will start from the likelihood of the logistic model because the dependent variable y has two values (0 = control, 1 = case). Recall that X is an $n \times p$ genotype matrix, and $\mathbf{y}$ has $n \times 1$ phenotypes. Let $\boldsymbol{\gamma}$ be a vector expressed as the variant weight, $p \times 1$, and let Γ be a diagonal matrix with the following elements: γ_j, $j = 1, \ldots, p$.

The joint probability density function gives the values of y as a function of $\boldsymbol{\beta}$, which is related to μ by a logit transform:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \sum_{j=1}^{p} \gamma_j x_{ij} \beta_j \quad i = 1, 2, \ldots, n, \tag{4.9}$$

where $\mu_i = E(y_i) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \gamma_j x_{ij} \beta_j}}{1 - e^{\beta_0 + \sum_{j=1}^{p} \gamma_j x_{ij} \beta_j}}$; we can re-write it with a different notation to include the intercept in the design matrix as $\mu_i = \frac{e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta\gamma}}}{1 - e^{\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta\gamma}}}$. The likelihood function expresses the value of β in terms of knowing the values for y. Thus, giving the observed data the likelihood of β is;

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} {\mu_i}^{y_i} (1-\mu_i)^{1-y_i} \tag{4.10}$$

To derive the score test, we will take the derivative of the likelihood function (4.10). We will take the log of the likelihood for simplicity.

$$\log L(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \log\left(\exp\left\{\beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j x_{ij}\right\}\right) + \log\left(1 + \exp\left\{\beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j x_{ij}\right\}\right)\right\} \tag{4.11}$$

To calculate the score function including the weight based on variants, we differentiate with respect to each β_j.

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \{ y_i \gamma_j x_{ij} - x_{ij} \gamma_j \mu_i \} = \sum_{i=1}^{n} \gamma_j x_{ij} (y_i - \mu_i), \tag{4.12}$$

where $j = 1, 2, \ldots, p$ and μ_i are the estimation under the null model. We can write this in matrix form:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \Gamma X^T(\mathbf{y} - \boldsymbol{\mu}) \tag{4.13}$$

Here, Γ is a $p \times p$ diagonal matrix of weight. Let us define $U(\boldsymbol{\gamma}) = \Gamma X^T(\mathbf{y} - \boldsymbol{\mu})$; then, its variance is $V(\boldsymbol{\gamma}) = \Gamma X^T D X \Gamma^T$, where D is a diagonal matrix with elements $\mu_i(1 - \mu_i)$.

The score test is in the following form:

$$S(\gamma) = U(\boldsymbol{\gamma})^T V(\boldsymbol{\gamma})^{-1} U(\boldsymbol{\gamma}). \tag{4.14}$$

However, the weight in this form is cancelled since

$$S(\boldsymbol{\gamma}) = (\Gamma X^T(y - \mu))^T (\Gamma X^T D X \Gamma^T)^{-1} (\Gamma X^T (y - \mu) = \mathbf{U}^T \underbrace{[X^T \underbrace{(\mu(1-\mu)}_{D} X]^{-1}}_{V} \mathbf{U}.$$

Then, we can modify the score test based on variant weights after

$$\Gamma(V^{-1/2}\mathbf{U}) \sim \mathcal{N}(0, \Gamma\Gamma),$$

which means that

$$S(\boldsymbol{\gamma}) = (V^{-1/2}\mathbf{U})^T \Gamma\Gamma V^{-1/2}\mathbf{U} = \mathbf{U}^T V^{-1/2} \Gamma\Gamma V^{-1/2}\mathbf{U}. \tag{4.15}$$

Construction of the Score Test Distribution for $S(\boldsymbol{\gamma})$

Let $\mathbf{U}$ be a vector of dimension p, which is a score function, and let V be the Fisher information matrix of dimension $p \times p$. Let Γ be a diagonal matrix of dimension $p \times p$ of weight. Recall that based on the asymptotic distribution theory, the distribution of $\mathbf{U}$ is as follows:

$$\mathbf{U} \sim \mathcal{N}(0, V), \tag{4.16}$$

where V is a covariance-variance matrix. Then, based on the properties of the normal distribution, we can write this as follows:

$$V^{-1/2}\mathbf{U} \sim \mathcal{N}(0, I_p).$$

Let Γ be a diagonal matrix; its dimension is $p \times p$ based on the variant weight with rank p. Using Theorems 4.5.1 and 4.5.3, where $\mathcal{W}$ is $\Gamma^T\Gamma$,

$$S(\boldsymbol{\gamma}) \sim \chi^2(p) \tag{4.17}$$

if $\Gamma^T\Gamma$ is an orthogonal projection on a space of rank p. This is one form of the score test which includes the variant weights. However, since the weight matrix is not an orthogonal projection, then, according to theorem 4.5.2, $S(\boldsymbol{\gamma})$ follows a mixture of χ^2 as

$$S(\boldsymbol{\gamma}) \sim \sum_{j=1}^{p} \lambda_j \chi^2(1), \tag{4.18}$$

where λs are the p real non-zero eigenvalues of the matrix $\Gamma^T\Gamma$, and $\chi^2(1)$ express under the summation the independent non-central χ^2 random variables, each with one degree of freedom. When there is no weight considered ($\Gamma^T\Gamma = I_p$), $\sum_{j=1}^{p} \lambda_j \chi^2(1)$ is equivalent to χ^2 with one degree of freedom. In the following sections, we will introduce variant weight scheme Γ.

4.8 Simulation

We have run simulation studies to examine the performance of the proposed score tests. In the simulations, we generated a genotype and trait values. We simulated $p = 200$ SNPs with a sample size of 1000 cases and 1000 controls: $n = 2000$. Each rare variant had a mutation rate or MAF uniformly distributed between 0.0005 and 0.05, while for a common variant, it was between 0.05 and 0.5. First, we generated a latent vector $\mathbf{z} = (z_1, \ldots, z_p)$ from a multivariate normal distribution. Note that the setting of the MAF for common and rare variants is changed when we evaluate the type I error and the power of the tests. For the power, we specify some RVs as causal variants by adjusting the odds ratio (OR). A large OR is associated with the rarest variants; the set of OR parameter settings is provided in (Table 4.1).

	OR				
Protective variants	0.66	0.5	0.4	0.33	0.25
Risk variants	1.5	2	2.5	3	4

Table 4.1: The set of parameter values for the effect size, represented by an odds ratio (OR)

The number of causal variants in the dataset varies from 10 to 30 with different MAFs. In a given model, 80% of the variants are rare, while the remaining 20% are common. To evaluate the type 1 error for the proposed test and weight, we simulated data under the null model ($\text{logit}[P(y_i = 1)] = \beta_0$) for $n = 2000$. For the disease status, y_i, of a subject i, is generated from the logistic regression model (4.1). For the null case, we used $\boldsymbol{\beta} = 0$, for non-null cases, we randomly selected the non-zero components of $\boldsymbol{\beta}$, while the remaining ones were all 0.

4.9 Variant Weights

Now, we will introduce two schemes of variant weights. More weight schemes will be presented in detail in Chapter 5.

4.9.1 Beta as a Variant Weight

The beta function was used by Wu *et al.* (2011) as a variant weight. The idea underlying its purpose is to up-weight the rare variants and down-weight the common ones to allow the rare variants to contribute more to the test. In this chapter, we will consider the beta used in Wu *et al.* (2011) while introducing the score test with variance-covariance matrix V.

4.9.2 Cauchy Function as a Variant Weight

This recently-proposed weight scheme can be used to up-weight the rare variants and down-weight the common ones. This function comes with a strong advantage based on its parameters: we can use it in different ways and for different purposes. For example, we can use a parameter-based estimate from the MAF instead of

having an arbitrary parameter, which is required in the beta function. More details on this weight will be presented in the next chapter. In this section, we will apply one scheme using this weight.

4.9.3 Type I Error

For a weighted score test using both weights (i.e. beta and Cauchy), we calculate the type I error rates for significance level $\alpha = 0.05$. Type I errors were calculated on the basis of 1000 replications. We can generate data with no genetic effects by fixing the $OR = 1$. The data are divided into two groups of sets: the first set, which includes extremely rare and moderately rare variants, has an MAF less than 0.01, and the other has an MAF larger than 0.01.

We evaluate the type I errors with a different percentage of inclusion for the second set, which is considered to have rare variants. We use the same scenario with two different sets, both of which are extremely rare and include moderately rare and common variants. The test based on variant weight $S(\boldsymbol{\gamma})$ has good control at different percentages of rare variant sets; the results are provided in Table (4.2). There is some concern about controlling the type I error rate when all variants in the data are considered extremely rare, with a maximum MAF of 0.005. This lack of control is likely the result of the rarity of variants among the 2000 individuals; see Table (4.3). When the MAF is very low, such as 0.0005, this means there are two variants among the 2000 individuals (i.e. there are 1998 zeros, and two elements are 1 or 2); hence, this affects the control over the type I error rate in the score test. The concern is when we have an MAF on the boundary (i.e. and MAF of less than 0.002).

	Large MAF (0.01,0.5)						
Tests	15%	30%	40%	50%	60%	70%	80%
S	0.05	0.05	0.04	0.04	0.03	0.05	0.04
Sw	0.05	0.05	0.04	0.035	0.05	0.04	0.05
Sw2	0.05	0.05	0.04	0.036	0.05	0.05	0.05

Table 4.2: We evaluate the type I error rates by expressing the effect of the amount of the number of rare variants in the data. The first non-causal variant has MAFs set to between 0.0005 and 0.01; then, we increase the amount variants in the other set, which ranges between 0.01 and 0.5 (common variants). Where S, Sw, and $Sw2$ are the standard score test, the score test with beta weight, and the score test with Cauchy weight, respectively. The type I error rate was evaluated at a significance level of 0.05.

We also evaluate the type I error rate at a very rare MAF 4.3:

	Large MAF (0.005,0.5)						
Tests	2%	20%	30%	40%	50%	60%	70%
S	0.025	0.05	0.03	0.06	0.04	0.04	0.04
Sw	0.025	0.05	0.02	0.05	0.05	0.04	0.04
Sw2	0.025	0.04	0.04	0.06	0.05	0.05	0.05

Table 4.3: We evaluate the type I error rates by expressing the effect of extreme rareness on the data. The first column, which represents most of the data, has the MAFs classified as an extremely rare variant; the MAFs are set between 0.0005 and 0.005. Then, we increase the amount of variants in the other set, which ranges between 0.005 and 0.5. S, Sw, and $Sw2$ are the standard score test, the score test with beta weight, and the score test with Cauchy weight, respectively. The type I error rates were evaluated at a significance level of 0.05.

4.9.4 Power of the Test

We simulate the data using different numbers of causal variants based on an OR of $(OR = 3)$ while increasing the percentage of causal variants $(5\%, 10\%, 15\%, 20\%, etc.)$. Then, we calculate the power of the score test with variant weights. The power

is calculated on the basis of 100 replications. Next, we can determine the power of the score test with variant weights. The power is calculated on the basis of 100 replications, and the result is then compared with the SKAT test results. We have many different scenarios to investigate the power of tests.

In the first scenario, we generate 60% of non-causal rare variants in the range of (0.0005–0.005), while other variants range between 0.005 and 0.05. We fix the causal variants at different values of MAF.

In the second scenario, we generate 60% of rare variants in a wide range (0.005–0.05), while other variants range between 0.0005 and 0.005. We fix the causal variants at different values of MAF. Both scenarios have 10% common variants.

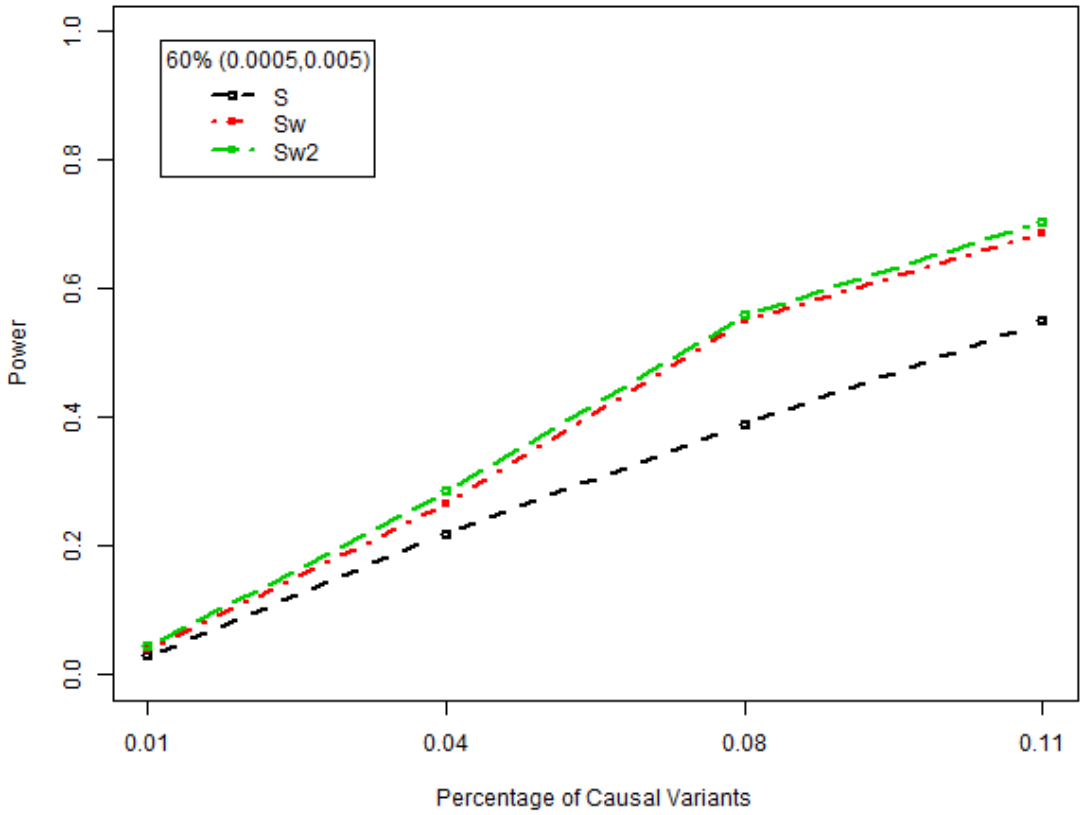


Figure 4.1: In this analysis, the MAF of non-causal variants is between 0.0005 and 0.005 and 60%, and the other 30% have an MAF between 0.005 and 0.05. A causal variant is between 0.0005 and 0.002, which is classified as extremely rare. On the X axis, we provide a range of the causal variant percentages in the generated data. S, Sw, and $Sw2$ are the standard score test with no weight included, with beta weight, and with $Cauchy(min(f), 0.01)$, respectively.

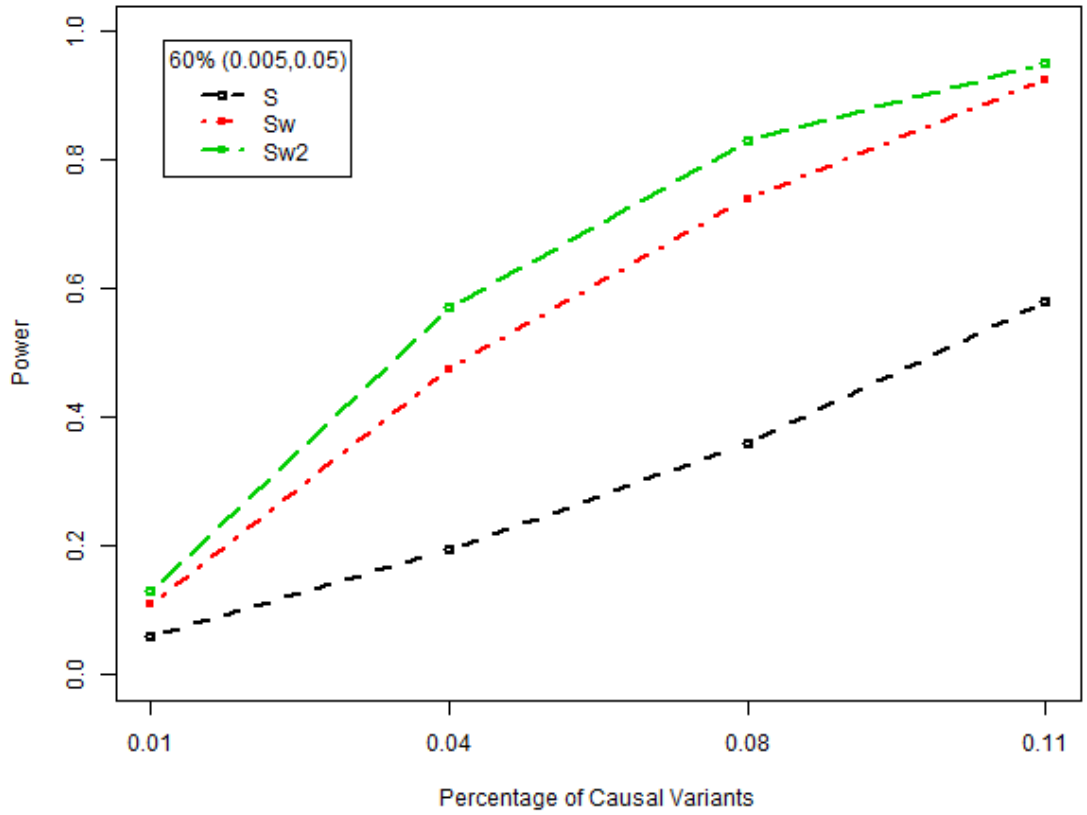


Figure 4.2: In this analysis, the MAF of non-causal variants ranges from 0.0005 to 0.005 for 30% of the generated variants. The rare variants, 60%, have an MAF between 0.005 and 0.05, and 10% are common variants between 0.05 and 0.5. The causal variant is between 0.0005 and 0.002, which is classified as extremely rare. On the X axis, we provide a range of the causal variant percentages in the generated data. S, Sw, and $Sw2$ are the standard score test with no weight included, with beta weight, and with $Cauchy(min(f), 0.01)$, respectively.

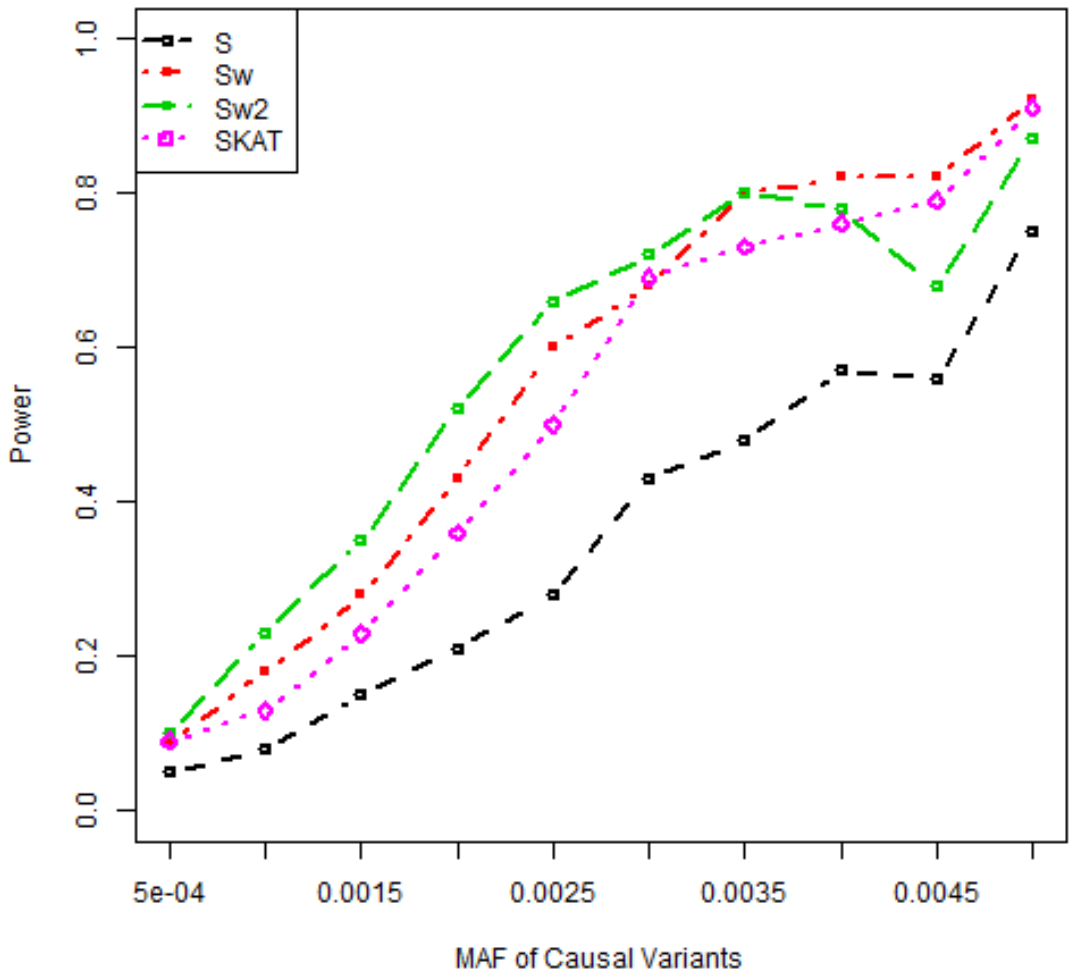


Figure 4.3: In this analysis, we consider the SKAT test, which is based on the variance component in the test comparison. The setting is the same as previously described (see Figures 4.1 and 4.2). S represents the standard score test, S_w expresses the score test with a beta weight, and S_{w2} expresses the score test with a Cauchy weight.

4.9.5 Conclusion

When the causal variants are very rare such that they have MAF= 0.0006 and the other variants are considered moderately rare or common, the score test performance is more powerful than that of SKAT. Therefore, when null variants are common with an MAF ranging between 0.05 and 0.5, the differences between SKAT and the performed tests is small compared to above settings. Using the Cauchy weight performs better in all settings of MAF on non-causal variants because it weighs the rare variants highly.

To confirm the results presented above, we can observe that when the non-causal variants tend to be moderate or common but the causal variants are extremely rare, $S(\lambda)$ performs slightly better.

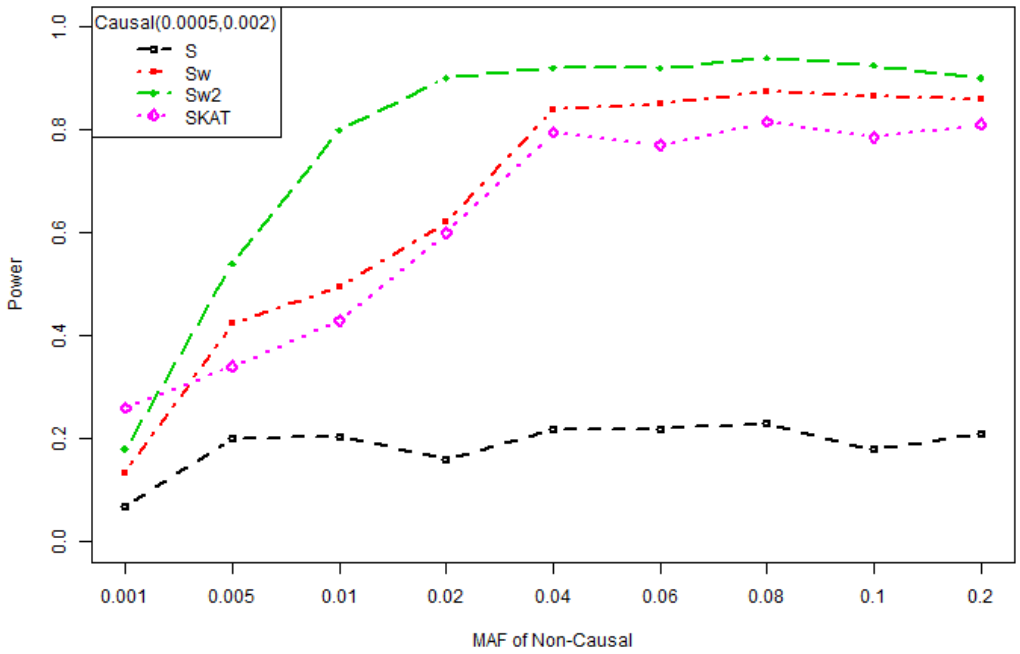


Figure 4.4: This figure shows the performance of power based on different forms of test S_w, which is the score test with variant weight $S(\lambda)$. We use the same weight here, which is beta with $(1, 25)$, in both test S_w and SKAT. We also compare performance of power based on different forms of test S_w with the Cauchy weight S_{w2}.

The power of $S = \mathbf{U}^T V^{-1} \mathbf{U}$ depends on the total of explained variation, $EV = \frac{Var[E(y|X)]}{Var(y)}$, the sample size, and rank V, and it is not sensitive to the direction of the SNPs' effect or the MAF. This is clear from its null distribution, $S \sim \chi^2_c$, where $c = \sum_j c_j = nEV$. This observation is consistent with a literature review by Newton-Cheh & Hirschhorn (2005).

While the weighted score test depends on the explained variation and the weight and sample sizes, it is neither sensitive to the direction of SNPs' effect nor to the MAF $\mathcal{F}_j$. $S_w \sim \sum \lambda \chi_1$, where λ represents the eigenvalues of $\Gamma^T \Gamma$.

The introduced form of the score test suffers from a singularity in the V matrix. The data must be independent. In the next section, we introduce a different

form of test statistic based on fixed and random effects, and we demonstrate that they are equivalent to each other in the form of the test.

4.10 Weighted Score Test Based on Score Function

4.10.1 Introduction

The proposed test in the previous section assumes independence among variants (SNPs); however, in this section, we propose a method that can accept a correlation structure in the data. Since real genetics data have a high correlation structure, the score test in the last section has a singularity issue when we take the inverse of the covariance-variance matrix, V. Therefore, to overcome these issues, we avoid taking the inverse of matrix V, which is accommodated in the reference distribution of the test. These methods perform well in terms of power and controlling the type I error.

4.10.2 Model

We assume that a group of SNPs, p, and a trait, $\mathbf{y}$, are under consideration. The objective is to test whether there is an association between y and one or more of the SNPs. For a random sample of n unrelated individuals, let y_i be a measured trait value for individual i, and let $\mathbf{y} = (y_1, \ldots, y_n)^T$. Let X_{ij} denote the SNP genotype for individual i as $i = 1, \ldots, n$ and $j = 1, \ldots, p$. For simplicity, we assume that x_{ij} denotes whether the rare allele is present ($x_{ij} = 1$) or absent ($x_{ij} = 0$) or present in two alleles ($x_{ij} = 2$); let $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})^T$. We also assume that the coding of SNPs is based on the additive model, and there is no adjustment for covariates. For the variant weight, as it is the first weight that we consider, let γ_j represent the weight value at SNP j, where $j = 1, \ldots, p$ and Γ is a diagonal matrix of dimension p.

$$\text{logit}P(y_i = 1) = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} \tag{4.19}$$

We test the null hypothesis

$$H_0 : y \quad \text{and} \quad X \quad \text{are independent,}$$

which is equivalent to $H_0 : \boldsymbol{\beta} = 0$. We propose methods for testing H_0 based on statistics from a (weighted) score test without taking the inverse of variance-covariance matrix V. We assume that y is defined so that an SNP with $\beta > 0$ is termed deleterious, while $\beta < 0$ is protective, and $\beta = 0$ is neutral. Both deleterious and protective SNPs are causal variants.

4.10.3 Test

We know from the previous section that $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ follows a normal distribution with a mean of zero, and variance-covariance $V = -E\big(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\big)$. It is written as $\mathbf{U} \sim \mathcal{N}(0, V)$. Based on a normal property, the variant weight on $\mathbf{U}$ can be incorporated by multiplying the diagonal matrix, Γ, by the vector, $\mathbf{U}$, which will also be incorporated into V. Thus,

$$\Gamma \mathbf{U} \sim \mathcal{N}(0, \Gamma V \Gamma^T).$$

Take the quadratic form of the test:

$$T(\boldsymbol{\gamma}) = \mathbf{U}^T \Gamma^T \Gamma \mathbf{U}.$$

4.10.4 Distribution of the Test

According to the theory of quadratic form of normal distribution mentioned in a previous section, we can consider the distribution of the test as a linear combination of chi-squared random variables 4.5.3;

$$T(\boldsymbol{\gamma}) \sim \sum_{j=1}^{p} \lambda_j \chi_1^2,$$

where λ_j, as a weight of the distribution, is equal to the eigenvalues of the variance-covariance matrix: $V(\boldsymbol{\gamma}) = \Gamma^T V \Gamma$. To approximate this distribution, we use a scaled chi-square and the method of the moment to matching the first two moments.

$$E(T(\boldsymbol{\gamma})) = E(\mathbf{U}^T\Gamma^T\Gamma\mathbf{U}).$$

Let $W = \Gamma^T\Gamma$. Then, we can re-write the equation as

$$E(T(\boldsymbol{\gamma})) = E(\mathbf{U}^T W\mathbf{U})$$

using the property of the expectation of the quadratic form

$$E(T(\boldsymbol{\gamma})) = tr(WV) + \mu_U W\mu_U,$$

where μ_U is the expectation of U. We can simplify it one more since $\mu_U = 0$

$$E(T(\boldsymbol{\gamma})) = tr(WV)$$

and the variance is

$$var(T(\boldsymbol{\gamma})) = var(\mathbf{U}^T W\mathbf{U}) = 2tr(WVWV) + 4\mu_U^T WIW\mu_U.$$

Since $\mu_U = 0$,

$$var(T(\boldsymbol{\gamma})) = var(\mathbf{U}^T W\mathbf{U}) = 2tr(WVWV).$$

We use the deviance approximation to get the p-value.

4.11 Variance Component

4.11.1 Introduction

Extensions of the GLM include models with random terms in the linear predictor and are called generalized linear mixed models (GLMMs). These are useful for accommodating the over-dispersion which is observed among outcomes for modelling the dependence between outcome variables implicit in repeated measures or longitudinal designs Stiratelli *et al.* (1984);Zeger *et al.* (1988) and for producing shrinkage estimates in multi-parameter problems. It is a traditional and often reasonable approximation to assume that the random error terms have a normal distribution and that the variance components are estimated from the data. When the outcomes are in binomial or Poisson form, a full maximum likelihood

analysis based on their joint marginal distribution necessitates numerical integration techniques to calculate the log likelihood, score equations, and information matrix (Breslow & Clayton, 1993).

Recent Bayesian methods avoid the need for numerical integration by using Gibbs sampling techniques or by taking repeated samples from posterior distributions. The Bayesian approach is flexible for a full evaluation of the uncertainty in estimated random effects. However, the drawbacks of this approach include the intensive computations and questions about when the sampling process has achieved equilibrium (Breslow & Clayton, 1993). Approximate procedures are exact methods for exploratory analyses and provide starting values for use with others.

There are two closely related approximate methods of inference in GLMMs:

- The penalized quasi-likelihood (PQL) method used by Green (1987) for semi-parametric regression analysis is available in hierarchical model inference, where the purpose is to shrink the estimation of the random effects Robinson (1991).

- Marginal quasi-likelihood (MQL) was proposed by Goldstein (1991) as an extension of GLMs in multilevel modelling (Goldstein, 1986). It is most appropriate when examining the marginal relationship between covariates and outcomes.

- PQL: The MQL and PQL methods, which were introduced by Goldstein (1991) and Breslow & Clayton (1993), are analogous to iteratively reweighted least squares for GLMs in that the model is linearized.

- MQL is the method of choice when interest is focused on the marginal relationship between covariables and responses, and the random effects model serves mainly to suggest a plausible covariance structure, as expressed in V, which enables one to obtain reasonably efficient estimating equations for the mean value parameters. In contrast, PQL is the procedure of choice for estimating parameters in a random model, particularly when the concentration is focused on the random effects Breslow & Clayton (1993).

The main difference between MQL and PQL is in the offset used. Since MQL sets the random effects to zero, the fixed-effects estimates are essentially marginal effects that are attenuated relative to the required conditional effects. The essential difference between the MQL estimating equations for the marginal model and the PQL equations for the hierarchical model is that the latter incorporate the random effect terms, $Z^T b$, in the linear predictor Breslow & Clayton (1993). The simulation conducted by Rodriguez & Goldman (1995) in his paper 'An Assessment of Estimation Procedures for Multilevel Models with Binary Responses' shows that the MQL and PQL can suffer from bias. Both fixed effect and variance components may be biased, especially when the response is binary.

4.11.2 The Score Test in GLMM

The score test can be called a variance component test. We initially introduce a logistic model with random effects:

$$\text{logit}P(y_i = 1) = \mathbf{z}_i^T \boldsymbol{\alpha} + \mathbf{x}_i^T \boldsymbol{\beta} \tag{4.20}$$

Let $\mathbf{x}_i^T \boldsymbol{\beta} = u_i$, where $u_i \sim \mathcal{N}(0, \tau^2 R)$. If $\tau = 0$, then we obtain the standard logistic model with fixed effects. The R matrix describes the dependence structure among the u_i values. We assume that the y_i values are mutually independent when the values of u_i are given.

4.11.3 Deriving the Score Test

To test the null hypothesis, $H_0 : \tau^2 = 0$ versus $H_a : \tau^2 > 0$, we use the marginal likelihood $L(\boldsymbol{\alpha}, \tau)$ obtained by integrating out the u_i values.

$$L(\boldsymbol{\alpha}, \tau) = e^{\left[\sum_{i=1}^{n} f_i(y_i|u_i, \boldsymbol{\alpha}, \tau)\right]}$$

To derive the score test for $\tau = 0$, it is necessary to calculate $\frac{\partial \ell(\boldsymbol{\alpha}, \tau)}{\partial \tau}$ and evaluate the resulting derivative at $\tau = 0$. However, it is difficult to evaluate this likelihood analytically. So, the marginal density of the i-th response vector y_i can be obtained as $f_{y_i}(y_i) = E_{u_i}[f_{y_i|u_i}(y_i|u_i)]$, where E_{u_i} denotes the expectation with respect to the distribution of u_i. Following Cox (1983), we can expand the

integrated log likelihood $\ell(\boldsymbol{\alpha}, \tau)$ (i.e.$f_{y_i|u_i}(y_i|u_i)$ using a Taylor series expansion of $\tau = 0$, taking expectations with respect to τ to obtain the marginal density of y_i using the Laplace method.

Taylor expanding $L(\boldsymbol{\alpha}, \tau)$ (quasi likelihood) with respect to vector u_i yields the following:

$$L(\boldsymbol{\alpha}, \tau) = \exp\Bigg[\prod_{i=1} f_i(0) + \sum_{i=1} u_i \frac{\partial f_i(0)}{\partial u_i} \prod_{i \neq j} f_j(0) + \frac{1}{2}\Bigg(\sum_{i=1} u_i^2 \frac{\partial^2 f_i(0)}{\partial u_i^2} \prod_{i \neq j} f_j(0) + \sum_{i=1} \sum_{i \neq j} r_i r_j \frac{\partial f_i(0)}{\partial u_i} \frac{\partial f_j(0)}{\partial u_j} \prod_{k \neq i,j} f_k(0)\Bigg) + o(u^2)\Bigg]$$

Let $l_i(u_i)$ be the log density; $l_i(u_i) = log[f_i(u_i)]$. The first and second derivatives satisfy the following:

$$\frac{\partial f_i(u_i)}{\partial u_i} = f_i(u_i)\big[\frac{\partial l_i(u_i)}{\partial u_i}\big]$$

$$\frac{\partial^2 f_i(u_i)}{\partial u_i^2} = f_i(u_i)\left[\frac{\partial^2 l_i(u_i)}{\partial u_i^2} + \left(\frac{\partial l_i(u_i)}{\partial u_i}\right)^2\right]$$

Take the following expectation:

$$L(\boldsymbol{\alpha}, \tau) = \prod_{i=1} f_i(0) \times \Bigg(1 + \frac{1}{2}\Bigg\{\sum_i R_{ii}\left[\frac{\partial^2 l_i(u_i)}{\partial u_i^2} + \left(\frac{\partial l_i(u_i)}{\partial u_i}\right)^2\right] + \sum_{i=1} \sum_{i \neq j} R_{ij} \frac{\partial l_i(0)}{\partial u_i} \frac{\partial l_j(0)}{\partial u_j}\Bigg\}\Bigg) + o(\tau^2) \quad (4.21)$$

If $\boldsymbol{\alpha}$ is known in our case, we do not have $\boldsymbol{\alpha}$, which is a fixed coefficient; the score test for $H_0 : \tau^2 = 0$ has the form

$$\frac{\frac{\partial log L(\boldsymbol{\alpha}, 0)}{\partial \tau^2}}{E\left[\left(\frac{\partial log L(\boldsymbol{\alpha}, 0)}{\partial \tau^2}\right)\right]^{1/2}}.$$

Let the likelihood under the null model $L(\boldsymbol{\alpha}, 0) = \prod f_i(0)$. The first derivative of the log likelihood with respect to τ^2 satisfies

$$\frac{\partial log L(\boldsymbol{\alpha}, 0)}{\partial \tau^2} = \frac{\partial}{\partial \tau^2}\Bigg\{\frac{1}{2}\Bigg(\sum_i R_{ii} \frac{\partial^2 \ell_i(0)}{\partial u_i^2} + \sum_{i=1} \sum_{i,j} R_{ij} \frac{\partial \ell_i(0)}{\partial u_i} \frac{\partial \ell_j(0)}{\partial u_j}\Bigg)\Bigg\}.$$

The first and second derivatives of $\ell_i(u_i)$ with respect to u_i are

$$\frac{\partial \ell_i(u_i)}{\partial u_i} = [y_i - \kappa_{1i}(u_i)]/a(\phi)$$

$$\frac{\partial^2 \ell_i(u_i)}{\partial u_i^2} = -\kappa_{2i}(u_i)/[a(\phi)]^2,$$

where $\kappa_{hi}(u_i)$ is the h^{th} moment of y_i. It is a function of u_i.

$$= \frac{1}{2[a(\phi)]^2}\left[\sum_{i=1}\sum_{i,j} R_{ij}(y_i - \kappa_{1i})(y_i - \kappa_{1i}) - R_{ii}\kappa_{2i}\right]$$

$$U = \frac{1}{2}[\mathbf{y} - \boldsymbol{\kappa}_1]^T R[\mathbf{y} - \boldsymbol{\kappa}_1] - tr(RV), \tag{4.22}$$

where V is the diagonal matrix of κ_{2i}, and $a(\phi) = 1$ in the case of a binary response with a canonical link function. This result and that of Lin (1997) are the same; however, in Lee's paper, he derived a score test for GLMMs in general, while we have y has a distribution from the exponential family with a canonical link.

To calculate the variance of the $\mathbf{U}$ score, we will take the expected square:

$$\begin{aligned} E(\mathbf{U}^2) &= E\left[\left(\frac{\partial logL(\boldsymbol{\alpha}, 0)}{\partial \tau^2}\right)^2\right] \\ &= \left[\frac{1}{2}\right]^2 E\left[\left\{[\mathbf{y} - \boldsymbol{\kappa}_1]^T R[\mathbf{y} - \boldsymbol{\kappa}_1]^2\right\} - \left\{\sum_i R_{ii}\kappa_{2i}\right\}\right]^2 \\ &= \frac{1}{4} var(\mathbf{U}^2) \end{aligned}$$

The variance of $\mathbf{U}^2$ can be expressed as a function of the second and fourth cumulates, κ_2 and κ_4, of $\mathbf{U}$ (Kendall & Stuart, 1977). We will focus on the first term on the right-hand side of (4.22). Let $C_i = (y_i - \kappa_{1i})$. The first moment of C_i under the null model is equal to 0, and the other moments are equal to the moments of y_i. Thus,

$$E\left[\left(C^T RC\right)^2\right] = E\left(\sum_i \sum_j \sum_k \sum_l R_{ij}R_{kl}C_iC_jC_kC_l\right)$$

Since $E(C_i) = 0$ and C_i are independent under the null model, we have

$$E\Big(C_iC_jC_kC_l\Big) = \begin{cases} \kappa_{4i} & \text{if } i = j = k = l \\ \kappa_{2i}\kappa_{2k} & \text{if } i = j, k = l, i \neq k \\ \kappa_{2i}\kappa_{2j} & \text{if } i = k, j = l \quad \text{or} \quad i = l, j = k, i \neq j \\ 0 & \text{Otherwise} \end{cases}$$

From this, it follows that

$$E\Big[\Big(C^TRC\Big)^2\Big] = \sum_i R_{ii}^2\kappa_{4i} + \sum_{j\neq k} R_{ij}R_{kk}\kappa_{2i}\kappa_{2k} + \sum_{i\neq j} R_{ij}^2\kappa_{2i}\kappa_{2j}$$

$$= \sum_i R_{ii}^2\kappa_{4i} + \sum_i\sum_k R_{ii}R_{kk}\kappa_{2i}\kappa_{2k} + 2\sum_i\sum_j R_{ij}^2\kappa_{2i}\kappa_{2j} - 3\sum_i R_{ii}^2\kappa_{2i}^2$$

$$= \sum_i R_{ii}^2(\kappa_{4i} - 3\kappa_{2i}^2) + 2\sum_i\sum_j R_{ij}^2\kappa_{2i}\kappa_{2j} + \Big(R_{ii}\kappa_{2i}\Big)^2.$$

Substituting this in (4.22), we obtain

$$E\left[\left(\frac{\partial logL(\boldsymbol{\alpha}, 0)}{\partial\tau^2}\right)^2\right] = \frac{1}{4}\Big[\sum_i R_{ii}^2(\kappa_{4i} - 3\kappa_{2i}^2) + 2\sum_i\sum_j R_{ij}^2\kappa_{2i}\kappa_{2j}\Big]$$
$$= \frac{1}{4}\sum_i R_{ii}^2(\kappa_{4i} - 3\kappa_{2i}^2) + 2tr(RVRV).$$

The test statistics are as follows:

$$S = \mathbf{q}^TH^{-1}\mathbf{q},$$

where

$$\mathbf{q} = \frac{1}{2}\Big[\mathbf{y} - \boldsymbol{\kappa}_1]^TR[\mathbf{y} - \boldsymbol{\kappa}_1 - tr(RV)\Big] \tag{4.23}$$

and

$$H = \Big[\frac{1}{4}\sum_i R_{ii}^2(\kappa_{4i} - 3\kappa_{2i}^2) + 2tr(RVRV)\Big].$$

Deriving the Variance Component Score Test (Lin 1997)

We derive the score test as suggested in Lin (1997). Recall model 4.20:

$$\text{logit}P(y_i = 1) = \mathbf{z}_i^T \boldsymbol{\alpha} + \mathbf{x}_i^T \boldsymbol{\beta}, \tag{4.24}$$

where Z is an $n \times q$ covariate matrix, and X is a genotype matrix with dimension $n \times p$. We assume that the random effects, $\boldsymbol{\beta}$, are generated from some distribution, F, with a mean of zero and covariance R, where τ is the variance component. We test the null hypothesis $H_0 : \tau = 0$, which is equivalent to $\boldsymbol{\beta} = 0$.

Integrated quasi-likelihood of (α, τ):

$$L(\alpha, \tau) = \exp\{\ell(\alpha, \tau)\} = \int exp\{\sum_{i=1}^{n} \ell_i(\alpha; \beta)\} dF(\beta; \tau),$$

where $\ell(\alpha; \beta) = \int \frac{a_i(y_i - u)}{v(u)}$.

To derive the score test for $\tau = 0$, it is necessary to calculate $\frac{\partial \ell(\alpha,\tau)}{\partial \tau}$ and evaluate the resultant derivative at $\tau = 0$. However, it is difficult to evaluate this likelihood analytically. So, the marginal density of the ith response vector, y_i, can be obtained as $f_{y_i}(y_i) = E_{u_i}[f_{y_i|u_i}(y_i|u_i)]$, where E_{u_i} denotes the expectation with respect to the distribution of u_i. Following Cox (1983), we can expand the integrated log likelihood $\ell(\alpha, \tau)$ (i.e. $f_{y_i|u_i}(y_i|u_i)$) using a Taylor series expansion of $\tau = 0$ and taking expectations with respect to τ to obtain the marginal density of y_i using the Laplace method. Taylor expansion gives the following:

$$exp\{\sum_{i=1}^{n} \ell_i(\alpha; \beta)\} = exp\{\sum_{i=1}^{n} \ell_i(\alpha; 0)\}\big(1 + \sum_{i=1}^{n} \frac{\partial \ell_i(\alpha; 0)}{\partial \eta_i} x_i^T \beta + \frac{1}{2}\beta^T \big[\{\sum_{i=1}^{n} \frac{\partial \ell_i(\alpha; 0)}{\partial \eta_i} x_i\} \{\sum_{i=1}^{n} \frac{\partial \ell_i(\alpha; 0)}{\partial \eta_i} x_i\} + \sum_{i=1}^{n} \frac{\partial \ell_i(\alpha; 0)}{\partial \eta_i} x_i x_i^T\big]\beta + \epsilon\big). \tag{4.25}$$

Write the integrated quasi-likelihood as $L(\alpha, \tau) = E(\exp\{\sum_{i=1}^{n} \ell(\alpha; \beta)\})$:

$L(\alpha; \tau) = \exp\big\{\sum_{i=1}^{n} \ell(\alpha; 0)\big\}\big\{1 + \big\{\frac{1}{2} tr\big(\big[\frac{\partial \ell_i(\alpha;0)}{\partial \eta_i} x_i\big\}\big\{\frac{\partial \ell_i(\alpha;0)}{\partial \eta_i} x_i^T\big\} + \frac{\partial \ell_i(\alpha;0)}{\partial \eta_i} x_i x_i^T\big] R\big) + ||\tau||\big\}$.

The marginal log likelihood $\ell(\alpha, \tau)$ can then be written as

$$\ell(\alpha,\tau)=\sum_{i=1}^{n}\ell(\alpha;0)+\frac{1}{2}tr\big[X^T\big\{\frac{\partial\ell_i(\alpha;0)}{\partial\eta_i}\frac{\partial\ell_i(\alpha;0)}{\partial\eta_i^T}\frac{\partial^2\ell_i(\alpha;0)}{\partial\eta_i\partial\eta_i^T}\big\}XR\big]+O||\tau||.$$

Let Δ and W be an $n\times n$ diagonal matrix with elements

$\delta_i=\frac{1}{\bar{g}(\mu_i)}$, $w_i=\big[V(\mu_i)\{\bar{g}(\mu_i)\}^2\big]^{-1}$, where $\mu_i=E(y_i)$ under H_0 and $g(\mu_i)=x_i^T\alpha$. First, we derive $\bar{g}(\mu_i)$ as

$$\bar{g}(\mu_i)=\frac{\partial}{\partial\mu_i}\big[\log\frac{\mu_i}{(1-\mu_i)}\big]$$

$$=\frac{\partial}{\partial\mu_i}\big[\log\mu_i-\log(1-\mu_i)\big]$$

$$=\frac{1}{\mu_i}+\frac{1}{(1+\mu_i)}=\frac{1-\mu_i+\mu_i}{\mu_i(1-\mu_i)}=\frac{1}{\mu_i(1-\mu_i)}.$$

Thus,

$$\delta_i=\mu_i(1-\mu_i)$$

and

$$w_i=\big[V(\mu_i)\{\bar{g}(\mu_i)\}^2\big]^{-1}=\big[\mu_i(1-\mu_i)\times\big\{\frac{1}{\mu_i(1-\mu_i)}\big\}^2\big]^{-1}=\mu_i(1-\mu_i).$$

Therefore,

$$-\frac{\partial^2\ell(\alpha;0)}{\partial\eta\partial\eta^T}=diag(w_i+e_i(y_i-\mu_i)),$$

where $e_i=0$ when the link function is canonical.

$U_\tau(\hat{\beta}_0)=\frac{\partial\ell(\alpha,\tau)}{\partial\tau}\Big|_{\tau=0,\alpha=\hat{\beta}_0}$

$$=\frac{1}{2}=tr\big[\big\{W\Delta^{-1}(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T\Delta^{-1}W-W\big\}ZZ^T\big]$$

$$=\frac{1}{2}=tr\big[(\mathbf{y}-\boldsymbol{\mu})W\Delta^{-1}XX^T\Delta^{-1}W(\mathbf{y}-\boldsymbol{\mu})^T-tr(WXX^T)\big]$$

Since our model has a canonical link function, and based on the derivation of Δ and W, $\Delta^{-1}W=I$, where I is the identity matrix.

We can re-write it as follows:

$$U_\tau(\hat{\beta}_0)=\frac{1}{2}=tr\big[(\mathbf{y}-\boldsymbol{\mu})XX^T(\mathbf{y}-\boldsymbol{\mu})^T-tr(WXX^T)\big].$$

Since the right-hand part does not depend on random y values, it is a constant and can be ignored. We can simplify it one more time as

$$U_\tau(\hat{\beta}_0) = \frac{1}{2} tr\big[(\mathbf{y} - \boldsymbol{\mu}) X X^T (\mathbf{y} - \boldsymbol{\mu})^T\big]. \tag{4.26}$$

$E(U_\tau(\hat{\beta}_0))$ is generally an increase function of τ Zhang & Lin (2008). For example, based on Zhang & Lin (2008), Figure 4.5 shows the expected score of $E(U_\tau(\hat{\beta}_0))$ versus τ for logistic-normal model where $n = 10$, $x_{ij} = 1$, $x_{ij} = 1$, and $\beta = 0.25$. Based on the arguments above, a large value of $U_\tau(\hat{\beta}_0)$ gives evidence against $H_0 : \tau = 0$; hence, H_0 is only rejected if $U_\tau(\hat{\beta}_0)$ is large.

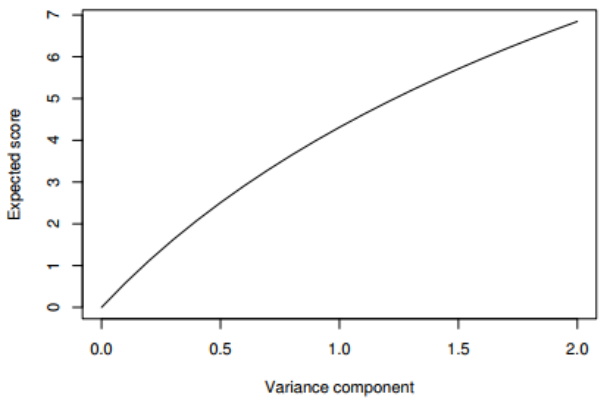


Figure 4.5: Expected score as a function of variance component τ (Zhang & Lin, 2008).

We can use equation (4.26) as a test statistic after omitting the trace part, which is the proposed test that will be used in subsequent chapters.

4.11.4 The Score Test with Variant Weights

Recall the following model 4.20:

$$\text{logit} P(y_i = 1) = \mathbf{z}_i^T \boldsymbol{\alpha} + \mathbf{u}_i, \tag{4.27}$$

where $u_i \sim \mathcal{N}(0, \tau^2 \tilde{R})$. If $\tau = 0$, we obtain the standard logistic model with a fixed effect. The $\tilde{R}$ matrix describes the dependence structure among the u_i

values and the variant weights, which are $X\Gamma^T\Gamma X^T$, where X is the genotype matrix. Γ is the diagonal matrix representing the variant weights. It can be expressed as $\boldsymbol{\beta} \sim \mathcal{N}(0, w_j\tau^2)$ under the model $logitP(y_i = 1) = \mathbf{z}_i^T\boldsymbol{\alpha} + \mathbf{x}_i^T\boldsymbol{\beta}$. We assume that the y_i values are mutually independent when the values of u_i are given.

To test the null hypothesis, $H_0 : \tau^2 = 0$ versus $H_a : \tau^2 > 0$, we use the marginal likelihood $L(\boldsymbol{\alpha}, \tau)$, which is obtained by integrating out the u_i values.

$$L(\boldsymbol{\alpha}, \tau) = e^{\left[\Pi_{i=1}^n f_i(y_i/u_i, \alpha, \tau)\right]}$$

To derive the test, we use the same derivation provided in the variance component section. The only difference between these derivations appears on the R matrix in the previous model, which has weight schemes here ($\tilde{R}$).

Thus, the score test is

$$E\left[\left(\frac{\partial logL(\boldsymbol{\alpha}, 0)}{\partial\tau^2}\right)^2\right] = \frac{1}{4}\sum_i \tilde{R}_{ii}^2(\kappa_{4i} - 3\kappa_{2i}^2) + 2tr(\tilde{R}V\tilde{R}V).$$

The global score test statistics are as follows:

$$S = \mathbf{q}^T H^{-1}\mathbf{q}, \tag{4.28}$$

where

$$\mathbf{q} = \frac{1}{2}\Big[(\mathbf{y} - \boldsymbol{\mu})^T\tilde{R}(\mathbf{y} - \boldsymbol{\mu}) - tr(\tilde{R}V)\Big],$$

where $\tilde{R} = X\Gamma\Gamma^T X^T$ and

$$H = \frac{1}{4}\Big[\sum_i \tilde{R}_{ii}^2(\kappa_{4i} - 3\kappa_{2i}^2) + 2tr(\tilde{R}V\tilde{R}V)\Big].$$

Instead of using the global test, S, it have been used only $\mathbf{q}$, ignoring the trace term because it does not involve y. All randomness of $\mathbf{q}$ comes from the ratio of the quadratic form in the first term of $\mathbf{q}$.

Moreover, the reason for ignoring the H part not involved in the test is the sufficiency of $\mathbf{q}$. If we assume the likelihood has only one local maximum, then $\mathbf{q}$ itself is a reasonable test statistic for testing H_0. A larger value of $\mathbf{q}$ indicates a departure from H_0. Also, using the global test, S, would treat both large and

small $\mathbf{q}$ values indifferently, so the test would lose power Zhang & Lin (2008). Based on the previous methods in this chapter, the score test derived from the fixed model is

$$\mathbf{q}_{fixed} = \frac{1}{2}\Big[(\mathbf{y}-\boldsymbol{\mu})^T\tilde{R}(\mathbf{y}-\boldsymbol{\mu})\Big],$$

and the score test based on the random effect is

$$\mathbf{q}_{random} = \frac{1}{2}\Big[(\mathbf{y}-\boldsymbol{\mu})^T\tilde{R}(\mathbf{y}-\boldsymbol{\mu}) - tr(\tilde{R}V)\Big].$$

Since we can ignore the trace part, $\mathbf{q}_{fixed}$ is equivalent to $\mathbf{q}_{random}$ in terms of the test form. From this point on and in subsequent chapters, we will consider the test to be based on $\mathbf{q}$. Since it can be used with different kinds of data (i.e. correlated or independent), there is no singularity issue.

4.11.5 Simulation

We evaluated both type I errors and the power of the proposed score tests with different weights, as we discussed in the context of a multi-locus association analysis with different numbers of SNPs. To obtain the genotype matrix, X, we generated $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ via the multivariate normal distribution, with a variance of 1 and a pairwise correlation between $z_{\grave{i}}$ and z_j at $0.5^{|\grave{i}-j|}$; $1 \leq \grave{i}, j \leq p$ between any two latent components. Next, we threshold each latent vector component to obtain a vector of binary variables, say $(\mathbf{d})$, which represents the haplotype. We generate two vectors of haplotypes d_1 and d_1 for each individual. Subsequently, we combine two independently generated haplotypes $(\mathbf{d})$ by taking the sum of $X = \mathbf{d}_1 + \mathbf{d}_2$, given vector X $(0/1/2)$, which represents the genotype. The details of this simulation will be explained in Chapter 5.

The threshold for component j, say c_j, was obtained such that $P(d = 1)$ was controlled to mimic rare or common variants.

We set $p = 100$ and 200 and $OR = 1$ in the model for type I error simulation (4.1) and $OR = 3$ to determine the power. This simulation was considered by Pan *et al.* (2014). To estimate the p-value, straight binomial proportions are used. Hence, they have the same standard error as any other binomial proportion, $\sqrt{(p(1-p)/n)}$, where p means the proportion of tests rejected and n represents the number of samples. Therefore, if $p = 0.05$ and $n = 2000$, the standard error

of the observed proportion is about 0.005, and we could say the uncertainty level is 1%.

4.11.6 Type I Errors and Power

To evaluate type I errors, we generated a genotype matrix as explained above, while we set the $OR = 1$. Three types of variants were considered in the simulation, based on a given MAF: extremely non-causal, moderately rare, and common. We used two scenarios to evaluate type I errors. For the first scenario, the maximum and minimum MAF rates were the same for each type of variant. We considered all values of MAF, from the MAF boundary of $1/n$ to the common variants at 0.5 (see Figure 4.6, data not scaled and Figure 4.7, data scaled).

In the second scenario, we generated the X_p in the same manner as that in the simulation above, so the data had different types of variants (e.g. rare and common). We randomly generated rare variants, ranging from the MAF boundary to $MAF = 0.01$, and common variants, ranging between 0.05 and 0.5. Then, we varied the second set, which considered common variants (see Figure 4.4). We conducted 1000 simulated datasets; the genotypes were randomly generated for each simulation. For both scenarios, we estimated the empirical type I error rate as a proportion of p-values less than the nominal level: $\alpha = 0.05$. The tests with both weights proposed here (beta and Cauchy) seemed to have satisfactory type I error rates that were well controlled at the specified nominal level of $\alpha = 0.05$. When the MAF is very low, such as 0.0005, this means there are two variants among 2000 individuals (there are 1998 zeros and two elements are 1 or 2), so this affects the control of the type I error rate in the score test. The concern is when we have an MAF on the boundary (i.e. an MAF of less than 0.002). We can see in Figure 4.7 that when there is only one or two variant(s) in the data set, the type I error rate is lower than the nominal level of 0.05.

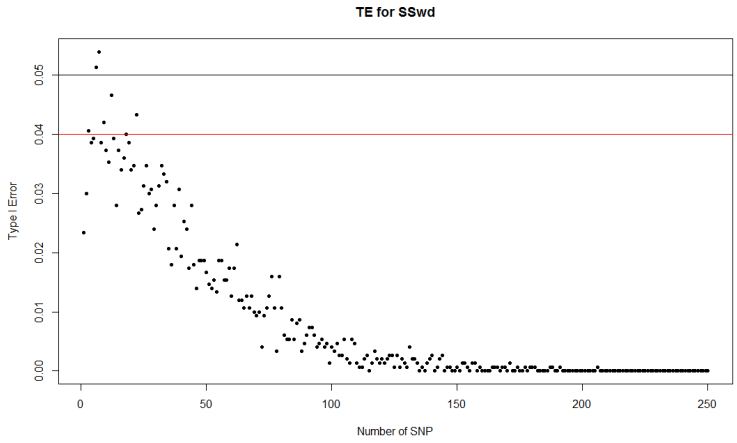


Figure 4.6: Type I error rate in scenario 1 which is the maximum and minimum MAF rates were the same for each type of variant. Data are not scaled. The horizontal lines highlight 0.04 and 0.06.

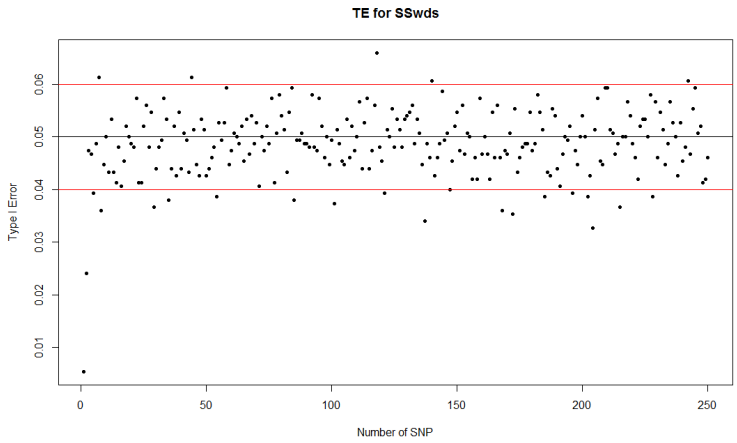


Figure 4.7: Type I error rate in scenario 1. Data are scaled. The horizontal lines highlight 0.04 and 0.06.

	MAF (0.05,0.5)				
Test	2%	20%	30%	40%	50%
Score test	0.05	0.05	0.04	0.06	0.05

Table 4.4: Type I error rate for the score test using scenario 2.

The power of using the variance component test will be studied in detail, using different weighting schemes, in subsequent chapters.

4.11.7 Conclusion

The score test with the proposed weight can control type I errors at different MAFs, except when the data are extremely rare, such as 0.0005 (when $n = 2000$).

The simulation in this chapter shows that the weighting scheme plays an important role. It is crucial to include weights in an analysis of rare variants since their MAF is lower than that of common variants, which dominate the association signal because of the difference in the MAF level. As a result of the impact of weight, the weighting scheme plays an important role in boosting the power (Lee *et al.* (2014), Wu *et al.* (2011)). In the following chapters, we will propose different weight schemes based on variants, as well as other novel schemes.

Chapter 5

Variant Weight Functions

5.1 Introduction

In the previous chapter, we proposed different association tests based on varying weights. We use a beta function of minor allele frequency (MAF) as a weight of association test. We believe that an appropriate weight can boost a test's power. Good weight choices can easily increase power Wu *et al.* (2011) in rare variant association. In this chapter, we will investigate alternative weights that can be used for rare variants.**We propose a new weighting scheme that we believe may initiate a new direction in the analysis of rare variants**. We also propose a number of weight schemes as a function of MAF. The scenarios in which these different schemes can be applied are explained in this chapter.

A beta function up-weights rare variants and down-weights common variants, which increases the contribution of rare variants to the test statistics. However, weights associated with rare variants that use a beta function with the parameters $(1, 25)$, such as in Wu *et al.* (2011), do not provide very different weights to MAF values in the extreme region. The result is that there is no difference in the weighting of extremely rare and moderately rare variants. The maximum weight value will be associated with singleton variants, which is 25. For example, when $n = 2000$, the rarest variant with a MAF of 0.00025 will be associated with the value of 25, while a MAF of 0.002 will be associated with 23. The magnitude of the difference is small. Therefore, when analysing data that are very rare (i.e., the maximum MAF value is less than 0.003), using a beta function as a

weight is almost the same as using an un-weighted scheme due to the assigned weight values. Therefore, one limitation of the weights proposed in the literature is that they depend on arbitrary parameters rather than on the MAF. When the parameter is based on the MAF or a function of the MAF, it can be modified based on the observed MAF.

In this section, we will propose a weighting scheme that addresses the above problem by using a heavy-tailed distribution to make the differences between different MAFs apparent even if they are very rare. Therefore, this weighting scheme (i.e., an adaptive choice of weight) is a new type of weighting scheme because it chooses weight parameters depending on the MAF (a function of MAF ($\mathcal{F}$)).

Another issue that arises in the association of rare variants is a large number of singletons can arise when next-generation sequencing is used. Variants that are observed only once in the dataset are difficult to distinguish from sequencing errors. The idea proposed here is a new weighting scheme, which, instead of up-weighting all rare variants so that the rarest variants are given the largest weights, up-weights rare variants under some constraints. We also proposed, as another property of the Gumbel function that all rare variants should be up-weighted, that singleton variants be given smaller weights than other rare variants. For these reasons, the smallest MAF will no longer be associated with the largest weight.

In this chapter, we will propose three different variant weight schemes based on the Cauchy function and two schemes based on the Gumbel function. We will also investigate the differences between them and evaluate the type I errors and the power of the tests. In each chapter of this thesis, we use the classification of variants by MAF described in Figure 5.1.

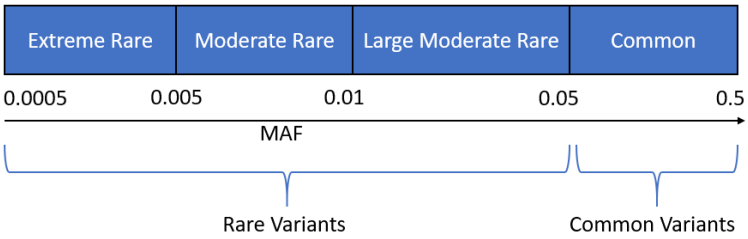


Figure 5.1: Terminology used for the levels of MAF in this thesis. Note that $n = 2000$ is the number of individuals, so a MAF of 0.005 is set as the threshold for extremely rare variants.

The structure of variant data can contain rare and common variants after a new technique has been implemented. In general, there are many scenarios in which the data structure ranges between very rare and common variants. The weight schemes that we propose in Chapters 5 and 6 can be useful in different scenarios by utilising either the adaptive weighting schemes or continuous weighting schemes described in Chapter 6. Figures 5.2 and 5.3 show different data scenarios.

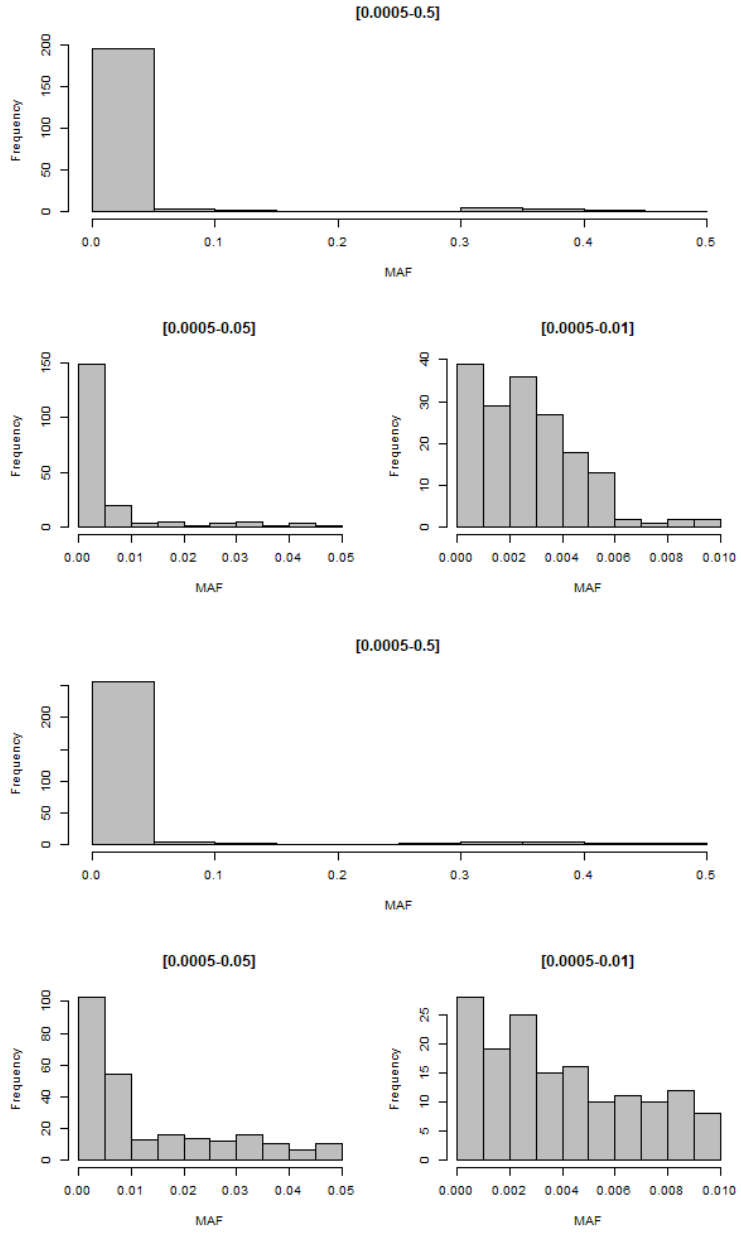


Figure 5.2: The top figure shows the simulated data has a large proportion of extremely rare variants, while the bottom figure shows there are large amount of extremely and moderately rare variants.

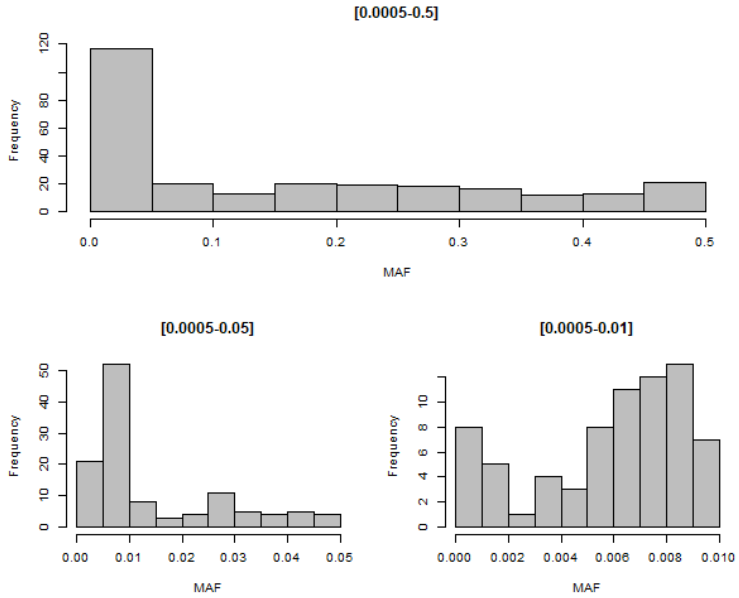


Figure 5.3: This figure shows the simulated data has a large proportion of common variants.

The causal variant distribution is unknown, so the proposed weight schemes in Chapter 5 and 6 differ regarding the causal variants distribution models. For example, some weighting schemes described in Chapter 5 assume causal variants are in rare variant regions, while in Chapter 6, we present weight schemes that can consider the distribution of causal variants within a wide range.

5.2 Simulation

We evaluated the type I errors and power of the proposed score tests with different weights, as we discussed previously in the context of multi-locus association analysis with a different number of SNPs. To obtain the genotype matrix X, we generated $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ via the multivariate normal distribution, with variance 1 and pairwise correlation between $z_{\grave{i}}$ and z_j at $0.5^{|\grave{i}-j|}$ and $1 \leq \grave{i}, j \leq p$ between any two latent components. Next, the threshold of each latent vector component is used to obtain a vector of binary variables, such as $(\mathbf{d})$, which represent haplotypes. We generate two vectors of haplotypes, $\mathbf{d}_1$ and $\mathbf{d}_2$, for

each individual. Then, we combine two independently generated haplotypes ($\mathbf{d}$) by taking the sum of $X = \mathbf{d}_1 + \mathbf{d}_2$ given a vector X (0/1/2), which represents genotype. The threshold for component j, c_j, was obtained using a method that controls $P(d = 1)$ to mimic the rare or common variants. Finally, given an odds vector $\boldsymbol{\beta}$, we generate the disease status y (0/1) for each X such that

$$\text{Logit}Pr(y_i = 1) = \frac{e^{Pr(y_i=1)}}{1 + e^{Pr(y_i=1)}} = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}, \tag{5.1}$$

where $\beta_0 = log(0.05/(1 - 0.05))$. For the null case, we set $OR = 1$; for non-null cases (i.e., causals), we set $OR = 3$ for extremely rare variants, $OR = 2$ for moderately rare variants, and $OR = 1.5$ for common ones.

Each allele of the haplotype is generated by dichotomising the marginal normal distribution. The cut-off is determined by the allele frequency that is randomly sampled from a uniform distribution between 0.0005 and 0.005 as extremely rare variants, 0.005 and 0.01 as moderately rare variants, 0.01 and 0.05 as large moderately rare variants, and 0.05 and 0.5 as common variants. We consider causal variants in the $0.03\% - 0.2\%$ range and the model with OR ranging from 1 (for the type I error rate) to 3.

We generated simulated data over a spectrum of MAFs, odds ratios, and proportions of extremely, moderately rare, and common variants. We set the number of variants (i.e., causal and non-causal) to $p = 200$ and 100 SNPs in all of the scenarios. Causal variants are set, in most cases, between 5% and 20% of the total variants. Each dataset will contain three types of variants (i.e., extremely and moderately rare variants and common variants), and the default percentages are 40% for ERV and 40% MRV and 20% CV. When these percentages change, it will be specified in the captions of the figures. Based on the proposed weighting schemes presented in sections (5.3.2) and (5.3.3), the percentage of variant types will differ to express the weight scheme's impact, and these percentages will be provided under each figure. For example, to express the impact of having a few extremely rare variants (ERV) compared to a large number of common ones, we will increase the common ones from 5% to 90%; see Table (5.2). Variants with $OR = 1$ are used in the standard logistic model with fixed effects (5.1) to simulate type I errors and $OR > 1$ to evaluate the tests' power.

Parameters	Parameter Values
Sample Size	n =2000 (# cases = # control = n/2)
Total Number of SNPs	100, 200
Proportion of Causal SNPs	[3%-20%]
Effect Size for Non-causal SNPs	OR = 1
Effect Size for Causal SNPs	OR = Unif(1.5,3) or Unif(1/2,1/4)
Percentage of Common Variants	Ranged from 10% to 90%

Table 5.1: The full set of parameters used in the simulation.

We generated data under various causal mechanisms and MAFs for causal and non-causal variants, as summarised in Table (5.2).

S	RE-C	RE-N	RM-C1	RM-N1	RM-C2	RM-N2	C-C	C-N
1	0.0005-0.005	0.0005-0.005	-	0.005-0.01	-	0.01-0.05	-	0.05-0.5
2	-	0.0005-0.005	0.005-0.01	0.005-0.01	-	0.01-0.05	-	0.05-0.5
3	-	0.0005-0.005	-	0.005-0.01	0.01-0.05	0.01-0.05	-	0.05-0.5
4	-	0.0005-0.005	-	0.005-0.01	-	0.01-0.05	0.05-0.5	0.05-0.5
5	0.0005-0.005	0.0005-0.005	0.005-0.01	0.005-0.01	0.01-0.05	0.01-0.05	0.05-0.5	0.05-0.5

Table 5.2: Summary of the four types of variants in the five scenarios utilised in the simulation study. For each variant, the MAF value was generated uniformly from the range given. RE-C and RE-N represent extremely rare causal and non-causal variants, respectively, while RM-C and RM-N represent moderately rare causal and non-causal variants, respectively, which are represented in two ranges. Finally, C-C and C-N represent common causal and non-causal variants, respectively.

To estimate p-values, straight binomial proportions are used. Hence, they have the same standard error as any other binomial proportion $\sqrt{(p(1-p)/n)}$,

where p here means the proportion of tests rejected and n the number of samples. Therefore, if $p = 0.05$ and $n = 2000$, the standard error of the observed proportion is about 0.005, and we could say the uncertainty is 1%.

Note: In some captions of the Figures that have the new simulation, we attached a table that illustrates the used parameters in that simulation. The following table (5.2) is an example:

	Causal	Non-Causal
ERV	.	✓50%
MRV	OR=2 (0.005-0.5)	✓30%
CV	.	✓20%

Table 5.3: We have included causal and non-causal variants in the simulated data, and these variants are classified into three categories: extremely rare variants (ERV), moderately rare variants (MRV), and common variants (CV). "✓" means we considered these variants, and '.' means we did not. The default percentage of non- causal variants are 50% for ERV and 30% and 20% for CV. If these percentages are different, it will be illustrated in the Figure's caption or in the table as depicted above.

Also, we have used McNemar's test to test the differences between the proposed weighting schemes and SKAT weighting scheme. In the analysis, we compared two proportions: the proportion of times that an association is detected (i.e. power) using one method compared to the proportion using another method. For example, given that the number of simulation is 1000, comparing 800/1000 with 750/1000 (if power is 80% and 75% respectively) and can test whether these differences are statistically significant. Since we use the same simulated datasets to analyse using both methods, and then the results are paired, an appropriate test would be McNemar's. We show the result of McNemar's test in detail in Figure 5.5 and we use that in the all figures that compared between two methods in this chapter. The two curves can be considered different if we can identify that the two methods differ in at least one point (such as at a MAF or percentage of causal variants).

		Method 2		Row Total
		Significant	Not Significant	
Method 1	Significant	a	b	a+b
	Not Significant	c	d	c+d
Column Total		a+c	b+d	n

McNemar's Test

The test assesses if a statistically significant change in proportions have occurred on a dichotomous trait at two time points on the same population. It is applied using a 2×2 contingency table with the dichotomous variable at event 1 and event 2.

The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same, i.e. $a + b = a + c$ and $c + d = b + d$.

Thus the null and alternative hypotheses are[1]

$$H_0 : \ b = c \, H_0 : \ b = c$$
$$H_1 : \ b \neq c \, H_1 : \ b \neq c$$

5.3 Cauchy

As mentioned in the introduction section, when using a weight function, its parameters are a function of the observed MAF, so adjusting the weight based on the MAF is suggested. Here, using a Cauchy function as a weighting scheme can provide this property. We will introduce three different schemes by modifying the parameters of a Cauchy function. We will investigate the impact of these weight schemes on different data scenarios. A general Cauchy function is given by

$$g(x) = \frac{1}{\pi s(1 + ((x - c)/s)^2)} \tag{5.2}$$

5.3.1 Cauchy Weight - Fixed Parameters

In this section, we introduce a weighting scheme that can manage up-weighting rare variants and down-weighting common ones. The data's distribution will not have an impact on the weight (see Figure 5.8). Therefore, increasing the number

of rare or common variants will not affect the magnitude of the weight because the parameter is fixed on this weighting scheme, which will outperform the beta function, especially when the causal variant is in the rare variant range.

The parameters of the weights in this scheme are functions of the MAFs. This weighting scheme is suggested if the data consist of rare variants and no common variants, or if only the rare variants are of interest.

The weight 5.2 is based on the Cauchy function with $c = min(\mathcal{F})$ as the location parameter of Cauchy density, where $\mathcal{F}$ is the MAF, and scale $s = 1/\sqrt{2n}$ as variants with $MAF \leq 1/\sqrt{2n}$ are considered rare, whereas variants with $MAF \geq 1/\sqrt{2n}$ are considered common (Tony Cai *et al.* (2011), Jeng *et al.* (2012), and Ionita-Laza *et al.* (2013)).

$$g(\mathcal{F}) = \frac{1}{(s \times (1 + ((\mathcal{F} - c)/s)^2)}. \tag{5.3}$$

We can rewrite the above equation as

$$g(\mathcal{F}) = \frac{20}{n} \times \left[1 + \left(\frac{n\mathcal{F} - 1}{20}\right)^2\right]. \tag{5.4}$$

Notably, in Figure 5.4, the Cauchy function up-weights the extremely rare variants more than moderate and common variants.

Given that in the simulation, the standard error is 0.005, and the type I error is controlled at a nominal level of 0.05, with 1% uncertainty, there is a difference between the Cauchy and SKAT weighing schemes in Figure 5.5. The difference between two weighting schemes is in the between MAF = 0.0005 to 0.005, indicating that the Cauchy weighting scheme outperforms the SKAT weighting scheme at this range of MAF, however, there is no significant difference between them when the MAF of the causal variants is between 0.005 and 0.03. Power estimates are based on $1,000$ simulation replicates. The differences above were tested using McNemar's test and the results are shown in Table 5.4. To clarify that the uncertainty that we calculated from McNemar's test is accurate, we repeated this analysis 100 times and performed the test also 100 times which gave us the same conclusion that we concluded here.

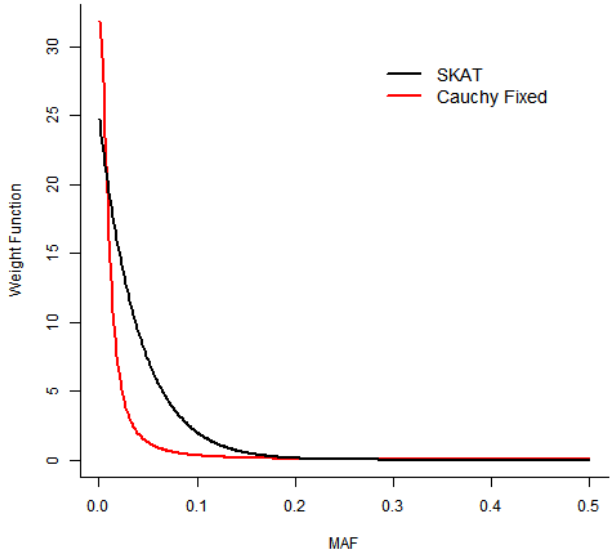


Figure 5.4: The weight using Cauchy function (fixed parameter) versus MAF.

MAF	P.values	Conclusion
0.0005	2.2e-16	Difference is significant
0.001	2.2e-16	Difference is significant
0.0015	2.2e-16	Difference is significant
0.002	2.2e-16	Difference is significant
0.0025	2.2e-16	Difference is significant
0.005	0.3832	Difference is not significant
0.0075	1	Difference is not significant
0.01	1	Difference is not significant
0.015	1	Difference is not significant

Table 5.4: The results of McNemar's test. The test was conducted at each MAF of causal variants.

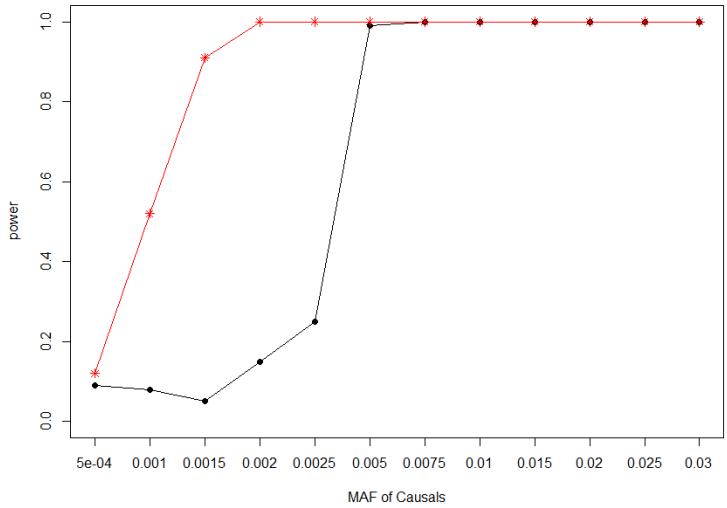


Figure 5.5: This figure shows the impact of increasing the MAFs of causal variants at $OR = 3$. The data is randomly generated so that 80% of the data are rare variants with a MAF between 0.0005 and 0.05, and the remainder are common variants. The red line with the star points shows the results of the test with the Cauchy weight, and the black line with round points shows the results of the test with no weights.

	Causal	Non-Causal
ERV	OR=3	✓
MRV	OR=2	✓
CV	OR=1.5	✓

Given the same parameters of uncertainty, the Cauchy weighting scheme outperforms SKAT when the causal variants located in the ERV range (0.0005 – 0.002), especially when the percentage of the causal in ERV range gets increased as in the top of Figure 5.6, however, the difference between them becomes smaller when the MAF of the causal variants increased as in the bottom of the Figure 5.6. According to the McNemar's test, the differences in both figures are significant. In Figure 5.7 (bottom one), there is no significant difference between Cauchy and SKAT weighting schemes in terms of the power while there is a significant difference in the top figure between them according to the McNemar's test.

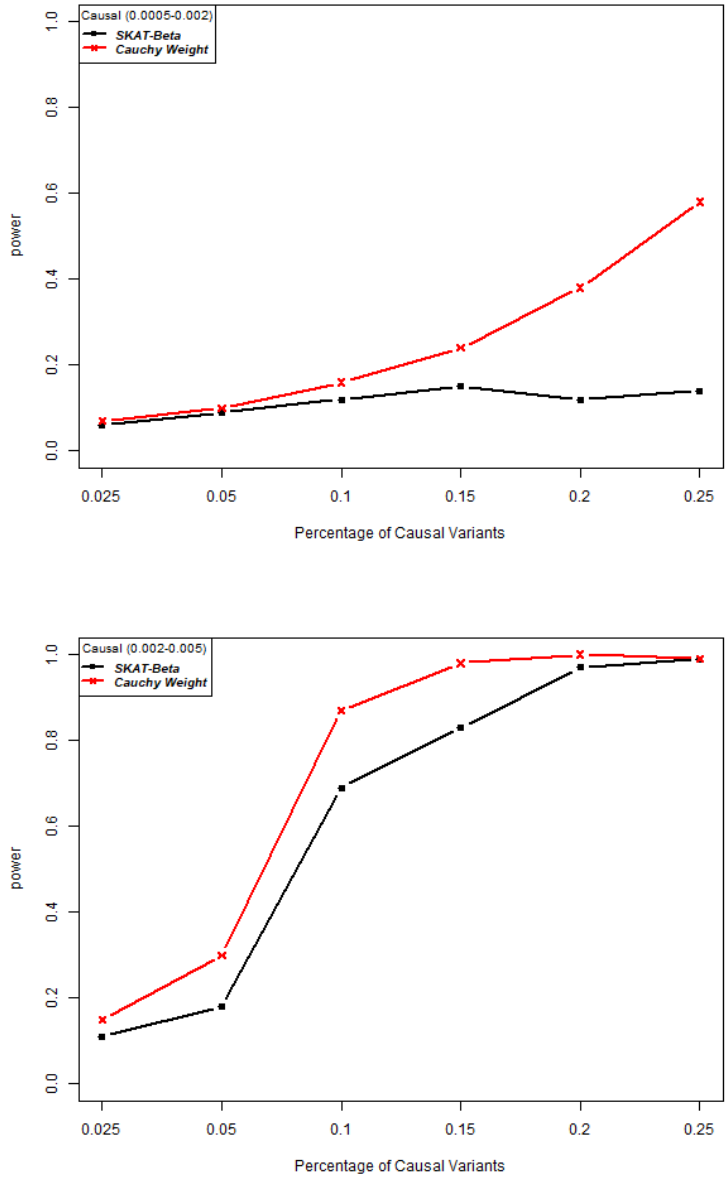


Figure 5.6: The analysis was conducted with ERVs as the causal variants, with $OR = 3$ and MAFs between $0.0005 - 0.002$ in the top figure, while the bottom one uses $OR = 2$ and MAFs between $0.002 - 0.005$, non-causal variants ranging between $0.0005 - 0.05$ as 50% between $0.0005 - 0.005$, 25% between $0.005 - 0.01$, and 25% between $0.01 - 0.05$. The causal variants increased from 2% to 25%.

	Causal	Non-Causal
ERV	OR=3 (0.0005-0.002)	✓
MRV	.	✓
CV	.	✓

	Causal	Non-Causal
ERV	OR=2 (0.002-0.005)	✓
MRV	.	✓
CV	.	✓

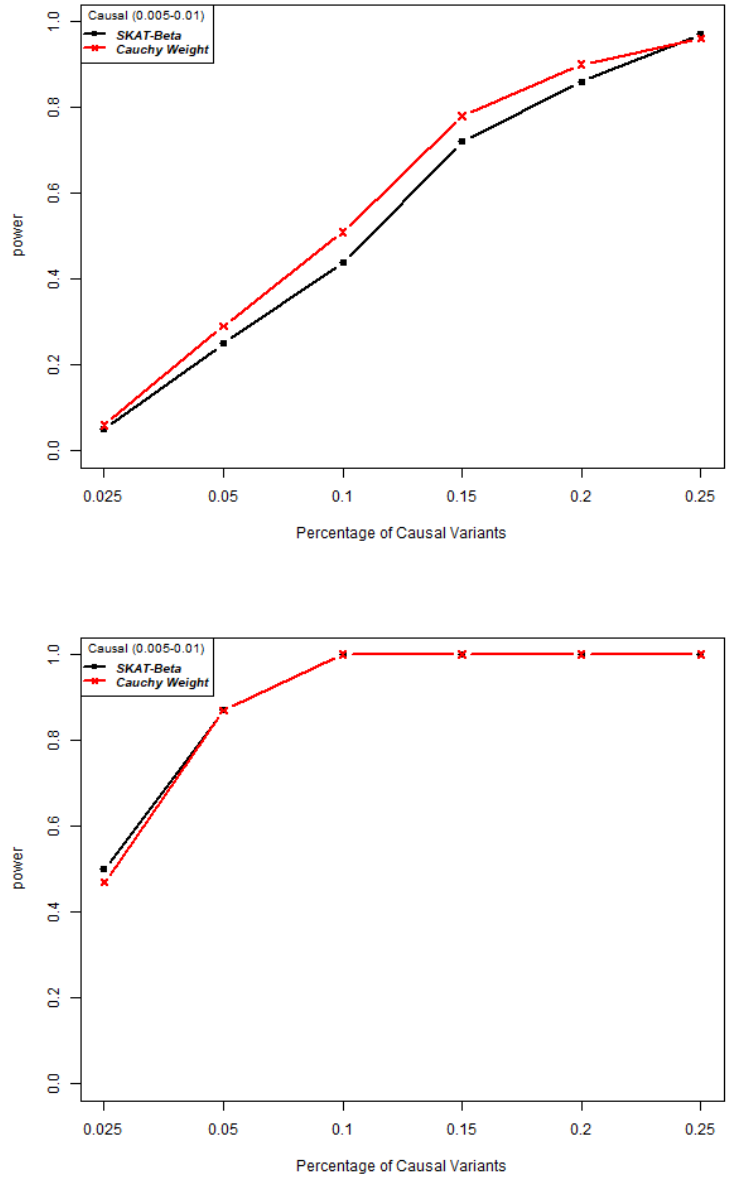


Figure 5.7: The analysis was conducted with MRVs as the causal variants, with $OR = 1.5$ and MAFs between $0.005 - 0.02$ in the top figure, while the bottom one has $OR = 2$, non-causal variants ranging between $0.0005 - 0.05$ as 50% between $0.0005 - 0.005$, 25% between $0.005 - 0.01$, and 25% between $0.01 - 0.05$. The causal variants increased from 2% to 25%.

	Causal	Non-Causal
ERV	.	✓
MRV	OR=1.5 (0.005-0.01)	✓
CV	.	✓

	Causal	Non-Causal
ERV	.	✓
MRV	OR=2 (0.005-0.01)	✓
CV	.	✓

Figure 5.8 shows a significant difference between Cauchy and SKAT weighting schemes according to the McNemar's test (the p values from McNemar's test are less than 0.05 at all the percentages of common variants). Figure 5.8 shows that increasing the number of common variants in the dataset has no impact on the power in case of the Cauchy weighting scheme while this will reduce the power in case of SKAT weighting scheme.

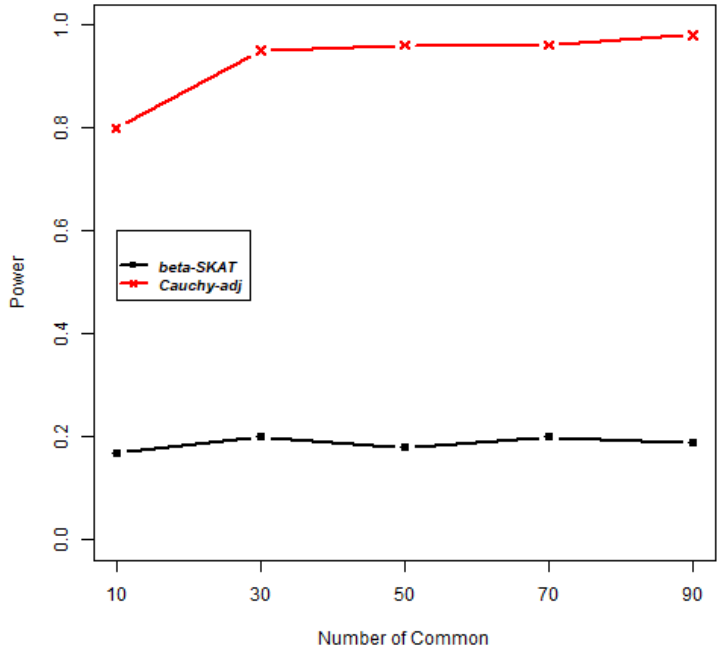


Figure 5.8: The causal variant in this analysis is located in the extremely rare data range with $OR = 3$, and the percentage of the causal variants is 20%. The rare data range from 0.0005-0.05, and the common variants range from 0.05-0.5. Rare data is fixed to 100 variants, and the number of common variants is increased from 10 to 100.

	Causal	Non-Causal
ERV	OR=3 (0.0005-0.002)	✓
MRV	.	✓
CV	.	✓increasing

5.3.2 Cauchy Adaptive Weight (1)

One variant weight we proposed was based on the Cauchy density function and uses fixed parameters based on MAF values. This weight assigns the largest values of weights to the smallest MAFs; however, it is best to consider both the MAF and the number of rare variants. Let $c = min(\mathcal{F}_j)$ and $s = \frac{\sum_{j=1}^{p} \mathcal{F}_j^r}{p^r}$, where $\mathcal{F}_j^r$ is the MAF value for rare variants excluding any $MAF > 0.05$, p^r is the number of rare variants in the data, and $\mathcal{F}_j$ is the minor allele frequency (MAF). We can rewrite the equation as

$$s = \frac{\sum_{j=1}^{p} \mathcal{F}_j I[\mathcal{F}_j \leq 0.05]}{\sum_{j=1}^{p} I[\mathcal{F}_j \leq 0.05]},$$

where I represents an indicator. Thus, the weight function will be

$$g(\mathcal{F}) = \frac{1}{\pi s(1 + ((\mathcal{F} - c)/s)^2)} \tag{5.5}$$

In this weighting scheme, the weight 5.5 is based on the frequencies of rare variants only; the common variants will not have an effect on the test's parameters since the parameters are based only on the data less than 0.05. Note that the common variants will have more weight only when there are more moderately rare variants in the data.

We evaluated the power of the score test with the Cauchy-adjusted weighting scheme 5.5. Figure 5.9 shows an evaluation of power versus the MAF of non-causal variants.

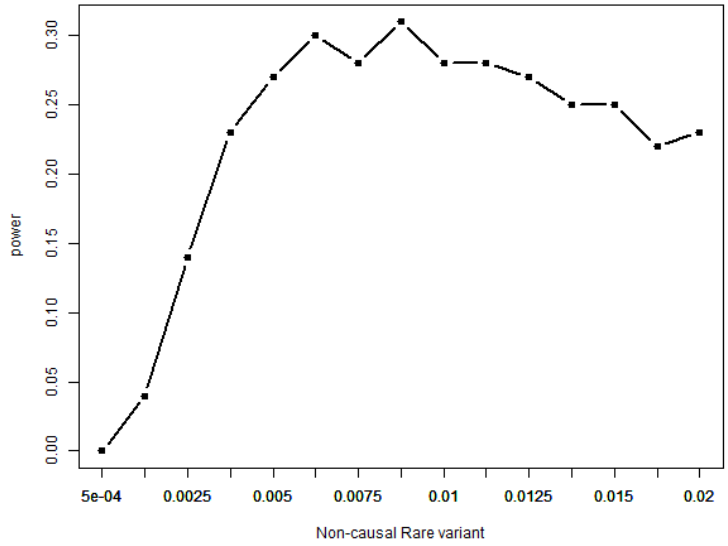


Figure 5.9: We fixed the simulation parameters so that the number of causal variants was 20 with MAFs (0.0005), $OR = 3$, and 100 non-causal variants, and we varied the MAF of non-causal variants from 0.0004 to 0.02.

The data that is required to successfully detect associations are suggested to be in the range from very rare to moderately rare $(0.0004, 0.01)$, as in Figure 5.10. The power of the test with this weighting scheme will increase along with the MAFs of causal variants. If there is a wide range of MAFs, it is required to have a large proportion of large moderately rare variants $(0.01 - 0.05)$ to detect the association when the causal is in this wide range of MAFs.

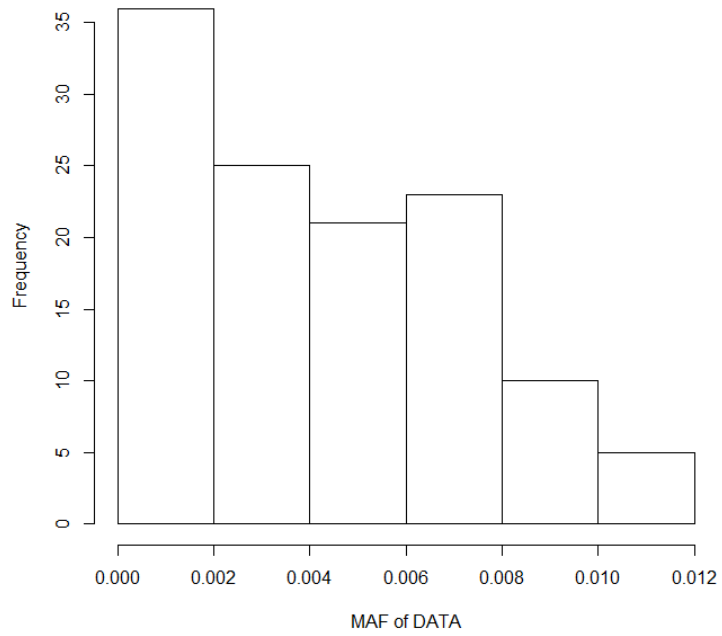


Figure 5.10: An example of the distribution of simulated data that works with the weight introduced in this section.

5.3.3 Cauchy Adaptive Weight (2)

Since SNP data often have different MAF distributions, a different weighting scheme that can be adjusted (i.e., is adaptive) and used with different datasets is needed so that if a few singleton variants are in the data, they will be associated with a lower weight. If we have a huge amount of data with moderately rare variants, why should we lose power just to consider one or two variants that are classified as extremely rare variants (ERV) by making them as large as possible? Instead, we consider them, but with lower weights (i.e., down-weighting the singletons in case they are genotype errors). As we saw in the Cauchy function with fixed parameters, the large moderately rare and common variants are not associated with a large weight. In this weighting scheme, we will adjust the weight based on the MAF, so the large moderately rare variants will have a large weight in case there are a large amount of this kind of variant in the data. This

will slightly down weight the extremely rare variants. Rare variants might have a systematic error when the variants are identified Johnston *et al.* (2015); if we worry about systematic errors in the ERV area, as is made evident in this section, the importance of ERVs are reduced according to the percentage of moderate and common variants in the data.

Figure 5.11 shows the difference between two simulated data sets based on the MAFs. The weights we propose will account for the distribution of the MAFs. Based on different simulated data, Figure 5.12 shows how Cauchy adaptive weighting scheme changes according to the distribution of the observed MAF.

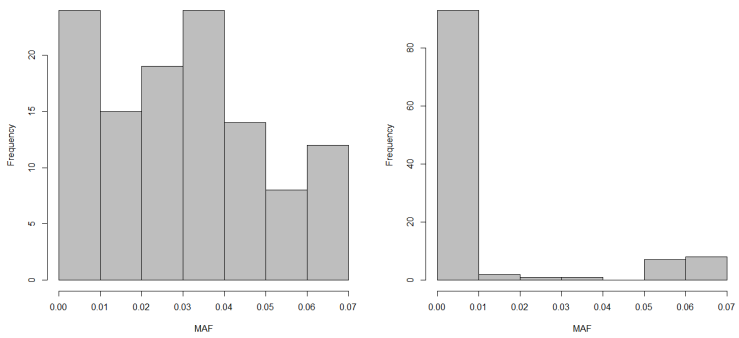


Figure 5.11: In this figure, we show the difference between two simulated datasets. The set on the right-hand side was comprised of more than 80% moderately rare variants, while in the set on the left-hand side, the moderately variants were less than 5% of the data.

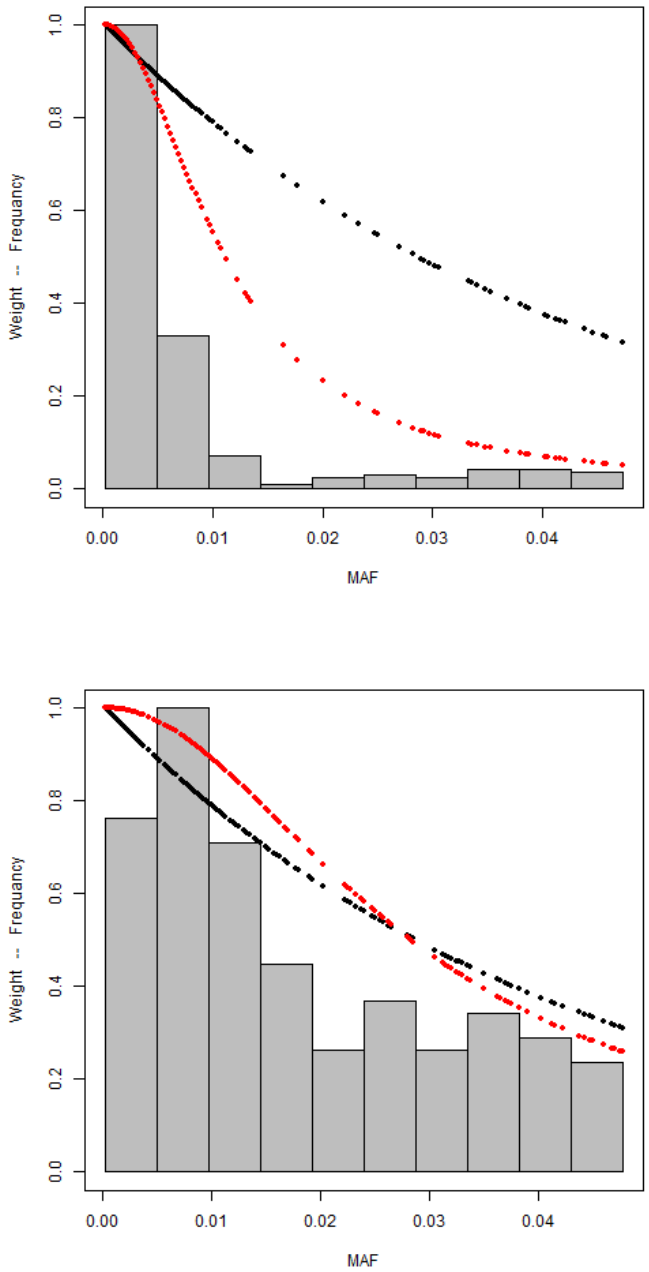


Figure 5.12: The adaptive weight, adjusted according to the distribution of the MAF of the data. In the top one, data has a large amount of extremely rare variants, while the figure below has a large amount of moderately rare and common variants. The red line represents the adaptive-Cauchy, while the black line represents the SKAT-weight.

Based on the assumption described above, we will use the Cauchy function as a variant's weight, with parameters $c = \min(\mathcal{F}_j)$; s is 75% of the data (upper quartile) divided by 10, plus $1/\sqrt{(2n)} = 0.01$, where $\mathcal{F}_j$ is the MAF. We will not allow the value of s to overflow 0.05; since the Cauchy weight will not up-weight the rare variants enough, it will be dominated completely by the common ones. Thus, we chose b to divide into 10 since the maximum of $\mathcal{F}$ will be 0.5, and by dividing it into 10, a value of 0.05 will be achieved. Thus, the weighting scheme presented in this section will account for the common variants. The s parameter will range between $0.01 and 0.05$.

$$g(\mathcal{F}) = \frac{1}{(s \times (1 + ((\mathcal{F} - c)/s)^2)}.$$

We can simplify it as

$$g(\mathcal{F}) = \frac{s}{(s^2 + (\mathcal{F} - c)^2)}, \tag{5.6}$$

$s = b + 0.01$, where b is the $Q3(\mathcal{F})$,75% percentile of the MAF dived by 10, and 0.01 is the result from $1/\sqrt{2n}$ where $n = 2000$.

Our proposed weight 5.6 allows for the development of a new weighting scheme in which, instead of up-weighting the rare variants only, the weights will be assigned after considering the frequency of the rare, moderate, and common variants. When there is a large number of extremely rare variants, they are assigned large weights; otherwise, they will be given low weights compared to moderately rare and common ones. Also, instead of choosing arbitrary parameters for the functions of the weighting scheme, we set the parameters as functions of the MAF $(\mathcal{F}_j)$.

The Cauchy function, in addition to up-weighting the rare variants, will not assign the largest weights to the rarest variants like the beta function; it will assign a lower weight after comparing them with the more moderately rare variants.

The figures (5.13) and (5.14) below show MAFs with different values of c and s, respectively.

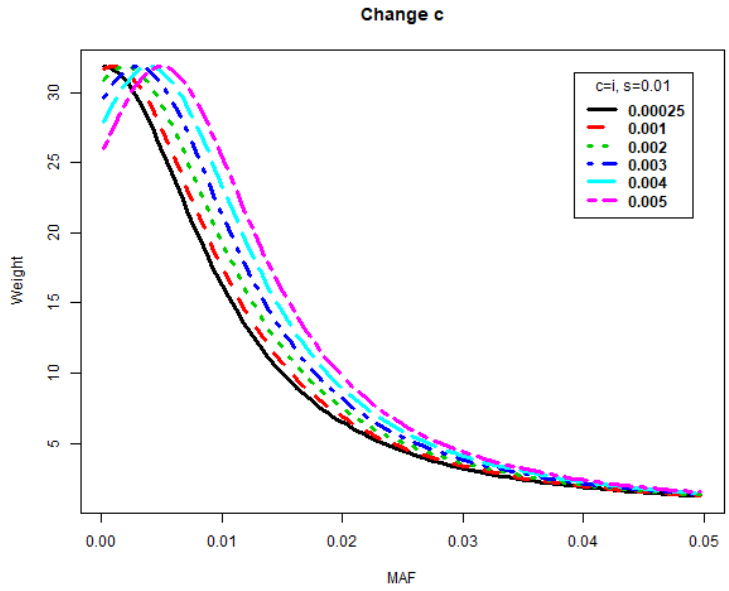


Figure 5.13: We show the MAF versus the weight using the Cauchy adaptive weight, with the s value fixed at 0.01, and the c value changed from 0.0005 to 0.005.

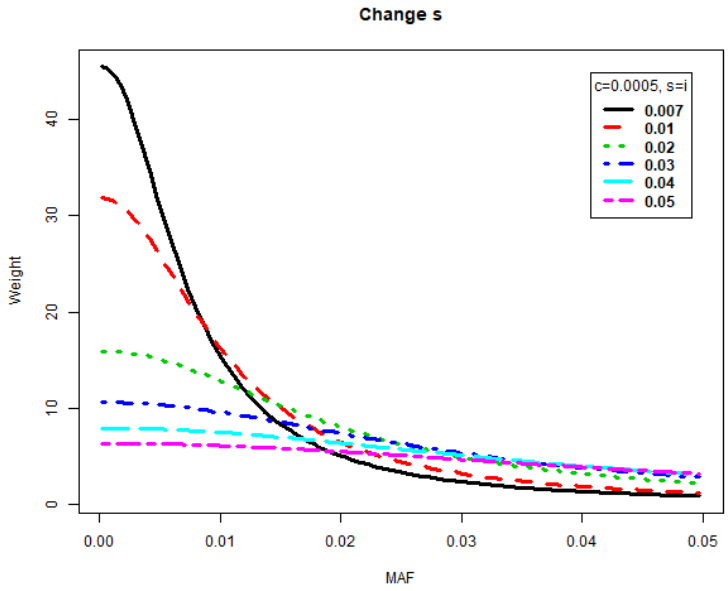


Figure 5.14: We show the MAF versus the weight using the Cauchy adaptive weight, with the c value fixed at 0.0005 and the s value changed from 0.007 to 0.05.

If there were more common variants in the sample than rare ones, then the weights for the rare variants were lower than those associated with the common ones if the latter's frequency was lower. So, when the number of common variants was much larger than the number of rare ones, the rare variants' importance was reduced, and the weight associated with them was also lower. This weight must be selected while considering the frequency of the rare variants because the power will decrease as the number of common variants increases. Note that the effect of this scenario will be obvious on the extremely rare variants (see Figure 5.15 and it shows there is a significant difference based on the McNemar's test between adaptive Cauchy and SKAT).

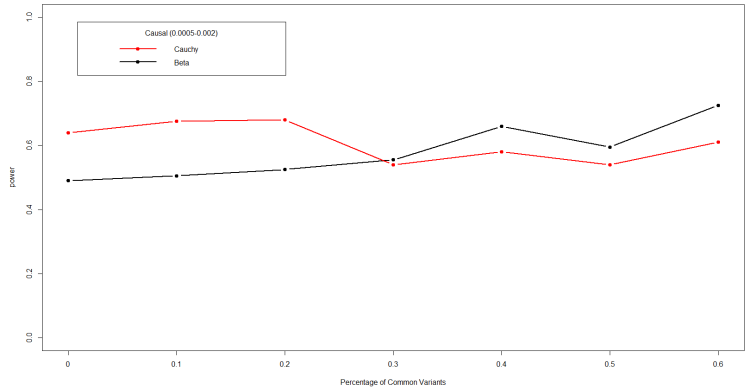


Figure 5.15: A comparison between Cauchy and beta SKAT functions, which shows the effect of increasing the common variants when the causal variants are ERV. Fifteen percent of the variants were causal with ($OR = 3$), MAFs $0.0004 - 0.002$, and 115 variants. The common variants are increased from 0% to 60% of the variants, and we reduce the number of rare variants. The red line is the Cauchy weight, while the SKAT beta is in black.

	Causal	Non-Causal
ERV	OR=3 (0.0005-0.002)	✓
MRV	.	✓
CV	.	increased from 0 to 60%

We will conduct the same analysis as above, but we are going to specify the causal variants to be moderately rare variants with MAFs between $0.005 - 0.01$ rather than extremely rare ones. Then, we can see the impact of increasing the number of common variants in the data. The effect will not be large as in extremely rare variants. Our goal is to reduce the importance of variants with very low minor allele frequencies (i.e., extremely rare variants) when half of the data is common. However, the importance of variants increases as the minor allele frequency increases.

When common variants comprise a large portion of the sample and the causal variants are moderately rare, then the importance of the moderately rare data are reduced somewhat only when the effect size is small. However, when the effect size is large, the power remains strong. Figures 5.16 and 5.17 illustrate this

relationship. Given that in the simulation, the standard error is 0.005, and the type I error is controlled at a nominal level of 0.05, with 1% uncertainty, there are no differences between the two weighting schemes in Figures 5.16 and 5.17, however, there is a slightly significant difference according to the McNemar's test (with p-values less than 0.05) when there are few common variants in the data while the causal variants are moderately rare with low size effect as in Figure 5.16 B.

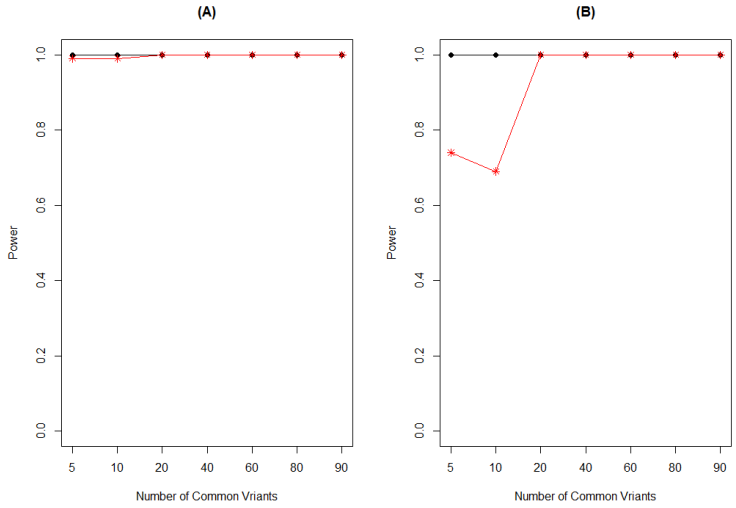


Figure 5.16: A comparison of Cauchy and beta SKAT that shows the effect of increasing the common variants when the causal variants are moderately rare (MRV) $(0.01-0.05)$. The parameters used in this analysis are 200 extremely rare variants with (15%) causal variants with $(OR = 3)$ in Figure A and $(OR = 2)$ in Figure B. The common variants increase from 5% to 90% out of 200. The red line with star points indicates the Cauchy weight, while the SKAT beta is represented by the black line with round points.

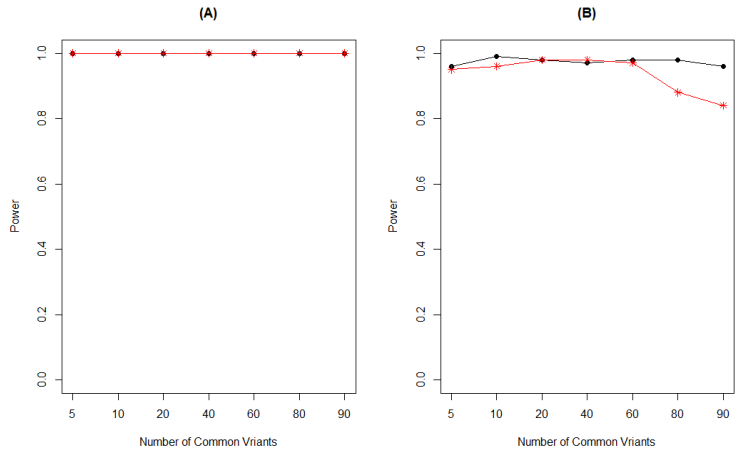


Figure 5.17: A comparison of Cauchy and beta SKAT that shows the effect of increasing the common variants when the causal variants are moderately rare. The parameters used in this analysis are 200 ERVs, and (15%) are causal. The MAF of causal variants is between 0.003 and 0.005 with ($OR = 3$) in Figure A and ($OR = 2$) in Figure B. The common and moderately rare variants range between $(0.005, 0.5)$ and increase from 5% - 90% of the variants. The red line with star points indicates the Cauchy weight, while the SKAT beta is represented by the black line with round points. Notably, power decreases when the proportion of common variants increases, especially when the effect size small.

Figure 5.18 shows that the power decreases when the proportion of common variants increases.

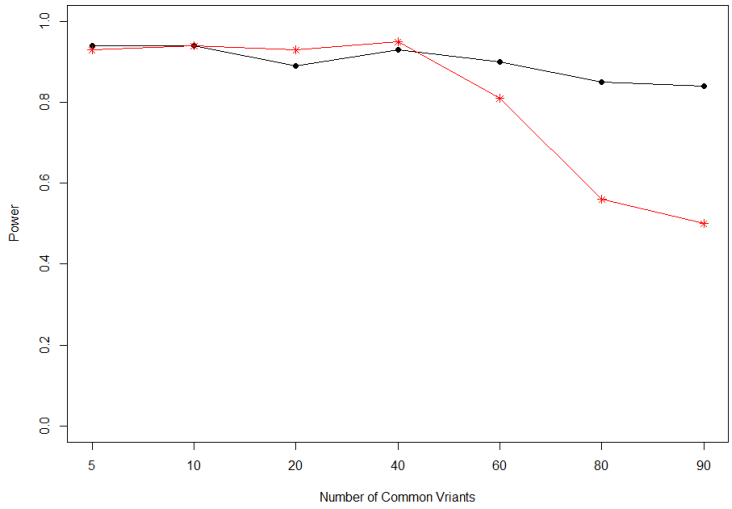


Figure 5.18: A comparison of Cauchy and beta SKAT that shows the effect of increasing common variants when the causal variants are ERV. The parameters used in this analysis are 200 ERVs, and (15%) are causal. The MAF values are between 0.001 and 0.002 with ($OR = 3$). The common variants range from 5% - 90% of the total variants. The red line with star points indicates the Cauchy weight, while the SKAT beta is represented by the black line with round points.

Figure 5.18 shows that there is a significant difference according to McNemar's test between the Cauchy adaptive and SKAT weighting schemes. We can see from Figure 5.18 that when common variants comprised more than 50% of the data, the power of the test decreased, especially when the causal variants had a small effect size. When the causal variants are ERVs in this kind of data, they contribute less to the test since they are rare, so it might be an error.

An increase in the moderately rare variants' frequency in the data will also require an adjustment in the weights to reduce the importance of ERVs but not as large as when common variants are more frequent in the data. If the percentage of moderately rare variants is much larger than ERVs in a given dataset, then the importance of the ERVs will be reduced and will not be as large as that of the common ones. Hence, in Figure 5.19, we only see a small reduction in the power because most of the data are moderately rare variants rather than common.

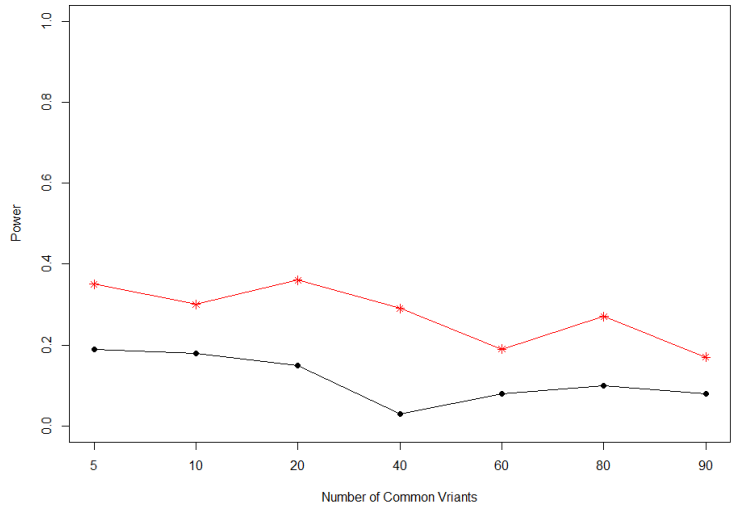


Figure 5.19: A comparison of Cauchy and beta SKAT that shows the effect of increasing the common variants when the causal variants are ERV. The parameters used in this analysis are 200 ERVs, and (15%) are causal. The MAF values are between 0.0005 and 0.001 with ($OR = 3$). The moderately rare variants range between $0.005 - 0.05$, increasing from 5% - 90% of variants. The red line with star points indicates the Cauchy weight, while the SKAT beta is represented by the black line with round points.

	Causal	Non-Causal
ERV	OR=3 (0.0005-0.001)	✓
MRV	.	increased from 5% to 90%
CV	.	✓

As we can see in Figure 5.19, the power starts decreasing when the proportion of moderate variants increases to more than 50% of the total, making a significant difference between SKAT and adaptive Cauchy according to the McNemar's test (all the p-values are less than 0.05). The next figure (Figure 5.20) illustrates the impact of the MAF of increasing common non-causal variants on the common causal variants, according to McNemar's test (all the p-values are less than 0.05), there is a significant difference between SKAT and adaptive Cauchy weighting schemes when the causal variants are common and the data has a large proportion

of the common variants. This difference becomes smaller when the size effect become small.

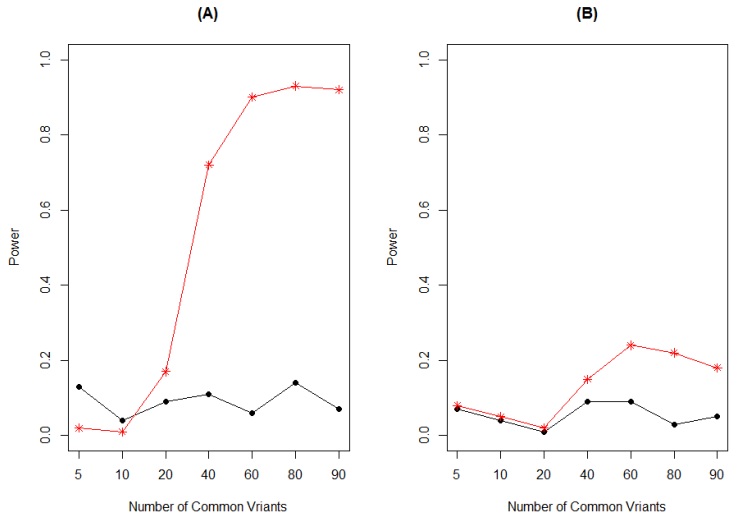


Figure 5.20: A comparison of Cauchy and beta SKAT that shows the effect of increasing the common variants when the causal variants are common. The parameters used in this analysis are $n = 2000$, and 120 common variants (15%) are causal. The MAF values are between $0.1 - 0.5$ with $OR = 2$ in Figure A and $OR = 1.5$ in Figure B. The first points in the plot result from data comprised of 90% ERVs and 5% moderately rare variants. The number of moderately rare variants was then increased from $5 - 100$ while decreasing the amount of ERVs. The common variants are also fixed as the causal ones. The red line with star points indicates the Cauchy weight, while the SKAT beta is represented by the black line with round points. Power decreases when the proportion of common variants increases.

Using a fixed parameter in a Cauchy weight will not account for the causal variants in the large moderately rare and common variant range. An adaptive weight will give the rare variants a higher weight and increase the power when there are more rare variants in the data. However, it will assign more weight when there are large proportions of moderately rare variants or common variants in the data, as expressed in Figure 5.21, since we add more moderate and rare variants in the data while we fix the MAF of causal variants to be moderately

rare variants, such as in Figure 5.21, and common variants, such as in Figure 5.22. Thus, according to McNemar's test, there is a significant difference between SKAT and adaptive Cauchy weighting schemes, when the causal variants are moderately rare variants and the data has a low percentage of common variants as in Figure 5.21. Therefore, under the same scenario, but when the causal variants are in a common region, then adaptive Cauchy will outperform SKAT, making a significant difference as in Figure 5.22.

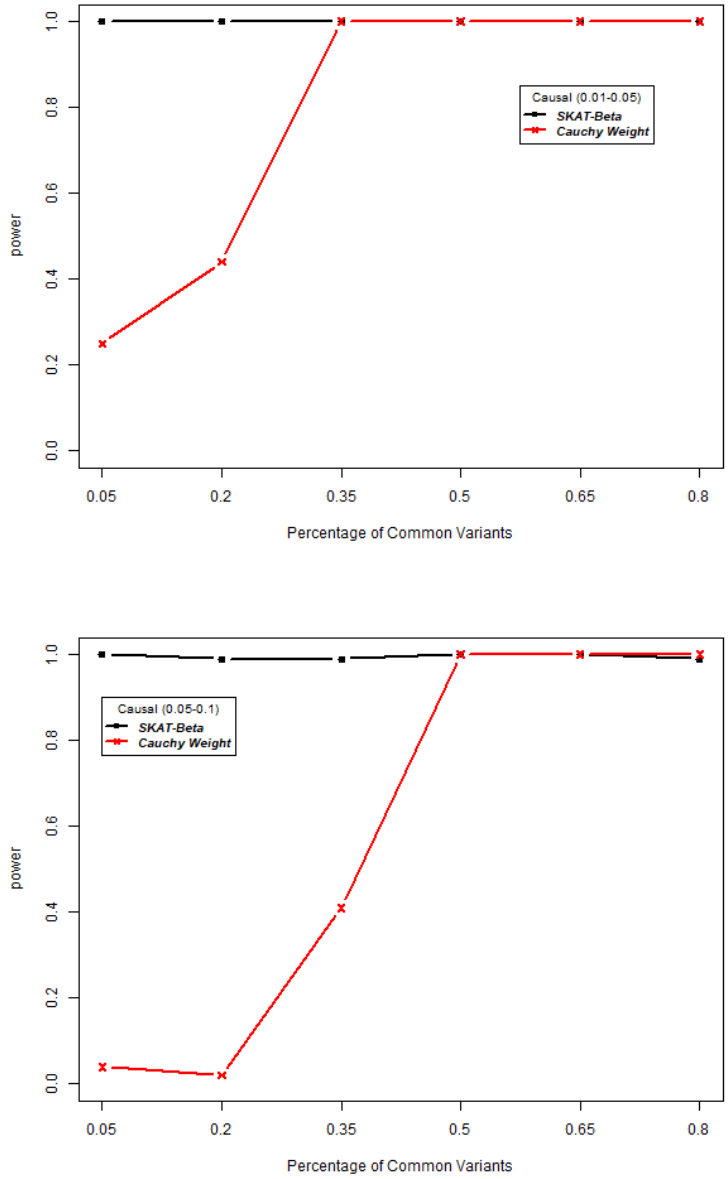


Figure 5.21: A comparison of Cauchy and beta SKAT that shows the effect of increasing the common variants when the causal variants are moderately rare. The parameters used in this analysis are $n = 2000$, and 120 common variants (15%) are causal. The MAF values are between $0.01 - 0.05$ in the top Figure, and the MAF of causal variants are between $(0.05 - 0.1)$ in the bottom Figure, with $OR = 1.5$ for both scenarios. The first points in the plot result from data comprised of 90% ERVs and MRVs and 5% common variants. The number of common variants is then increased from 5% to 80% while decreasing the amount of ERVs.

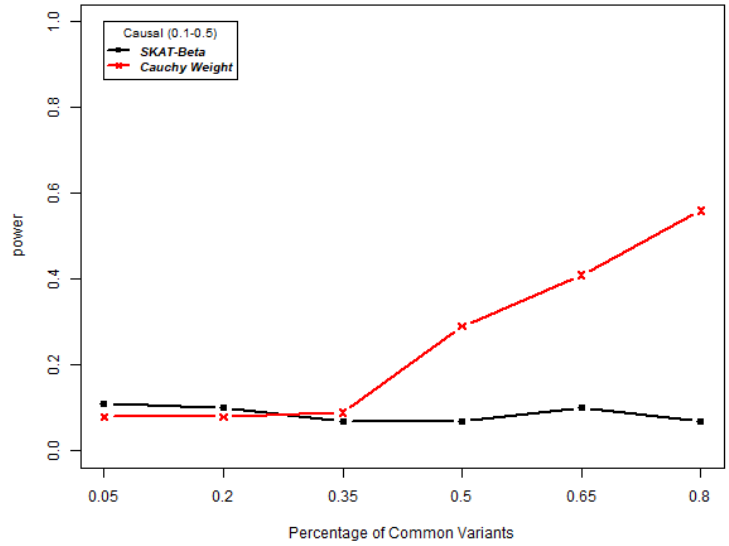


Figure 5.22: A comparison of Cauchy and beta SKAT that shows the effect of increasing the common variants when the causal variants are moderately rare. The parameter used in this analysis is $n = 2000$, and 120 common variants (15%) are causal. The MAF values are between $0.1 - 0.5$ with $OR = 1.5$. The first points in the plot result from data comprised of 90% ERVs and MRVs and 5% CVs. The number of common variants is then increased from $5\% - 80\%$ while decreasing the amount of ERVs. The common variants are also fixed as causal ones.

5.4 Gumbel Function

We will introduce new functions that can be used for rare variant association studies. These new functions are based on a Gumbel function. However, we made some modifications, so they will be appropriate for up-weighting rare variants.

Additionally, an issue that arises in rare variant associations is a large number of singletons can arise when using next-generation sequencing; there is great concern about this kind of rareness inducing systematic errors. The idea proposed here is a new weighting scheme, which, instead of up-weighting all rare variants so that the rarest variants are given the largest weights, up-weights the rare

variants under some constraints. We also propose, as another property of the Gumbel function, that all rare variants should be up-weighted, but singleton variants should be given smaller weights than other rare variants. Based on this premise, the smallest MAF will no longer be associated with the largest weight, as shown in Table 5.5. We can also use a Gumbel function as an adjusted (adaptive) weight for different datasets. In this section, we propose two different schemes of variant weights based on the Gumbel function. We investigate the differences between them and evaluate type I errors and the power of the test.

MAF	1/n	4/n	20/n	200/n
Weight	37	93	62	1

Table 5.5: The Gumbel-based weight after down weighting the singleton, which will be discussed later in this section

5.4.1 Gumbel Fixed Parameter

A Gumbel function can be used to control how much the rare variants are up-weighted. In the previous functions, we up-weighted these variants, but controlling the threshold was not clear or straightforward. In this function, by modifying the scale parameter, we can fix the threshold, so this function can give higher weights to the rarest variants. This is a function that can be used as a weight based on variants.

$$g(\mathcal{F}) = \frac{1}{s} e^{(-ze^{z})}, \tag{5.7}$$

where $z = \frac{(\mathcal{F}-\mu)}{s}$,$\mathcal{F}$ is the MAF, μ is the location parameter, and s is the distribution scale ($s > 0$). If we fix μ as the minimum of the MAF, which is based on the data, then the smallest MAF will have the highest weight. We can change the threshold for rare variants using the scale parameter $s = 0.05$ as follows:

$\begin{cases} MAF < s & \text{high weight} \\ MAF > s & \text{low weight} \end{cases}$

Figure 5.23 shows the comparison between SKAT and Gumbel weighting schemes. We can see Gumbel function put large weight in the ERV region while lower in the common region.

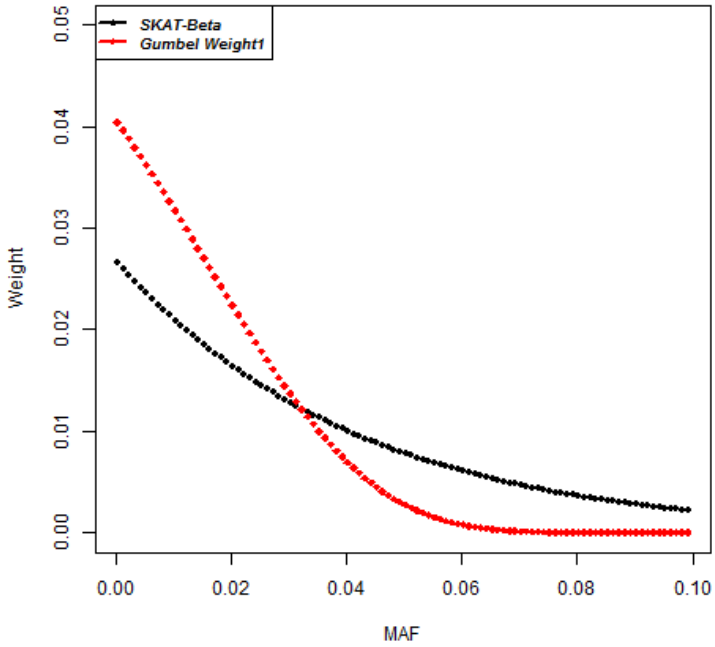


Figure 5.23: The weight of Gumbel function versus the beta-SKAT function.

The Gumbel function (5.7) performs better with a large amount of large moderately rare variants $(0.01 - 0.05)$ than beta-SKAT, as shown in figure (5.24).

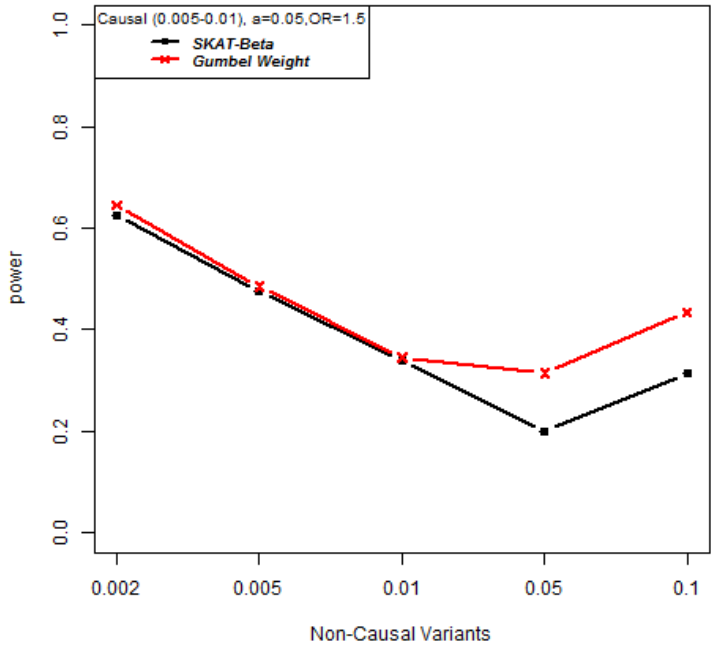


Figure 5.24: This figure shows the impact of non-causal variants on the score test using beta-SKAT and Gumbel in terms of power. The causal variants are fixed to be 10 variants out of 100 with MAF $0.005 - 0.01$. The OR for causal variants is fixed to be 1.5. We generate the MAF for non-causal variants from a uniform distribution, and we fix the minimum parameter to be 0.0005 while we vary the maximum as represented on the X-axis $(0.002, 0.005, 0.01, 0.05, 0.1)$

Figures 5.25 and 5.26 show a comparison of Gumbel and beta weight schemes in terms of score test power when the causal variants are in different MAF settings. Given that in the simulation, the standard error is 0.005, and the type I error is controlled at a nominal level of 0.05, with 1% uncertainty, and according to McNemar's test, there is a significant different between SKAT and Gumbel weighting schemes, when the causal variants are large percentage extremely rare variants as in the top figure of the Figure 5.25. on the other hand there is no significant difference between them when the MAF of causal variants are larger as in the bottom figure of the Figure 5.25. According to McNemar's test there is no significant difference when the causal variants are moderately rare variants.

However, there is a significant difference between these methods when the causal variants are between 0.01 and 0.05 as in the top figure of the Figure 5.26.

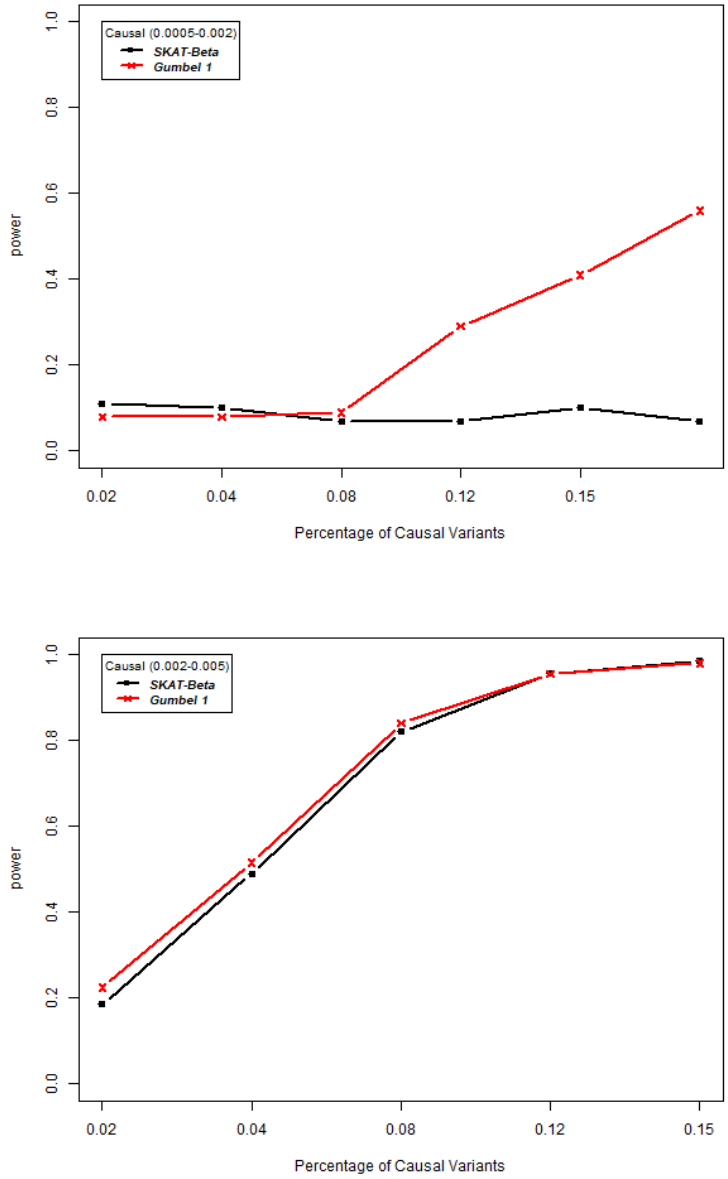


Figure 5.25: A comparison of the Gumbel function and beta-SKAT. There are 200 variants with different MAFs of extremely and moderately rare and common variants. We fixed the causal variants at MAF $0.0005 - 0.002$ with $OR = 3$ in the top figure, while in the bottom one, the causal variants are fixed at MAF $0.002 - 0.005$ with $OR = 2$.

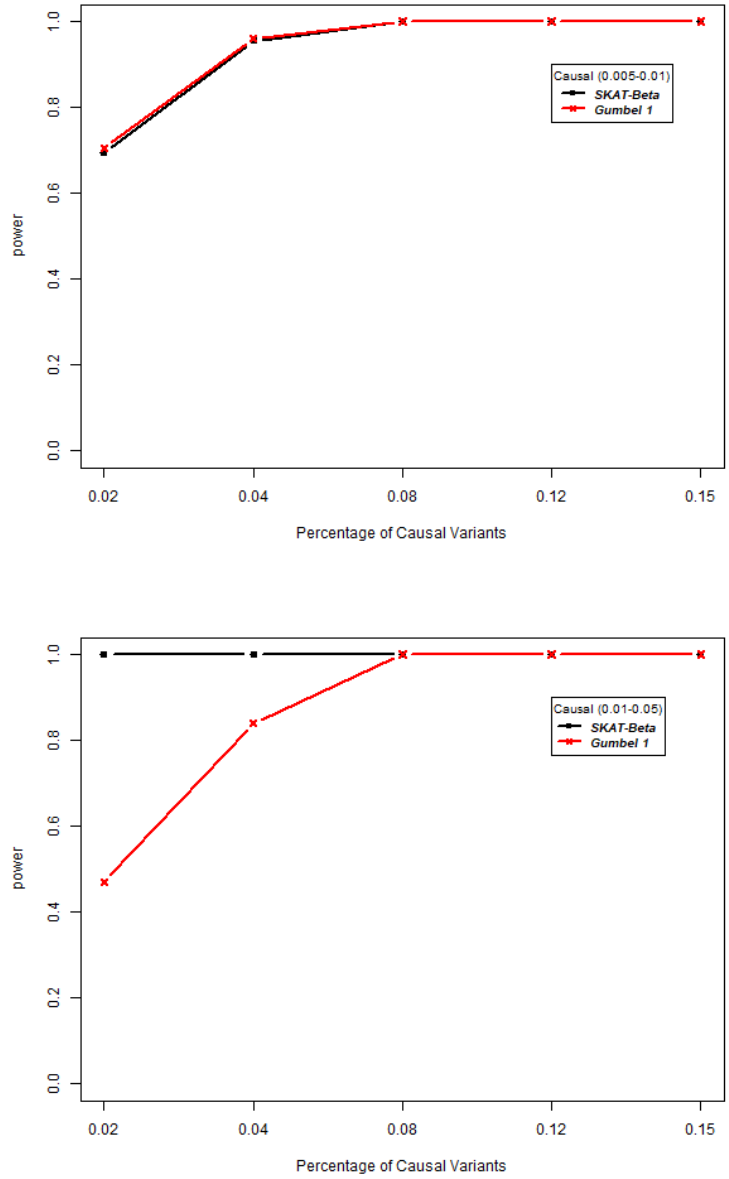


Figure 5.26: A comparison of the Gumbel function and beta-SKAT. There are 200 variants with different MAFs of extremely and moderately rare and common variants. We fixed the causal variants at MAF $0.005 - 0.01$ with $OR = 2$ in the top figure, while in the bottom one, the causal variants are fixed at MAF $0.01 - 0.05$ with $OR = 2$.

5.4.2 Adaptive Gumbel

Since the data structure sometimes has few common variants, and the weights proposed in the literature cover most of the MAF region, we have to pay a cost; we must lower the detection of causality in the rare regions to detect the signal elsewhere. The question is why we have to pay this cost when there are few common variants in the data. This is the same idea a Cauchy, meaning more weight will be associated with the most frequent MAFs in the data. Based on this kind of data, we propose a weighting scheme that can be adjusted based on the data:

$$g(\mathcal{F}) = \frac{1}{a} e^{-z} e^{-e^{-zn}} \tag{5.8}$$

where $z = \frac{\mathcal{F}-c}{a}$, $\mathcal{F}$ is the MAF, c is the minimum of MAF, $a = t + \frac{b}{2}$, b is the 75% quartile of the data $Q_3(\mathcal{F}_j)$, and t is the threshold for the extremely rare variants $t = (\sqrt{2n})^{-1}/2$ and n number of individuals.

This weighting scheme (5.8) is suggested if the data are mostly rare or extremely rare with a low proportion of common variants. However, if there is a large number of common variants or rare variants greater than 0.01 are included in the data, the detection of causal variants in extremely rare areas will be low unless there is a large number of them in the data; then, the detection will remain high. Thus, the adaptive Gumbel weighting scheme is changing based on the distribution of the observed MAF (see Figures).

This weight is adjusted according to the distribution of the observed MAF. It focuses on the extremely rare and rare region by up-weighting them significantly (see Figures 5.28 and 5.29). This up-weight is not applied to the singleton variants and gives more weight to the moderately rare variants if there is a large proportion of them in the data (see Figure 5.27).

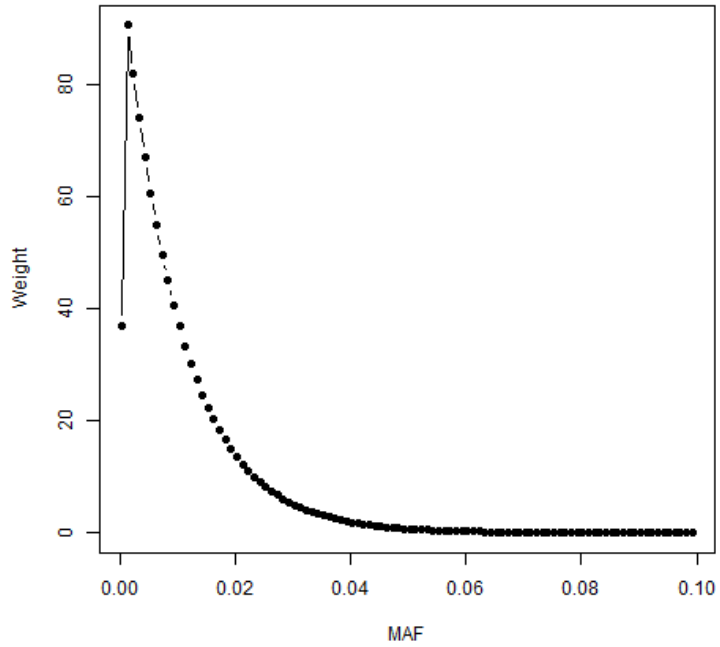


Figure 5.27: The weight using the un-singleton Gumbel function, which shows a singleton SNP that occurs one time among 2000

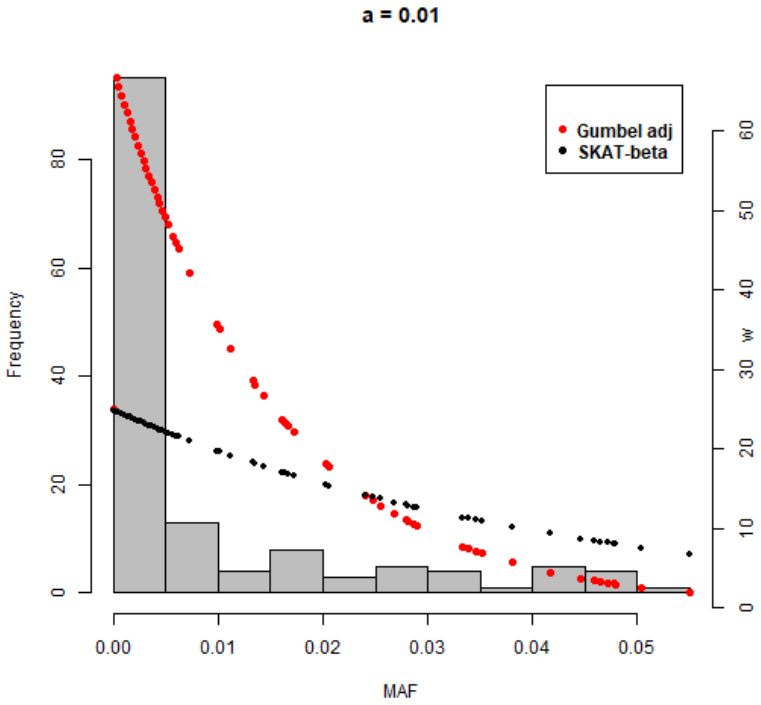


Figure 5.28: A distribution of MAFs ($\mathcal{F}$) showing most of the data is in the ERV range.

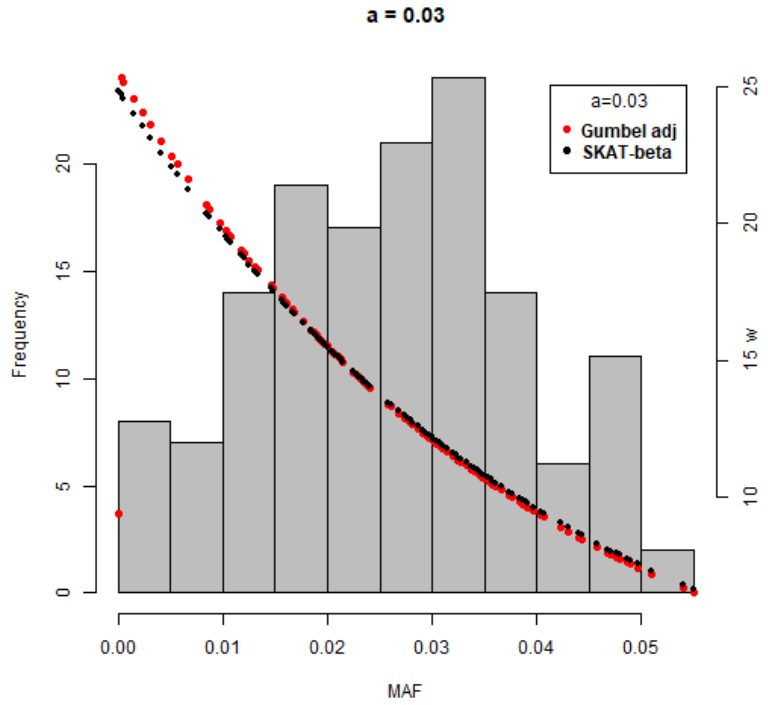


Figure 5.29: A distribution of MAFs ($\mathcal{F}$) showing data ranges in the different areas of MAF.

We evaluate the power of the test statistic using this weight with two datasets: one with a large proportion of ERVs and one with many moderately rare variants and some common. The figure illustrates the difference between these two data sets based on the strength of their detection of the extremely rare variants. When the data is 75% rare SNPs, the detection of extremely rare variants will be large, and it will be reduced when the number of common variants reaches 75% of the data; see (Figure 5.30).

Given that, in the simulation, the standard error is 0.005, and the type I error is controlled at a nominal level of 0.05, with 1% uncertainty, there is significant difference between SKAT and adaptive Gumbel weighting schemes, Figure 5.30 illustrates the impact of changing the parameter a in the Gumbel function, which is responsible for adjusting the weight. Parameter a, as illustrated here, is estimated based on 75% of the data. Figure 5.31 shows the impact of the weight when a large proportion of the data is less than 0.01 and few common variants

are included; these weighting schemes make a significant difference in terms of the power between the score test with these different weighting schemes. When the causal is in the extremely rare region and non-causal data are all moderately rare, the importance of these extremely rare causal variants will decrease, as is shown in Figure 5.32. However, if the moderately rare or common variants are present but represent a low proportion of the data and the causal variants are in the extreme region, then the power of the test will increase compared to SKAT since the SKAT weight gives a very large weight to moderate variants even if they are a very small portion of the dataset; see Figure 5.33.

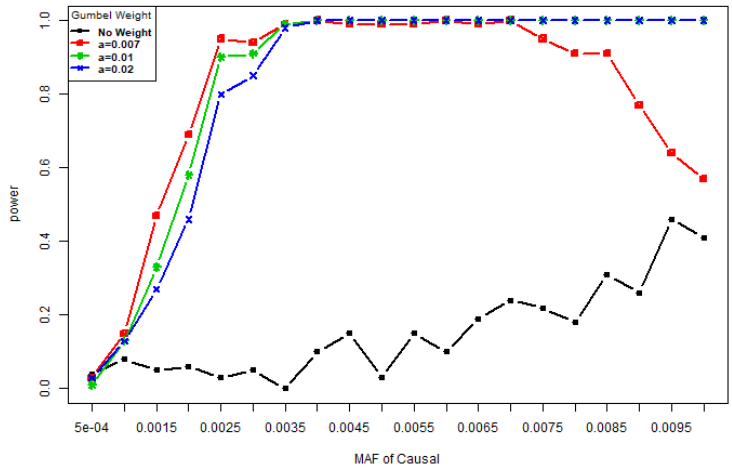


Figure 5.30: We show the impact of parameter a, which will measure the frequency of data. In this figure, the analysis is conducted using 200 SNPs classified as 33% for extremely and moderately rare variants and common ones. We fix the parameter a at 0.007,0.01 and 0.02. The effect size is fixed at 3 with 7% causal rare variants.

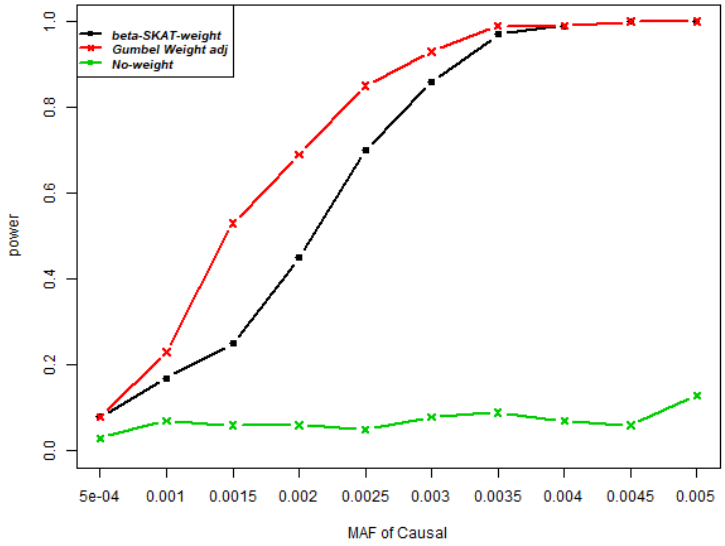


Figure 5.31: This is a comparison of a beta weight and the adaptive-Gumbel weight. 80% percent of the data is less than 0.01, and 20% is between 0.01 and 0.5. Then, we change the MAF of causal variants on X-axis while we fixed the percentage of causal variants to be 7% and the effect size to be $OR = 3$.

	Causal	Non-Causal
ERV	OR=3 (0.0005-0.005)	✓40%
MRV	.	✓40%
CV	.	✓20%

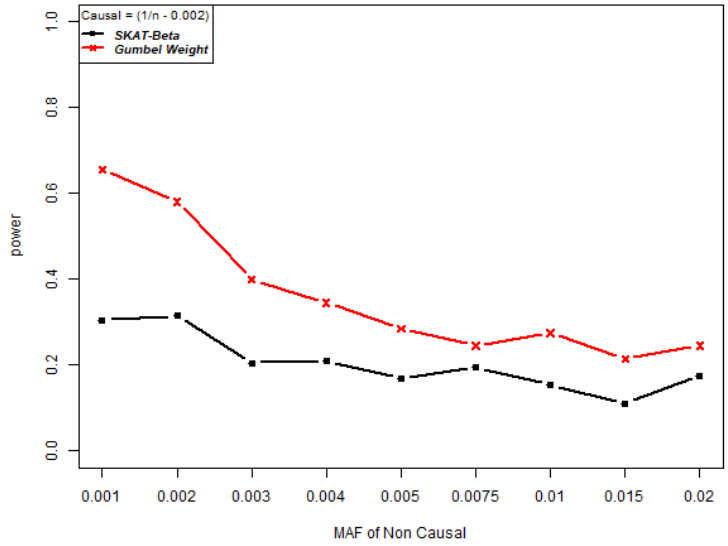


Figure 5.32: In this figure, we fix the causal variants to be in the extremely rare variants region at MAFs between 0.0005-0.0025 and fix the number of non-causal variants to be 200; then, we change the non-causal variants' MAF.

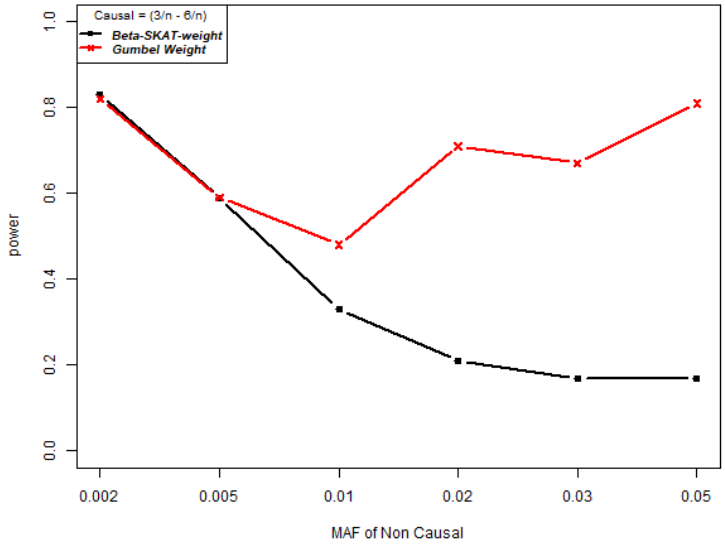


Figure 5.33: In this figure, we fix the causal variants to be in the extreme region; the number of SNPs is 200, and 8% are causal variants with MAFs ranging between 0.001-0.003. The X axis represents the non-causal MAFs; we change the cutoff of the non-causal variants. To clarify, the first points generate non-causal variants between $0.0005 - 0.001$, while the last one generates non-causal between $0.0005 - 0.05$. There are 10% common variants.

When most of the data is extremely rare, using this weight can help detect the association with a large effect size in the moderately rare region, as shown in Figure 5.34.

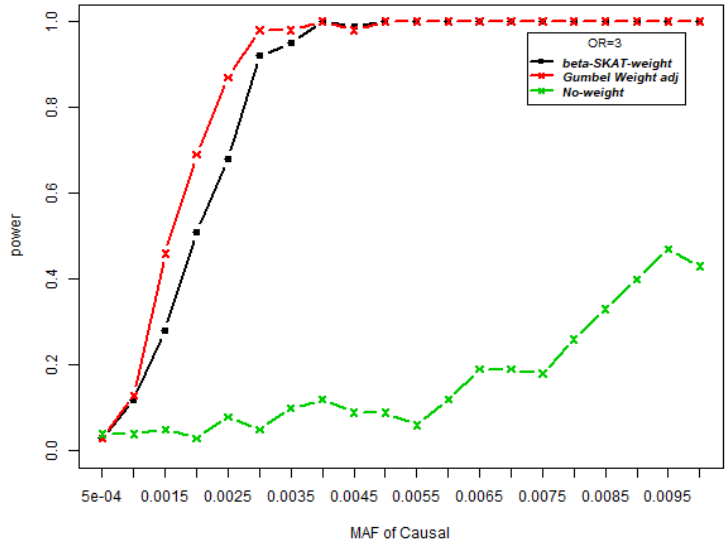


Figure 5.34: In this analysis, we fix the percentage of the rare variants with MAFs less than 0.01 to be 80% of the data. The percentage of causal variants is 7%. Then, we change the MAF of the causal variants.

	Causal	Non-Causal
ERV	OR=3 (0.0005-0.005)	✓40%
MRV	OR=3 (0.005-0.01)	✓40%
CV	.	✓20%

5.5 Relationship Between the U Vector and Variants' Weights

The score test statistics will reject the null hypothesis when the value of **U** is large. To illustrate, let $X_{j'}$ be the positions of causal variants, the values of the vector U_j that correspond to the position of causal variants be $U_{j'}$, and δ be a threshold that can identify the largest values of U_j. We will choose the causal variants to be 10 rare variants, so we will take the ten largest values of U_j. Thus,

if

$$\sum_{j'=1}^{p'} I(u_{j'} > \delta) = \sum_{j=1}^{p} I(u_j > \delta),$$

all the causal variants contribute to the score test, and the null hypothesis will be rejected, which is the ideal case. Hence, the weighting scheme that can address this scenario by up-weighting the data that is assumed to be causal while down-weighting the other data is preferable. Moreover, this is the golden point that can explain the difference between varying weights (see Figure 5.35). Figures 5.36 and 5.37 and Table 5.6 shows the different between Cauchy weighting scheme and SKAT weighting schemes.

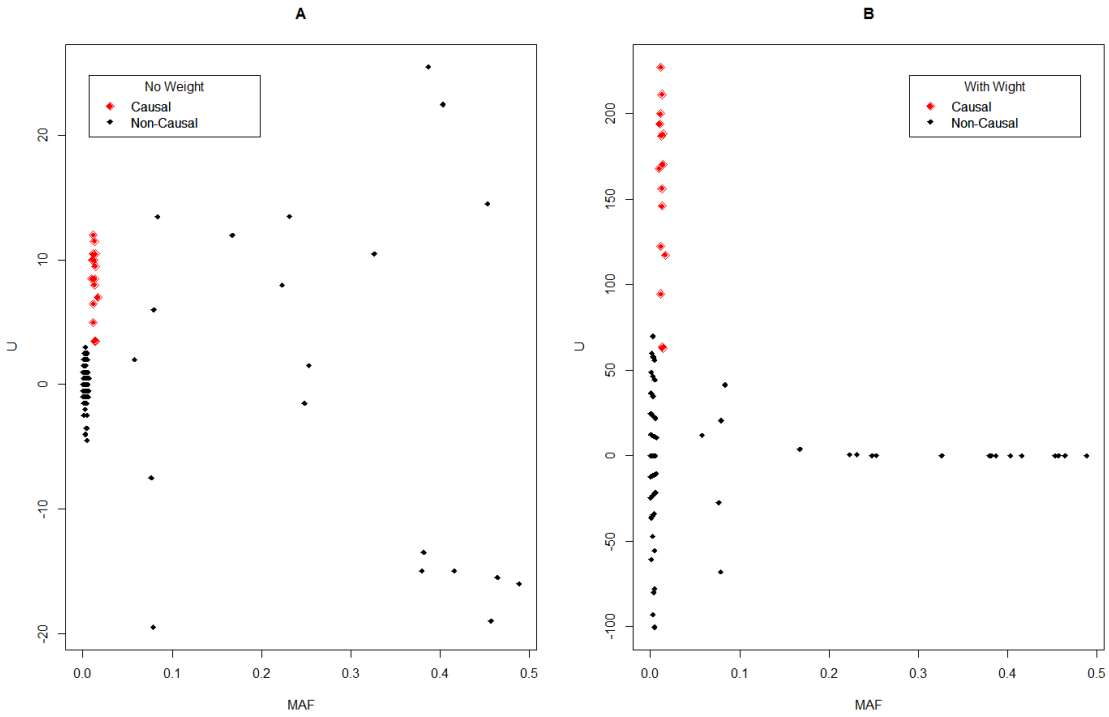


Figure 5.35: The difference between **U** vectors with and without weights, which expresses the impact of the weights on the causal rare variants that cause the weighting schemes to differ in terms of their impact on the **U** vector.

	MAF			
Weight	0.0005	0.001	0.005	0.5
Beta(1,25)	24.7	24.4	22	0
Cauchy(min(MAF),0.01)	31.8	31.7	26.4	0.01

Table 5.6: Illustration of the difference between weights based on the beta and Cauchy functions. In rare regions (e.g., $0.0005, 0.001$), we can see the differences between weight values for the MAFs using beta are small, while the differences between values using Cauchy are large.

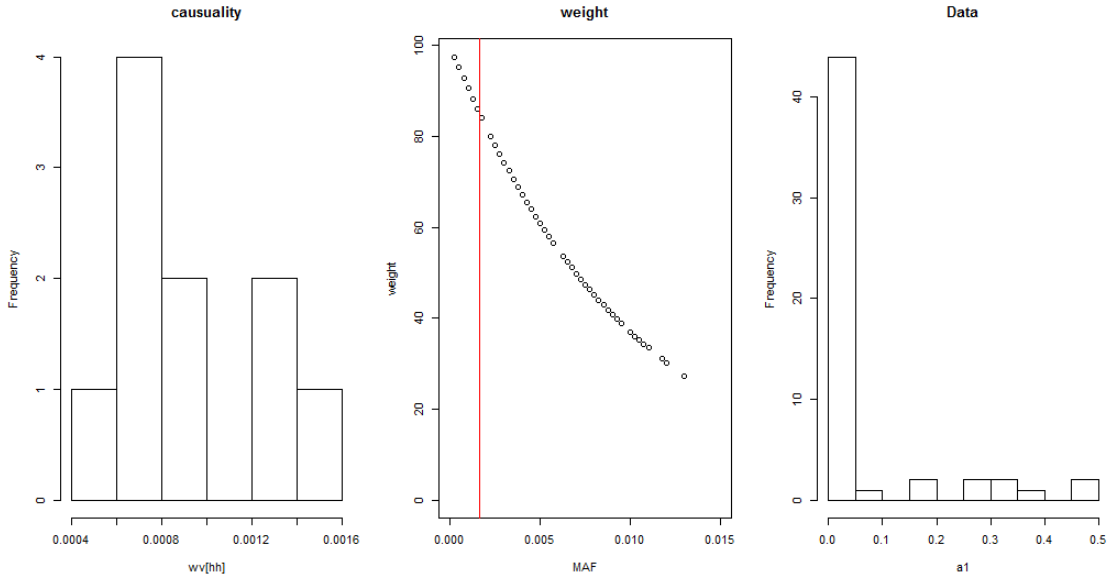


Figure 5.36: The shape of the beta function versus the distribution of the data and the causal variants in terms of MAF. The red line indicates the $MAF = 0.002$.

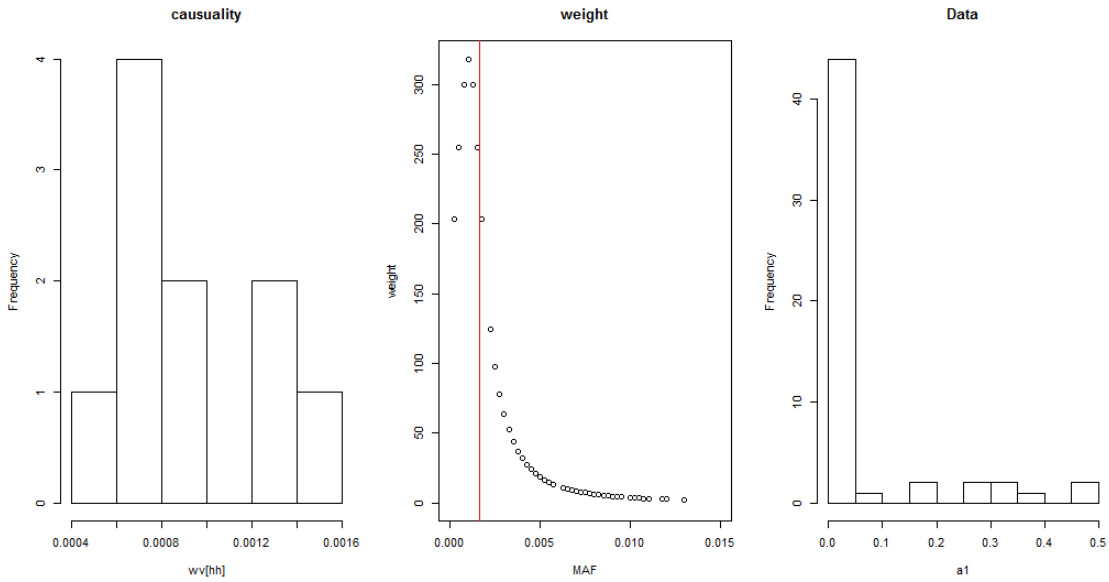


Figure 5.37: The shape of the Cauchy function with an $3/n$ parameter versus the distribution of the data and the causal variants in terms of MAF. The red line indicates the $MAF = 0.002$.

Type I Errors

To evaluate type I errors, we begin by generating a genotype matrix as explained above in the simulation section and setting the $OR = 1$. We consider three types of variants in the simulation based on a given MAF: non-causal extremely rare, moderately rare, and common variants. We use two methods to evaluate type I errors. First, we fix the maximum and minimum MAFs to be the same for each type of variant; see Figures 5.38 and 5.39. We consider all MAF values from $1/n$ (the boundary of MAF) up to $(0.5n)/n$. Then, we generate the X_p via the scenario above so that the data has different types of variants (i.e., rare and common). We randomly generate rare variants ranging from the MAF boundary to $MAF = 0.01$ and common variants ranging between 0.05-0.5. Next, we fix the percentage of rare and common variants to be 50% for each set and modified the rare dataset by changing the threshold of rare variants from $1/n$ to $30/n$; see Table (5.7). For each scenario, we conduct 1000 simulated datasets with randomly generated genotypes for each simulation and estimate the empirical type I error

rate as the proportion of p-values less than the nominal level $\alpha = 0.05$. The results show the test using this weighting scheme had satisfactory type I error rates except when all the variants have very low minor allele frequency. Rarity is a concern in the control of type I errors; see Figures 5.38 and 5.39. We can also see in Table 5.7 that when the threshold equals 1/n, which means we generate a dataset and divide it to two sets, one is considered a common set with MAFs between $0.05 - 0.5$ and the other between $1/n - 1/n$, meaning half of the datasets have very low MAF. The Cauchy density is a good function to use to avoid the issue in beta functions; in a beta function, the ERVs' weighting is not as large as it should be. Also, the parameters can be based on the data, and they can be fixed.

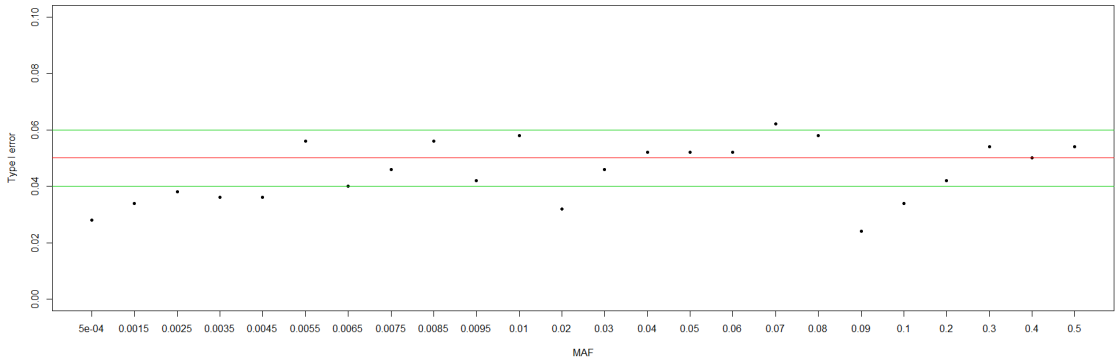


Figure 5.38: Type I error for tests using the Cauchy density as a variant weight function 1. The green and red lines highlight the following values: 0.04, 0.05, and 0.06.

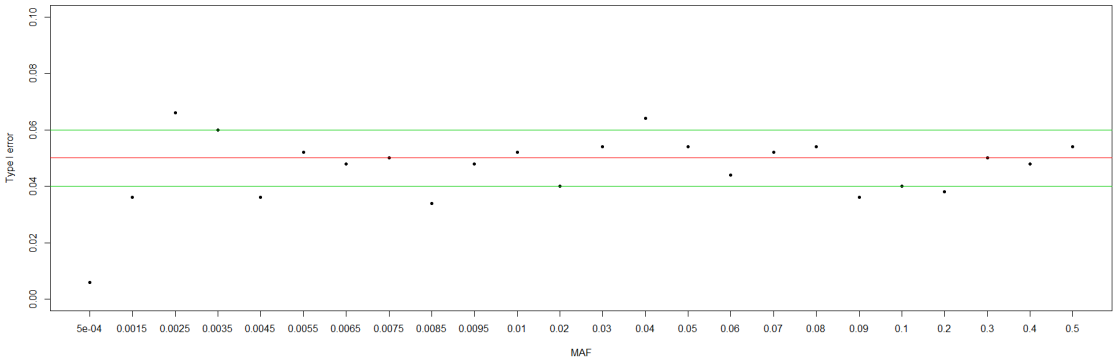


Figure 5.39: Type I error for a score test using the Cauchy density adjusted function. The green and red lines highlight the following values: 0.04, 0.05, and 0.06.

	Threshold of rare			
	1/n	4/n	20/n	30/n
Cauchy 1	0.035	0.05	0.06	0.05
Cauchy 2	0.04	0.05	0.05	0.05
Cauchy 3	0.05	0.045	0.035	0.045
Gumbel 1	0.035	0.045	0.06	0.05
Gumbel 2	0.045	0.05	0.045	0.05

Table 5.7: Type I errors for a score test with different weights.

5.6 Conclusion

In this chapter, we expressed new ideas based on variant weight schemes designed for use in rare variant association studies. The adjusted weights can be adjusted depending on the data (i.e., they are adaptive weights). They down-weight SNPs that occur once in the sample, and the weight with fixed parameters can analyse moderately rare and extremely variants more effectively than common ones. However, the weighting schemes presented in this chapter cannot detect associations when the causal is within a large MAF range, especially when the weight has a fixed parameter. Furthermore, the proposed weight schemes presented in

this chapter focus on rare variants only (i.e., less than 0.01), which is not as effective at detecting causal variants in the common range; see Figure (5.40). In the next chapter, we will introduce new weighting schemes that can be extended to the proposed weight scheme and detect causal variants in rare and common MAF ranges.

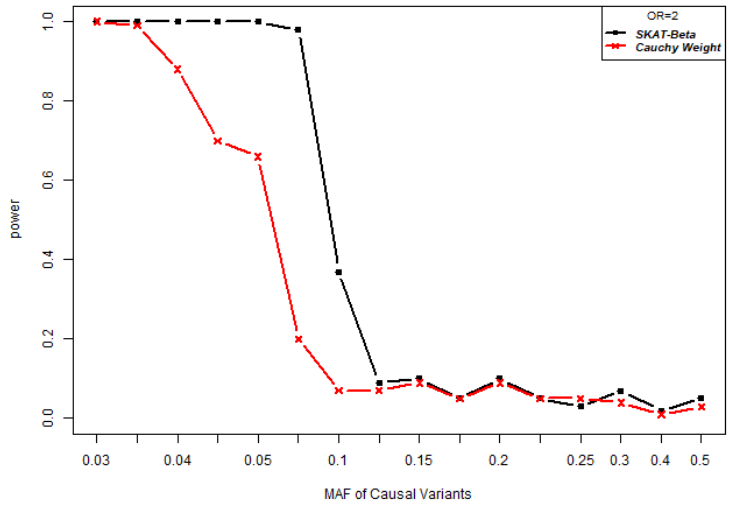


Figure 5.40: This figure shows the proposed weight and SKAT-beta weight, which cannot detect causal variants when they are in the MAF common range. We simulate 100 variants, all distributed equally according to their MAF, so we have extremely rare, moderately rare, and large frequency rare variants and common variants of 25% for each category; the causal variants are fixed at $OR = 2$ and are all located in the common range.

Chapter 6

Weight Functions for the Continuous Spectrum of MAF

6.1 Introduction

In Chapter 5, we proposed different weighting schemes. Most of these schemes allow for signal of association detection within a range of minor allele frequencies ($\mathcal{F}$). For example, using the parameter $(1, 25)$, the beta function is effective for rare data up to $MAF = 0.05$, and some Cauchy functions have good power for rare data up to $MAF = 0.03$. In these cases, the focus is on signals in the rare MAF range rather than the complete range of MAFs. We divide the MAF range into extremely rare, moderately rare, and common variants. In this chapter, we extend the weighting scheme to consider the entire range of MAFs ($\mathcal{F}$); by using these weights, we are able to detect the causality across the whole MAF range from extremely rare to common variants (i.e., causal variants uniformly distributed within the MAF range; see Figure 6.1).

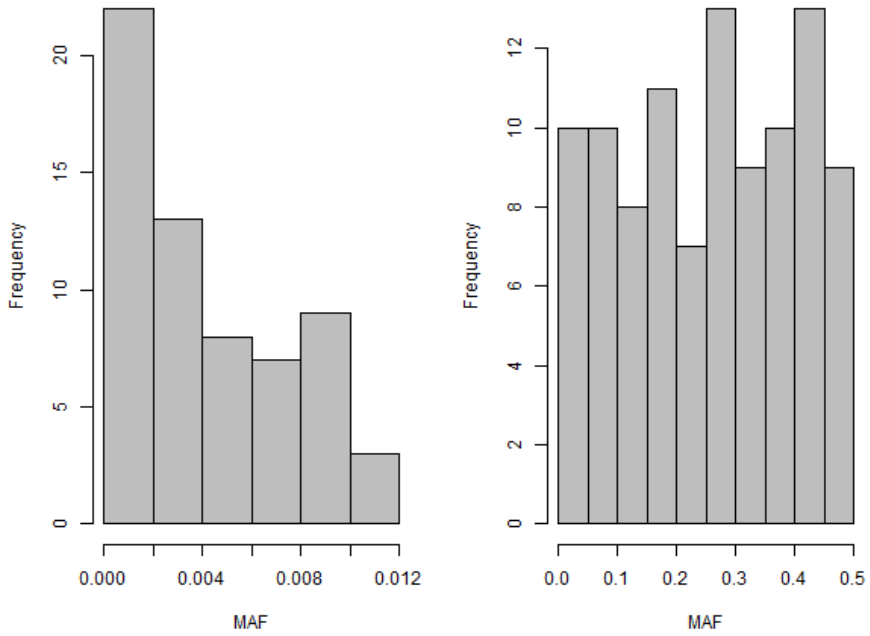


Figure 6.1: Examples of the distribution of the causal variants that can be detected using the weighting schemes described in Chapter 5 (left Figure) and this chapter (right Figure).

When data contain rare and common variants, prior research has largely used methods that use the beta function as a weight and focus only on rare variants while ignoring common regions (the functions are not continuous for the entire range of MAFs), such as the VT method of Price *et al.* (2010) and the SKAT method (Wu *et al.*, 2011). In this chapter, we propose functions that can address this limitation; we use continuous functions because they may be able to detect associations in any range of MAFs. Specifically, we include the common range by using different weighting schemes and exploring the possibility of applying Cauchy (with adjustment), Levy, beta, and Burr functions. Many of these weights perform well in the extremely rare and moderately rare ranges, as well as in common regions with small effect sizes. Notably, we compare our proposed weighting scheme with that of the beta function proposed by Wu *et al.* (2011).

This chapter is separated into two parts. The first part introduces different weighting schemes that can help detect associations across the whole range of MAFs. We consider a combined weighting scheme in the second part that can improve the detection of the association signal by taking advantage of two weighting schemes. This score test method is applicable when $p > n$. However, in this thesis, we do not discuss the issue of high dimensionality; we only consider cases where $p < n$.

6.1.1 Motivation

In this section, we introduce four functions used as weighting schemes at the variant level in the analysis of rare variant association. All these functions share the concept of up-weighting rare variants and down-weighting common variants. However, the down-weighting of common variants does not dominate the signal of association. We introduce a Cauchy function that has two properties. The first property is that it has two parameters that need to be specified, and both are functions of MAF. The second property is that the weight can be adaptive depending on the MAF distribution. We also introduce a Levy function in which only one parameter is specified and consider a weighting scheme that can up-weight rare variants, especially extremely rare ones.

6.2 Simulation 1

For the simulation, we follow the same settings used in Chapter 5. We set $p =$ 200 and 100 SNPs in all of the scenarios. Each dataset contains three types of variants (i.e., extremely rare, moderately rare, and common) simulated based on pre-specified threshold values using a uniform distribution. Based on weighting schemes proposed in section (6.3.2) and (6.3.3), the percentage of variant types vary to express the weighting scheme's impact. These percentages are provided under each Figure. For example, to express the impact of having a low number of extremely rare variants (ERV) compared to a large number of common ones, we increase the common ones from 5% to 90%. The variants' effect sizes are $OR = 1$ to simulate a type I error and $OR > 1$ to evaluate the tests' power.

See Table (6.2) for additional details. To estimate p-values, straight binomial proportions are used. Hence, they have the same standard error as any other binomial proportion $\sqrt{(}p(1-p)/n)$, where p means the proportion of tests rejected and n the number of samples. Therefore, if $p = 0.05$ and $n = 2000$, the standard error of the observed proportion is about 0.005, and we could say the uncertainty is 1%.

Parameters	Parameter Values
Sample Size	n =2000 (cases = control = n/2)
Total Number of SNPs	100, 200
Proportion of Causal SNPs	[3%-20%]
Effect Size of Non-causal SNPs	OR = 1
Effect size of Causal SNPs	OR = Unif(1.5,3) or Unif(1/2,1/4)
Percentage of Common Variants	Ranged from 10% to 90%

Table 6.1: The full set of parameters used in the simulation.

We generated data under various causal mechanisms and MAFs for causal and non-causal variants, as summarised in Table (5.2).

In the section related to combining weights (6.8), we set the number of variants to be $(100 - 200)$, classified as extremely rare, moderately rare, and common variants; the percentages are 40% for both extremely and moderately rare and 20% for common variants, but the causal variants are fixed at 7%.

We evaluate power using different weighting schemes and show the impact of each weighting scheme on the power of score test outcomes among the minor allele frequencies from extremely rare to common variants.

Part I

Weighting Schemes

6.3 Cauchy Function

6.3.1 Adjusted Cauchy Weight

The adjusted Cauchy weight is based on the ratio between the different MAFs, with the previously proposed weight up-weighting low MAFs and down-weighting high ones. However, by doing so, the common variants have low weights (sometimes even approaching zero) as in Wu *et al.* (2011), which dominate any signals of association in the common variant region. When the weights for low and high MAFs are applied, the ratio between the maximum and minimum weights becomes very large. Unfortunately, the signals of association are not detectable if some causal variants are more common.

It is well known that if we divide the weight into its sum, there will be no impact on the test. The weight depends on the ratio between its values for different variants.

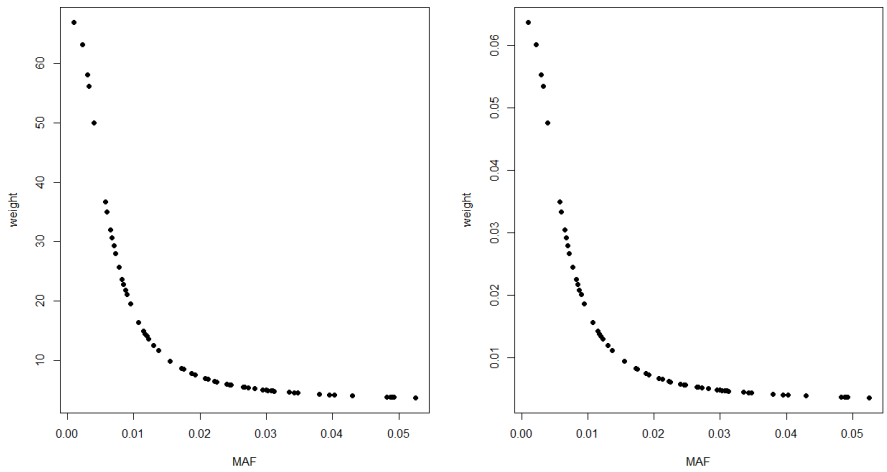


Figure 6.2: The weight in different values; the first one is the original weight w, and the second one is divided into its sum $w/\sum(w)$.

The **U** vector, which is the score function, is affected by the weight, so if the ratio between the weights of the rare and common variants is large, the ratio between the **U** values for the rare and common variants will also be large. By

using the previous weight discussed in Chapter 5 or the weight introduced by Wu *et al.* (2011), if the causal variants are in the common range, it will not be possible to detect it because it will be dominated by the non-causal variants in the rare region. To overcome this limitation, we made a weight adjustment so that we can detect associations in the common variant range. We added a constant, which was a function of the weight, to the weight itself. To calculate this adjustment, let w be a weight, m the maximum of w, and $a = 0.05$ to control the magnitude of the constant. Then, the adjusted weight will be

$$w_{adj} = am + w. \tag{6.1}$$

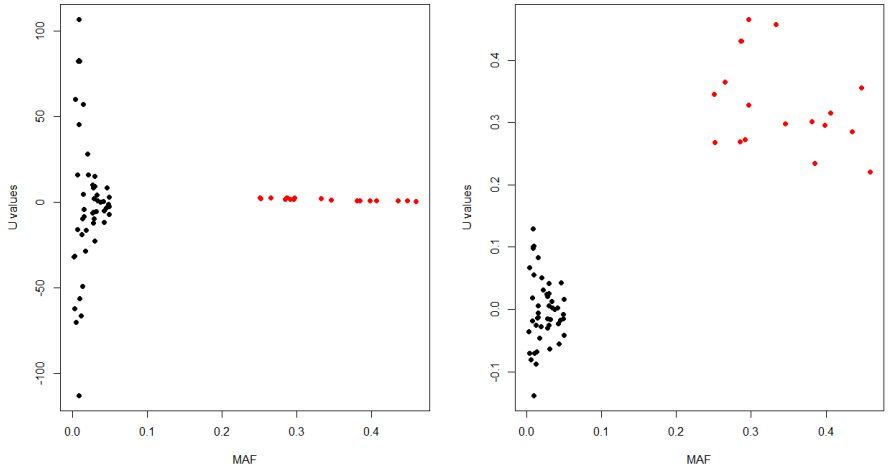


Figure 6.3: An illustration of the **U** vector versus the MAF; the left panel shows the **U** with no adjustment to the Cauchy weight; the right shows the Cauchy weight with the adjustment. The red dots indicate the **U** values associated with common causal variants. In these Figures, we simulate 100 variants with different MAFs ranging between 0.0005 and 0.5, while we fix the causal variants to be common $(0.1 - 0.5)$.

If the constant in equation 6.1 is large, it will dominate the association signal in the rare range; we find that a 0.05 multiple of the maximum weight is optimal for various proposed Cauchy weights. This value balances the weight and facilitates

the detection of the signal in different MAF ranges; see Figure (6.4) for additional details. Therefore, by using this adjustment, the causal variants can be detected at any range of MAF ($\mathcal{F}$); Figure (6.4) shows the benefit of using an adjustment when the causal variants are in the MAF common range.

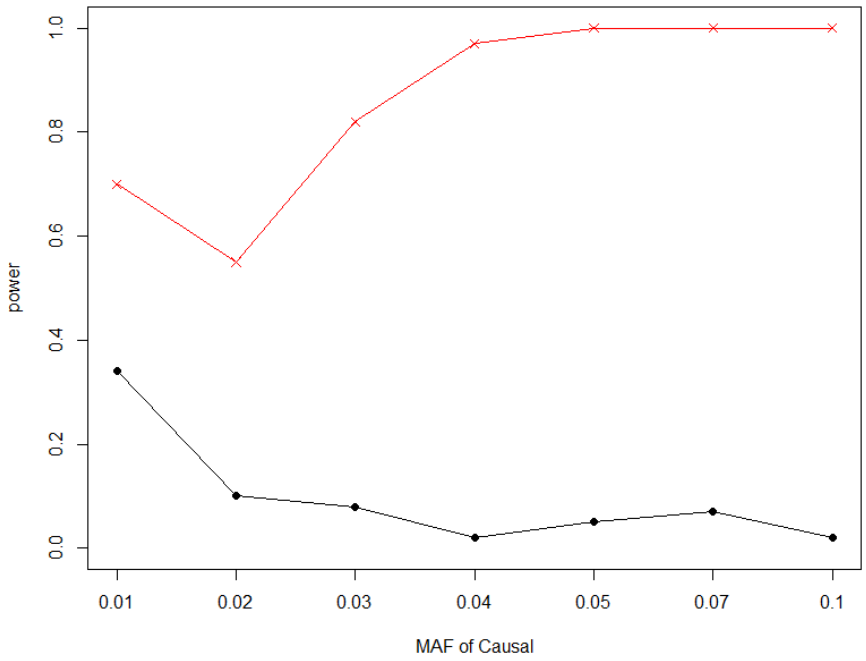


Figure 6.4: Illustration of the impact of the adjusted weight when the MAFs of the causal variants are greater than 0.01 (i.e., common variants). The red line represents Cauchy with an adjustment, and the black line represents no adjustment.

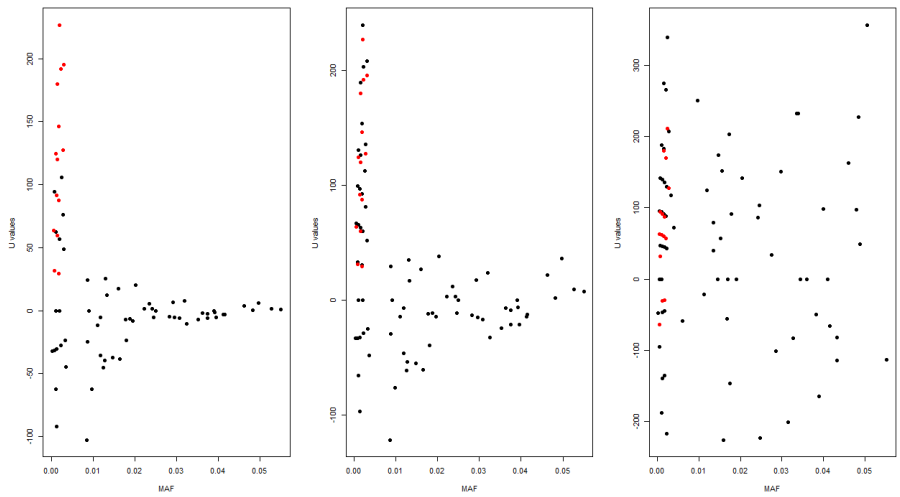


Figure 6.5: Illustration of the difference between the adjusted weights when a is equal to 0.05 and 0.5, including a comparison with no adjustment. The red dots are the causal variants.

6.3.2 Cauchy Adaptive Weight Scheme 1

In Chapter 5, we proposed a weighting scheme based on the distribution of the MAF by fixing one parameter of the Cauchy function and varying the other parameter based on MAF. In this section, we consider the same idea but include the Cauchy adjustment, so the common region can be included in the region that can have a non-zero weight.

Let $g(\mathcal{F}) = cauchy(\mathcal{F}, \min(\mathcal{F}), b + 0.01)$, where $\mathcal{F}$ is the MAF, $\min(\mathcal{F})$ is the minimum possible MAF, and b is the third quartile ($Q_3(\mathcal{F})$) divided by 10 as illustrated in detail in the previous chapter 5.3.3. Then, take the adjusted weight as in equation 6.1;

$$w(\mathcal{F})_{adj} = g(\mathcal{F}) + 0.05 \ \max(g(\mathcal{F})) \tag{6.2}$$

If we assume the causal variants have very low MAFs (i.e., extremely rare variants), then if there are more rare variants than common ones, the association signal increases, and if there are more common variants than rare ones, the association signal decreases, especially in the extremely rare variant region since

the frequency of extremely rare variants in the data is low. The idea behind this weighting scheme is reducing the weight in the extremely rare variant region (range) of MAF when the number of variants in this region is very low (few variants), which may occur due to a systematic error.

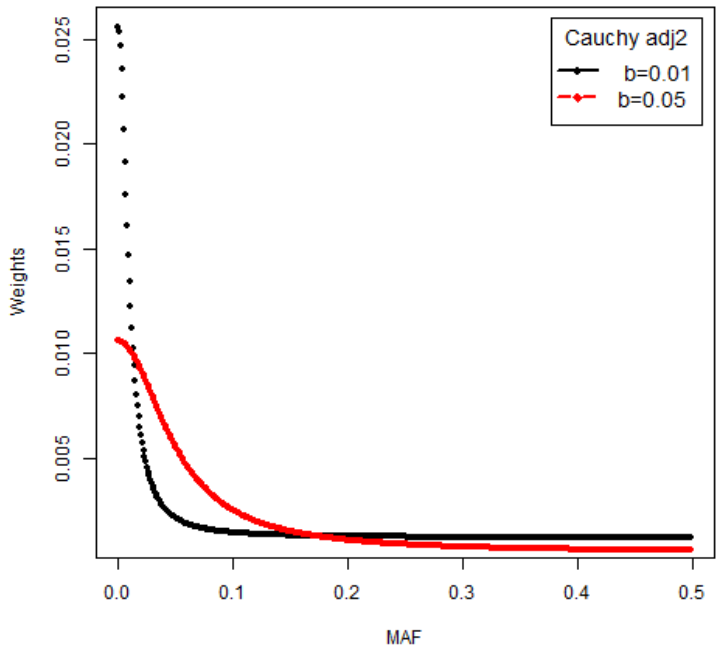


Figure 6.6: The weight based on MAF. The red line shows the weight based on data has 50% common variants between $0.05 - 0.5$ so that $b = 0.05$, while the black line shows the weight for data with only 5% common variants ($b = 0.01$).

In Figure 6.7, the impact of having a large proportion of common variants on the association's detection will be larger when the causal variants are extremely rare in terms of MAF and reduced when the MAF increases. Figure 6.7 shows the impact of increasing the number of common variants in the data, which will lower the detection rate for signals of association.

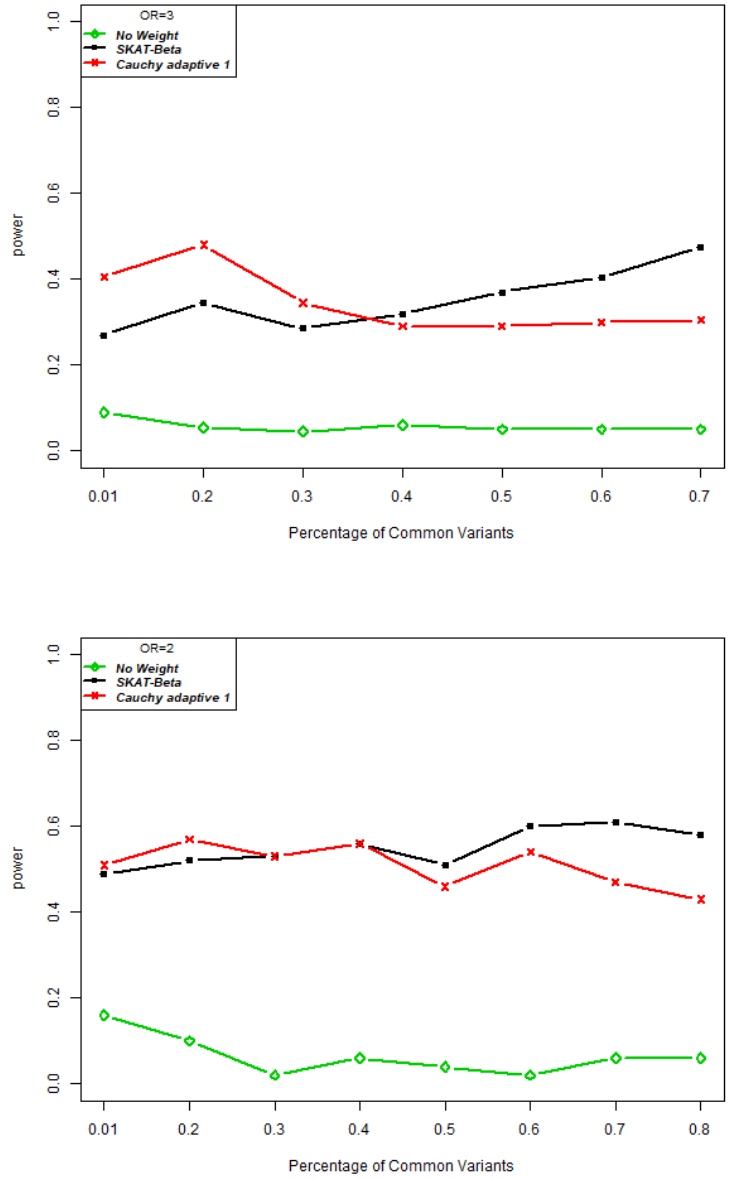


Figure 6.7: This Figure shows the impact of increasing numbers of common variants in the data, which lowers the ability to detect the signal of association. The causal variants (10%) in this analysis are located in the extremely rare range; in the top Figure, causal variants have MAFs between $0.0005 - 0.0015$ and $OR = 3$, and in the lower one, the MAFs are $0.002 - 0.005$ and $OR = 2$. The number of rare variants is fixed to 100, while we increase the number of common variants from 1% to 80%. Therefore, the endpoint in the horizontal axis shows 70% of the data are common.

When the causal variants are in the extremely rare range and most of the data are rare, this weighting scheme is very effective at detecting association signals. When the causal is still in the moderately rare range but with a large MAF, then the importance of causality is not the same as when the causal variants are in the extreme range of MAF (see Figure 6.8). In short, this weighting scheme can be adjusted based on the frequency of the data. For example, the variant $x1$ at MAF $f1$ has a large weight $w1$ when a large proportion of variants have MAFs close to $f1$.

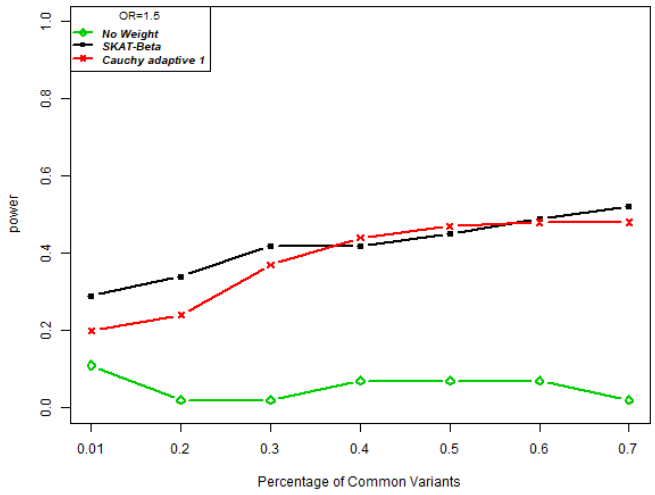


Figure 6.8: The Figure shows the impact of increasing the proportion of common variants in the data, which increases the signal of the association since the causal variants are in the moderate range of MAF. The causal in this analysis is located in the moderately (large MAF) rare data range between 0.005 and 0.01, and the percentage of the causal variants is 10%. The number of variants is fixed at 100. In the first analysis, all the data are rare, and then we increase the number of common variants from 1 to 80 and decrease the number of rare variants.

	Causal	Non-Causal
ERV	.	✓
MRV	OR=1.5 [0.005-0.01]	✓
CV	.	✓Common variants increased from 1%-70%.

6.3.3 Cauchy Adaptive Weight Scheme 2

The third function considered is similar to the previous one. It is also based on a Cauchy function. However, the two parameters of the Cauchy are related to the data, and it is not fixed. Both of them are based on the distribution of the MAF, so the rarest variants do not have the largest weight. Let $\mathcal{F}$ represent the minor allele frequency, and $\mathcal{F}^*$ is the minor allele frequency excluding the common one (more than 0.05). Then, the weight function is $g(\mathcal{F}) = Cauchy(\mathcal{F}, a, b+c)$, where a is the 25% quartile ($Q_2(\mathcal{F}^*)$) of MAF less than 0.05, b is the 75% quartile of MAF less than 0.05 ($Q_3(\mathcal{F}^*)$) and $c = \sqrt{(2n)^{-1}/2}$. Thus, the weight after the adjustment, as in equation 6.1, is

$$w_{adj} = g(\mathcal{F}) + 0.05 \ \max(g(\mathcal{F})) \tag{6.3}$$

This weighting scheme can be adjusted based on the type of data available. In this scheme, we do not give the highest weights to the extremely rare variants (ERVs); instead, we assign weights to rare SNPs in the order of frequency, weighing the most frequently occurring rare data the highest. We adjust these weights by adding a constant, which helps give the common variants a small weight, facilitating detection if the causal variants are located in the common region.

The next Figures show the changes in weight based on the data and MAFs.

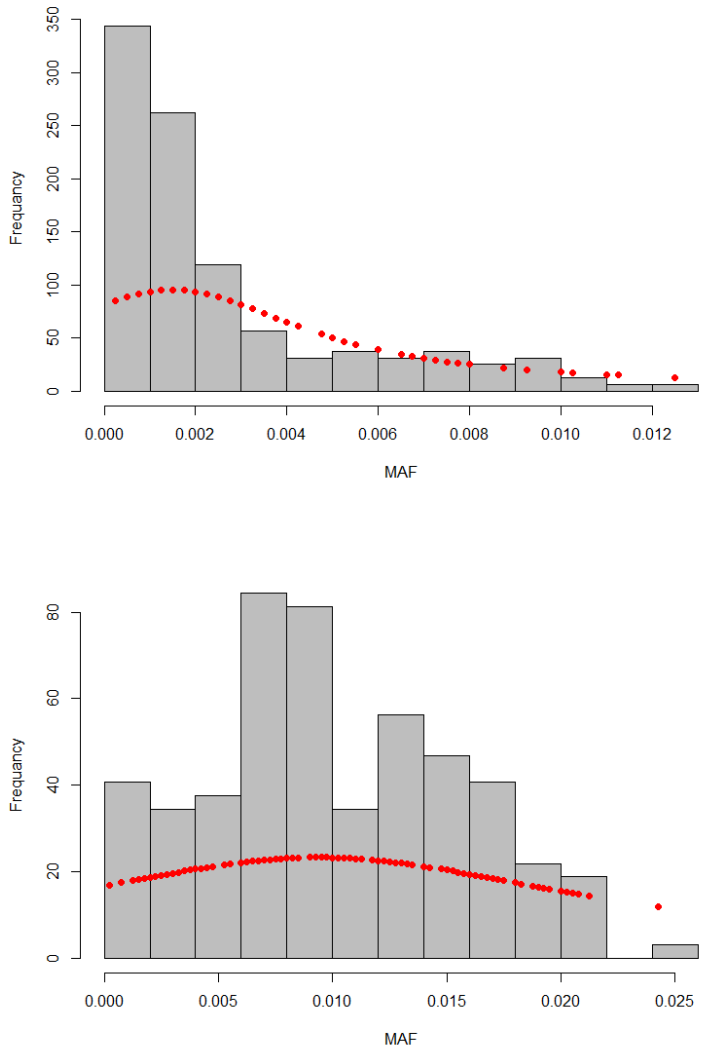


Figure 6.9: The top Figure shows the data distribution where most of the variants are extremely rare in terms of MAF. The bottom one presents a data distribution where most of the variants have moderately rare MAFs, and some are close to the common range. The weights are indicated by the red dots.

When most of the data are rare, having a causal variant in the same range is most detectable as a real association; however, if the causal variant is in a different range, then the association is less detectable. Therefore, if the data sample has more common variants compared to the number of rare variants,

then the association is not easily detectable. Having more common and fewer rare variants thus reduces the importance of causal variants that we assume are extremely rare (see Figure 6.10).

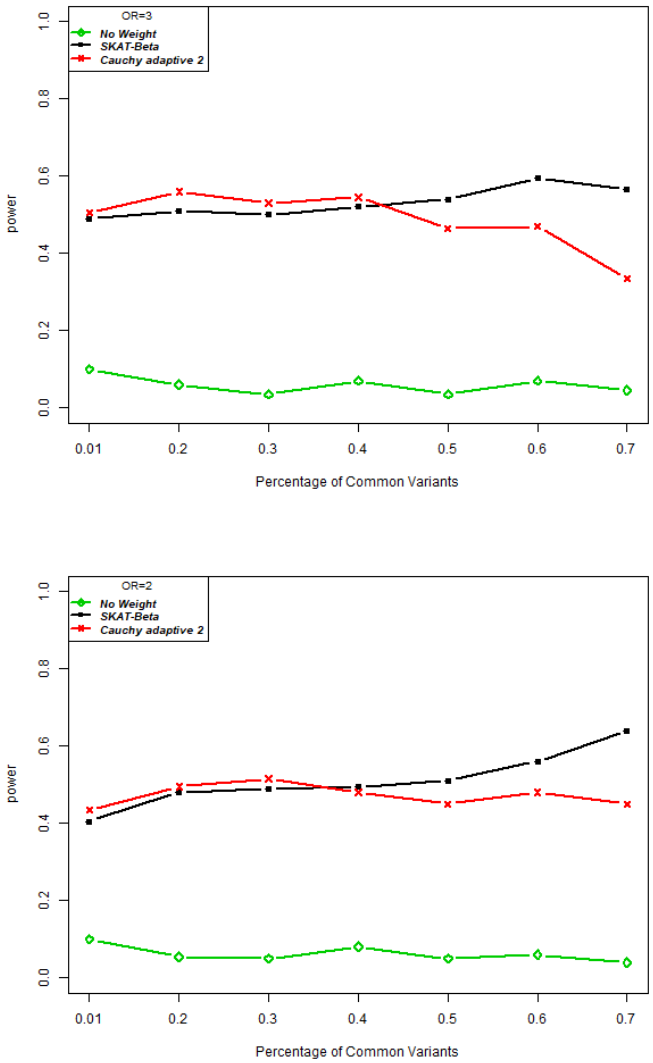


Figure 6.10: Impact of having a larger number of common variants in the data when causality is located in the extremely rare range, fixed at MAF $(0.0005, 0.002)$ with $OR = 3$ for the top Figure and MAF $(0.002, 0.005)$ with $OR = 2$ for the bottom one. The horizontal axis line represents the number of common variants, increasing from 10% to 70%; there are 100 rare variants. The vertical axis shows the power of the test, which equates to the detection of association.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.002]	✓
MRV	.	✓
CV	.	increase from 1%-70%.

	Causal	Non-Causal
ERV	OR=2 [0.002-0.005]	✓
MRV	.	✓
CV	.	increase from 1%-70%.

Figure 6.11 shows the impact of the parameter a in the weighting scheme, which is associated with the 25 quartile of the data. In Figure 6.11, we fix the causal's MAF to be 0.001 and vary the parameter a, so on the horizontal axis, we can see that when the 25th quartile of data is between $0.0005 - 0.0035$, the power is higher than when it is less. Thus, when extremely rare variants occur more often in the data, they have a higher weight.

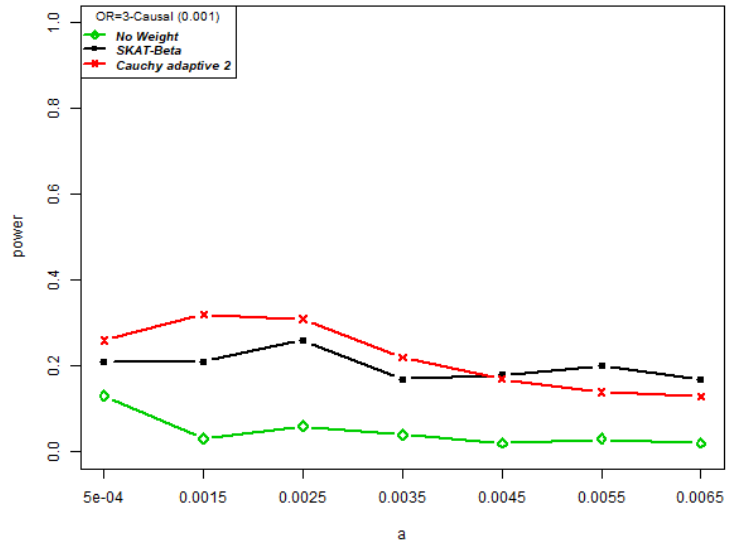


Figure 6.11: In this analysis, we generate 200 variants: 50% extremely rare, 30% moderately rare, and 20% common variants. We fix the causal to be extremely rare variants (0.001). We change the parameter a of Cauchy to take values between $(0.0005 - 0.006)$.

However, if the causal variants are in the ERV range when the amount of extremely rare data increases, then the detectability of the association likewise increases.

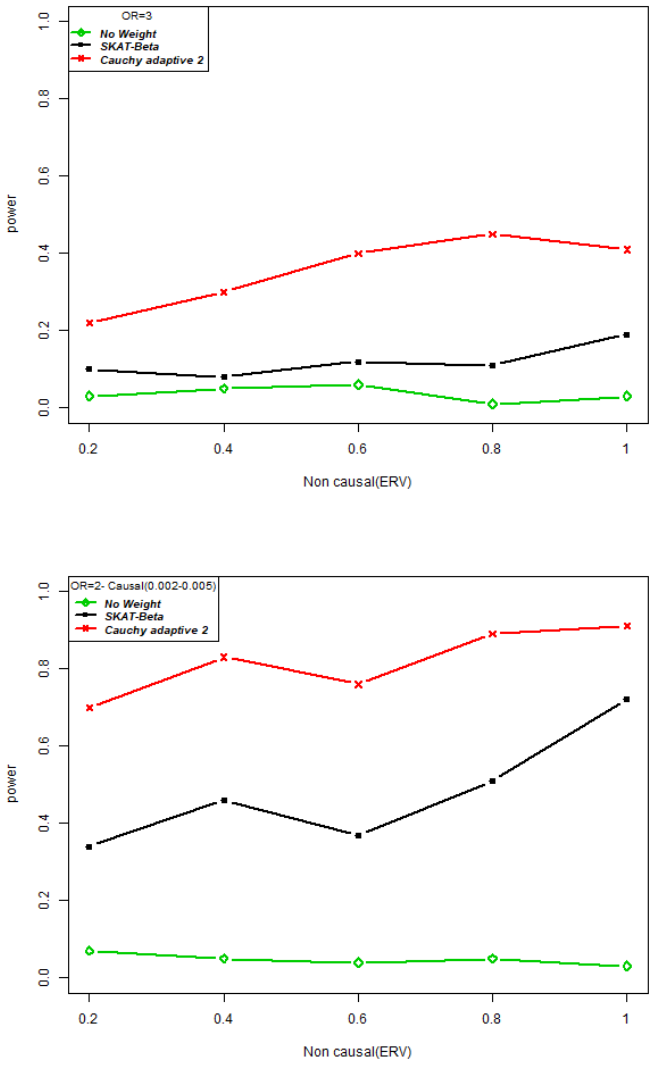


Figure 6.12: Impact of having a larger number of rare variants in the data when the causality is located in the extremely rare range, fixed at MAF $(0.0005, 0.002)$ for the top Figure and at MAF $(0.002, 0.005)$ for the lower one. The horizontal axis line represents the number of rare variants, increasing from 20% to 100%; 100 variants are included in the analysis. The vertical axis shows the power of the test, which equates to the detection of an association.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.002]	increased 20%-100%
MRV	.	✓
CV	.	✓

	Causal	Non-Causal
ERV	OR=2 [0.002-0.005]	increased 20%-100%
MRV	.	✓
CV	.	✓

6.3.4 Cauchy (Fixed)

In this section, we modify the Cauchy function to accommodate all MAF regions. It performs well with extremely rare variants and can detect association signals throughout the MAF scale. The parameters a and s are fixed.

$$\begin{aligned} g(\mathcal{F}) =& \big[(s(0.5 + \frac{(\mathcal{F} - a)}{s}))\big]^{-1} \\ =& \big[\frac{s}{2} + (\mathcal{F} - a)\big]^{-1}, \end{aligned}$$

where $s = \frac{1}{\sqrt{(n)}}$, so that when $n = 2000$ then $s = 0.02$; this threshold classifies the rare and common variants the same way as in other works, such as Tony Cai *et al.* (2011), Jeng *et al.* (2012), and Ionita-Laza *et al.* (2013), and a is the minimum MAF ($\mathcal{F}$). Then, the weight is

$$g(\mathcal{F}) + 0.05 \ \max(g(\mathcal{F})) \tag{6.4}$$

This weighting scheme (6.4) can cover the entire MAF range ($\mathcal{F}$) when the effect size is small, except for the $(0.4, 0.5)$ region. Hence, by using the Cauchy adjustment explained above, we can say it covers all MAF regions under these circumstances. When s is small, the extremely rare variants are highly weighted, meaning the association signal is easily detected when the causal variants are in the extremely rare region; however, this reduces the power of detection when the causal variants are less rare $(0.01 - 0.05)$.

We consider another weight function that is based on fixed parameters. The weighting scheme is based on a Cauchy function with an adjustment to put reasonable non-zero weights on the common variants. Let $g(\mathcal{F}) = cauchy(\mathcal{F}, \min(\mathcal{F}), \frac{1}{\sqrt{(2n)}})$, where $\min(\mathcal{F})$ is the minimum possible of the MAF. Then, let

$$w_{adj} = g(\mathcal{F}) + 0.05 \ \max(g(\mathcal{F})) \tag{6.5}$$

This weighting scheme is not affected by increasing the number of common variants in the data and can detect signals of association in any MAF region. Figure 6.13 shows that increasing common variants in the data has no impact.

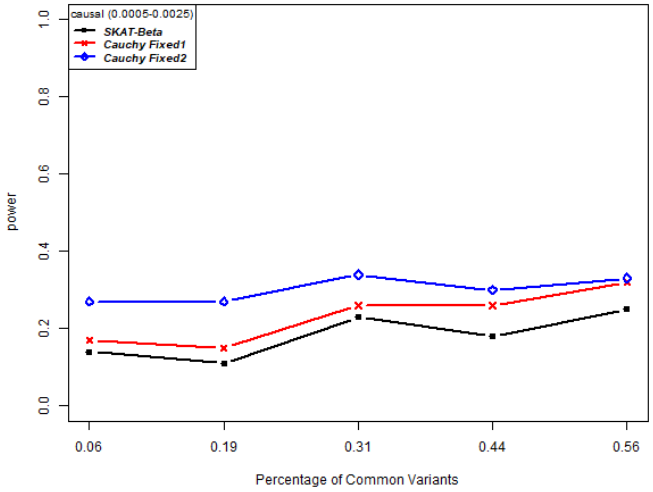


Figure 6.13: The causal in this analysis is located in the extremely rare data range $0.0005 - 0.0025$ with $OR = 3$, and the percentage of causal variants is 9%. The rare data is fixed to be 100 variants, and we increase the percentage of common variants from 6% to 56%. The MAF of rare variants ranges between 0.0005 to 0.05, while the common ones are between $0.05 and 0.5$.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.0025]	✓
MRV	.	✓
CV	.	Increase from 6% to 56%

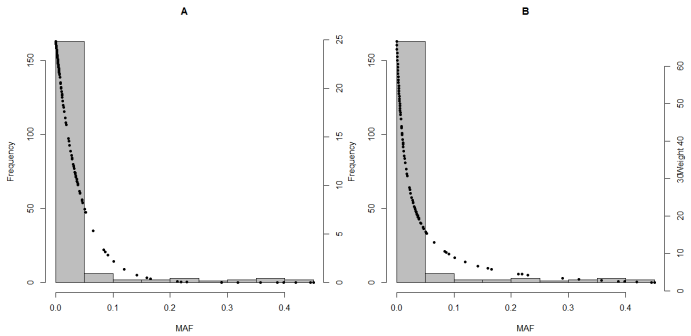


Figure 6.14: Comparing beta (A) and Cauchy-fixed1 weight (B).

Figure 6.15 shows the impact of this weight scheme in terms of test power. The Cauchy-fixed (1) and Cauchy-fixed (2) weight schemes clearly perform better than beta when the causal variants have extremely rare MAFs.

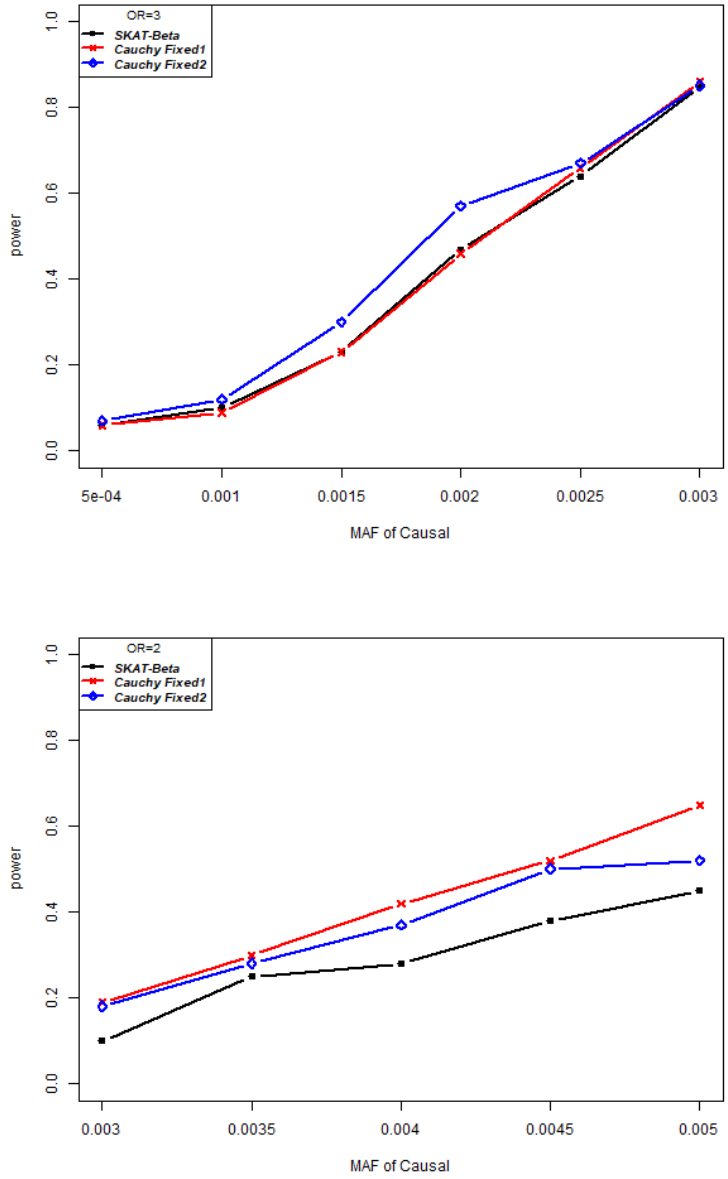


Figure 6.15: The Figure depicts a comparison of beta-SKAT, Cauchy-fixed1 and Cauchy-fixed2 when causal variants are in the extremely rare region. There are 10 causal variants with MAFs ranging between $(0.0005 - 0.0025)$ with OR=3 in the top Figure and $0.003 - 0.005$ with OR=2 in the bottom one, among 200 non-causal variants. The data contains 40% non-causal extremely rare, 40% moderately rare, and 20% common variants.

With moderately rare variants and common variants, we see small differences between beta and Cauchy-fixed 1; however, Cauchy-fixed 1 is still able to detect the association in the common region using the Cauchy adjustment. The bottom Figure in 6.16 shows the weights with no adjustment, but by using the Cauchy adjustment, the detection of association in common areas is larger (see the top Figure in 6.16). Additionally, Cauchy-fixed (2) is more powerful than beta-SKAT in most of the MAF regions except 0.02-0.1.

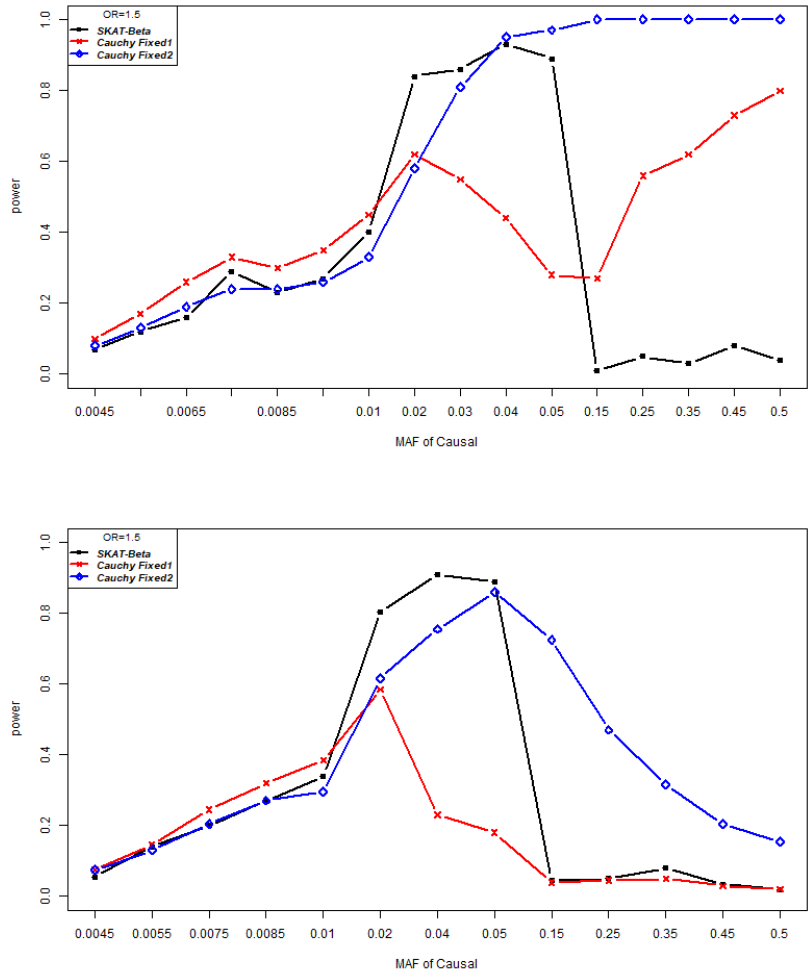


Figure 6.16: The Figure shows a comparison of beta-SKAT, Cauchy-fixed 1, and Cauchy-fixed 2 when the causal variants are in moderately rare and common regions. The top Figure includes the Cauchy adjustment, and in the bottom one, there is no Cauchy adjustment. There are 10 causal variants with MAFs ranging between $(0.0045 - 0.5)$ among 200 non-causal variants; the data contains 40% non-causal extremely rare, 40% moderately rare, and 20% common variants.

	Causal	Non-Causal
ERV	OR=1.5 [0.0005-0.005]	40%
MRV	OR=1.5 [0.005-0.05]	40%
CV	OR=1.5 [0.05-0.5]	20%

This weighting scheme performs well when the causal variants are extremely rare. Figure 6.17 shows the impact of increasing the number of extremely rare causal variants $(1/n - 3/n)$.

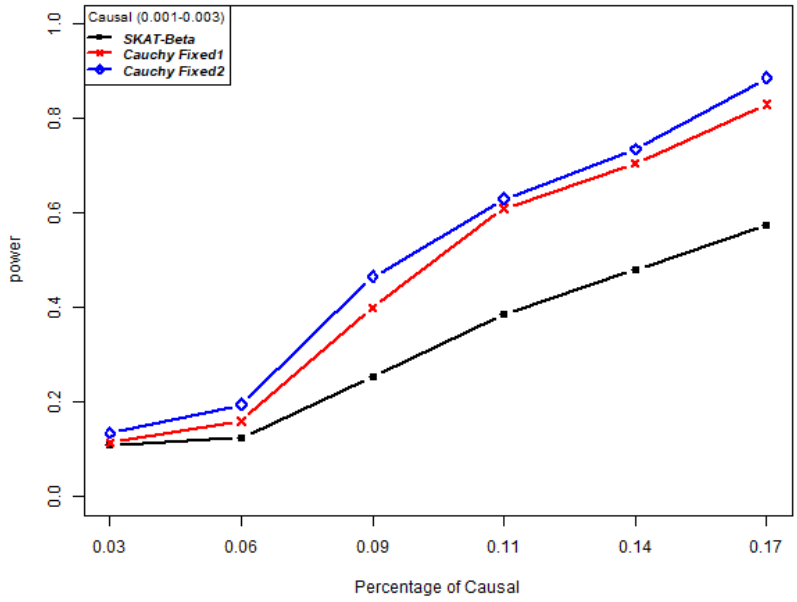


Figure 6.17: The Figure presents a comparison of beta-SKAT, Cauchy-fixed 1, and Cauchy-fixed 2 when the causal variants are extremely rare. The MAF of causal variants is fixed at $(0.0005 - 0.003)$, and we increase the amount of these variants as shown on the X-axis. In the analysis, there are 200 non-causal variants, 40% non-causal extremely rare, 40% moderately rare, and 20% common variants. The OR is fixed at 3 for causal variants.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.003] Increase 3% to 17%	✓40%
MRV	.	✓40%
CV	.	✓20%

6.4 Levy Weight Scheme

Choosing the right parameters in weighting scheme functions is critical. In Chapter 5, the Cauchy weight and Gumbel use parameters that are estimated from the data (i.e., MAF), so all the parameters are a function of the observed MAFs. Each function has at least two parameters. It is a good idea to reduce the number of parameters to avoid choosing arbitrary parameters.

In the Levy function, we only use one parameter, and it is also a function of the observed MAF, which is the minimum observed MAFs. Therefore, this kind of weighting scheme reduces the number of parameters, and it is a function of the observed MAF. The Levy weighting scheme is a recommended weight for rare variants since it can be used to up-weight rare variants and down-weight common ones. This weight up-weights the extremely rare variants more than Cauchy. It has fixed parameters, so it is not adjustable for different data distributions. A Levy weighting scheme function with some modification is introduced as follows:

$$\sqrt{\frac{s}{\mathcal{F}}}exp\{\frac{-s}{2\mathcal{F}}\}, \tag{6.6}$$

where $s = 1/n$ and $\mathcal{F} = MAF$.

This function (6.6) can be used to up-weight rare variants; we suggest setting $s = 1/n$, which is a function of MAF because it increases the weight of rare variants while still giving acceptable non-zero weights to variants with MAF $1\% - 5\%$. Care must be taken that the end of the weight range does not get too close to zero (i.e., no lower than 0.01). As mentioned previously in this chapter, one limitation of using the beta function is that since the weights for variants with MAFs of $1\% - 5\%$ are close to zero (less than $0,01$), the ratio between rare variants and common variants is very large, so it is difficult to detect any signals of common variants. Another advantage of this weight is that by modifying s, we can down-weight the singleton, so when $s = 1/n$, the smallest MAF 0.0005 is associated with a large weight, and when $s = 2/n$, $1/n$ has a low weight as shown in Figure 6.18.

The weights of variants with MAFs between $1\% - 5\%$ are not affected when the sample size increases. For example, when we increase the sample size to

50,000, the ratio of the minimum to maximum MAFs is 0.01; however, in the beta weight, it will be close to zero.

We show a comparison of the beta and Levy weight schemes (6.6) below using a MAF range from $0.0005 - 0.5$. We will illustrate the Levy weight scheme's performance when the causal variants are extremely rare and the performance of this weight for MAFs with different effect sizes.

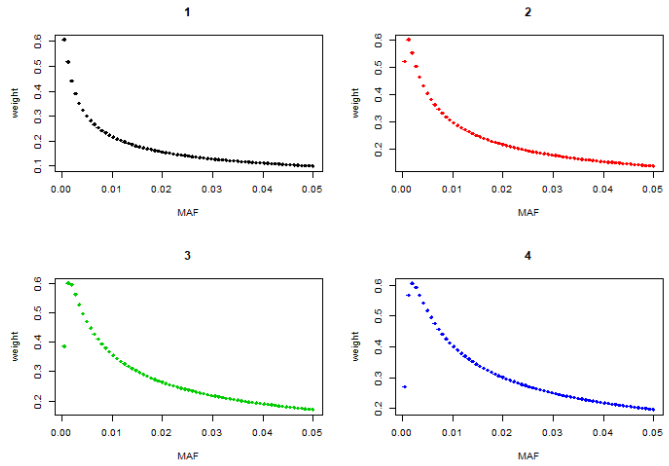


Figure 6.18: Levy comparisons using different s values: $s = 0.0005, 0.001, 0.0015$, and 0.002, respectively.

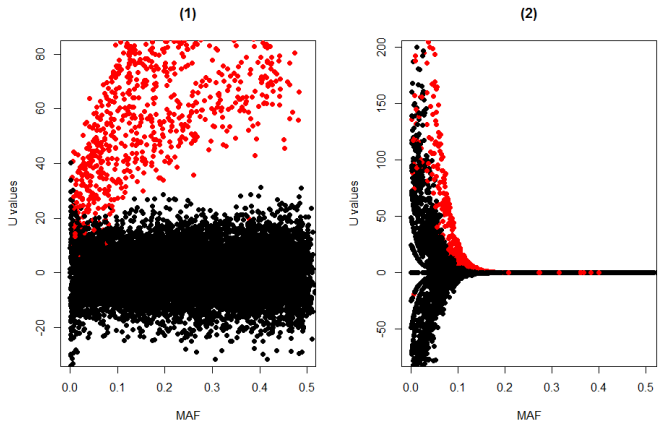


Figure 6.19: Comparison of the beta proposed by Wu *et al.* (2011)(right) and Levy weight (left) in terms of the **U** vector. We sample the effect size of the causal variants between 1.2-3; the red dots show the **U** values associated with causal variants. The left Figure shows the **U** values can take large values across the MAF, while it is penalised to be very small in the beta weight.

Here, we demonstrate the effect of weights on the **U** vector and differences between the beta function by Wu *et al.* (2011) and with a Cauchy adjustment in terms of **U**. First, we show that the weight affects the **U** vector. For simplicity, consider a model in which y is normal:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i \quad \text{for} \quad i = 1, \ldots, n, \tag{6.7}$$

where $e_i \sim (0, \sigma^2)$, and X is a matrix with the elements $x_{ij} = 0$ and 1, which correspond to independent variants (SNPs).

Let $P(x_{ij} = 1) = p_j$ and $m_j = \sum_{i=1}^{n} x_{ij}$. Consider the score statistics $S = (S_1, \ldots, S_p)^T$ with

$$S_j = \sum_{i=1}^{n} (y_i - \bar{y}) x_{ij} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j) y_i,$$

which arises from the maximum likelihood theory for testing $H_0 : \boldsymbol{\beta} = 0$. Since y is normal, the distribution of S_j given the genotypes X is $S_j \sim N(m_j(1 - m_j/n)\beta_j, m_j(1 - m_j/n)\sigma^2)$, where $m_j = \sum_{i=1}^{n} x_{ij}$. For any given sample, the m_j

are treated as fixed values, and for simplicity, we consider the case where m_j is equal to its expected value np_j so that

$$S \sim N(\mu, \Sigma),$$

where $\mu = (np_1(1-p_1)\beta_1, \ldots, np_p(1-p_p)\beta_p)$ and $\Sigma = diag(np_1(1-p_1)\sigma^2, \ldots, np_p(1-p_p)\sigma^2)$. Considering the model (6.7), the total variation of y explained by set of SNPs p is

$$\Upsilon = \frac{var(E(y_i|x))}{var(y)} = \sum_{j=1}^{p} p_j(1-p_j)\beta_j^2/\sigma^2 = \sum_{j=1}^{p} \Upsilon_j$$

Where Υ_j is the explained variation by a set of SNPs p. We will introduce the effect on power. Consider the test by Madsen & Browning (2009). Let w be a weighting scheme. The distribution of

$$T_w = w^T S.$$

is

$$T_w \sim N(\sum_{j=1}^{p} w_j p_j(1-p_j)\beta_j, n\sum_{j=1}^{p} w_j^2 p_j(1-p_j)\sigma^2)$$

$$T_w^2/(\sum_{j=1}^{p} w_j^2 p_j(1-p_j)\sigma^2) \sim \chi^2_{1,s}$$

where

$$s = \frac{n(\sum_{j=1}^{p} w_j p_j(1-p_j)\beta_j/\sigma)^2}{\sum_{j=1}^{p} w_j^2 p_j(1-p_j)}$$

Based on the result above, the power of a linear statistic, as one can see in the result above, depends on signs of effects (effect directions of β) and weights. So, the optimal weight may boost power.

Next, we focus on the effect of different weighting schemes on the vector $\mathbf{U}$. In the Figure 6.19 above, we can see the difference between two weighting schemes; the first one is the beta function used by Wu *et al.* (2011), and the second one is a Cauchy function with the adjustment proposed in this thesis.

Let w be the weight and w_1 the weight proposed by Wu *et al.* (2011), and w_2 is a Cauchy function with the adjustment proposed in this thesis. We define the $\mathbf{U}$ vector, as well as the weights and their effect;

$$\mathbf{U}_1 = W_1 X^T(\mathbf{y} - \boldsymbol{\mu})$$

$$\mathbf{U}_2 = W_2 X^T(\mathbf{y} - \boldsymbol{\mu})$$

where W_1 and W_2 are the diagonal matrixes for the beta weight by Wu *et al.* (2011) and a Cauchy weight with the adjustment, respectively. To illustrate the effect of weighting schemes on vector $\mathbf{U}$, we consider a matrix X 3×2000, so we have three values of MAF ($\mathcal{F}$) fixed at $0.0005, 0.005, and 0.5$ to express the extremely rare, moderately rare, and common variants, respectively.

Using the weighting schemes w_1 and w_2,

$$\begin{cases} min(w_1) \approx 0 \\ min(w_2) > 0 \end{cases}$$

Then, $\mathbf{U}$ has three values according to the given MAF. By using w_1, the $\mathbf{U}$ values are defined as

$$\begin{cases} \mid U_1 \mid > 0 \quad if \quad \mathcal{F} = 0.0005 \\ \mid U_1 \mid > 0 \quad if \quad \mathcal{F} = 0.005 \\ \mid U_1 \mid \approx 0 \quad if \quad \mathcal{F} = 0.5 \end{cases}$$

Using w_2, the $\mathbf{U}$ values are defined as

$$\begin{cases} \mid U_2 \mid > 0 \quad if \quad \mathcal{F} = 0.0005 \\ \mid U_2 \mid > 0 \quad if \quad \mathcal{F} = 0.005 \\ \mid U_2 \mid > 0 \quad if \quad \mathcal{F} = 0.5 \end{cases}$$

Thus, by using the Cauchy weighting scheme, we are putting a weight greater than zero on the common region, which will allow us to detect the signal of association.

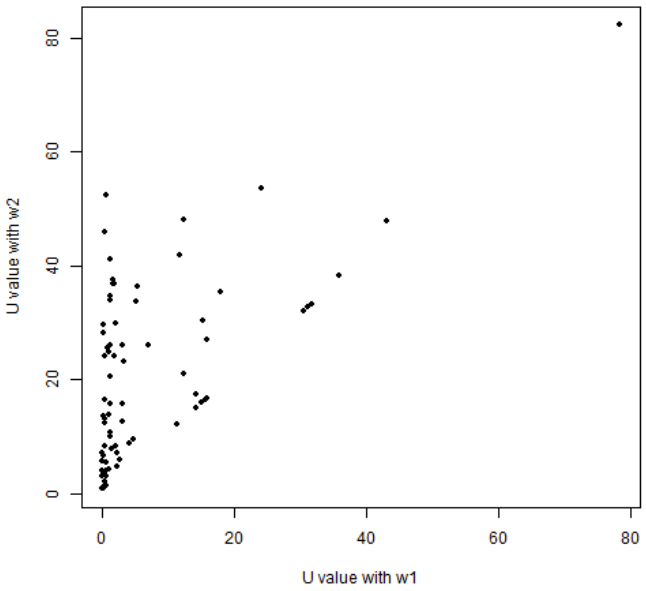


Figure 6.20: Comparison of **U** values using Cauchy with no adjustments to include common variants and Cauchy with an adjustment. We sample the effect size of the causal variants between 1.2-3. w_1 is the beta-SKAT weight, and w_2 is the proposed weight that can consider the common region.

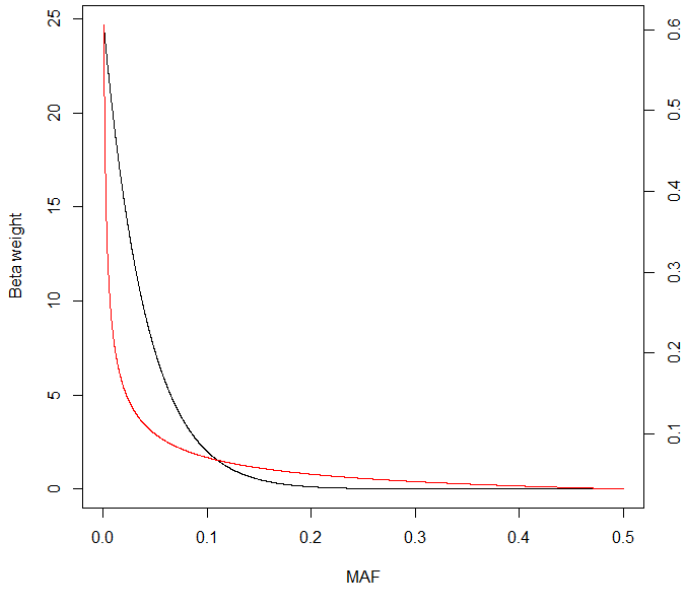


Figure 6.21: The weight for both SKAT-beta with parameters $(1, 25)$ (black line) and the Levy function (red line).

When the causal variants are in the extremely rare range, the Levy weight function performs better than the beta weight, as shown in Figure 6.22.

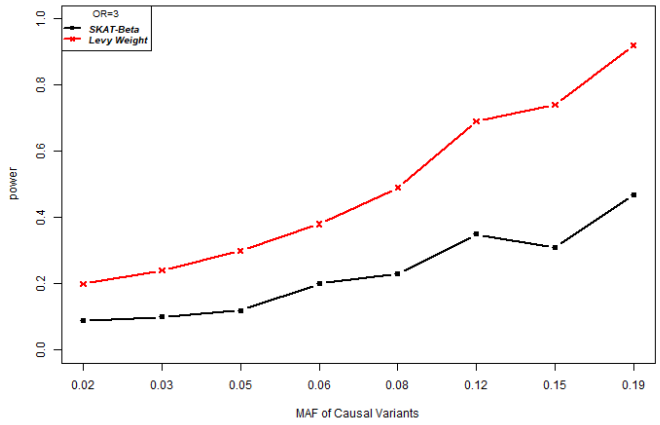


Figure 6.22: In this Figure, we show the Levy weight scheme's performance when the causal variants are very rare (i.e., extremely rare variants). We generate 100 SNPs ranging from extremely rare to common variants with causal variants classified as extreme, which have MAFs between 0.0005 and 0.002 with $OR = 3$. Then, we vary the percentage of causality along the horizontal axis.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.002] Increase from 2% to 19%	✓40%
MRV	.	✓40%
CV	.	✓20%

In Figure 6.23, we show the performance of the Levy compared to the beta function in the ERV region when MAFs are between 0.0005 and 0.005 with $OR = 3$. When we reduce the OR to 2, then we can see that in the top Figure in Figure 6.24, the Levy weight scheme still outperforms the SKAT weight scheme, especially in the extreme region; SKAT also appears to be worse in the common area compared to Figure 6.24.

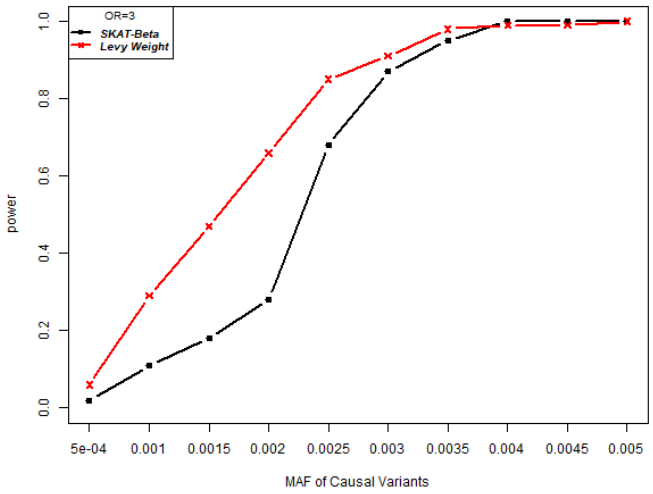


Figure 6.23: In this Figure, we show a comparison of the Levy function and beta-SKAT in terms of the MAF of causal variants. There are 10 causal variants among 100 variants. The OR of the causal variants is fixed at 3, and the causal variants are in the MAF extreme range. The non-causal variants are extremely rare, moderately rare, and common with percentages of 60%, 30%, and 10%, respectively.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005]	✓60%
MRV	.	✓30%
CV	.	✓10%

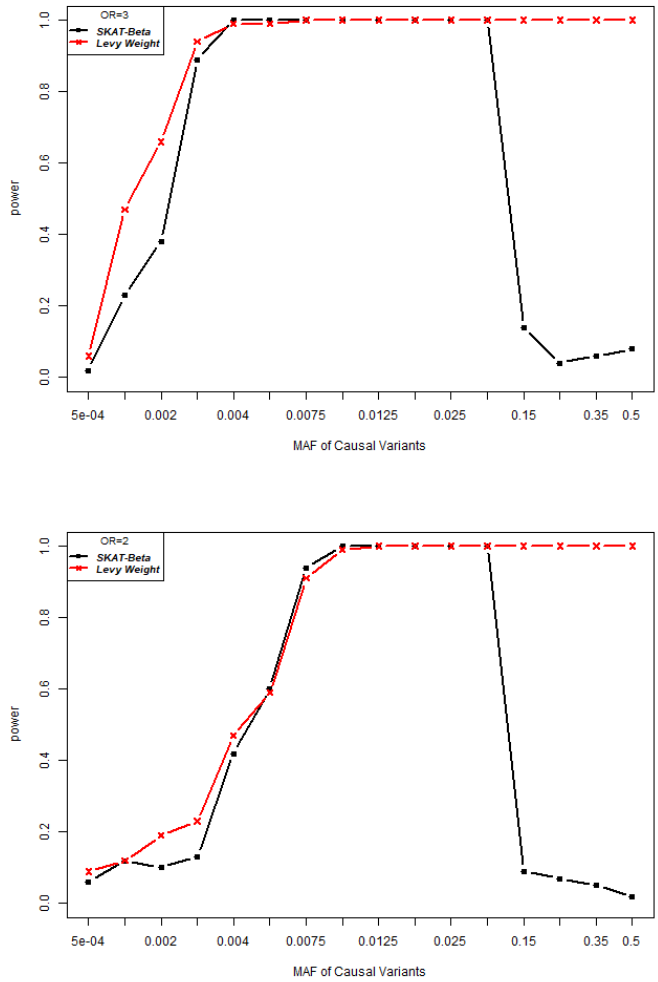


Figure 6.24: This Figure presents a comparison of the Levy function and beta-SKAT in terms of the MAF of causal variants. There are 10 causal variants among 100 variants. The OR of the causal variants is fixed at 3 in the top Figure and OR= 2 in the bottom Figure; the causal variants range between extremely rare and common. The non-causal variants are extremely rare, moderately rare, and common with percentages of 60%, 30%, and 10%, respectively.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005]	✓60%
MRV	OR=3 [0.005-0.05]	✓30%
CV	OR=3 [0.05-0.5]	✓10%

	Causal	Non-Causal
ERV	OR=2 [0.0005-0.005]	✓60%
MRV	OR=2 [0.005-0.05]	✓30%
CV	OR=2 [0.05-0.5]	✓10%

The beta weight scheme performs slightly better than the Levy weight under very small effect sizes and when the causal variants happen to be between 0.01 and 0.04, but the SKAT weight scheme is worse in common regions when effect sizes are smaller: ($OR = 1.5$) and ($OR = 1.3$) in the top and bottom of Figure (6.25), respectively.

Finally, the Levy weight function performs better in the area of common variants greater than MAF= 0.05. Hence, we can say that this Levy function weighting scheme can detect association signals in all MAF regions; see Figure (6.25).

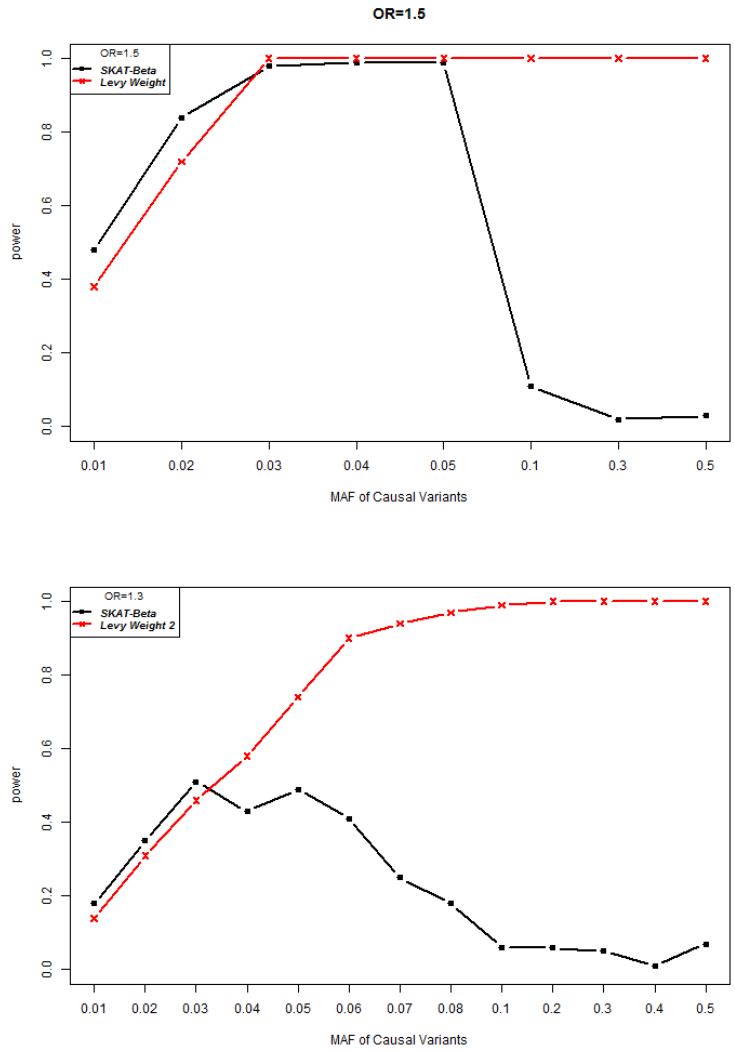


Figure 6.25: In this Figure, we show a comparison of the Levy function and beta-SKAT in terms of the MAF of causal variants. There are 10 causal variants among 100 variants. The OR of causal variants is fixed at 1.5 for the top Figure and 1.3 for the bottom one, and the causal variants range between largely moderately rare to common $(0.01 - 0.5)$. The non-causal variants are extremely rare, moderately rare, and common with percentages of 60%, 30%, and 10%, respectively.

	Causal	Non-Causal
ERV	.	✓60%
MRV	OR=1.5 [0.01-0.05]	✓30%
CV	OR=1.5 [0.05-0.5]	✓10%

	Causal	Non-Causal
ERV	.	✓60%
MRV	OR=1.3 [0.01-0.05]	✓30%
CV	OR=1.3 [0.05-0.5]	✓10%

6.5 Beta Weight Scheme

In the previous weighting scheme, we saw that there is a drawback when the data has a large percentage of moderately rare variants. The Cauchy and beta functions(with parameters 1 and 25) have a low rate of association signal detection when the causal variants are extremely rare and when the data contains a large number of moderately rare variants.

If we review the Cauchy and beta weights, we see that these weighting schemes assign large weights to the moderately rare variants, so they detect any association in moderately rare regions with very small effect sizes. However, they dominate the signal of extremely rare variants and require more attention because it is difficult to detect the signal even with large effect sizes. Thus, by assigning large weights to the moderately rare variants, they dominate the signal in the extremely rare range, especially if the data contains a large amount of moderately rare variants. The moderately rare variants with a high probability of detecting association signals with a reasonable effect size need sufficient weights but not large enough that they will affect the association of extremely rare variants (ERVs).

As stated previously, if the number of moderately rare variants is low in comparison to ERVs, the Cauchy and Levy functions effectively detect association signals in the extremely rare region. However, when the data contains a large number of moderately rare variants, the distribution of MAFs ($\mathcal{F}$) can be uniform. In this scenario, we must use another weighting scheme that can address the domination of the signal in the ERV region if the causal variant happens to be there. This will also adequately address the detection of the association signals in the moderately rare and common variant regions.

We propose two weighting scheme functions that perform better in the above scenario: (1) beta with 0.5 and 1 and (2) the Burr function introduced in the next section. The weighting scheme that we propose performs very well in the above situation by extending the difference between the largest value (associated with the smallest MAF, $\mathcal{F}$) and the moderately rare variants, as shown in Figure 6.26. Therefore, when the data are distributed equally across all three MAF ranges (i.e., extremely rare, moderately rare, and common) or the proportion of

extremely rare variants is low compared to that of the moderately rare variants, we use the beta function proposed in this section.

The function of beta with parameter 0.5 and 1;

$$g(\mathcal{F}) = \mathcal{F}^{(0.5-1)} \times (1-\mathcal{F})^{(1-1)} \times 1 \tag{6.8}$$

We can re-write 6.8 as;

$$g(\mathcal{F}) = \mathcal{F})^{(-0.5)} \tag{6.9}$$

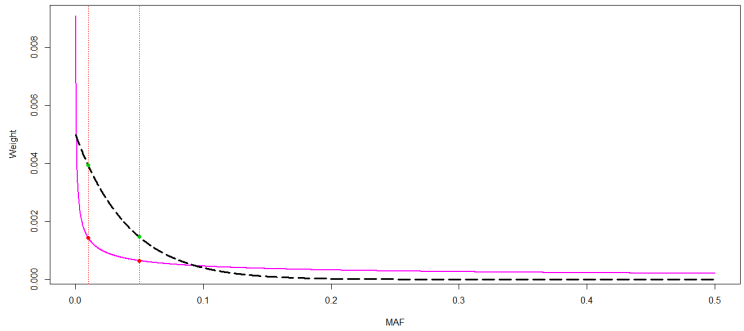


Figure 6.26: The beta function with parameter 0.5 and 1 (red line) versus 1 and 25 (dashed black line). The red dashed horizontal lines are the MAF at 0.01 and 0.05, respectively. The dots indicate the weight at MAFs of 0.01 and 0.05 for both weights.

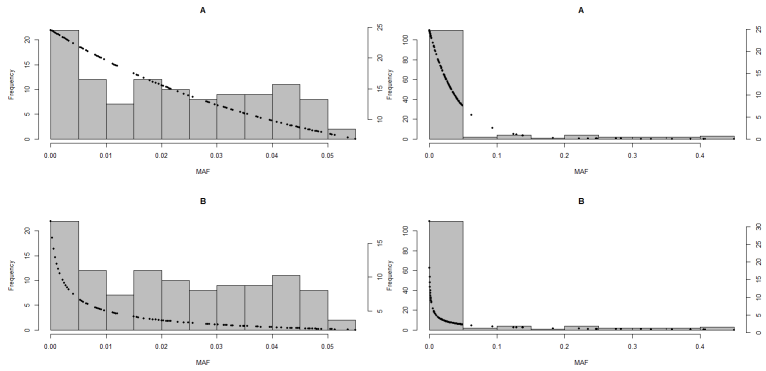


Figure 6.27: Comparison of beta-SKAT and $(\mathcal{F})^{(-0.5)}$ weighting schemes.

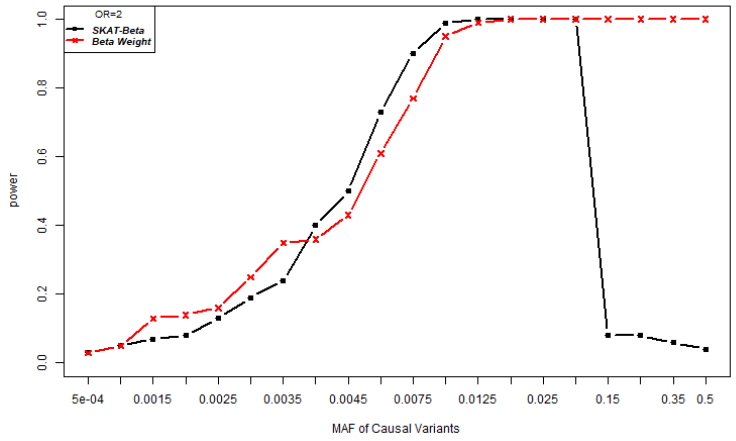


Figure 6.28: Comparison of beta-SKAT and $(\mathcal{F})^{(-0.5)}$ in all MAF regions. There are 10% causal variants with an OR fixed at 2 and 200 extremely rare, moderately rare, and common variants with percentages of 50%, 40%, and 10%, respectively.

	Causal	Non-Causal
ERV	OR=2 [0.0005-0.005]	✓50%
MRV	OR=2 [0.005-0.05]	✓40%
CV	OR=2 [0.05-0.5]	✓10%

In cases where the causal variants are rare, when the number of causal ERVs increases, the power of the proposed beta also increases.

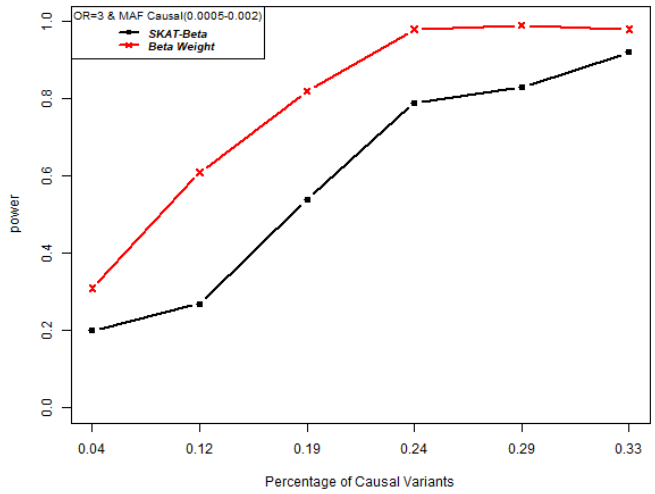


Figure 6.29: Comparison of beta-SKAT and $(\mathcal{F})^{(-0.5)}$ where the causal variants are in the extremely rare variant region. There are 10% causal variants with OR fixed at 3 and MAF fixed at $0.0005 - 0.001$. There are 100 extremely rare, moderately rare, and common variants with percentages of 50%, 40%, and 10%, respectively. The causal variants are increased from 5 to 60 (i.e., $(4\% - 33\%)$).

	Causal	Non-Causal
ERV	OR=2 [0.0005-0.002] Increased from 4% to 33%	✓50%
MRV	.	✓40%
CV	.	✓10%

6.30 shows the weight impact under small effect size $OR = 1.5$ when the causal is moderately rare.

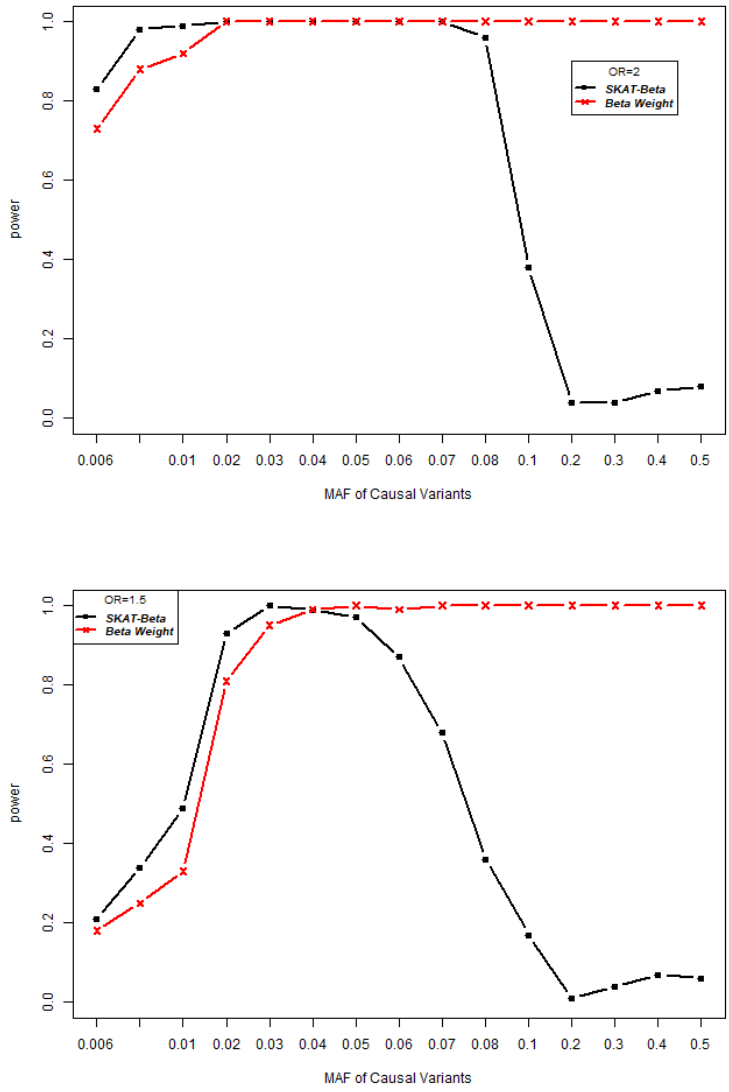


Figure 6.30: Comparison of beta-SKAT and $(\mathcal{F})^{(-0.5)}$ when the causal variants are in the moderately rare variant region. There are 10% causal variants with OR fixed at 2 in the top Figure and OR = 1.5 in the bottom one, and the MAF varies between 0.006 and 0.5. There are 100 extremely rare, moderately rare, and common variants with percentages of 50%, 25%, and 25%, respectively.

	Causal	Non-Causal
ERV	.	✓50%
MRV	OR=2 [0.006-0.05]	✓25%
CV	OR=2 [0.05-0.5]	✓25%

	Causal	Non-Causal
ERV	.	✓50%
MRV	OR=1.5 [0.006-0.05]	✓25%
CV	OR=1.5 [0.05-0.5]	✓25%

When the data are distributed uniformly, as in, for example, Figure 6.31, and the causal variants are in the extreme region, this weighting scheme (6.9) outperforms other weights. Figure 6.32 shows the power of the test using this weight when the causal variant happened to be within the boundaries of MAF $(0.0005 - 0.002)$.

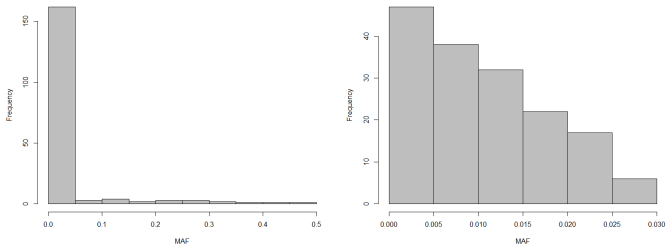


Figure 6.31: This an example of the MAF distribution ($\mathcal{F}$); there are many variants distributed along the MAF scale.

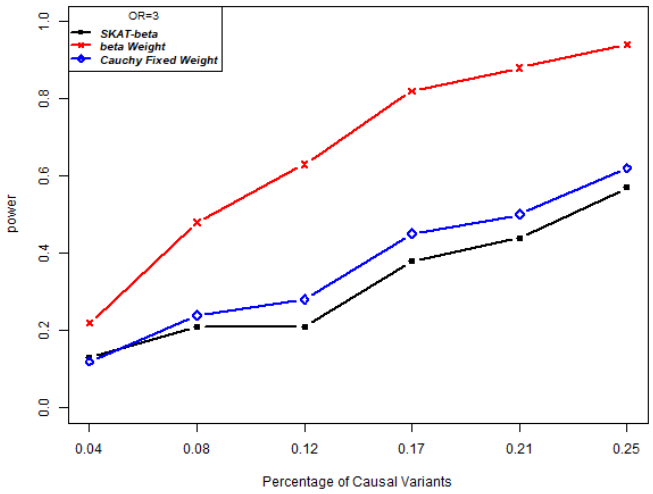


Figure 6.32: This Figure presents a comparison of our proposed beta-function, beta-SKAT, and a Cauchy fixed weight. There are 100 variants in total. The OR of causal variants are fixed at 3, and the MAF of causal variants are fixed at $0.0005 - 0.002$. The non-causal variants are extremely rare, moderately rare, and common with percentages of 33%, 33%, and 33%, respectively.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.002] Increase from 4% to 25%	✓33%
MRV	.	✓33%
CV	.	✓33%

6.6 Burr Weight Scheme

In this section, we introduce the Burr weight scheme. The one limitation of the beta function, introduced by Wu *et al.* (2011), is that it gives moderately rare variants larger weights, which affects the signal of association in ERVs and reduces the power of detection. The Burr function is based on up-weighting rare variants so that the smallest MAFs ($\mathcal{F}$) are given the largest weights, while the moderately rare variants are assigned weights that can detect the signal of association and still keep an adequate non-zero in the common ranges.

$$g(\mathcal{F}) = \frac{0.5\mathcal{F}^{0.5}}{\mathcal{F}[1+\mathcal{F}^{0.5}]^2} \tag{6.10}$$

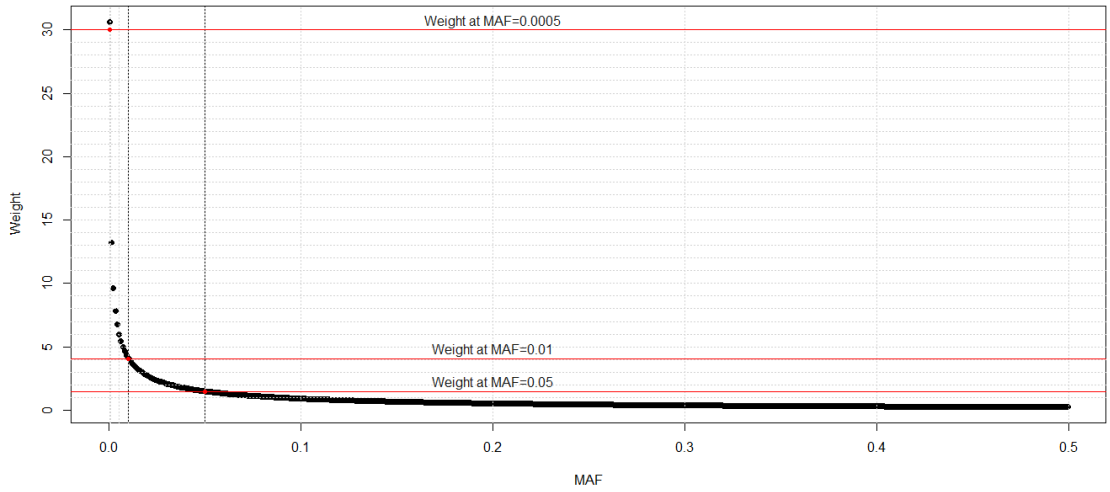


Figure 6.33: The Burr weight versus the MAF.

The weight given to the moderately rare variants' range is not very large compared to the weight assigned to the same area in the beta function by Wu *et al.* (2011), as we can see in Figure 6.34. Therefore, the weight given to ERVs is larger than the weight of the moderately rare variants, so there is a large difference between them (i.e., a small ratio).

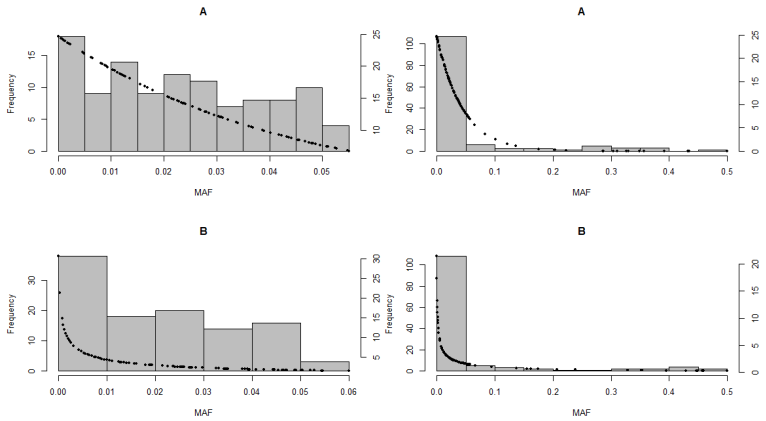


Figure 6.34: Beta with 1,25 and the Burr function applied on data have variants in most of the MAF range. We reduce the impact of moderately rare variants on the signal of extremely rare variants, so the signal of moderately rare variants can be detected.

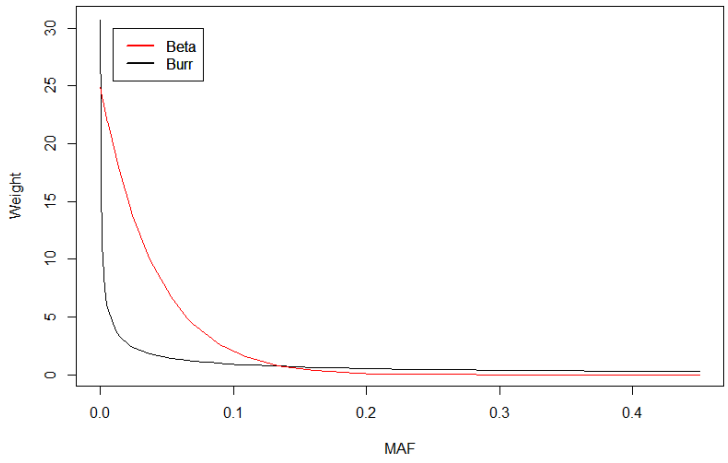


Figure 6.35: The behaviour of the beta function and Burr weight among MAFs.

The impact of having moderately rare variants can be seen in the plot of vector **U** versus MAFs with different weights using the SKAT and Burr functions.

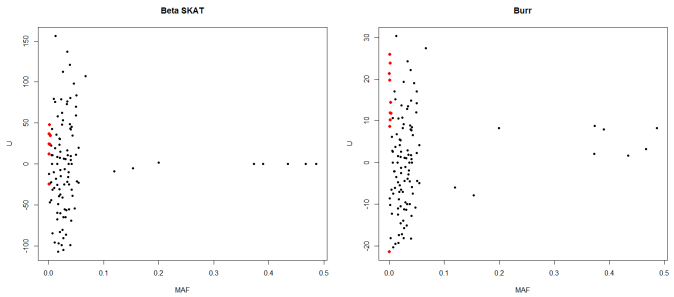


Figure 6.36: Comparison of **U** values on the score test using two weighting schemes: beta-SKAT and Burr. The causals are indicated by red dots, while non-causal variants are represented by black dots.

The Burr weight scheme (6.10) outperforms the beta function because the moderately rare variants are assigned sufficient but not outsized weights; see Figure (6.37).

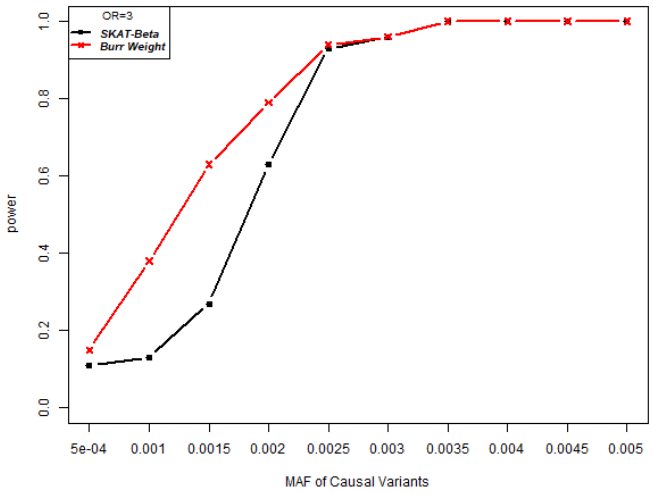


Figure 6.37: The power of detecting association in the extremely rare variants range. The analysis is conducted using data with 80% rare variants (most of them extremely rare) and 20% common variants.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005]	✓50%
MRV	.	✓30%
CV	.	✓20%

The Burr weight performs well in the moderately rare and common variant range; see Figure (6.38).

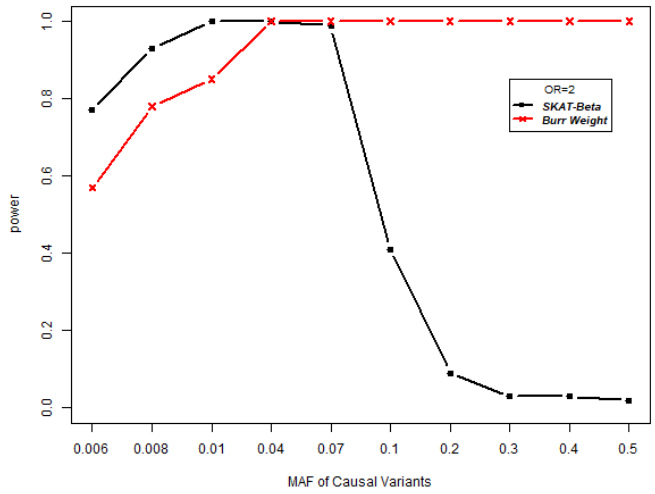


Figure 6.38: The power of detecting the association in the moderately rare and common variant range. The analysis is conducted using data with 80% rare variants (most of them extremely rare) and 20% common variants.

	Causal	Non-Causal
ERV	.	✓50%
MRV	OR=2 [0.005-0.05]	✓30%
CV	OR=2 [0.05-0.5]	✓20%

However, when the effect size is very small ($OR = 1.3$), the beta function is a more effective choice with MAFs of 1% and 5%; see Figure (6.39).

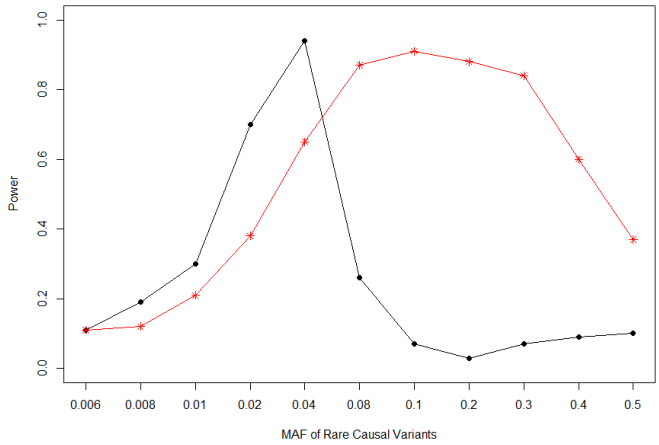


Figure 6.39: The power of detecting the association in the moderately rare and common variant range. The analysis is conducted using data with 80% rare variants (most of them extremely rare) and 20% common variants. The black line represents SKAT, and the red line with stars is the Burr weight.

	Causal	Non-Causal
ERV	.	✓50%
MRV	OR=1.5 (0.005-0.05)	✓30%
CV	OR=1.5 (0.05-0.5)	✓20%

When the number of ERVs increases, the test power also increases, so tests using this weighting scheme still outperform tests using the beta function; see Figure (6.40).

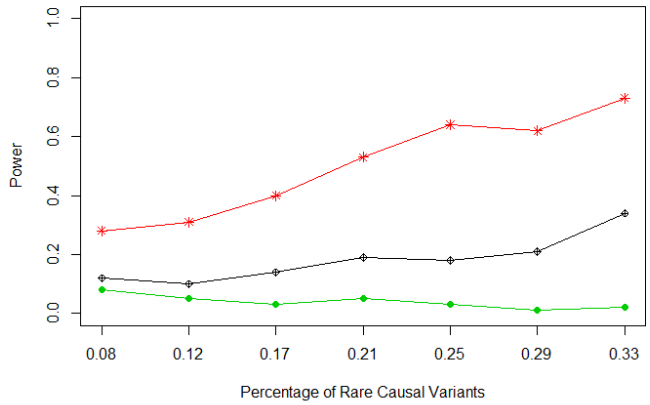


Figure 6.40: The power of detecting the association in the extremely rare variant region with MAF $(0.0005 - 0.002)$ and OR fixed at 3. The analysis is conducted using data with 80% rare variants (50% of them extremely rare) and 20% common variants.

The next two Figures, 6.41, and 6.42, illustrate improvements in the test power when the data is distributed across the range of MAFs rather than clustered in the extremely rare range. It is the reason for considering this kind of weighting scheme.

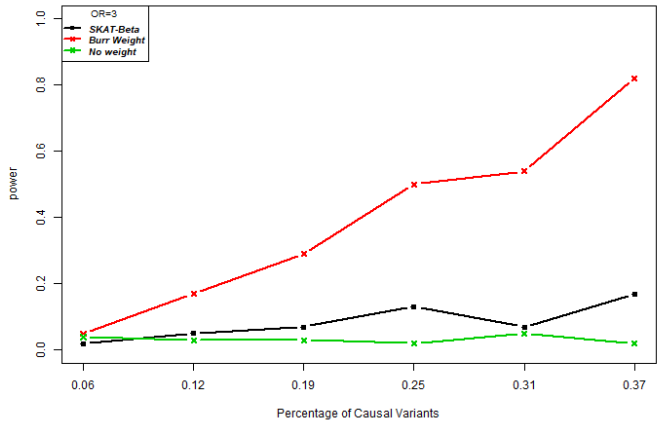


Figure 6.41: When the data is distributed as shown in Figure 6.31 (i.e., 40% ERV, 40% MRV, and CV 20%), and the causal variants have MAFs fixed at $0.0005 - 0.001$ with OR=3, the X-axis is the percentage of these kinds of variants in the data. The red line is the test using the Burr weight, the black one is the SKAT weight, and the green line is with no weight.

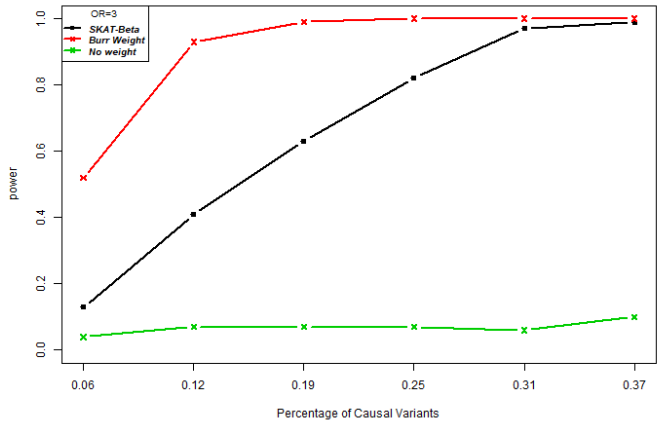


Figure 6.42: When the data is distributed as depicted in Figure 6.31 (i.e., 40% ERV, 40% MRV and CV 20%), and the causal variants have MAF $0.001 - 0.004$ with OR=3, the X-axis is the percentage of these kinds of variants in the data. The red line is the test using Burr weight, the black one is the SKAT weight, and the green line is no weight.

6.7 Type I Errors

To evaluate type I error rates for score tests using these weights, we conduct different simulations with varying settings. The first one uses data that contains differing amounts of rare variants: 30%, 50%, and 80%. The second scenario involves fixing the MAF for a specific value and conducting an evaluation of extremely rare to common MAFs. To evaluate the type I error rate for the proposed score test with all weight schemes included in this chapter, we conduct simulations under the null model $(\text{logit}P(y_i = 1) = \beta_0)$ using various settings. First, we use data that has different proportions of rare variants: 30%, 50%, and 80%. Second, we fix the MAFs for specific values and evaluate each one, beginning with extremely rare MAFs and progressing to the most common ones. We use 1000 simulated data to evaluate the type I error rate, and the results based on significance level $\alpha = 0.05$ are shown in Tables (6.2) and (6.3), which show the tests using the weighting schemes had satisfactory Type I error rates except when the all the variants have very low minor allele frequencies. Controlling type I errors is a concern due to rarity; see Table 6.3.

Weight Function	MAF		
	30%	50%	80%
Cauchy Fixed 1	0.05	0.05	0.035
Cauchy Fixed 2	0.05	0.055	0.05
Cauchy adap. 1	0.055	0.045	0.05
Cauchy adap. 2	0.04	0.05	0.04
Levy	0.05	0.05	0.035
Beta (0.5,1)	0.045	0.05	0.045
Burr	0.05	0.06	0.05

Table 6.2: Type I error results of the first scenario. We change the percentage of rare data with MAF $[0.0005 - 0.002]$ from 30% to 80%

Weight Function	MAF								
	0.0005	0.005	0.01	0.05	0.1	0.2	0.3	0.4	0.5
Cauchy Fixed 1	0.015	0.04	0.05	0.05	0.05	0.06	0.05	0.05	0.05
Cauchy Fixed 2	0.01	0.035	0.04	0.045	0.05	0.05	0.05	0.05	0.06
Cauchy adap. 1	0.02	0.04	0.045	0.06	0.05	0.05	0.06	0.05	0.05
Cauchy adap. 2	0.015	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.045
Levy	0.02	0.04	0.05	0.05	0.045	0.05	0.05	0.05	0.05
Beta (0.5,1)	0.02	0.035	0.045	0.05	0.05	0.05	0.05	0.045	0.05
Burr	0.025	0.045	0.05	0.05	0.06	0.045	0.05	0.05	0.05

Table 6.3: Results of the second scenario of Type I errors. The MAF fixed at all the SNPs in the sample are from 0.0005 to 0.5.

Part II

Combined Weighting Schemes

6.8 Combined Effects of Two Different Weighting Schemes

Both rare and common genetic variants have been discovered due to recent progress in sequencing technologies. Genome-wide association studies (GWASs) can be used to test for the effect of common variants, while sequence-based association studies can test the cumulative effect of both rare and common variants on disease risk. Many group-wise association tests, including burden tests and variance-component tests, have been proposed for this purpose. Although such tests do not exclude common variants from their evaluation, they focus mostly on testing the effect of rare variants by up-weighting rare-variant effects and down-weighting common-variant effects and can, therefore, lose serious power when both rare and common variants in a region affect trait susceptibility. There is some evidence that the allelic range of risk variants at a given locus might include rare, novel, low-frequency, and common variants.

In this chapter, we proposed different weighting schemes that can include most of the MAF range. However, we can improve the weighting schemes by combining two of them. Here, we introduce a variance component score test to evaluate the cumulative effect of two different weights that can be effective on rare and common variants. The proposed tests are computationally efficient. We evaluate these tests on data simulated under comprehensive scenarios and show that when compared to different weights, they can achieve substantial increases in power.

6.8.1 Description of the Combination Method

To test for the joint effect of two weights for variants in a genetic region, we combine score test statistics as a weighted sum. Recall model 8.1; then, as we defined in Chapter (4), the score statistic will be

$$S(\boldsymbol{\gamma}) = \mathbf{U}^T \Gamma \Gamma \mathbf{U} \tag{6.11}$$

Next, we define the combined test statistics as

$$T(\boldsymbol{\gamma}) = \phi S_1(\boldsymbol{\gamma}) + (1-\phi) S_2(\boldsymbol{\gamma}) \tag{6.12}$$

where $S_1(\boldsymbol{\gamma})$ is the score test with the first variant's weight, and $S_2(\boldsymbol{\gamma})$ is the score test with the second variant's weight.

Because both $S_1(\boldsymbol{\gamma})$ and $S_2(\boldsymbol{\gamma})$ follow a mixture of chi-square distributions, the distribution of $T(\boldsymbol{\gamma})$ will be

$$T(\boldsymbol{\gamma}) \sim \sum_{j=1}^{p} \lambda_j \chi_1^2 \tag{6.13}$$

where the λ eigenvalues of $\Gamma_1 X^T D X \Gamma_1 + \Gamma_2 X^T D X \Gamma_2$ according to the theorems 4.5.1 and 4.5.3.

Since the $S_1(\boldsymbol{\gamma})$ and $S_2(\boldsymbol{\gamma})$ share the same matrix X, the combined test statistics given in equation 6.12 are equivalent to score test $S(\boldsymbol{\gamma})$ with the combined weight as

$$w^* = \phi w_1 + (1 - \phi) w_2 \tag{6.14}$$

We select ϕ such that the two weights contribute equally to the test statistic, so we choose $\phi = 0.5$.

A combination of two variant weights will have the advantages of these two weights. If the first weight performs better in the extremely rare variant region than the moderately rare variant region, it is helpful to combine it with another variant weight scheme that performs better in the moderate region. We illustrate the performance of this type of combination in the next two Figures. We combine Cauchy (fixed parameters) and beta weight schemes, which combines the benefits of two weights, taking advantage of the Cauchy weight's performance in the extremely rare variant region and the beta weight's performance in the moderately rare variant region; see Figures (6.43, 6.44)

6.9 Power

1. **Cauchy and beta-SKAT**

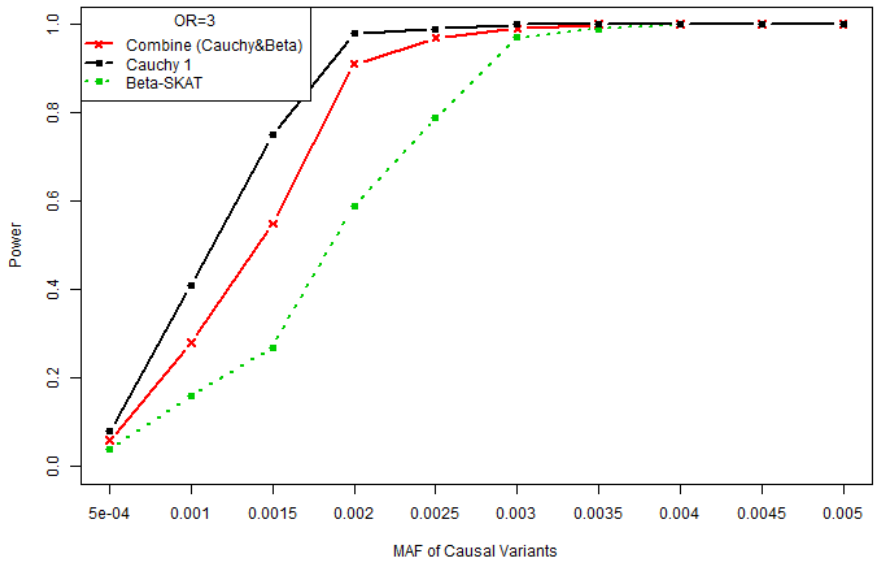


Figure 6.43: We generate data with extremely and moderately rare variants and common variants (i.e., 40% ERV, 40% MRV, and CV 20%) where the number of variants was fixed at 200; 7% of the variants are causal with $OR = 3$. We combine beta-SKAT and Cauchy weights. The MAF of causal variants is between 0.0005 and 0.005.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005]	✓40%
MRV	.	✓40%
CV	.	✓20%

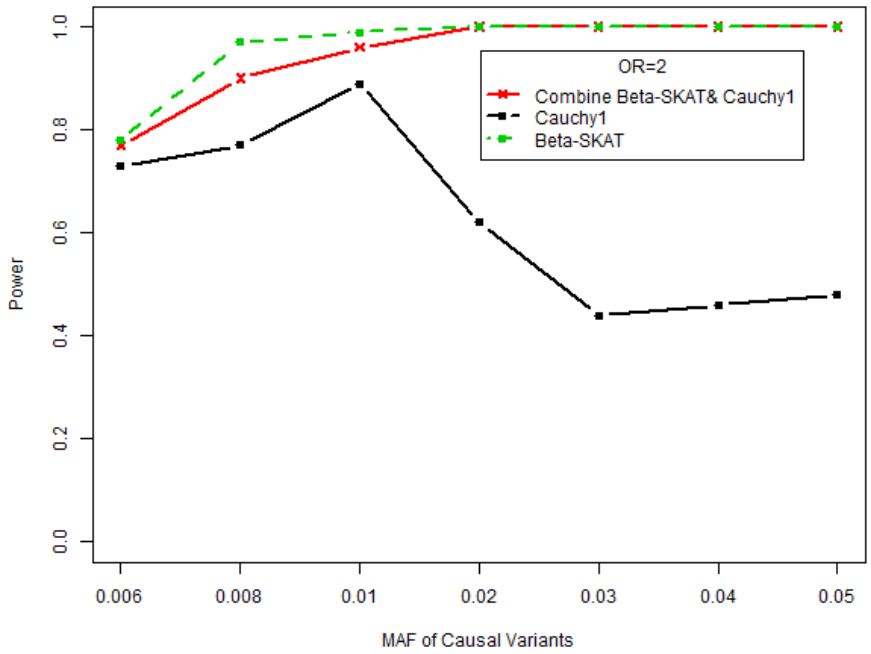


Figure 6.44: We generate data with extremely rare, moderately rare, and common variants (i.e., 40% ERV, 40% MRV, and CV 20%) where the number of variants is fixed at 200; 7% of the variants are causal with $OR = 2$. We combine beta-SKAT and Cauchy weights. The MAF of causal variants is between 0.006 and 0.05.

	Causal	Non-Causal
ERV	.	✓40%
MRV	OR=2 [0.006-0.05]	✓40%
CV	.	✓20%

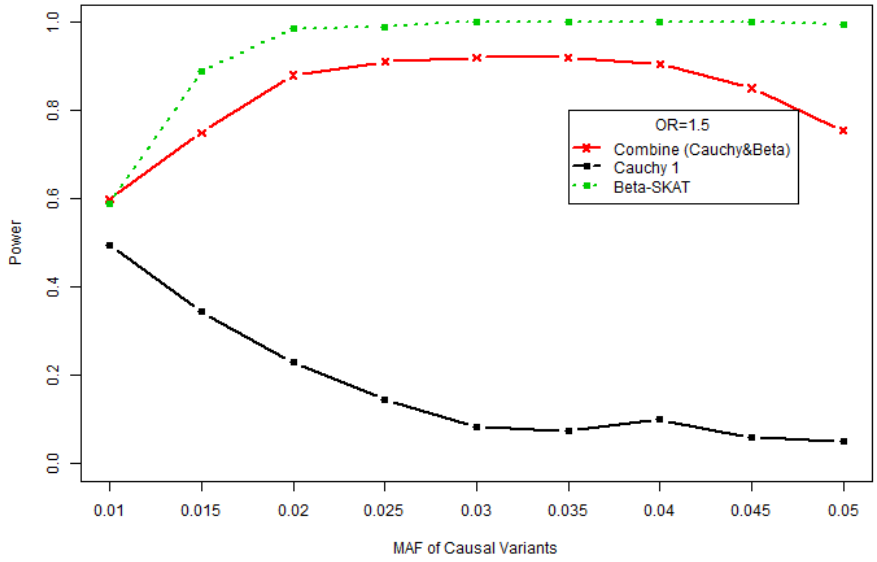


Figure 6.45: We generate data with extremely rare, moderately rare, and common variants (i.e., 40% ERV, 40% MRV, and CV 20%) where the number of variants is fixed at 200, and 7% are causal variants with $OR = 1.5$. We can see the combined weight schemes outperform the Cauchy weight, and the beta weight outperforms the combined weight scheme and Cauchy. The MAF of causal variants is between 0.01 and 0.05.

	Causal	Non-Causal
ERV	.	✓40%
MRV	OR=1.5 [0.01-0.05]	✓40%
CV	.	✓20%

2. **Cauchy and Burr**

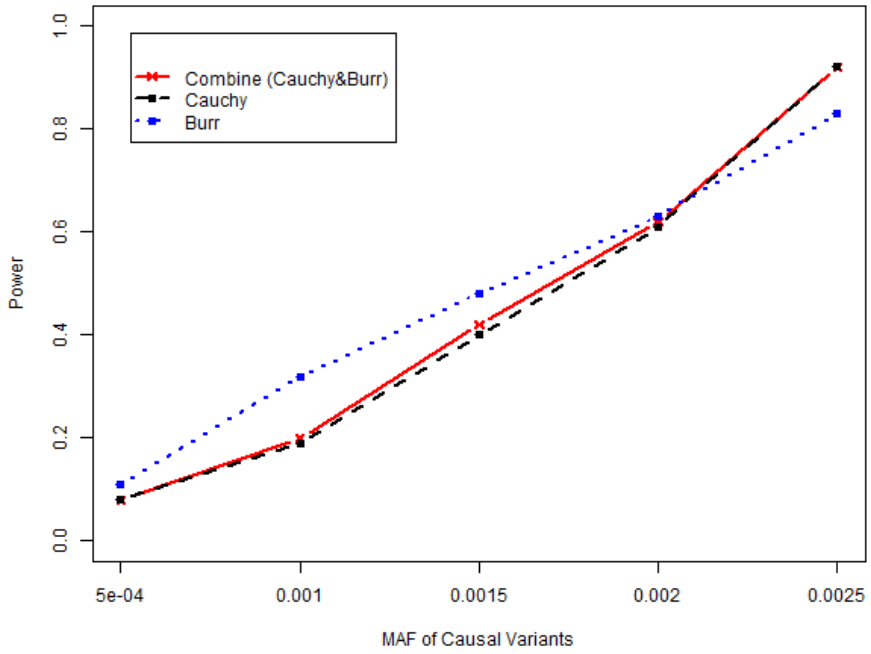


Figure 6.46: We generate data with extremely rare, moderately rare, and common variants (i.e., 40% ERV, 40% MRV, and CV 20%) where the number of variants is fixed at 200, and 7% of the variants are causal with $OR = 3$. We combine Burr and Cauchy weight schemes. The MAF of causal variants is between 0.0005 and 0.003.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.0025]	✓40%
MRV	.	✓40%
CV	.	✓20%

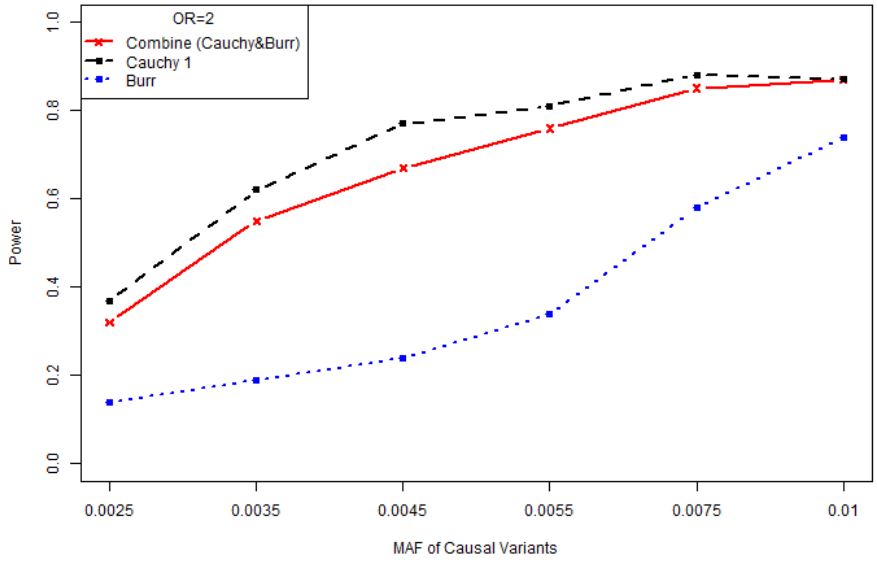


Figure 6.47: We generate data with extremely rare, moderately rare, and common variants (i.e., 40% ERV, 40% MRV, and CV 20%) where the number of is variants fixed at 200, and 7% of the variants are causal with $OR = 2$. We combine Burr and Cauchy weight schemes. The MAF of causal variants is between 0.002 and 0.01.

	Causal	Non-Causal
ERV	.	✓40%
MRV	OR=2 [0.0025-0.01]	✓40%
CV	.	✓20%

3. **Cauchy Fixed 1,2**

   We introduce two functions of Cauchy that have fixed parameters (Cauchy fixed weight schemes). Cauchy 1 is suggested for rare variants, and Cauchy 2 is suggested for the whole MAF range, but it has low detection in the rare region, especially for extremely rare variants when compared to Cauchy 1. Combining them will reduce the limitations of both functions as shown in Figures (6.48, 6.49 and 6.50)

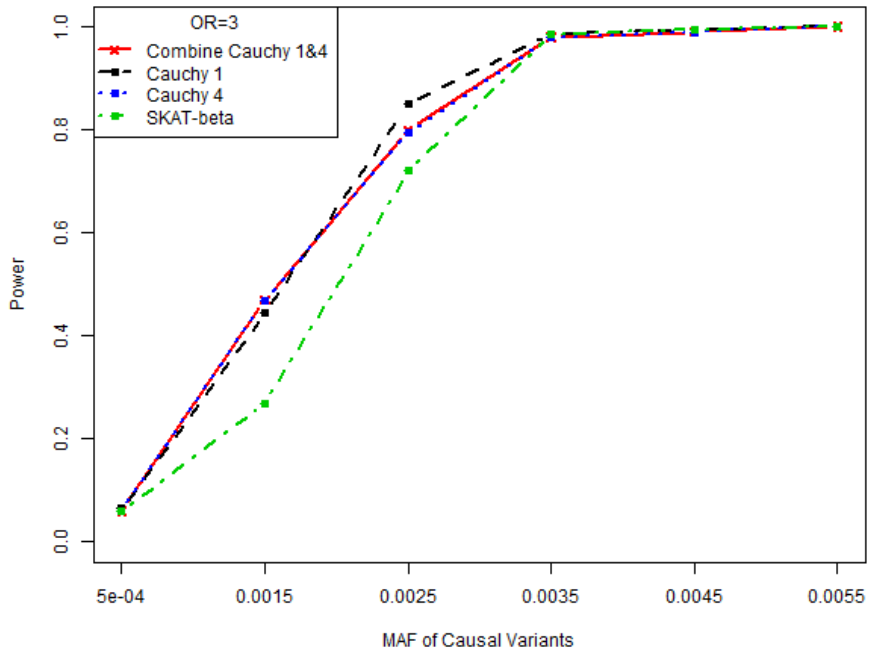


Figure 6.48: We generate data with extremely rare, moderately rare, and common variants (i.e., 40% ERV, 40% MRV, and CV 20%) where the number of variants is fixed at 200, and 7% of the variants are causal with $OR = 3$. We combine the Cauchy fixed 1 and 2 weight schemes. The MAF of causal variants is between 0.0005 and 0.005.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005]	✓40%
MRV	.	✓40%
CV	.	✓20%

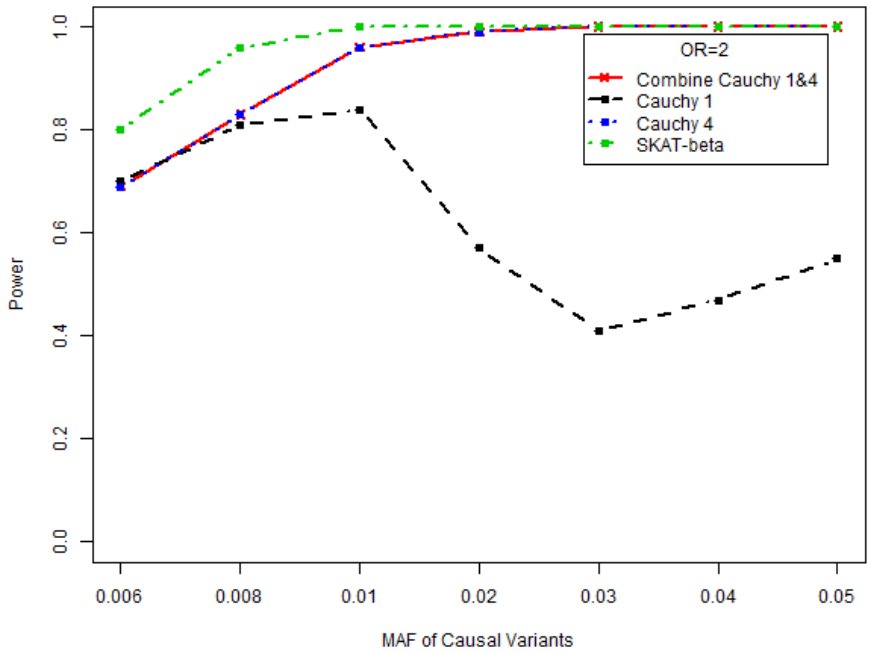


Figure 6.49: We generate data with extremely rare, moderately rare, and common variants (i.e., 40% ERV, 40% MRV, and CV 20%) where the number of variants is fixed at 200, and 7% of the variants are causal variants with $OR = 2$. We combine Cauchy fixed 1 and 2 weight schemes. The MAF of causal variants is between 0.005 and 0.05.

	Causal	Non-Causal
ERV	.	✓40%
MRV	OR=2 [0.006-0.05]	✓40%
CV	.	✓20%

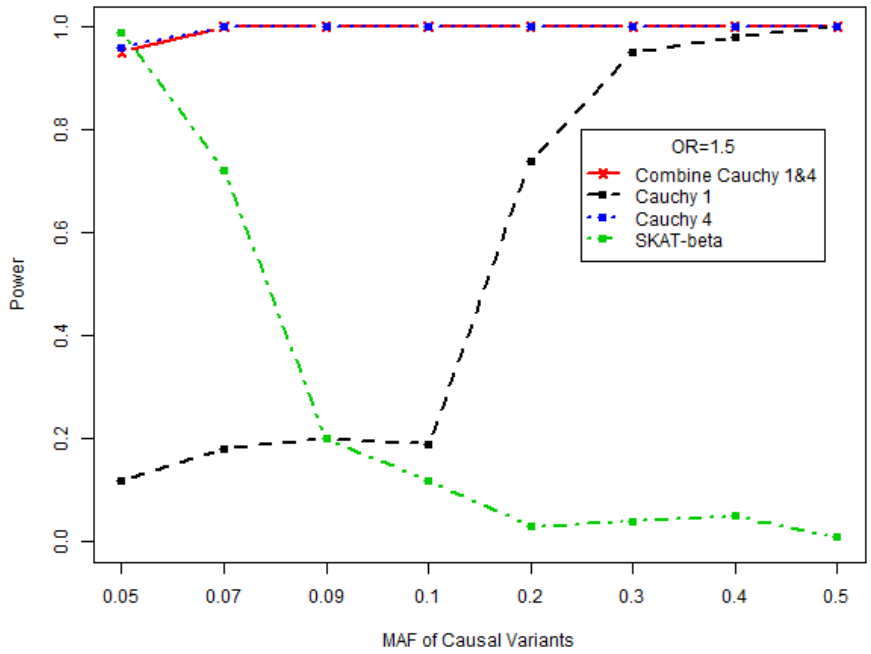


Figure 6.50: We generated data with extremely rare, moderately rare, and common variants (i.e., 40% ERV, 40% MRV, and CV 20%) where the number of variants is fixed at 200, and 7% of the variants are causal with $OR = 1.5$. We combine Cauchy fixed 1 and 2 weight schemes. The MAF of causal variants is between 0.05 and 0.5.

	Causal	Non-Causal
ERV	.	✓40%
MRV	.	✓40%
CV	OR=1.5 [0.05-0.5]	✓20%

6.9.1 Type I Errors

To evaluate the type I error rate for the proposed score test with all weight schemes included in this chapter, we conduct simulations under the null model $(\text{logit}P(y_i = 1) = \beta_0)$ using various settings. We use data that has different proportions of rare variants: 30%, 50%, and 80%. We conduct

1000 simulated data to evaluate the type I error rate, and the results, based on significance level $\alpha = 0.05$, are shown in Table (6.4)

	MAF [0.0005-0.002]		
Weights	30%	50%	80%
Cauchy and Beta	0.04	0.05	0.05
Cauchy and Burr	0.035	0.05	0.04
Cauchy Fixed 1 and 2	0.045	0.04	0.04

Table 6.4: Type I error results for the first scenario. We change the percentage of rare data with MAF $[0.0005 - 0.002]$ from 30% to 80%

6.10 Conclusion

Weighting schemes proposed in some studies only focus on rare variants. In this chapter, we have studied new weighting schemes that can be used for a continuous spectrum of MAFs. The chapter introduced two main points. It first introduced some functions that can be used for a continuous spectrum of MAFs. Then, it introduced the idea of combining two weighting schemes to combine the benefits and help avoid any weakness in these functions.

Chapter 7

Incorporating Information into the Variant Weight

7.1 Introduction

In the previous chapters, we introduced a variant weighting scheme. The main idea is to up-weight the rare variants while down-weighting the common variants. Therefore, we also proposed a new variants weighting scheme which performs this task to a degree that allows detecting the association of both rare and common variants -continues function-. There is a concern about genotyping errors in rare variant association due to low minor allele frequencies Daye *et al.* (2012). The accuracy of association study depends on the quality of variant calling. Suboptimal variant calling will affect the true association, and it will reduce the power to identify them. The threshold criteria to filter out low-quality variants is crucial because it depends on the choice of threshold, and every removed variant is a potentially missed causal variant. In this chapter, we will introduce a new weighting scheme to accommodate new information such as sequencing information on the variants level. This weight scheme has the same idea of up-weighting the rare variants and down-weighting the common ones; however, the weight will not only be a function of MAF but also a function of extra information. In previous chapters, we covered a weight based on MAF ($\mathcal{F}$) only $g(\mathcal{F})$. In this chapter, we will extend the weighting scheme to be a function of MAF ($\mathcal{F}$) and extra information

$\mathfrak{q}$, so instead of $g(\mathcal{F})$, we will extend it to be $g(\mathcal{F}, \mathfrak{q})$. We will consider two different functions. The first one can take any weighting scheme from the previous method and adjust the weight that is associated with any variants based on extra information, such as quality based on the variants level. The second one is a function that can up-weight the rare variants and down-weight the common one and incorporate the other information -quality- in the same function using the Burr function.

The first function would be in this form:

$$\omega = g_1(\mathcal{F}) \times g_2(\mathfrak{q}) \tag{7.1}$$

while the second one would be in this form:

$$\omega = g(\mathcal{F}, \mathfrak{q}) \tag{7.2}$$

where $\mathcal{F}$ is the MAF, and $\mathfrak{q}$ represents the extra information based on the variant level. We will introduce both functions in the following sections.

This is one example of sequencing information which is a Phred quality score; it is related to the error in the figure 7.1. Phred quality scores Q are related to the base-calling error probabilities P via

$$Q = -10logP$$

and

$$P = 10^{-Q10}$$

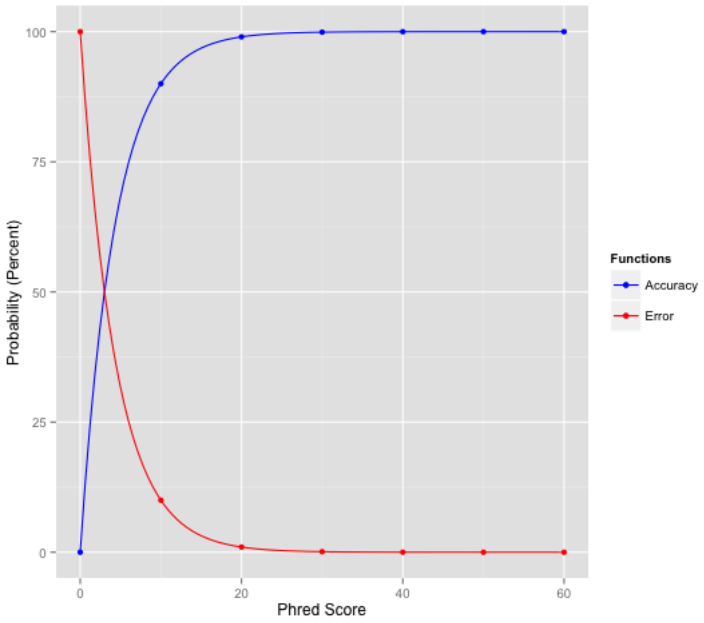


Figure 7.1: The Phred quality score versus the probability of error.

7.2 Simulation

We assume that there are n individuals and p genotypes. We simulate a genotype data X to mimic the allele frequencies in a different scenario dataset which is explained in detail in Chapter 5. Note that the number of variants used in the simulations are fixed at 200 SNPs, unless we specified another scenario in the caption. These SNPs are classified as 60% for ERV 30% MRV and 10% for CV. The variant quality scores are simulated based on sampling with replacement from real data (not filtered) by using the *vcfR* package in R program Knaus & Grünwald (2017).

We simulate causal variants and assume, in most scenarios, all the causal variants are risk variants (one direction). Therefore, we simulate causal with risk and protective variants. We assign an identical OR for each variant and set $OR = 1$ if we need to evaluate the type I error probability. The quality score simulation has three scenarios. First, we permit all positive Phred-scaled quality scores in the simulation. However, the causal variants' Phred-scaled quality score

will be restricted to values greater than 10 and refer to this filter as Q10. This scenario is equivalent to causal variants associated with correct variant calls with a probability of 90% or larger.

In the second scenario, variants are simulated in the usual manner such that quality scores are not restricted to any Phred-scaled score (i.e. the correct variant call probability can be any value—relaxed filtering of variants), and the quality scores of causal variants follow the same distribution as those non-causal variants.

In the third scenario, variants are filtered in the usual manner such that quality scores are restricted to Phred-scaled scores > 10 (correct variant call probability $> 90\%$), and the quality scores of causal variants follow the same distribution as those non-causal ones (i.e. we remove all the data that are less than a threshold [say Q10]). To estimate p-values, straight binomial proportions are used. Hence, they have the same standard error as any other binomial proportion $\sqrt(p(1-p)/n)$, where p here means the proportion of tests rejected and n the number of samples. Therefore, if $p = 0.05$ and $n = 2000$, the standard error of the observed proportion is about 0.005, and we could say the uncertainty 1%.

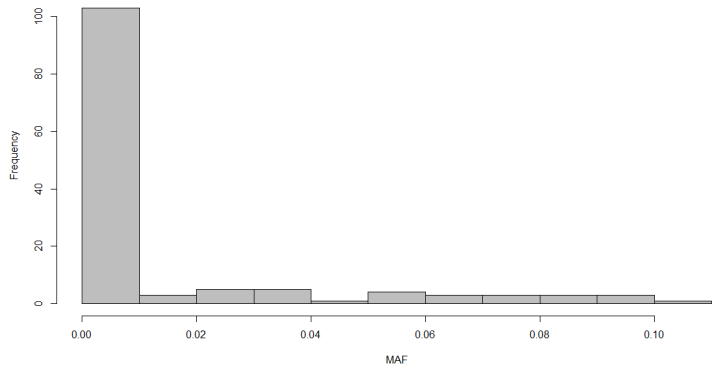


Figure 7.2: Example of the distribution of the observed MAFs that we use in the simulation.

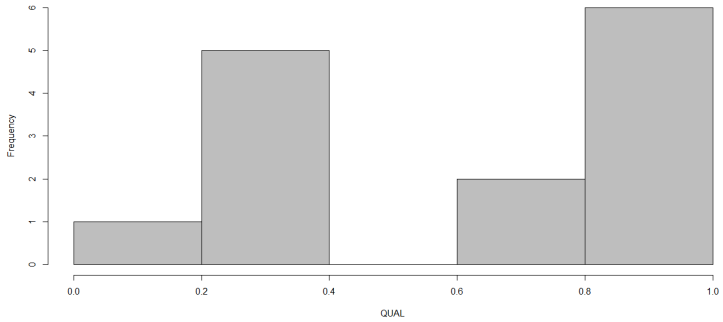


Figure 7.3: The distribution of the quality that is associated with the causal variants, which has the same distribution of the original Phred-scaled quality scores from a real VCF output call.

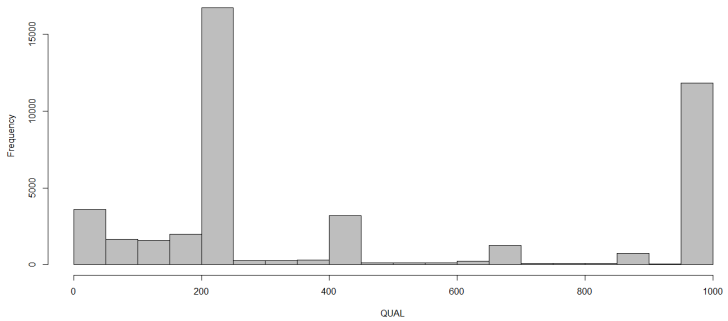


Figure 7.4: The Phred-scaled quality score from unfiltered data.

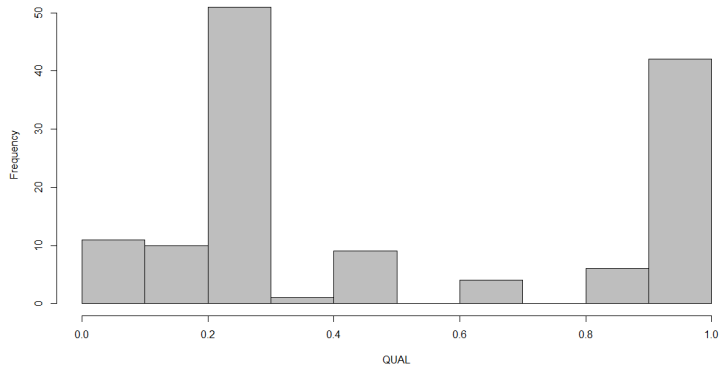


Figure 7.5: The sample phred-scaled quality score from the original unfiltered Phred-scaled quality score.

7.3 Weight Scheme I

As previously mentioned, we will incorporate new information such as sequencing information to take part in the weighting scheme which is based on the variant level. Since the previous weighting schemes are based only on MAF, it will not be helpful to decide or distinguish whether this site has an accurate variant allele. Combining and incorporating other information can help increase the probability that the site is correct to call variants or SNPs. This weighting scheme is designed especially for rare variants. We can incorporate extra information into the variant's weight using the following function. We will use the quality measure as the extra information.

$$\omega_1 = g_1(\mathcal{F}) \times g_2(\mathfrak{q}) \tag{7.3}$$

Then, we will assign $g_1(\mathcal{F})$ to be any of the variant functions as we described in previous chapters and $g_2(\mathfrak{q})$ to be the new information measure which ranges between $[0:1]$. Thus, the function will be

$$\omega_j = w_j(\mathcal{F}) * \mathfrak{q}_j^b \tag{7.4}$$

where $w_j(\mathcal{F})$ is the variant weight that can be used to up-weight rare variants and down-weight common ones, and $\mathfrak{q}_j$ is any extra normalized information normalized (i.e. its values range between 0 and 1). b is pre-specified.

$$\omega_j = \begin{cases} w_j(\mathcal{F}), & b = 0 \\ w_j(\mathcal{F}) * \mathfrak{q}_j^b, & \text{and} \quad b > 0 \end{cases}$$

where j is the index for variant j.

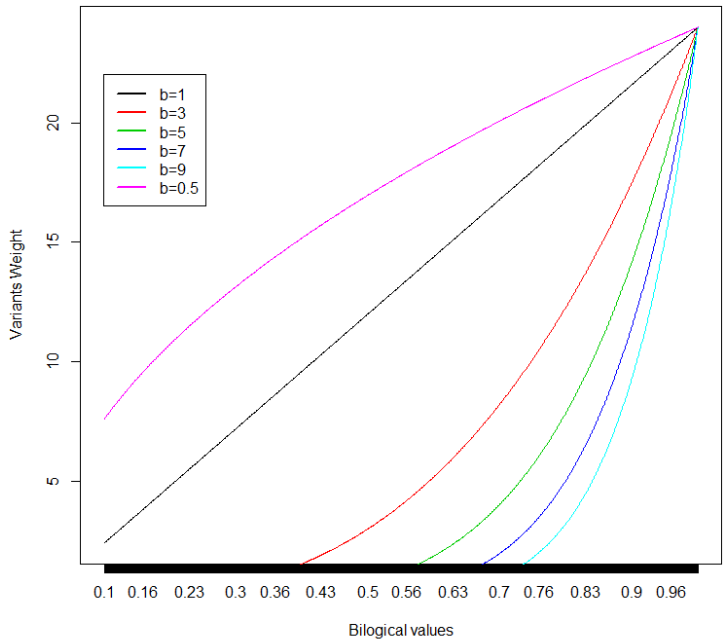


Figure 7.6: The weight can be adjusted using the b parameter, which will allow for acceptable sequencing information. For example, if $b = 1$, we will accept $\mathfrak{q} < 0.5$ to contribute large in the weight. Another way involves using $b = 7$ to down-weight any variants which have $\mathfrak{q} < 0.5$. In this figure, we chose $w = 24$, as we can see if $\mathfrak{q} = 1$, and then $\gamma = 24$

.

The following figures, 7.7 and 7.8, illustrate the impact of parameter b on the test. When b increases, we are increasing the threshold of accepted level from the quality. When the b increases, we down-weight more variants which are associated with low quality. At this figures, the simulation was conducted based

on the setting that the causal variants are extreme between $(0.0005, 0.001)$. Then, we fixed the quality for the figure 7.7 to be between $(0.7, 1)$, and for figure 7.8, the quality is fixed at a high level between $(0.9, 1)$.

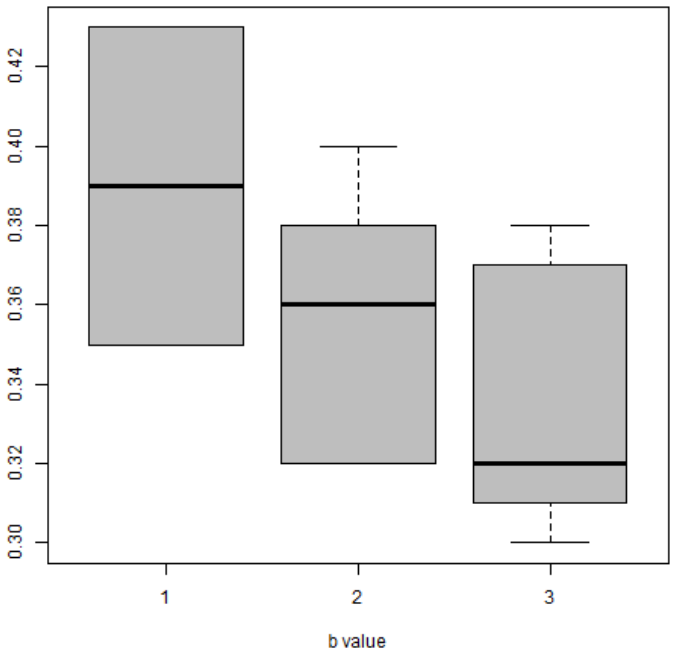


Figure 7.7: This figure shows the impact of b on the power. In this analysis, there are 100 variants, and 15% of them are causal. The causal variants are associated with high quality $(0.7-1)$. Note that the quality of some of the causal variants will be low so that as b increases, some of the causal variants which are associated with low quality will be down-weighted. The horizontal and vertical axes represents the b value and the power, respectively.

As we can see in Figure 7.8, since the quality of causal variants are large, the impact of b is low.

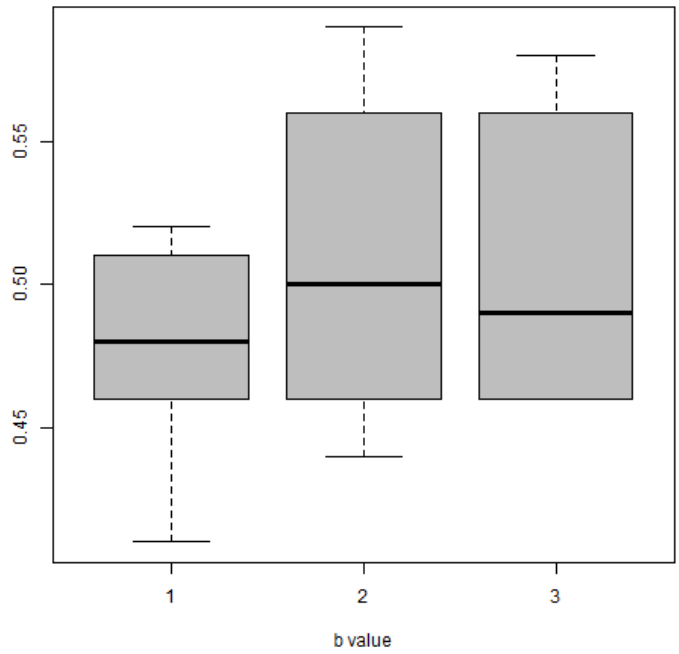


Figure 7.8: This figure shows the impact of b. In this analysis, there are 100 variants, and 15% of them are causal. The causal variants associated with high quality $(0.9 - 1)$.

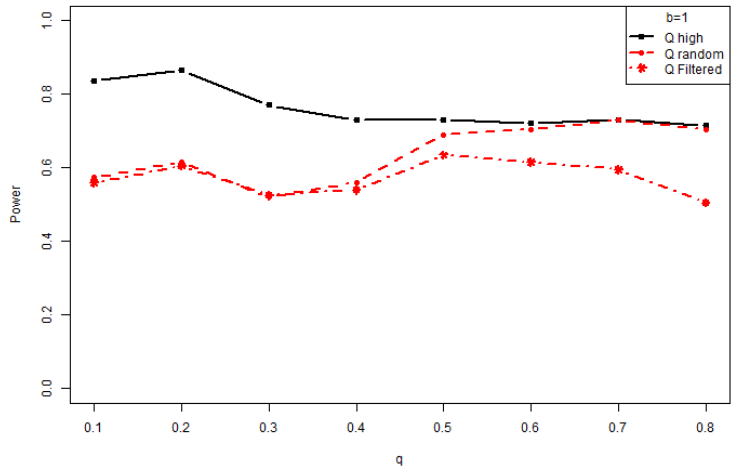
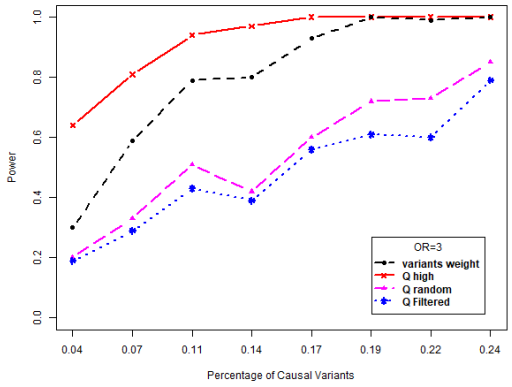


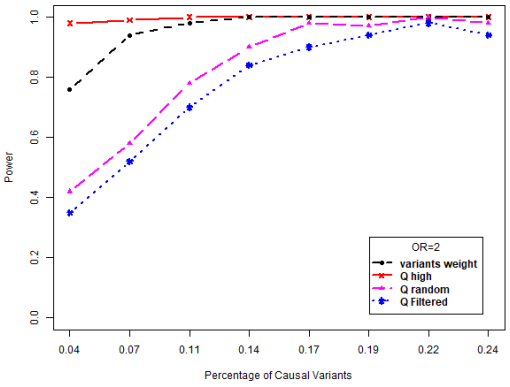
Figure 7.9: In this figure, we show the impact of quality on the power of the test. The first weight (Q high) considered 50% of causal variants and has high quality ($q > 0.6$), and (Q random) represents the quality without specifying any values of the causal variants that range $[0, 1]$; the last one was when we removed the SNPs that are associated with low quality. The horizontal axis reflects the minimum points of a uniform parameter from which we generate the quality. As we can see, when the value of $q = 1$, it means all the SNPs have high quality, and when $q = 0.1$, it means we are allowing low quality in the simulation. In this analysis, there are 200 non-causal SNPs (variants), 10% of which are causal.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005] 10%	✓60%
MRV	.	✓30%
CV	.	✓10%

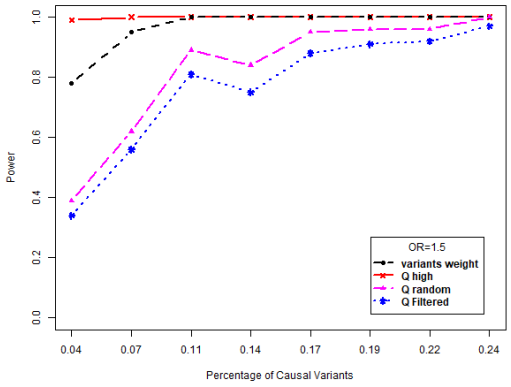
In Figures 7.10 and 7.11, we show the impact of incorporating quality with three scenarios that were illustrated in simulation section. We can see the variants with high quality will be equal or higher than variants' weight without including quality information. The variant weight that we use for comparison is *Burr function*, which was presented in chapter 6; however, we can choose any variant weight. In Figure 7.10, the causal variants are in the same direction and in different directions in Figure 7.11.



(a)



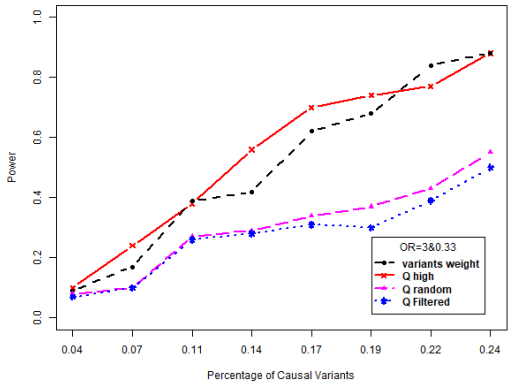
(b)



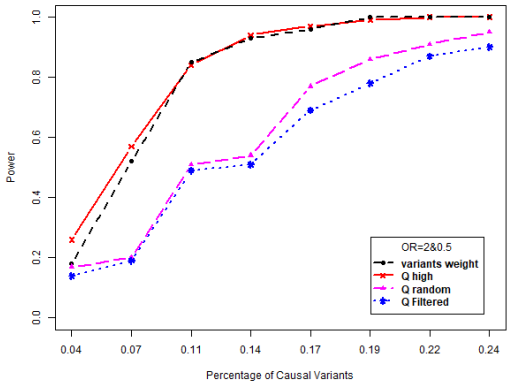
(c)

Figure 7.10: In this figure, we illustrate the impact of incorporating a quality score on the power of the test. We evaluate it in the three scenarios. There are 200 variants of the data including common and rare. We increase the percentage of causal variants from 4% to 24%. The causal variant in the first figure (a) is from the extremely rare range $(0.0005 - 0.005)$ with $OR = 3$; the second one (b)is from the moderately rare range $(0.005 - 0.05)$ with $OR = 2$; and the last figure (c) is from common variants ranging $(0.05 - 0.5)$ with $OR = 1.5$.

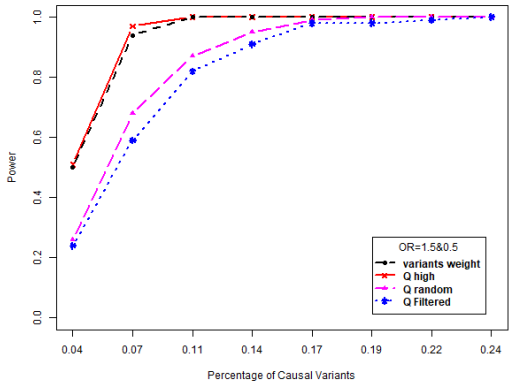
	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005] in Figure A	✓60%
MRV	OR=2 [0.005-0.05] in Figure B	✓30%
CV	OR=1.5 [0.05-0.5] in Figure C	✓10%



(a)



(b)



(c)

Figure 7.11: In this figure, we illustrate the impact of incorporating a quality score on the power of the test. We evaluate it in the three scenarios. There are 200 variants of the data including common and rare. We increase the percentage of causal variants from 4% to 24%. The causal variant in the first figure (a) is from the extremely rare range $(0.0005 - 0.005)$ with $OR = 3$ and $OR = 0.33$; the second one (b) is from the moderately rare range $(0.005 - 0.05)$ with $OR = 2$ and $OR = 0.5$; and the last figure (c) is from the common variants range $(0.05 - 0.5)$ with $OR = 1.5$ and $OR = 0.5$.

	Causal	Non-Causal
ERV	OR=3 and 0.33 [0.0005-0.005] in Figure A	✓60%
MRV	OR=2 and 0.5 [0.005-0.05] in Figure B	✓30%
CV	OR=1.5 and 0.5 [0.05-0.5] in Figure C	✓10%

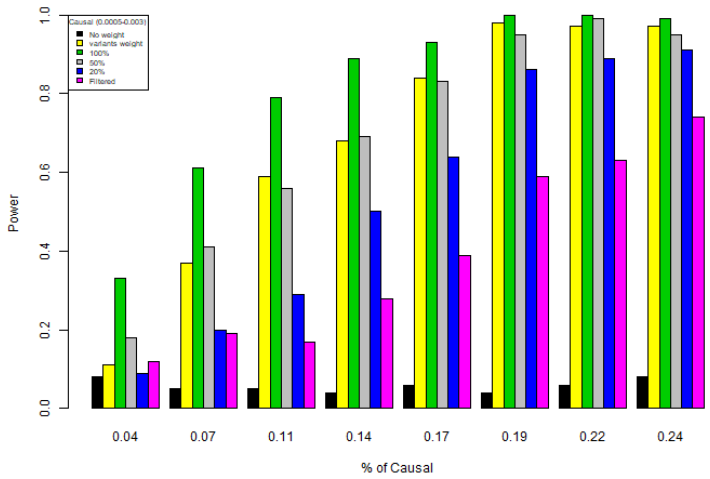


Figure 7.12: In this figure, we illustrate the impact of associating different values of quality with the causal variants. We compare the variants' weight in three scenarios to incorporate quality based on variants: 200% restrict the causal to have a high-quality score; 50% of the causals will have high quality; and the remainder can take very low to high quality same as 20%. Also, we compare it with data after removing the variants which are associated with low quality (i.e. less than $q = 0.4$).

7.4 Weight Scheme II (Burr Function)

In the previous weight scheme (7.3), we used separate functions, $g_1(\mathcal{F})$ and $g_2(\mathfrak{q})$, to incorporate two weight schemes which are the variants' weight based on any previously-described function and quality information. In this section, we can use the Burr function as a weight function that can take into account the quality measure or any extra information. This weighting scheme not only up-weights the rare variants based on MAF but also takes into account quality information based on variants or any prior sequencing.

$$\omega_2 = \mathfrak{q}\, s\, \frac{(\mathcal{F}/m)^{(s-1)}}{(m(1+(\mathcal{F}/m)^s)^{(\mathfrak{q}+1)})} \tag{7.5}$$

where equation 7.5 define as $Burr(\mathcal{F}, m, s, \mathfrak{q})$, $\mathcal{F}$ is the MAF, and $\mathfrak{q}$ is the extra information. m is fixed at 1 and $s = 0.5$; these two parameters were chosen

to meet the following requirements: up-weight the rare variants, down-weight the common variants, and incorporate the extra information. Since $m = 1$, we can re-write the equation 7.5 with some modification:

$$\omega_2 = (\mathfrak{q})\, s\, \frac{(\mathcal{F})^{(s-1)}}{((1+(\mathcal{F})^{s})^{((\mathfrak{q})+1))}} \tag{7.6}$$

The impact of $\mathfrak{q}$ will be large when the MAF is very rare, so we ensure that the common variants associated with high quality will not dominate the signal of association in the rare variants region. Figure 7.13 and Table 7.4 illustrate the impact of $\mathfrak{q}$ on the weight (7.6) under different values of MAF. In other words, when the variants are considered to be extremely rare (which has very low MAF), the ratio between weights with high quality and low quality is larger than if the variants are common 7.14, which shows the impact of q value and ratio between weights with large and low quality.

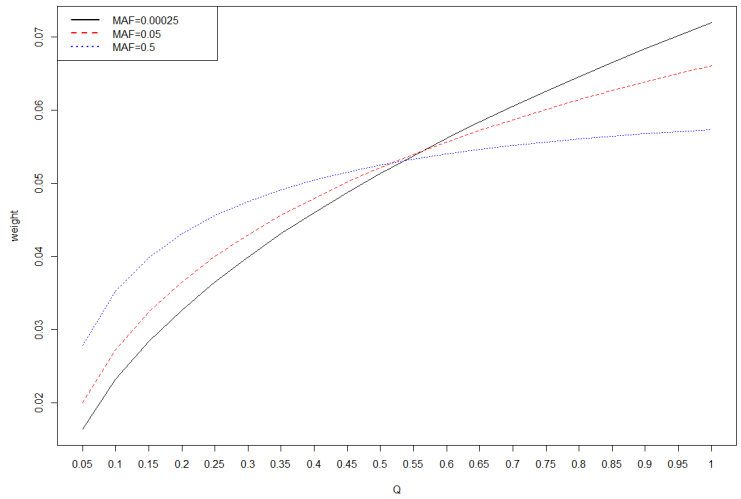


Figure 7.13: The impact of $\mathfrak{q}$ on the weight with three values of MAF $(0.00025, 0.05, 0.5)$. We can see that when the data is common, the impact is smaller than when the MAF is rare.

As we can see in Figure 7.13, rare variants with low quality has lower weight than common variants with low quality. In other hand, rare variants with high quality has larger weight than common variants with high quality and this is the

other benefit from this weighting scheme and the reason to consider this weighting scheme.

MAF	0.0005	0.005	0.05	0.4
$\mathfrak{q} = 1$	21	6	1.4	0.29
$\mathfrak{q} = 0.5$	15	5	1.1	0.26
$\mathfrak{q} = 0.01$	2	0.7	0.2	0.06

Table 7.1: The different values that will be associated with the variants based on MAF and $\mathfrak{q}$.

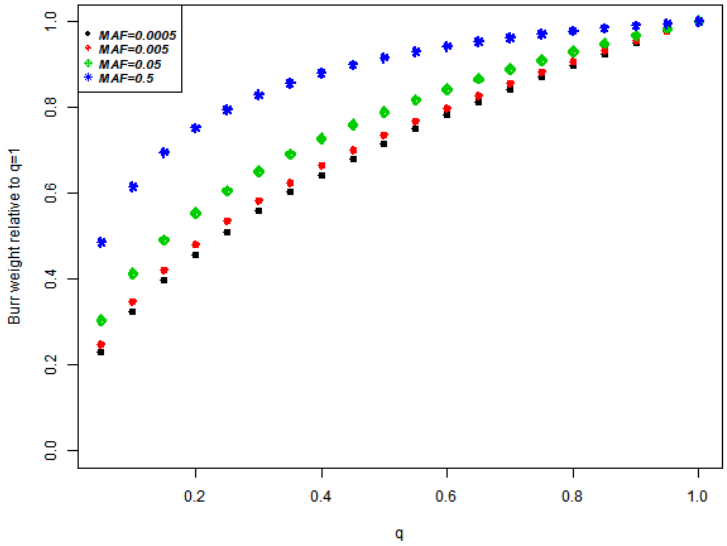
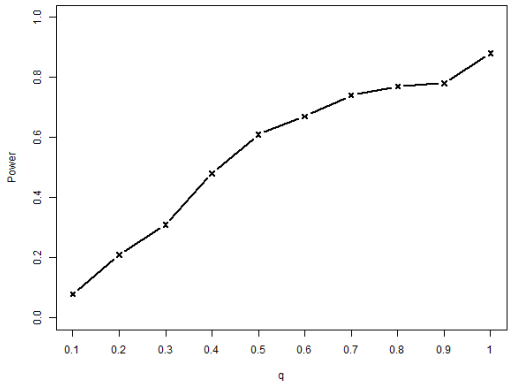


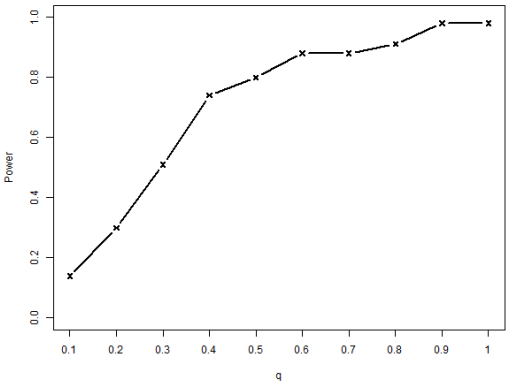
Figure 7.14: The impact of MAF on the weight in terms of the ratio between weights with large and low quality (Burr weight relative to $q = 1$). Relative ratio $= \frac{\text{Burr wight with q= (0-1]}}{\text{Burr wight with q= 1}}$, and the last points in all of the scenarios equal 1 (lines end up with 1) since all of them equal to $= \frac{\text{Burr wight with q= 1}}{\text{Burr wight with q= 1}}$. We can see that when the data is common, the impact is smaller than when the MAF is rare. The x-axis is the value of q, and the y-axis is the Burr weight relative to $q = 1$; the final point equals 1.

This weighting scheme (7.6) takes into the account the quality on the variants' level. Figure 7.15 shows the impact of quality call on the power of the test; when

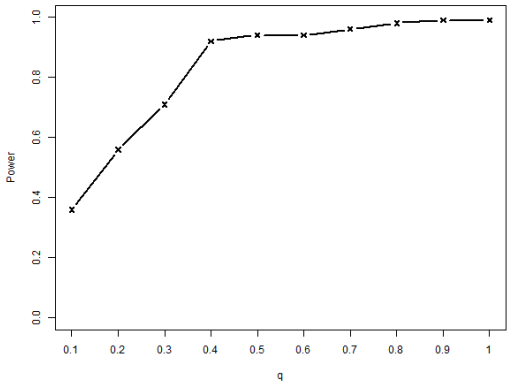
the quality is high the, power increases and decreases when the quality is low. The low quality decreases the power more when the variants are extremely rare.



(a)



(b)



(c)

Figure 7.15: In this figure, we fixed the causal variants to be in the extreme range, which will be less than 0.005 in (a), moderate rare in (b), and common variants in (c). Then, we express the impact of $\mathfrak{q}$ on the power to detect the signal. There are 200 variants containing rare and common variants and, and 10% of them are causal variants with $OR = 3$ in (a), $OR = 2$ in (b), and $OR = 1.5$ in (c). Thus, we vary the quality from 0.1 to 1.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005] in Figure A	✓60%
MRV	OR=2 [0.005-0.05] in Figure B	✓30%
CV	OR=1.5 [0.05-0.5] in Figure C	✓10%

Therefore, we can see the impact of $\mathfrak{q}$ based on the MAF of the causal variants. Low quality will affect the power when the causal variant is extremely rare, which is acceptable because our concern is in the extreme range of MAF. We can say that as the MAF increases, the impact of $\mathfrak{q}$ decreases (see Figure 7.15). In Figure 7.16, we show a comparison of the Burr weight with high quality and with low quality. Since the impact of $\mathfrak{q}$ is large in extremely rare variants, we show in Figure 7.17the comparison of three scenarios which are presented in the simulation section where the MAF of causals are in extremely rare variants as in Figure 7.17 (a), moderately rare variants in (b), and common variants in (c).

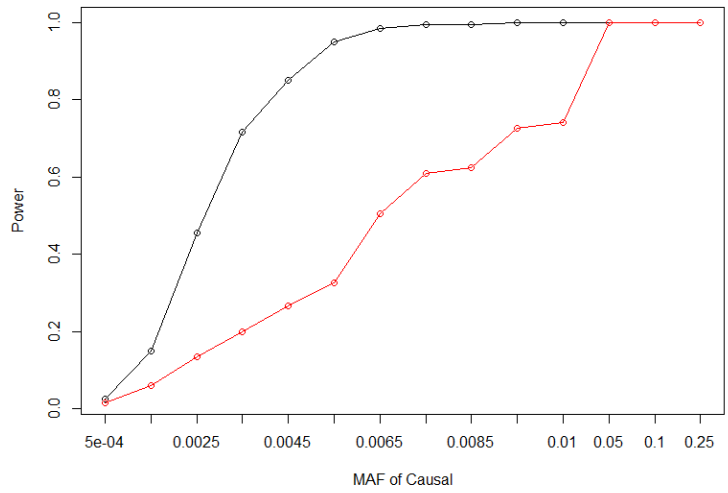
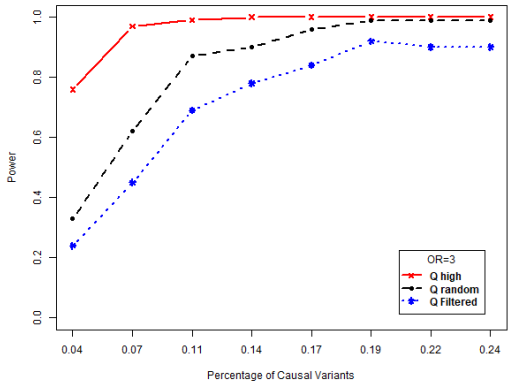
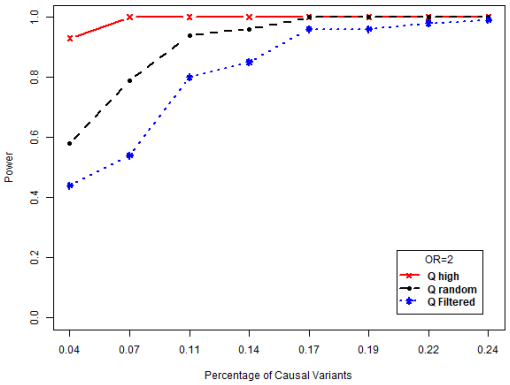


Figure 7.16: The impact of $\mathfrak{q}$ when the MAF of the causal variants gets larger from extremely rare to common variants, fixing the OR=2 at all the type of variants. The black line is the result with $\mathfrak{q} = 1$, and the red line is the result with low quality fixed at $\mathfrak{q} = 0.5$. Thus, we compare the result of the test using high quality and low quality in term of MAF of causal variants.

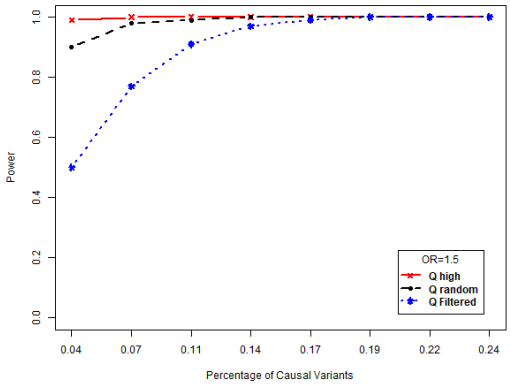
	Causal	Non-Causal
ERV	OR=2 [0.0005-0.005]	✓60%
MRV	OR=2 [0.005-0.05]	✓30%
CV	OR=2 [0.05-0.5]	✓10%



(a)



(b)



(c)

Figure 7.17: The causal variants are fixed to be extremely rare variants ($0.0005 - 0.002$) with odds ratio 3. The data has a 60% range between $0.0005 - 0.005$ and 30% moderately rare variants and 10% common variants. The three simulation scenarios are presented in this figure. (Q high) means all the causal variants have high quality, while (Q high 50%) means 50% have high quality. (Q random) means the causal variants have the same distribution as the non-causal variants.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.002] in Figure A	✓60%
MRV	OR=2 [0.005-0.05] in Figure B	✓30%
CV	OR=1.5 [0.05-0.5] in Figure C	✓10%

The result above is applicable to any MAF of a causal. The quality not only helps detect a true association but also improves the detection of the signal of association when effect sizes are small. Figure 7.18 shows the impact of incorporating the quality under the three scenarios when we change the MAF of causal from extremely rare to moderate and common variants.

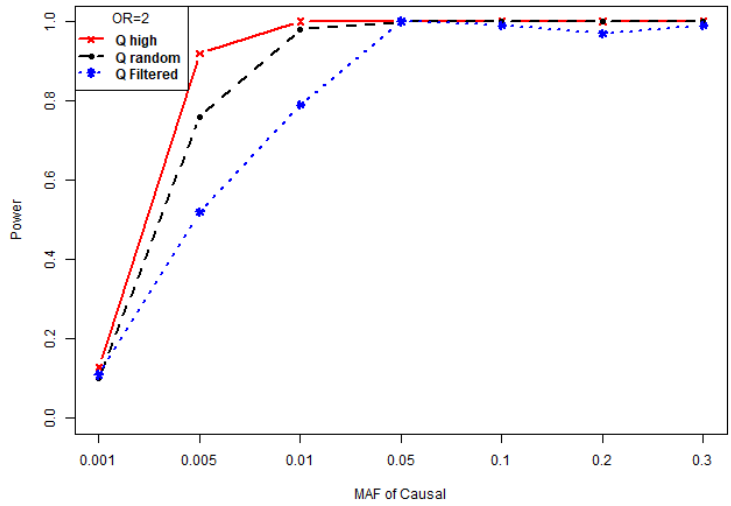


Figure 7.18: The causal variants are 10% and their MAFs vary from extremely rare variants to common with odd ratio fixed at 2. The data has 60% range between $0.0005 - 0.005$ and 30% moderately rare variants and 10% common variants . The three simulation scenarios are presented in this figure. (Q high) means all the causal variants have high quality while (Q random) means the causal have the same distribution as the non-causal have.

	Causal	Non-Causal
ERV	OR=2 [0.0005-0.005]	✓60%
MRV	OR=2 [0.005-0.05]	✓30%
CV	OR=2 [0.05-0.5]	✓10%

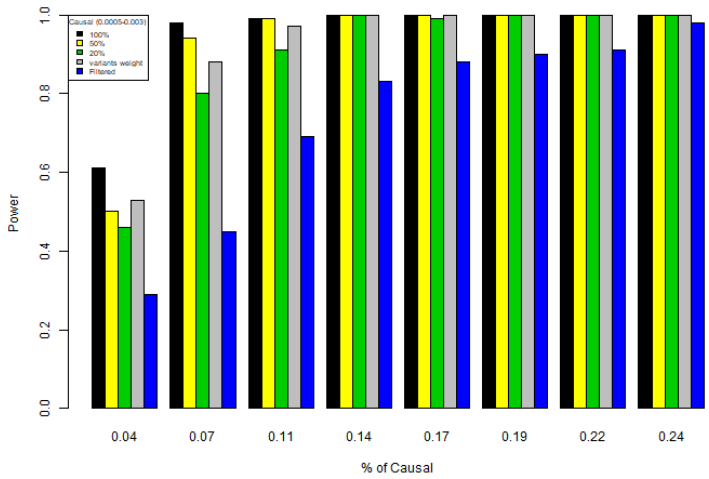


Figure 7.19: In this figure, we illustrate the impact of associating different values of quality to the causal variants. We compare the variants' weight in three scenarios to incorporate quality based on variants as follows: 200% restrict the causals to have a high-quality score, and 50% of the causals will have high quality, while the remaining causals can take very low to high quality same as 20%. Also, we compare it with the data after removing the variants which are associated with low quality (i.e. less than $q = 0.4$).

Type I Error

Based on the above simulation, we generate data in no genetic effects by fixing the $OR = 1$ for causal variants. We split the data into two sets; the first one has a MAF between 0.0005–0.02, and we have data which is rarer and have MAF between 0.0005–0.002 for the other percentage of the data. Then, we will vary the percentage of the first data set which considered rare but not very low frequency.

	MAF (0.0005,0.01)				
Tests	2%	20%	30%	40%	50%
High Quality	0.05	0.05	0.04	0.06	0.05
Random Quality	0.05	0.05	0.05	0.05	0.05
Filtering Low quality	0.05	0.04	0.04	0.06	0.05
Burr Function as variant weight (q=1)	0.05	0.04	0.04	0.05	0.04

Table 7.2: Type I error for the score test using the Burr function in different scenarios.

7.5 Conclusion

In this chapter, we introduced a new weighting scheme that can incorporate sequencing information based on variant j. The weighting scheme in this chapter not only incorporates sequencing information on the variant level but can also up-weight the rare variants and down-weight common variants. We use different simulation settings to show the impact of including the Phred-quality score measure of variant call (QUAL) in the test, which will help keep information and avoid removing information based on a pre-specified threshold such as thresholds that we apply to filter out the genotype with low quality.

As we can see in the results of the simulation, the variant weight which is based on the SKAT weight is dropped in when the causal variants are in the common region, while using the Burr function with its specified parameters will help detect the causal variants even in common regions, especially when the effect size is large. The simulation can be improved in future work so that we can simulate the distribution of errors.

Chapter 8

Score Test with Individual Weights

8.1 Introduction

Using next-generation sequencing technologies, it is now possible to efficiently sequence individuals at a sufficient depth of coverage to determine rare variants. Nevertheless, each sequencing platform has characteristic error profiles, and sample collection, target amplification, and library preparation are additional processes whereby errors are introduced and propagated. Many studies account for these errors by using ad hoc quality thresholds McCrone & Lauring (2016).

Next-generation sequencing studies have a significant number of systematic artifacts that lead to sequencing errors. Therefore, because rare variants are, by definition, seen infrequently, it is difficult to distinguish between errors and real variants (Johnston *et al.*, 2015). Few researchers have addressed this problem.

The remedy seen in different sequencing methods uses an ad hoc method based on a threshold quality control (QC) procedure to reduce the error rate in variant calls; for example, Phred scores are filtered when $Q < 20$ or 30 as part of the quality control process. However, choosing the threshold is important, and it introduces some problems. A high threshold will lead to losing important information and remove correctly called variants. Conversely, a large number of incorrectly called variants will remain if the threshold is too low. Incorporating

more information related to the sequencing or quality in the score test is a new weighting scheme that will help up-weight high-quality variants.

In this chapter, we propose a new weighting scheme for use in rare variant association studies. Since these studies have not previously dealt with incorporating individual-level weighting schemes, the weighting scheme is a novel approach. All the weighting schemes considered in the literature are based on up-weighting rare variants to contribute more to the test relative to common variants. Thus, research has tended to focus on weighting schemes based on variants rather than individual-level weights. We incorporate individual and variant weights to have the benefit of down-weighting common variants since only using an individual weight will result in common variants dominating the results, which is demonstrated in the first section of this chapter.

This chapter is organised as follows. First, we derive the score test by incorporating the individual weighting scheme. Next, we demonstrate the limitations of only using the individual weight and how these are resolved when we incorporate both an individual and a variant weighting scheme. Finally, 8.2 introduces the novel individual-variant weighting scheme.

8.1.1 Derive The Score Test with Individual Weight and Its Distribution

Consider the model;

$$\text{logit}\Big\{P(y_i = 1)\Big\} = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} \qquad i = 1, \ldots, n \tag{8.1}$$

Where y is $n \times 1$ vector of response 0 and 1, and $\mathbf{x}_i^T$ is $n \times 1$ rows of genotype matrix $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$.

To derive the score test based on the individual weight, we derive the score test from the log likelihood of model 8.1. Let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ represent a $n \times 1$ vector of weight and Ψ represent an $n \times n$ diagonal matrix; its elements are ψ. The standard likelihood function is given by

$$L(\beta \mid y_i) = \prod_{i=1}^{n} \mu_i^{y_i}(1 - \mu_i)^{1-y_i} \tag{8.2}$$

where

$$\mu_i = \frac{e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}}. \tag{8.3}$$

Next, we introduce the individual-level weight. Let ψ_i be the weight for individual i, where $i = (1, \ldots, n)$. In this case, the individual likelihood becomes

$$L(\beta \mid y) = \prod_{i=1}^{n} [\mu_i^{y_i}(1 - \mu_i)^{1-y_i}]^{\psi_i}. \tag{8.4}$$

It is more convenient to work with a log-likelihood.

$$\begin{aligned} \ell(\beta \mid y) &= \sum_{i=1}^{n} \psi_i [y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)] \\ &= \sum_{i=1}^{n} \left[\psi_i y_i \left(\log \frac{\mu_i}{(1 - \mu_i)} \right) + \psi_i \log(1 - \mu_i) \right]. \end{aligned} \tag{8.5}$$

The substitution for μ_i can be written as

$$\ell(\beta \mid y_i) = \sum_{i=1}^{n} \left\{ \psi_i y_i \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) + \psi_i \log \left(1 - \frac{e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}} \right) \right\} \tag{8.6}$$

To find the critical points of the log-likelihood function, we differentiate with respect to each β_j.

$$\begin{aligned} \frac{\partial \ell(\beta \mid y)}{\partial \beta_j} &= \sum_{i=1}^{n} \psi_i y_i x_{ij} + \psi_i \frac{\partial}{\partial \beta_j} \log \left[\frac{1 + e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}} - \frac{e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}} \right] \\ &= \sum_{i=1}^{n} \psi_i y_i x_{ij} - \psi_i \frac{e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}} \frac{\partial}{\partial \beta_j} \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) \end{aligned} \tag{8.7}$$

Given that $\frac{\partial}{\partial \beta_j} \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) = x_{ij}$, Equation (8.7) simplifies to

$$\sum_{i=1}^{n} \psi_i y_i x_{ij} - \frac{e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j}} x_{ij}\psi_i. \tag{8.8}$$

Using equation (8.3), (8.8) will be

$$\frac{\partial \ell(\beta \mid y)}{\partial \beta_j} = \sum_{i=1}^{n} \psi_i y_i x_{ij} - \mu_i x_{ij} \psi_i \tag{8.9}$$

We can simplify it as follows:

$$\frac{\partial \ell(\beta \mid y)}{\partial \beta_j} = \sum_{i=1}^{n} \psi_i x_{ij}(y_i - \mu_i) \qquad j = 1, \ldots, p$$

which can be written in a matrix form

$$\frac{\partial \ell(\beta \mid y)}{\partial \beta_j} = X^T \Psi(\mathbf{y} - \boldsymbol{\mu}). \tag{8.10}$$

Let $U(\boldsymbol{\psi}) = X^T \Psi(\mathbf{y} - \boldsymbol{\mu})$, and its covariance-variance is the fisher information (minus the second derivative of the likelihood) $V(\boldsymbol{\psi}) = X^T \Psi^T D \Psi X$ where Ψ is the weight diagonal matrix $n \times n$, and D is a diagonal matrix with elements $\mu_i(1 - \mu_i)$.

If we assume the score function follows a normal distribution as in Pawitan (2001),

$$X^T \Psi(\mathbf{y} - \boldsymbol{\mu}) \sim \mathcal{N}(0, X^T \Psi^T D \Psi X)$$

then

$$(X^T \Psi^T D \Psi X)^{-1/2} X^T \Psi(\mathbf{y} - \boldsymbol{\mu}) \sim N(0, I_p)$$

Since the left part is equal to $\mathbf{U} V^{1/2}$, and we need to prove that the square of it follows a chi-square distribution with p degree of freedom, we will work on the left side of the equation above and square them.

$$\left\{ (X^T \Psi^T D \Psi X)^{-1/2} (X^T \Psi(\mathbf{y} - \boldsymbol{\mu})) \right\}^T \left\{ (X^T \Psi^T D \Psi X)^{-1/2} (X^T \Psi(\mathbf{y} - \boldsymbol{\mu})) \right\}$$

Then;

$$\left\{ [X^T \Psi(\mathbf{y} - \boldsymbol{\mu})]^T [X^T \Psi^T D \Psi X]^{-1/2^T} \right\} \left\{ [X^T \Psi^T D \Psi X]^{-1/2} [X^T \Psi(\mathbf{y} - \boldsymbol{\mu})] \right\}$$

We can say
$[X^T \Psi^T D \Psi X]^{-1/2} \times [X^T \Psi^T D \Psi X]^{-1/2} = [X^T \Psi^T D \Psi X]^{-1}$, if $[X^T \Psi^T D \Psi X]^{-1}$ is symmetric . Because Ψ is a diagonal matrix, it is symmetrical. D is also a diagonal matrix and symmetrical, and $\Psi^T D \Psi$ is diagonal since it is just a multiplication of diagonal matrices, so it is symmetrical. Any matrix multiplied by its transpose is symmetrical $X^T X$. Then we can say that this part $X^T \Psi^T D \Psi X$ is

symmetrical. Since this part $[X^T\Psi^T D\Psi X]^{-1/2^T}$ is symmetrical, we can say these two parts $[X^T\Psi^T D\Psi X]^{-1/2}[X^T\Psi^T D\Psi X]^{-1/2}$ are equal to $[X^T\Psi^T D\Psi X]^{-1}$.

$$(\mathbf{y}-\boldsymbol{\mu})^T\Psi^T X(X^T\Psi^T D\Psi X)^{-1}X^T\Psi(\mathbf{y}-\boldsymbol{\mu}) \sim \chi^2_p$$

which can be written as

$$U^T(\boldsymbol{\psi})V(\boldsymbol{\psi})^{-1}U(\boldsymbol{\psi}) \sim \chi^2_p \tag{8.11}$$

Hence, the test in equation 8.11 will follow a χ^2_p with p degrees of freedom. This form of test can be written in another form when we treat β as a random effect, which is discussed in Chapter 4.

Recall that $\mathbf{U} \sim \mathcal{N}(0, V)$; we know that from a previous chapter that $\mathbf{U} = X^T(\mathbf{y}-\boldsymbol{\mu})$ and $V = X^T DX$ where D is a diagonal matrix with elements $\mu_i \times (1-\mu_i)$. By including an individual weight, $U(\boldsymbol{\psi}) = X^T\Psi(\mathbf{y}-\boldsymbol{\mu})$ and $V(\boldsymbol{\psi}) = X^T\Psi D\Psi X$. From the normal distribution properties and the theorem in Chapter 4, we can construct the test based on an individual weight as follows:

$$S(\boldsymbol{\psi}) = U(\boldsymbol{\psi})^T U(\boldsymbol{\psi}) \tag{8.12}$$

Which will be equivalent in form to the variance component test $(\mathbf{y}-\boldsymbol{\mu})^T\Psi X\Psi^T D\Psi X\Psi(\mathbf{y}-\boldsymbol{\mu})$. Using theorems 4.5.1 and 4.5.2, the distribution of the $S(\psi)$ is a mixture of χ^2 as

$$S(\boldsymbol{\psi}) \sim \sum_{j=1}^{p} \lambda_j \chi^2_1 \tag{8.13}$$

where the χ^2 variate is distributed independently of every other variate, and the λ's are the p real non-zero latent roots of the matrix $V(\boldsymbol{\psi})$.

8.1.2 Individual Weighting Scheme

In variant weights, the weight is usually a function of the MAF, which can be used to up-weight rare variants and down-weight common ones. To incorporate the individual weighting scheme, we will build this weighting scheme using a function that can relate high quality to individuals who have extremely rare variants. The individual weight is not associated with parameters (variants); it is associated

with individuals. Consequently, we need to identify individuals who have one or more rare variants among p positions. In this weight, we will take into account the weight of all the positions for i individuals. Therefore, we are interested in external information based on individuals or will take a marginal weight based on external information, for example, quality data, via a matrix $n \times p$. Because the individual weight will contribute to the test statistics, choosing a weighting scheme or function to meet the following conditions is critical. The individual weight is associated with all individuals, and then a weight function is applied that up-weights individuals who have extremely rare variants. We also use a threshold to identify the extremely rare variants and up-weight these individuals. The weight function has to set a threshold between rare and extremely rare variants. We choose the threshold in this function $1/2\sqrt{2n}$ since it can be the threshold between large minor allele frequency or moderately rare variants and extremely rare variants; if we have 2000 individuals, the threshold will be 0.007 so that any individual with an extremely rare variant will be up-weighted using a beta function with parameters 2, and 1. Based on these conditions, we use intensive simulations that evaluate type I errors and power of test, as well as the signal of causality at different MAFs to arrive at the following function:

$$\psi_i = \begin{cases} \frac{{w_i}^{2(\upsilon-1)}\Gamma(\upsilon+1))}{\Gamma(\upsilon))} & if \quad \sum_{j=1}^{p} x_{ij} < 1/2\sqrt{2n}, \qquad i = (1, \ldots, n) \\ w_i & otherwise \end{cases}$$

where $\upsilon = 2$ and w_i is the quality based on individuals; its values are between 0 and 1. It can be expressed as the row sums of genotype quality. If the quality equals 1, then the individual who has extremely rare variants will be associated with the weight equal to 2, which is the result of a beta function with parameters $(2, 1)$, and the other individuals will have only the quality weight, which is between $(0, 1)$.

The value υ associated with the individual who has extremely rare variants will not change the power when we increase it from 2 to a larger number because the contribution to the test using this weight will be the same. However, there will be an effect when the causal is in the moderately rare variant region. Hence, to minimise this cost while benefitting from the individual weight, we will use the

value $v = 2$. Figure 8.1 shows the difference in power with different values of v (up-weighting value). This component expresses the threshold that can be used to express extremely rare variants.

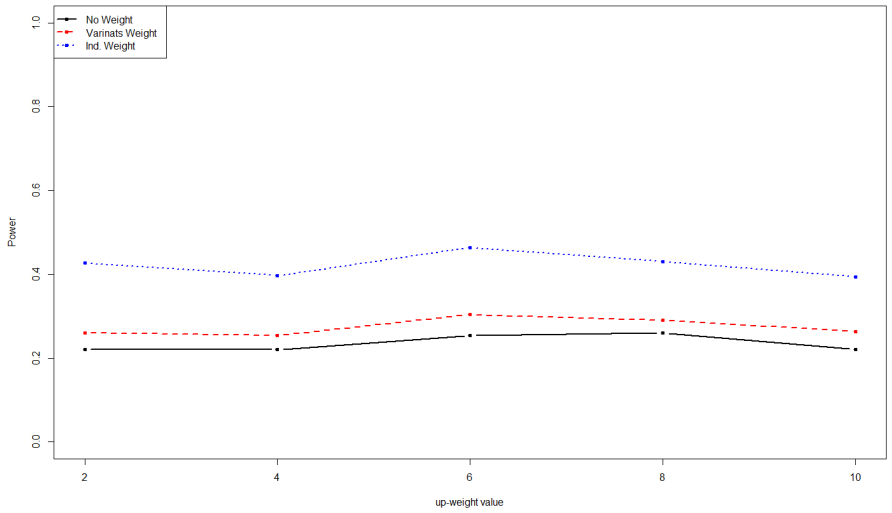


Figure 8.1: In this figure, we show the impact of increasing the value associated with individuals who have one or more extremely rare variants. The causal variants are extreme $0.0005 - 0.001$, and the individuals who have these variants are up-weighted.

8.1.3 Simulation

We evaluate type I errors and the power of the proposed score tests with different weights as discussed previously in the context of multi-locus association analysis with a different number of SNPs. To obtain a genotype matrix X, we follow the same simulation setting explained in section (1) of Chapter 5.

In most scenarios, we generate 100 variants. Each dataset will contain three types of variants (i.e., extremely rare, moderately rare, and common). The amount of each type of variant will vary depending on the context. In this chapter, we generate 100 variants with different MAFs, which are presented in Table 8.1. Each dataset contains extremely rare, moderately rare, and common variants (40%, 40%, and 20%, respectively).

Variants Type	MAF (n=2000)
Extreme Rare Variants	[0.0005 - 0.005]
Moderate Rare Variants	[0.005 - 0.01]
Large Moderate Rare Variants	[0.01 - 0.05]
Common Variants	[0.05 - 0.5]

Table 8.1: The complete set of types of variants in the simulated data.

For the individual weight, we simulate a quality based on an individual using a beta distribution with two different scenarios of the parameter: $(0.5, 0.2)$ to include low, moderate, and high quality and $(1, 0.03)$ for high quality in most of the individuals. The second scenario mimics the distribution of genotype quality from real data in VCF output, which is explained in Chapter 2, and mimic the distribution of genotype quality in Patel *et al.* (2014). In some scenarios, we specify high quality for individuals who have extremely rare variants.

To incorporate individual-level information as a weight, we have three scenarios to simulate. Based on the individual weighting scheme explained in section 8.1.2, the simulation will be separated into two parts, first using an individual who has extremely rare variants (ERV) and then using an individual who does not. The simulation setting is based on two pieces of information associated with each individual, which is the simulated quality and read depth based on each genotype sampled from real data.

First, we simulate a quality based on individuals using a beta distribution with two different scenarios of parameters. Scenario (A) is beta with $(0.5, 0.2)$ to allow the inclusion of low, moderate, and high quality. Scenario $(B)$is beta with $(1, 0.03)$ for high quality in most of the individuals. The details are shown in Table (8.2). In the third scenario, we sample the read depth from real data, the *TYR* gene, which was introduced in Chapter 2. The read depth distribution from 23 individuals at $40,727$ positions is shown in Figure (8.2). Then, we sum up the rows of the read depth matrix and use it for the simulation after normalization. The power when using a read depth simulation will be discussed in section (8.2).

To estimate p-values, straight binomial proportions are used. Hence, they have the same standard error as any other binomial proportion $\sqrt{(}p(1-p)/n)$, where p means the proportion of tests rejected and n the number of samples.

Therefore, if $p = 0.05$ and $n = 2000$, the standard error of the observed proportion is about 0.005, and we could say the uncertainty is 1%.

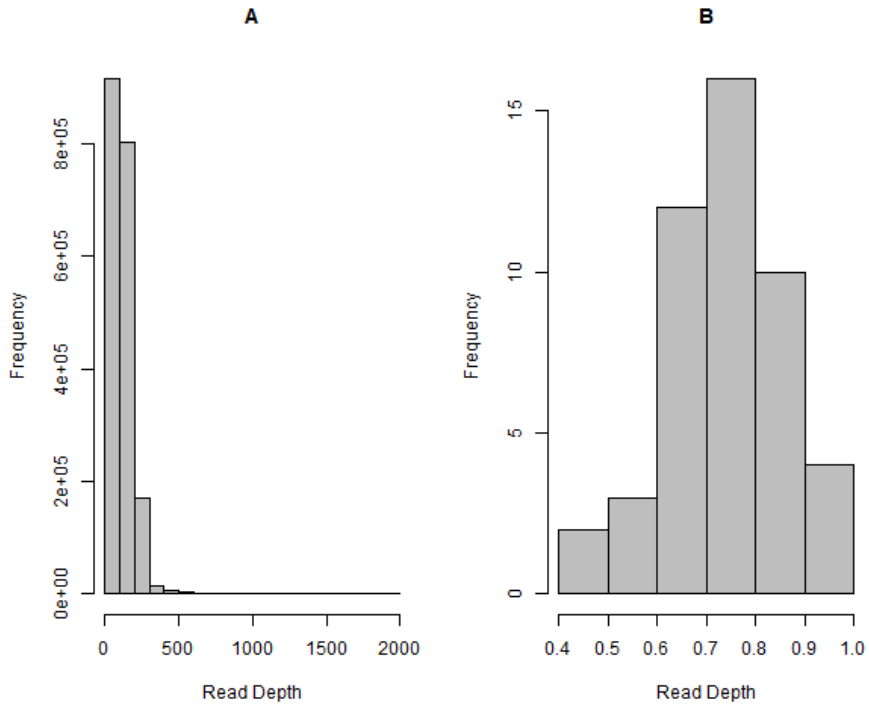


Figure 8.2: The histogram of read depth values at each genotype in Figure A and after we sum up all read depth at each individual and normalised them, which is shown in Figure B.

Scenario		Weight	
		Individual without ERV	Individual with ERV
A	A1	beta(0.5,0.2)	beta(0.5,0.2)
	A2		beta(1,0.03)
B	B1	beta(1,0.03)	beta(1,0.03)
	B2		beta(0.5,0.2)
C	C1	Read Depth (Real Data)	Read Depth (Real Data)
	C2		Unif(0.8,1)
	C3		Unif(0.2,0.5)

Table 8.2: The three scenarios that we use in this chapter to simulate the individuals' information.

8.1.4 Type I Errors

To test for type I errors, we first generate datasets under the null model ($\text{logit}[P(Y_i = 1)] = \beta_0$) . We use two settings to test for type I errors. First, we fix the MAF at each variant. We consider all MAF values, starting from the boundary of the MAF, which is $1/n$, to the MAF of common variants, which is 0.5; see Figure (8.3). Second, we generate X via the scenario described above so that the data have different types of variants (extremely rare, moderately rare, and common variants), we randomly generate extremely rare variants ranging from the boundary of MAF to $MAF = 0.005$, moderately rare variants between $0.005 - 0.05$, and common variants ranging between 0.01 to 0.5. We conduct 1000 simulated datasets; the genotypes are randomly generated for every single simulation. Since the common range controls type I errors because of a large MAF, and we separated the data into two parts: (1) moderately rare and common variants and (2) extremely rare variants. Then, we increase the extremely rare variants in the data from 6% to 50% while we decrease the first part using the same percentage; see Table (8.3).

Generating the individual weight for type I error is randomly chosen using beta a distribution with parameters 0.09 and 0.03 so that all the individuals will be associated with a value from 0 to 1. For both scenarios, we estimate the type I error rate as the proportion of p-values less than the nominal level $\alpha = 0.05$.

We can see type I errors are controlled at level 0.05 for both scenarios. However, for extremely rare variants, controlling type I errors is a concern.

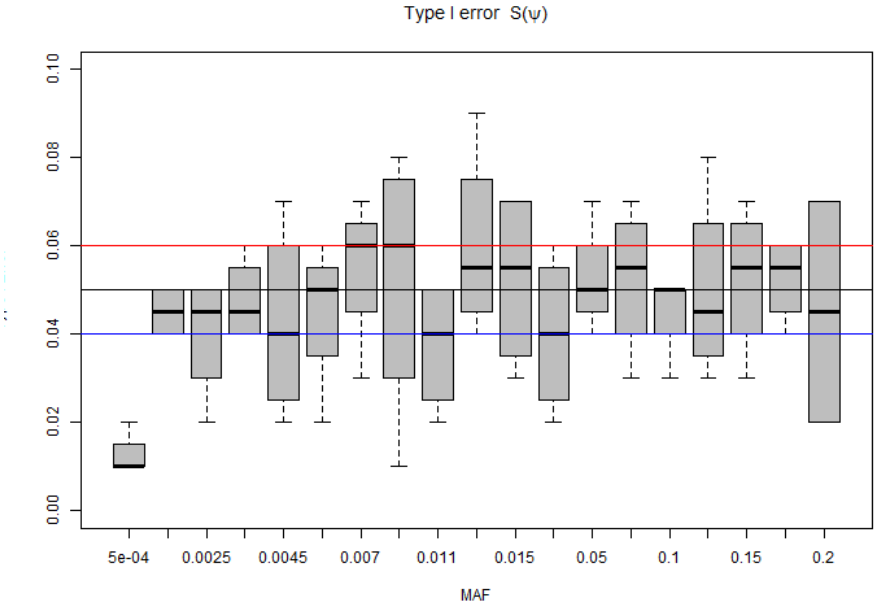


Figure 8.3: Type I error in the individual weight test $S(\boldsymbol{\psi})$

	MAF [0.0005,0.005]					
Tests	6%	10%	20%	30%	40%	50%
$S(\boldsymbol{\gamma})$	0.05	0.05	0.04	0.06	0.05	0.045
$S(\boldsymbol{\psi})$	0.04	0.035	0.04	0.05	0.04	0.04

Table 8.3: Type I error for test with the variant weight ($S(\boldsymbol{\gamma})$), and the individual weight ($S(\boldsymbol{\psi})$).

8.1.5 Power

To evaluate the power with the individual weight, we use three types of variants in the simulation based on a given MAF: non-causal rare variants, causal rare variants, and non-causal common variants, which will be excluded only in the individual weighting scheme.

First, we generate a genotype matrix as explained above (simulation 1). Then, we set the $OR = 3$ for extremely rare causal variants, which is appropriate for the detection of extremely rare variants and set $OR = 2$ for causal moderately rare variants. For non-causal variants, we generate non-causal variants with no effect ($OR = 1$) at different MAFs. In the individual weight only, we exclude common variants since they will dominate the power.

To include the individual weight, we randomly generate it three ways: (1) use the row sum of a genotype matrix from real data (we use the TYR gene's quality matrix, (2) base it on a uniform distribution with parameter 0 and 1, and (3) base it on a beta distribution with parameters 1 and 0.03. Next, we apply the individual weight function, which is assigned the weight after applying the beta function with parameters 2 and 1 as a weight to any individual who has rare variants (the individual can be one case or a control group).

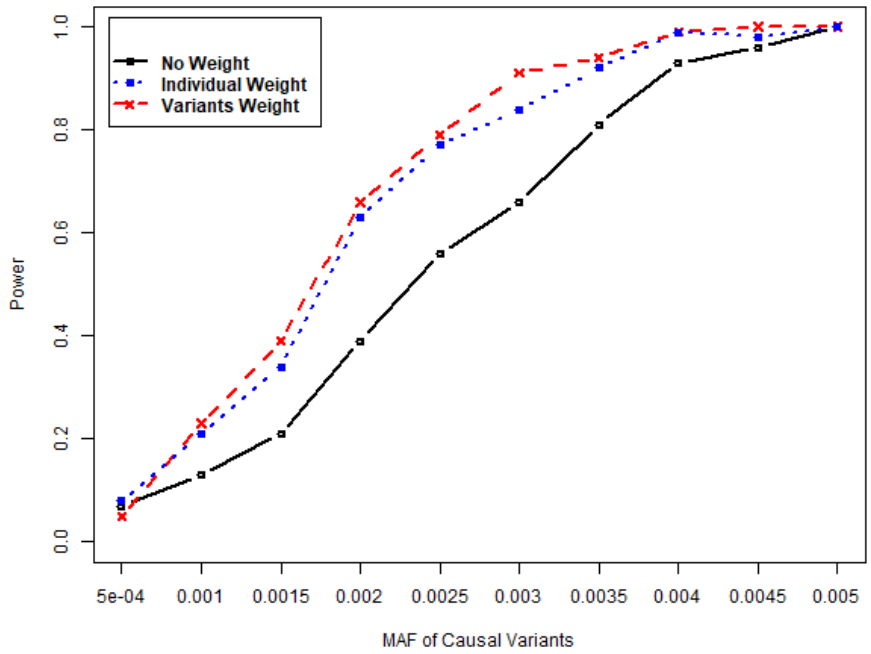


Figure 8.4: Evaluate the power of score test with the individual weight applied. This is the result of scenario $A2$. All individual quality levels follow a beta with parameters $(0.5, 0.2)$, and the quality for an individual who has extremely rare variants are associated with high quality and follow the beta distribution with parameters $(1, 0.03)$ to allow high quality in these individuals. Only extremely and moderately rare variants are included in this analysis. The horizontal axis represents the MAF of causal variants, which is between $(0.0005 - 0.005)$ and in the extremely rare variant range (ERV).

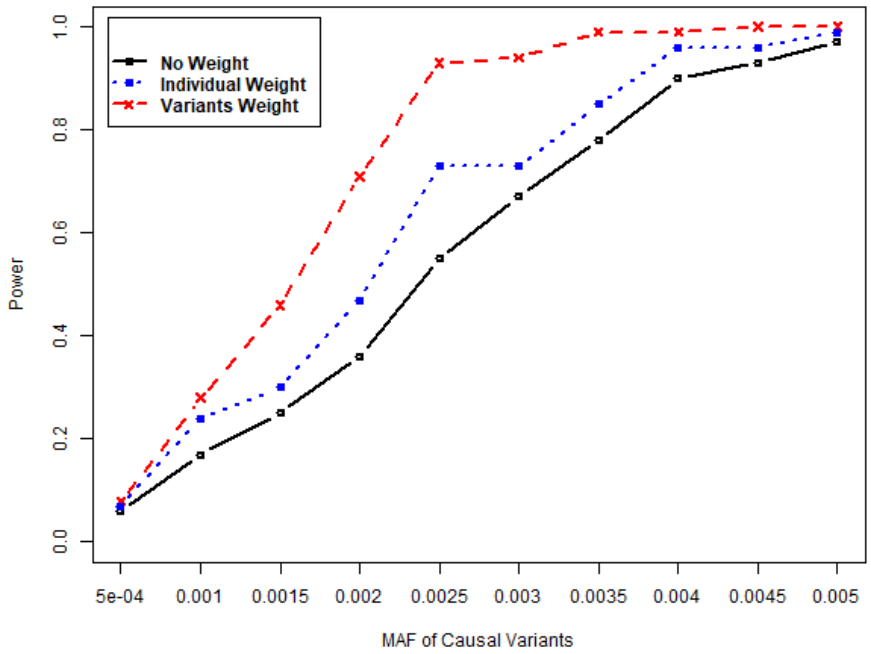


Figure 8.5: Evaluate the power of score test with the individual weight applied. This is the result of scenario $A1$. All individual quality levels follow beta with parameters $(0.5, 0.2)$, and the quality for an individual who has extremely rare variants are associated with quality levels following the beta distribution with parameters $(1, 0.03)$ to allow high quality in these individuals. Only extremely and moderately rare variants are included in this analysis. The horizontal axis represents the MAF of causal variants, which is between $(0.0005 - 0.005)$ and in the extremely rare variant range (ERV).

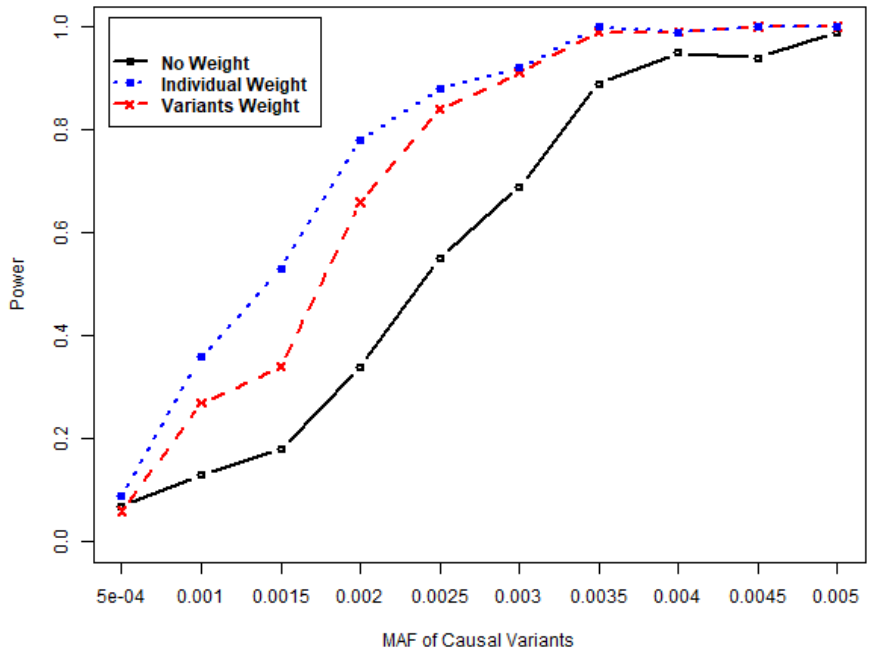


Figure 8.6: Evaluate the power of score test with the individual weight applied. This is the result of scenario $A2$. All individual quality levels follow beta with parameters $(0.5, 0.2)$, and the quality for an individual who has extremely rare variants is associated with quality following the beta distribution with parameters $(1, 0.03)$ to allow high quality in these individuals. We also specify that an individual with a causal variant has quality between $(0.8, 1)$. Only extremely and moderately rare variants are included in this analysis. The horizontal axis represents the MAF of causal variants, which is between $(0.0005 - 0.005)$ and in the extremely rare variant range (ERV).

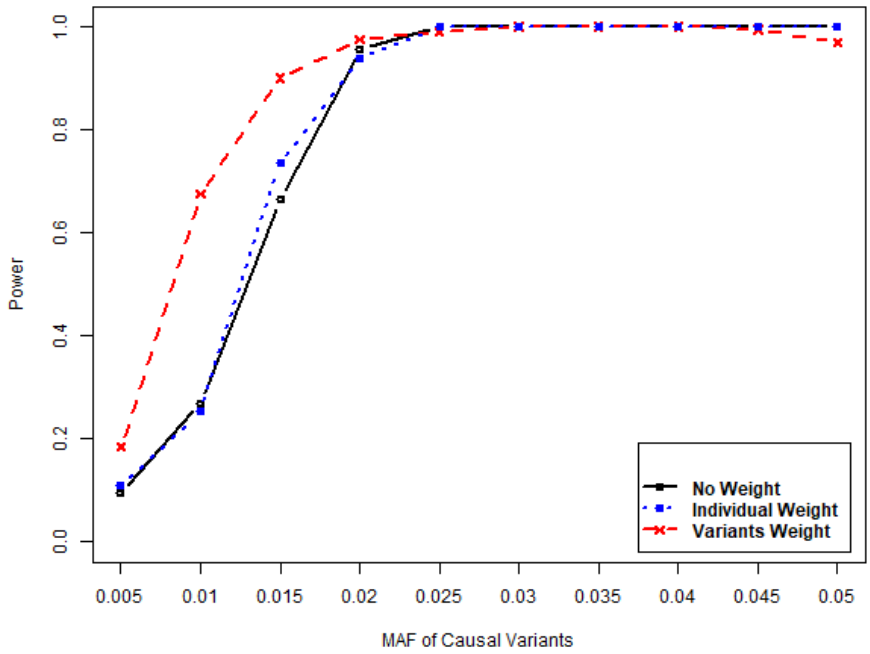


Figure 8.7: Evaluate the power of score test with the individual weight applied. This is the result of scenario $A2$. All individual quality levels follow beta with parameters $(0.5, 0.2)$, and the quality for an individual who has extremely rare variants is associated with quality following the beta distribution with parameters $(1, 0.03)$ to allow high quality in these individuals. Only extremely and moderately rare variants are included in this analysis. The horizontal axis represents the MAF of causal variants, which is between $(0.005-0.05)$ and in the moderately rare variant range (MRV).

8.1.6 Discussion

The test can control type I errors at different MAFs except in the case of extreme rarity (i.e., the boundary of MAF, such as 0.0005 when the $n = 2000$). In the second simulation, when we generate the genotype randomly and use the individual weight, we find that type I errors are controlled at a nominal level, 0.05. However, when the data has extremely rare variants such that 0.0005 when $n = 2000$, then we have a conservativeness issue for the type I error. Thus, the

estimated type I error rates for these two approaches is the same.

The simulation shows that the power can be increased if the individuals' rare variants are casual, which already have a high weight. One major drawback of this approach is that using an individual weight alone will not help down-weight common variants, so the signal of association in rare variants will be lowered. Data with large MAFs will contribute significantly to the test and dominate the signal from the rare region. Therefore, for a better result, the common variants should be excluded from the data or individual and variant weighting schemes should both be applied, which is discussed in the next section where we incorporate variant and individual weights into the score test.

8.2 Variants and Individuals Weight.

8.2.1 Introduction

The benefit of incorporating an individual weight into the test is that extremely rare variants can be identified more confidently. However, including common variants in the test will decrease the power. We can resolve this issue by incorporating variant and individual weights into the score test. By including the variant weight, we down-weight the common variants with a high probability of detecting signals of association in the common variants region yet retain the benefits gained from the individual weight.

8.2.2 Derive the Test with Variant and Individual Weights.

Recall that $\mathbf{U} \sim N(0, V)$. When we multiply the variant weight Γ on $\mathbf{U}$, then we have

$$\Gamma\mathbf{U} \sim N(0, \Gamma V \Gamma^T) \tag{8.14}$$

Take the quadratic form of the test

$$\mathbf{U}^T\Gamma^T\Gamma\mathbf{U} \sim \sum_j \lambda_j \chi_1^2 \tag{8.15}$$

Considering the individual weight Ψ, the test statistic's components (score and covariance-variance matrix) are

$$U(\boldsymbol{\psi}) = X^T \Psi(\mathbf{y} - \boldsymbol{\mu}) \tag{8.16}$$

$$V(\boldsymbol{\psi}) = X^T \Psi D \Psi X$$

Then, combining equations 8.14 and 8.15 with the equation 8.16 result in the following test;

$$\Gamma U(\boldsymbol{\psi}) \sim N(0, \Gamma V(\boldsymbol{\psi}) \Gamma^T) \tag{8.17}$$

So, the distribution of this test is a mixed chi-square;

$$U(\boldsymbol{\psi})^T \Gamma^T \Gamma U(\boldsymbol{\psi}) \sim \sum_j \lambda_j \chi_1^2$$

Where λ_j is the eigenvalues of $\Gamma V(\boldsymbol{\psi}) \Gamma^T$.

8.2.3 Type I Errors and Power

To test for type I errors, we follow the same simulation settings as for the individual weight. The results, which are shown in Figure (8.8) and in Table (8.4), these weighting schemes have satisfactory Type I error rates except when all the variants have very a low minor allele frequency.

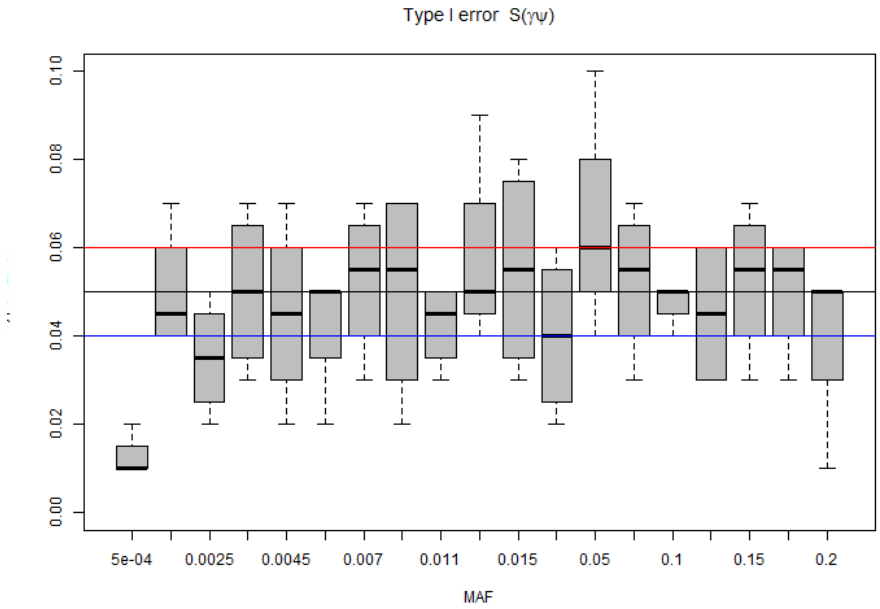


Figure 8.8: Type I error rates for variant and individual weights $S(\boldsymbol{\gamma\psi})$.

	MAF [0.0005,0.005]					
Tests	6%	10%	20%	30%	40%	40%
$S(\boldsymbol{\gamma})$	0.05	0.05	0.04	0.06	0.05	0.04
$S(\boldsymbol{\psi})$	0.04	0.035	0.04	0.05	0.04	0.035
$S(\boldsymbol{\gamma\psi})$	0.05	0.04	0.04	0.06	0.05	0.05

Table 8.4: Type I error for $S(\boldsymbol{\gamma})$, $S(\boldsymbol{\psi})$, $S(\boldsymbol{\gamma\psi})$ score test with a variant weight, individual weight, and combined variant and individual weight, respectively.

To evaluate the power, we generate a genotype matrix as explained above in section 8.1.3 and set $OR = 3$, which is appropriate for the detection of very rare variants. We use three types of variants in the simulation based on a given MAF, which are extremely rare, moderately rare, and common variants; all of them can be causal or non-causal. We also follow same individual weight simulation discussed in the previous section.

Assuming all individuals have a large weight will boost the power of test when the causal happens to be an extremely rare variant (ERV); see Figures (8.9,8.10).

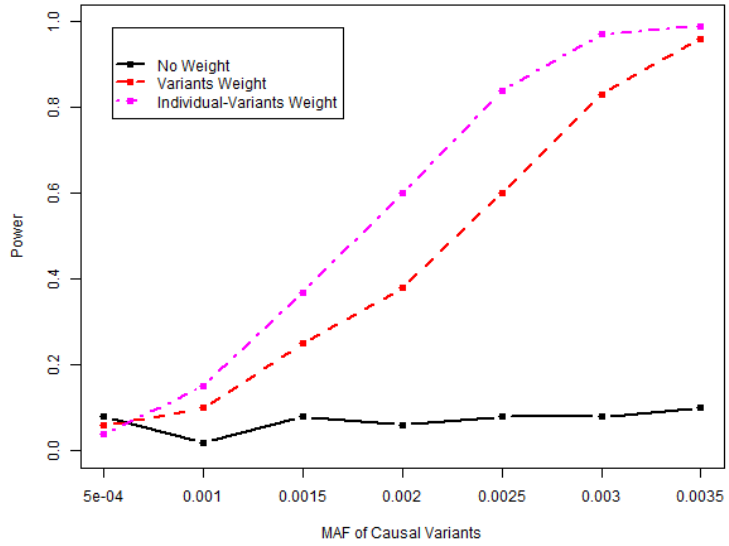


Figure 8.9: This figure shows the individual weight' impact on the power when the causal is in the extremely rare region. We generate 100 extremely rare, moderately rare, and common variants. We fix the percentage of causals at 10% and vary the MAF of causal variants from 0.0005 to 0.0035. We generate the individual weight from beta with 1 and 0.03 as parameters. We generate the individual high weight that is associated with the extremely rare region ($MAF < 0.007$) from beta with 1 and 0.03.

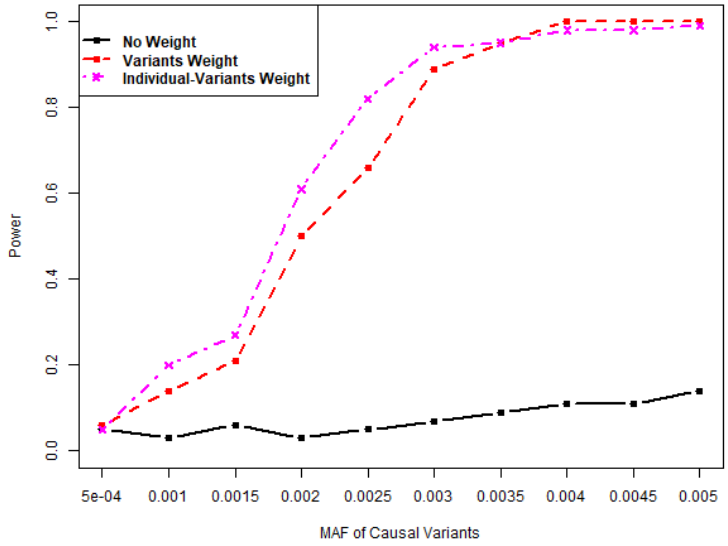


Figure 8.10: We generate 100 extremely rare, moderately rare, and common variants. We fix the percentage of causals at 10% and vary the MAF of causal variants from 0.0005 to 0.005. We generate the individual weight from beta with 1 and 0.03 as parameters. We generate the individual high weight that is associated with the extremely region ($MAF < 0.007$) from beta with 1 and 0.03.

In Figures 8.11, 8.12, and 8.13, we compare tests with low and high weights for individuals according to the previously explained scenarios. In Figure 8.11, individual quality is simulated according to scenario $A1$ and $A2$; $A1$ means the quality of individuals is simulated using beta with parameters 0.2 and 0.5, which allows some individuals to have low quality and most to have high quality. However, in scenario $A2$, the quality of individuals who have extremely rare variants is simulated from beta with parameters 1 and 0.03, which allows most of the specified individuals to have high quality. In Figure 8.12, the quality for all individuals will be simulated from beta with parameters 1 and 0.03 (scenario $B1$), although in $B2$, an individual with an ERV will have quality simulated from beta with parameters 0.2 and 0.5.

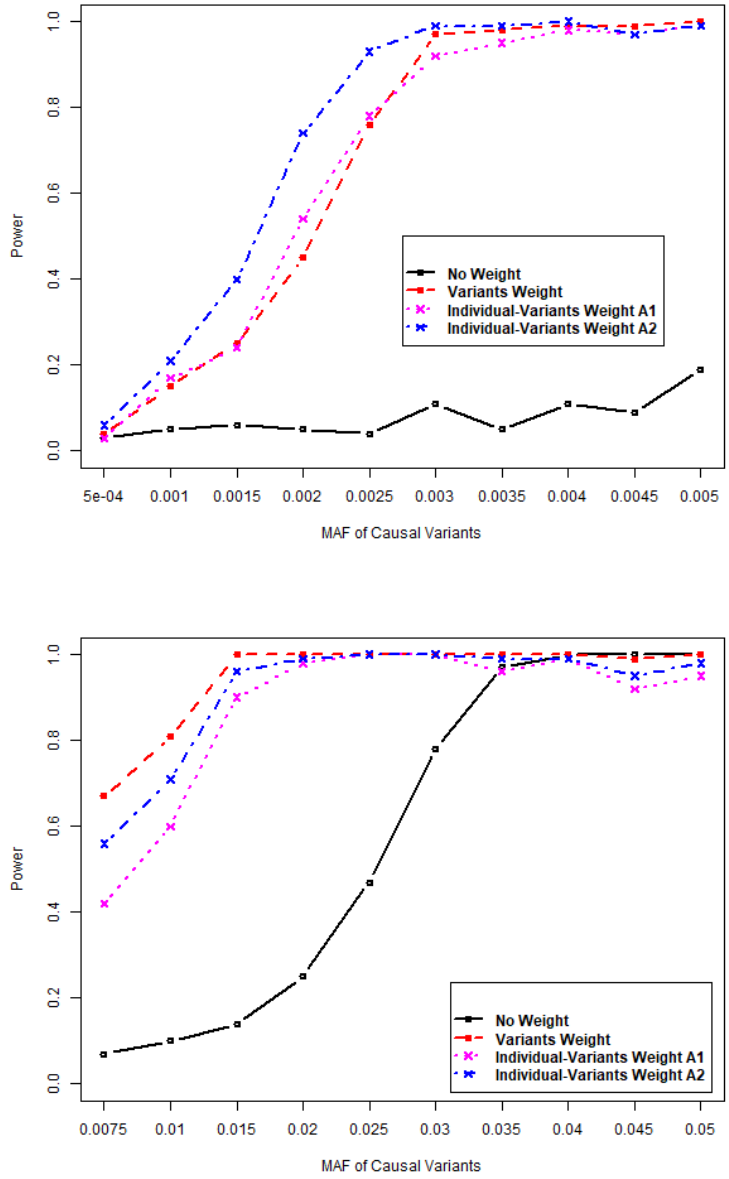


Figure 8.11: We generate the individual weight from beta with 0.5 and 0.2 as parameters. We generate the individual weight is that associated with the extreme region ($MAF < 0.007$) from beta with 0.5 and 0.2 (scenario $A1$), which is the same distribution used for other individuals, or from beta with 1 and 0.03 (scenario $A2$). We generate 100 extremely rare, moderately rare, and common variants. For the top figure, we fix the percentage of causals at 10% with $OR = 3$ and vary the MAF of causal variants from 0.0005 to 0.005, and for the bottom one, we fix the percentage of causals at 5% with $OR = 2$ and vary the MAF of causal variants from 0.007 to 0.05.

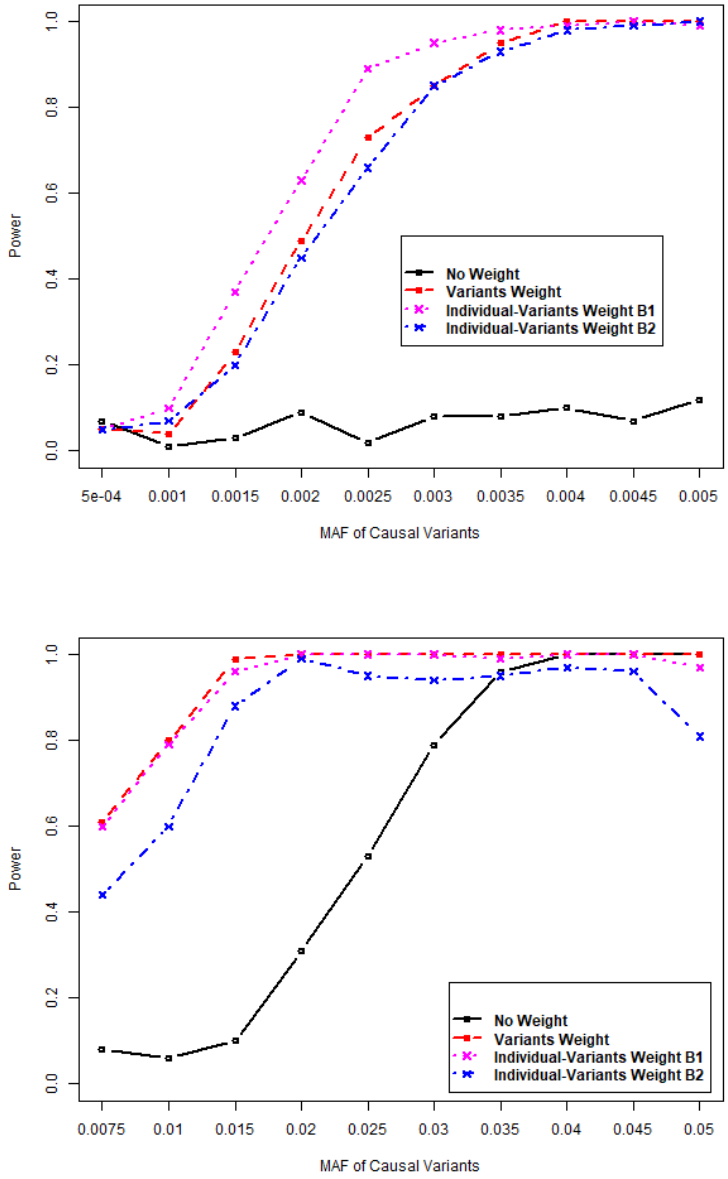


Figure 8.12: We generate the individual weight from beta with 1 and 0.03 as parameters. We generate the individual weight that is associated with extreme region ($MAF < 0.007$) from beta with 1 and 0.03 (scenario $B1$), which is the same distribution used for other individuals, or from beta with 0.5 and 0.2 (scenario $B2$). We generate 100 extremely rare, moderately rare, and common variants. For the top figure, we fix the percentage of causals at 10% with $OR = 3$ and vary the MAF of causal variants from 0.0005 to 0.005, and for the bottom one, we fix the percentage of causals at 5% with $OR = 2$ and vary the MAF of causal variants from 0.007 to 0.05.

In Figure 8.13, we use the read depth from real data. We study the impact of having a low read depth on the individuals who have ERVs at one or more positions (scenario $C3$) and compare the power with individuals who have high read depth values (scenario $C2$). In Figure 8.14, we show an increased percentage of causal variants when the causal occurs in ERV under the scenarios $C1 - C3$.

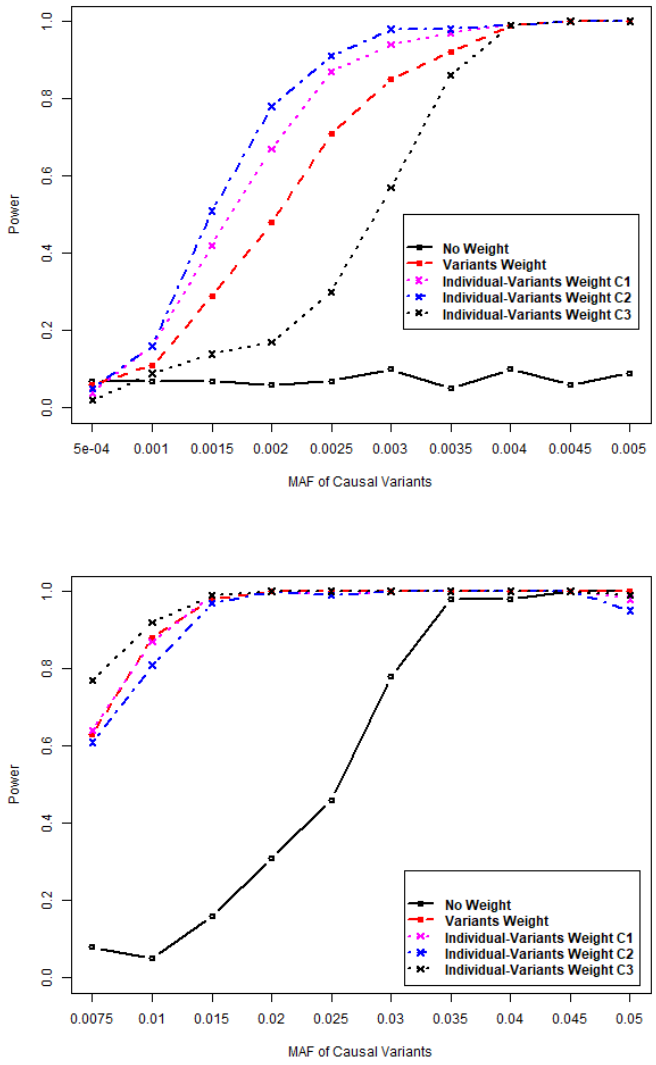


Figure 8.13: We sample the individual weight from a real read depth. We generate the individual weight that is associated with the extreme range ($MAF < 0.007$) from the same distribution of read depths (sampled from real read depth) (scenario $C1$) or from a uniform distribution with parameters $(0.8, 1)$ (scenario $C2$). For scenario $C3$, we generate the weight for individuals who are associated with extremely rare variants from a uniform distribution $(0.2, 0.5)$. We generate 100 extremely rare, moderately rare, and common variants. For the top figure, we fix the percentage of causals at 10% with $OR = 3$ and vary the MAF of causal variants from 0.0005 to 0.005, and for the bottom one, we fix the percentage of causals at 5% with $OR = 2$ and vary the MAF of causal variants from 0.007 to 0.05.

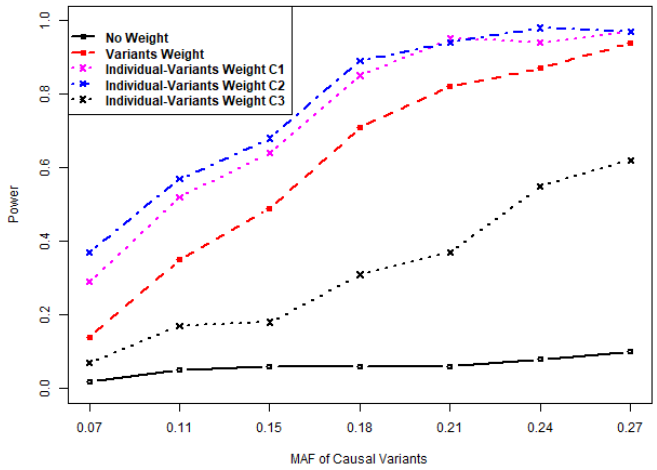


Figure 8.14: We sample the individual weight from a real read depth. We generate the individual weight that is associated with the extreme range ($MAF < 0.007$) from the same distribution of read depths (sampled from real read depth) (scenario $C1$) or from a uniform distribution with parameters $(0.8, 1)$ (scenario $C2$). For scenario $C3$, we generate the weight for individuals who are associated with extremely rare variants from a uniform distribution $(0.2, 0.5)$. We generated 100 extremely rare, moderately rare, and common variants. We fix the MAF of causal variants to be in the ERV range $0.0005 - 0.0025$ and vary the percentage of them from 7% to 27%.

8.2.4 Conclusion

The test controls type I errors; however, there is concern regarding its very low-frequency control. Incorporating individual-level information in the association study is a new weighting scheme that can help up-weight true causal variants. The test is becoming more powerful, and it outperforms the SKAT-weight when causals are in the extremely rare variant range, and individuals with these variants also have high quality. We conclude that the individual weight based on up-weighting individuals that have extremely rare variants will increase the power and reduce the type I error rate to a nominal level. This is the first study to investigate an individual weighting scheme. However, further research should be

undertaken to explore which functions or information can be used in individual weighting schemes.

Chapter 9

Score Test with Cell Weight

9.1 Introduction

Rare variants are statistically challenging due to limited sampling and possible sequencing errors for low-frequency alleles, producing spurious singletons. The inflated singleton count seriously affects statistical analysis and inference in association studies. Genotyping errors commonly occur and could reduce the power and bias of statistical inference in genetics studies. In addition to genotypes, some automated biotechnologies also provide a quality measurement of each individual genotype (GQ-genotype quality) and provide quality for variant calls. These kinds of measurements can be found in the different outputs of variant calls such as VCF output.

A genotype quality score can serve as a good measurement of genotyping accuracy. It cannot tell us whether the genotype call is correct, but as we defined genotype quality in Chapter 2, it gives an estimate of the likelihood of error. We will use genotype quality (GQ) scores rather than simple quality (QUAL) scores, which were introduced in chapter 7. The QUAL scores in VCF output reflect the SNP caller's estimate of how likely there is to be a polymorphism at a given site, while GQ scores are an estimate of how likely the called genotype is to be correct.

Since error rates also correlate with the minor allele frequencies of SNPs, with rare or novel variants much harder to call correctly than common ones Wall *et al.* (2014), it is important to incorporate such quality information in the weighting scheme. Furthermore, we develop here a weight at individual i and

variant j; we call this weighting scheme a cell weight since it is based on the matrix cell. This is the first time we incorporate a quality measure based on the cell weighting scheme. In chapters 5 and 6, we only weighted the variants based on MAF with different functions (i.e. up-weight rare variants and down-weight the common ones), and we extended it in Chapter 7 to include and incorporate other information based on variant quality. Then, in Chapter 8, we introduced joint weight (i.e. individual and variants' weight). In this Chapter, we will consider the use of quality information based on genotype. The idea behind the new weighting scheme as the novel scheme is that we account for both MAF and individual genotyping quality. The difference between cell weighting and the other weighting scheme is that variant weight is based only on up-weighting rare variants and down-weighting common ones, and in chapter 7 in addition to the up-weighting and down-weighting scenario, we incorporated external information based on variant quality. We can say that all the weight schemes in Chapters $5, 6,$ and 7 are based only on the variant level, so the information in the genotype cell is not taken into account. In this chapter, we account for the information based on the cell level. This information can be any information, such as read depth or genotype quality.

9.2 Model and Test

Recall the logistic model;

$$\text{logit}P(y_i = 1) = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} \tag{9.1}$$

Where $y_i = 1, \ldots, n$ and X_i is the row of genotypes. Let $\tilde{X} = X \times \Omega$, then $U(\Omega) = \tilde{X}(\mathbf{y} - \boldsymbol{\mu})$. The test statistics with cell weight is

$$S(\Omega) = U(\Omega)^T U(\Omega) \tag{9.2}$$

where Ω represents the weights based on individuals i and variants j where $i = (1, \ldots, n)$and $j = (1, \ldots, p)$. Using the theorem 4.5.1 and 4.5.3, the distribution of the test (9.2) will be a scaled chi-square as

$$S(\Omega) \sim \sum_{j=1}^{p} \lambda_j \chi_1^2 \tag{9.3}$$

where λ_j is the eigenvalues of the matrix $\tilde{X}^T D\tilde{X}$ where $\tilde{X} = X \times \Omega$, and D is a diagonal matrix with elements $\mu_i(1-\mu_i)$.

9.2.1 Cell Weight Scheme

We are creating a weight which is a function of MAF and all genotype quality levels. The cell weight will be based on two parameters. The first parameter is related to the MAF, and the second parameter related to the genotype quality at individual i and position (SNP) j. This weight will give low-frequency SNPs high weight conditional on the value of genotype quality and at the same time down-weight common SNPs. The Pareto function is one of the functions that work well, for two reasons. First, it up-weights rare variants and down-weights common ones. Second, it will be effective at detecting the association in low frequency (rare) and large frequency (common) variants. Although it up-weights the rare regions, it can still detect the association in the common regions (continuous spectrum of MAF), which was covered in chapter 6. Finally, the impact of lower quality at rare SNPs will be larger than the lower quality at the common SNPs to avoid down-weighting the common variants more when they are associated with moderate quality. To achieve this, we found Pareto is the appropriate function. Let $\mathcal{F}$ be the MAF, $\mathfrak{q}$ the shape parameter, and $\mathfrak{b}$ the scale parameter of the Pareto distribution function, and then the function of Pareto is

$$\Omega = g(\mathcal{F}) = \frac{\mathfrak{q}\mathfrak{b}^{\mathfrak{q}}}{(\mathcal{F}+\mathfrak{b})^{\mathfrak{q}+1}}. \tag{9.4}$$

Where $\mathcal{F}$, the MAF ranges between $1/n$–0.5, $\mathfrak{q}$ lies between $0-1$ and represents the quality of the genotype, and $\mathfrak{b}$ is a scale parameter which is fixed at $0.08+\mathcal{F}$. We choose 0.08 because it is the appropriate number for up-weighting the rare variant and down-weighting the common one (without 0.08, the very rare variants will have a very large weight; see Figure 9.1). We choose plus MAF because we need to keep the value of $g(\mathcal{F})$ increasing as the MAF increases in the rare and common regions so that the largest value for the scale will be 0.58 when the minor allele frequency is 0.5. Equation 9.4 can be simplified as

$$\Omega = g(\mathcal{F}) = \frac{\mathfrak{q}(0.08+\mathcal{F})^{\mathfrak{q}}}{(2\mathcal{F}+0.08)^{\mathfrak{q}+1}} \tag{9.5}$$

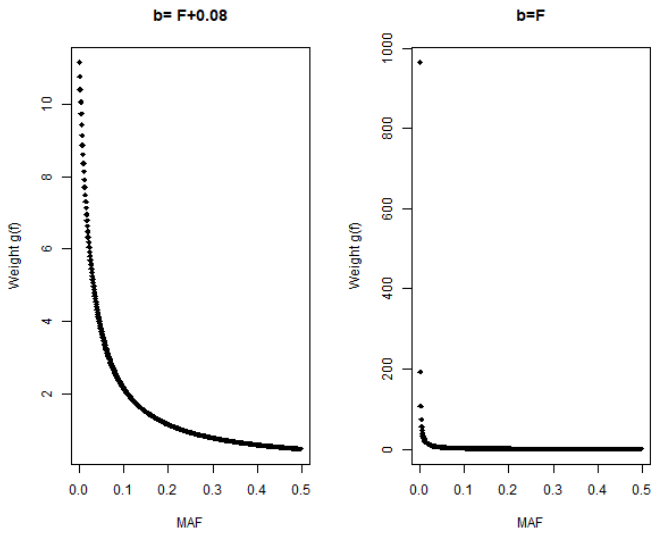


Figure 9.1: The MAF versus the weight using the Pareto function with the inclusion of 0.08 and without. The quality is fixed at 0.9.

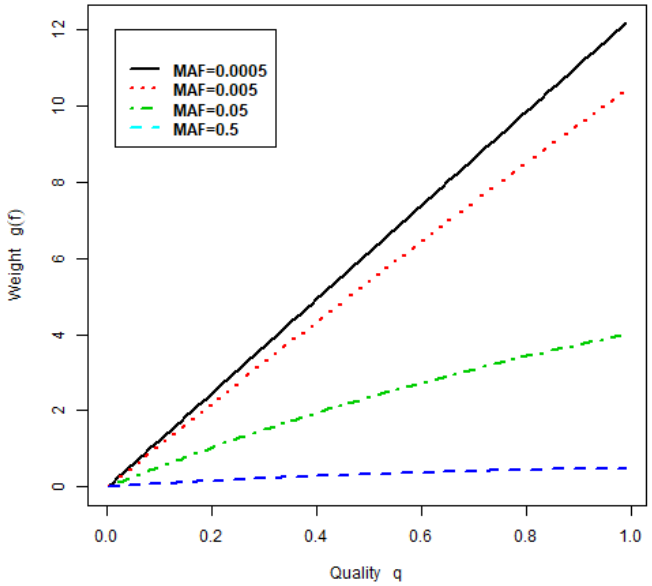


Figure 9.2: The quality versus the weight (g) with respect to the MAF.

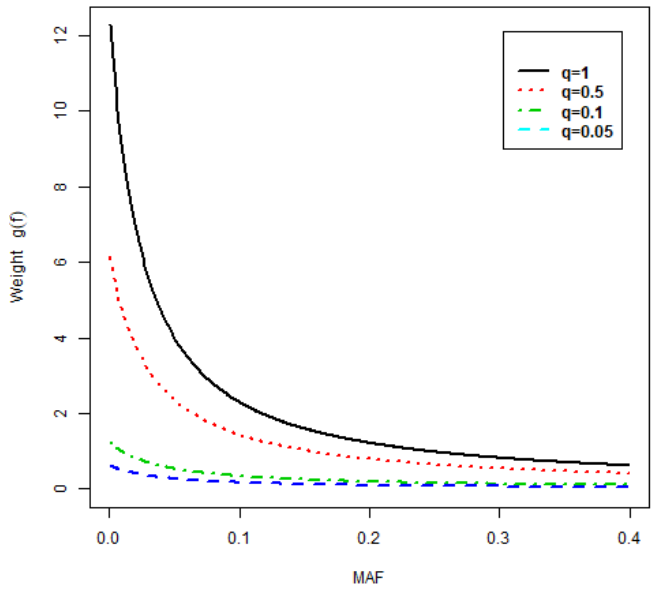


Figure 9.3: The MAF versus the weight using the Pareto function with respect to the quality.

9.3 Simulation

For the genotype, we generated simulated data over a spectrum of MAFs and odds ratios. We also provided scenarios where non-causal variants and variants having effects with different magnitudes are included. As in chapter 5, we first generated d latent variables from the multivariate normal distribution with the autoregressive covariance structure $\Sigma = \rho^{|i-j|}$, where $\rho = 0$ was used to generate independent variants. The latent variables were then used to produce a haplotype at a given MAF by dichotomizing variables at a specified quantile. Two independent haplotypes thus generated were combined to obtain the underlying genotypes x_{ij}, with which we generated dichotomous phenotypes for a case-control study under the logistic regression model

$$\text{logit}P(y_i = 1) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

at given effect sizes β_j or odds ratios $\exp(\beta_j)$. More details regarding this simulation are illustrated in Chapter 5.

We conducted simulation studies to evaluate the performance of the test while incorporating the quality measure. The simulation study is based on ten scenarios. All of these scenarios are presented in Table 9.1.

- In the first scenario, using the weighting scheme introduced in this Chapter, we compare the score test with genotype data assuming large genotype quality and the same genotype with low quality.
- In the second scenario, we simulate genotype quality dependent on the MAF.
- Then, we compare data that has low and high quality with the same data that filters the low-quality genotype (i.e. remove genotypes with quality less than 20%).
- Finally, we sample the genotype quality from the real genotype quality that is associated with the gene *PARP*. The study provides sequencing data for 2014 individuals over 300 rare variant sites, amounting to 604200 genotype quality scores. We sampled from these empirical quality scores directly to generate simulated data.

We will use these simulations to assess the power and evaluate the impact of the quality (i.e. whether it is large or low). First, for high quality, we simulate quality Q_{ij} from the beta distribution with parameters 1 and 0.03 as shape1 and shape2, respectively $A1$. We choose these parameters to mimic the distribution of genotype quality in Patel *et al.* (2014) and to associate high quality with most of the data. The cells that have causal variants will be associated with high quality which is distributed from uniform with parameters 0.8 and 1 to insure high quality will be associated with causal variants; although it is unrealistic, it is just for comparison. We also simulate low-quality associated with a large proportion of cells, from a beta distribution with parameters 0.09 and 0.03 as shape 1 and 2 of beta distribution, respectively $A2$, to allow more cells with moderate to very low quality, and the cells that have causal variants will be associated with low

quality distributed from uniform with parameters 0.2 and 0.6 see; see Figure 9.4 for both scenarios. Additionally, we simulate the same as above, but the causal variants will follow the same distribution as other cells; scenario $A4$ and scenario $A5$ respectively represent large and low quality.

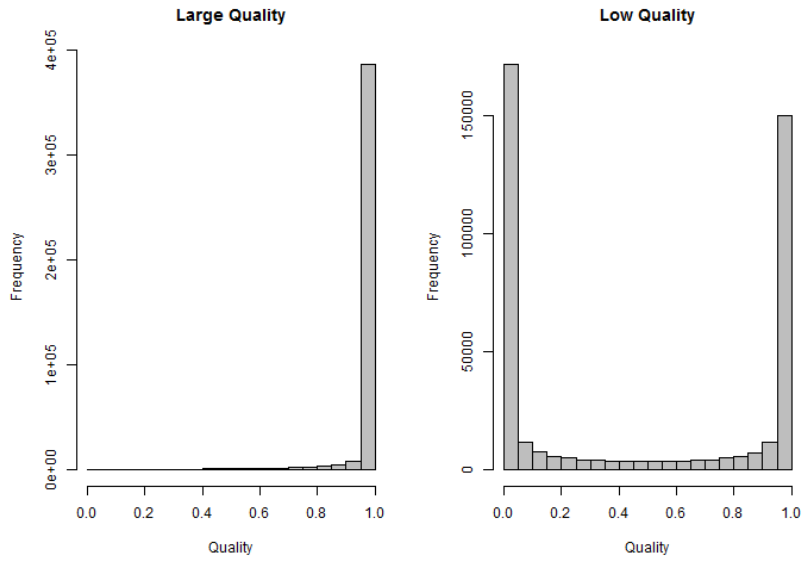


Figure 9.4: Differences between quality simulation in scenarios $A1$ for left figure and $A2$ for the right one.

Then, we present another simulation to assess the power and evaluate the impact of the quality whether large or low. In the above scenarios, we generate quality independently from the MAF. In this scenario, we generated a quality dependent on the MAF. Thus, we will have two scenarios: when the MAF gets small, the quality increases $B1$, and conversely, when the MAF gets small, the quality gets small $B2$. We use a gamma function with parameter 1 and 20 for the first scenario to link the MAF to the quality.

In the next scenario (C), we simulate the genotype with low quality using beta distribution and remove the lowest quality (we use threshold 20%), so scenario $C1$ represents the power with low quality and the $C2$ while filtering lower genotype with very low quality.

Lastly, we sampled genotype quality from real genotype quality and compared it with other scenarios. Additionally, To estimate p-values, straight binomial proportions are used. Hence, they have the same standard error as any other binomial proportion $\sqrt{(}p(1-p)/n)$, where p here means the proportion of tests rejected and n the number of samples. Therefore, if $p = 0.05$ and $n = 2000$, the standard error of the observed proportion is about 0.005, and we could say the uncertainty 1%.

	Quality for Cell	
Scenario	Non Causal	Causal
A1	Beta (1,0.03)	Unif(0.8,1)
A2	Beta (0.09,0.03)	Unif(0.2,0.6)
A3	Beta (1,0.01)	
A4	Beta (1,0.03)	
A5	Beta(0.09,0.03)	
B1	Gamma(1,20)/20	
B2	Beta (0.09,0.03)	all variants less than 0.005 Unif(0.2,0.6)
C1	Beta (0.09,0.03)	
C2	Beta (0.09,0.03)	Filter $GQ > 0.2$ >20%
D	Real quality	

Table 9.1: Scenario simulations considered in this Chapter. See text for details.

Type I error

To evaluate type I errors, we first generate datasets under the null model (logit$[P(y_i = 1)] = \beta_0$). We fix the MAF at each variant as shown in Table 9.2. We use different scenarios to evaluate Type I errors. As we can see, when all the variants have very low minor allele frequency, controlling type I errors is a concern due to rarity. When the MAF is very low such as 0.0005 which means there are two variants among 2000 individual (there are 1998 zeros and two elements are 1 or 2), then this will effect controlling the type I error rate in score test. The concern is when we have MAF on the boundary (MAF less than 0.002).

Scenario	MAF Fixed at 100 variants									
	0.0005	0.002	0.004	0.006	0.008	0.01	0.05	0.1	0.3	0.5
A1	0.035	0.045	0.04	0.05	0.05	0.045	0.05	0.05	0.05	0.05
A2	0.025	0.04	0.06	0.045	0.06	0.05	0.04	0.05	0.05	0.05
A3	0.03	0.05	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.05
A4	0.03	0.04	0.05	0.03	0.06	0.04	0.04	0.04	0.04	0.05
A5	0.02	0.03	0.05	0.03	0.04	0.06	0.04	0.05	0.05	0.05
D	0.03	0.04	0.03	0.04	0.03	0.04	0.06	0.05	0.04	0.04
B1	0.035	0.035	0.05	0.035	0.04	0.05	0.06	0.05	0.05	0.05
C1	0.02	0.035	0.04	0.035	0.045	0.05	0.05	0.05	0.05	0.05

Table 9.2: Type I error for score test with cell weight and significance level of 0.05.

Power

We compare variant weights by Wu *et al.* (2011) (Beta-SKAT) and the cell-weighting scheme that we proposed here. We can see that the cell-weighting scheme under the assumption that there is high quality in the data has an advantage over the weighting scheme by Wu *et al.* (2011) because it can detect the association among the MAFs (see Figure 9.5). We can see the drop in variants weight based on SKAT. In this Chapter, we will compare cell weights with the variant weights, which is based on the Beta-SKAT weighting scheme, and we will frequently see that SKAT does not cover all MAFs (the power of the test drop in the common region).

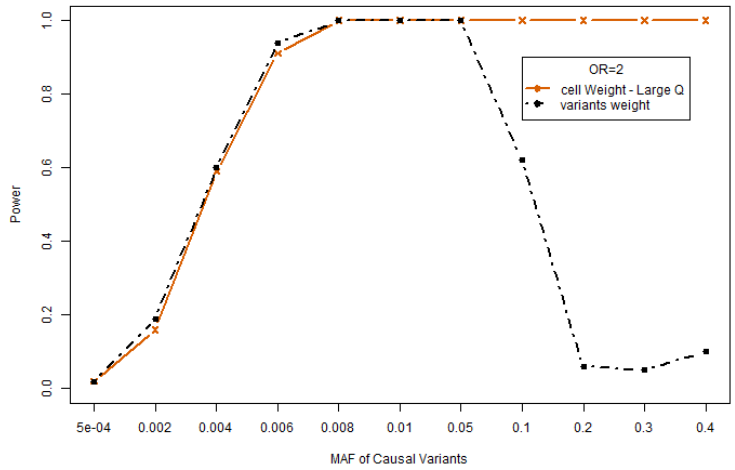


Figure 9.5: There are 200 variants used in this analysis, 80% are rare (40% ERV and 40% MRV), and 20% are common. We ranged the MAF of causal variants between $0.0005 - 0.4$, and the percentage of these is 7% where the OR is fixed at 2.

	Causal	Non-Causal
ERV	OR=2 [0.0005-0.005]	✓40%
MRV	OR=2 [0.005-0.05]	✓40%
CV	OR=2 [0.05-0.4]	✓20%

We show the results of the simulations in different scenarios above. We will apply the simulations above with different MAFs and effect sizes represented by OR. The first result is the comparison of cells that have large and low quality, especially at the cell that is considered to be causal. In Figure 9.7, we show the result of scenarios $A1$ and $A2$; the causal variants in scenario $A1$ have high quality, while in scenario $A2$, they have low quality. We consider in Figure 9.6, $OR = 3$ and the MAF for causal variants between $(0.0005 - 0.01)$ to focus on extremely and moderately rare variants, while in Figure 9.7, $OR = 2$ for top figure and $OR = 1.5$ for bottom figure and MAF for the causal variant are between $0.0005 - 0.5$ for both figures.

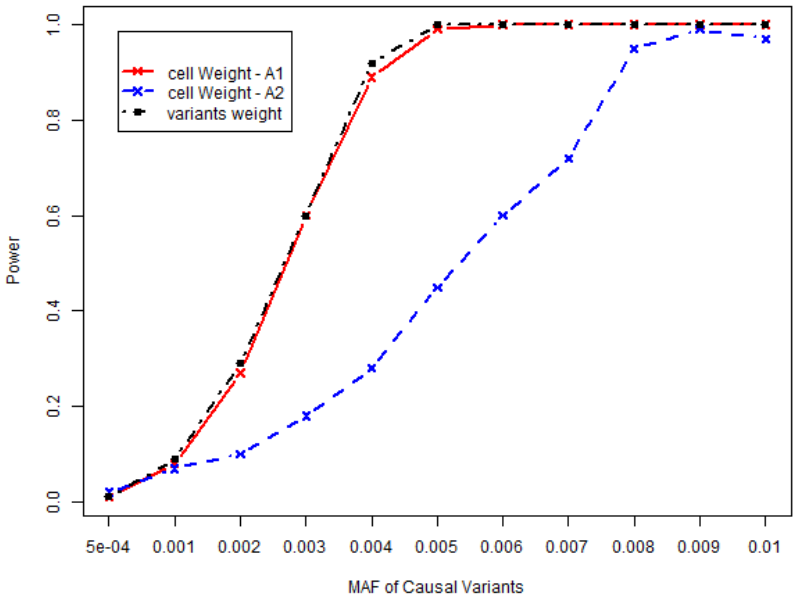


Figure 9.6: This figure shows the differences between causal variants with low and large cell-quality. In this figure, we associate all the causals with high quality in Scenario $A1$, and we associate them with low quality in Scenario $A2$. There are 200 variants used in this analysis 30% for each category (i.e. ERV, MRV, and large MRV), and 10% are common variants. The effect size for causal variants is fixed at $OR = 3$, and the horizontal axis represents the MAF for causal variants $(0.0005 - 0.01)$.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005]	✓30%
MRV	OR=3 [0.005-0.01]	✓30%
large MRV	.	✓30%
CV	.	✓10%

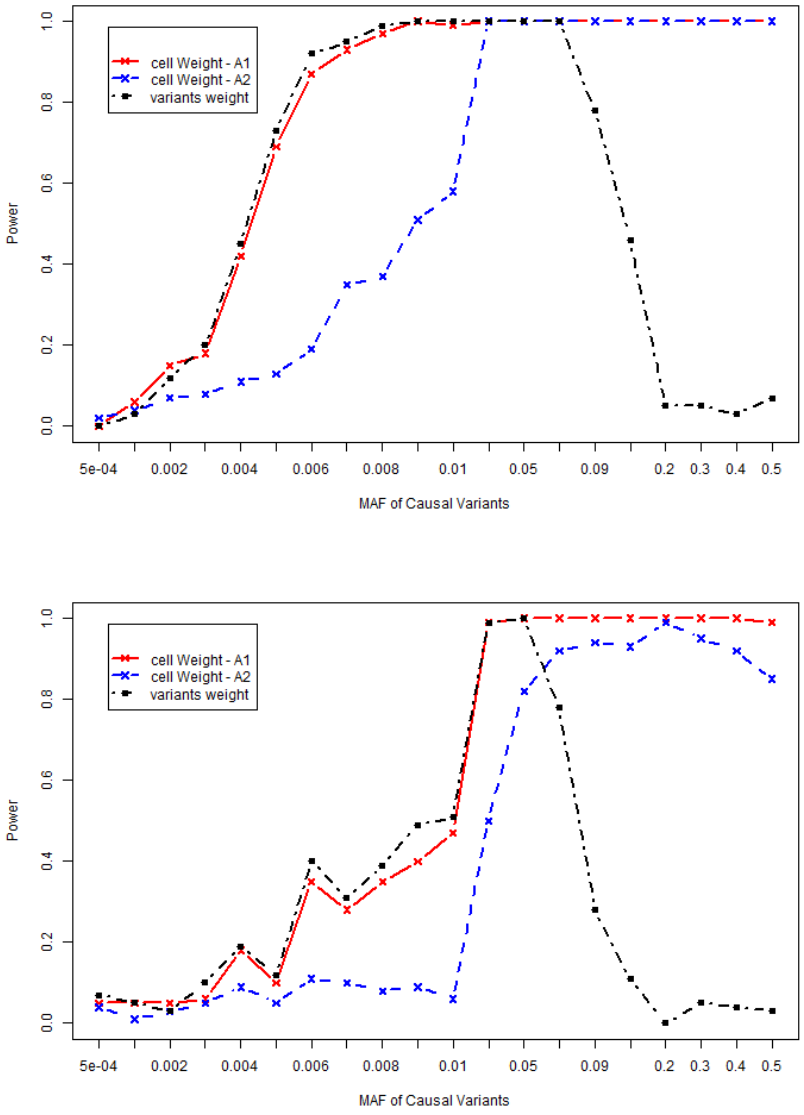


Figure 9.7: This figure shows the difference between causals with low and large cell-quality. In this figure, we associate all the causals with high quality in Scenario $A1$, and in Scenario $A2$, we associate them with low quality. There are 200 variants used in this analysis 30% for each category (ERV, MRV, and large MRV), and 10% are common variants. The effect size for causal variants is fixed at $OR = 2$ for the top figure and for $OR = 1.5$ for the bottom one. The horizontal axis represents the MAF for causal variants, which are between $(0.0005 - 0.5)$.

In Figures 9.8 and 9.9, we follow the same strategy as in scenario $A1$ and $A2$. However, in scenarios $A4$ and $A5$, we did not specify the causal cells to be large or low; it follow the same distribution for all cells.

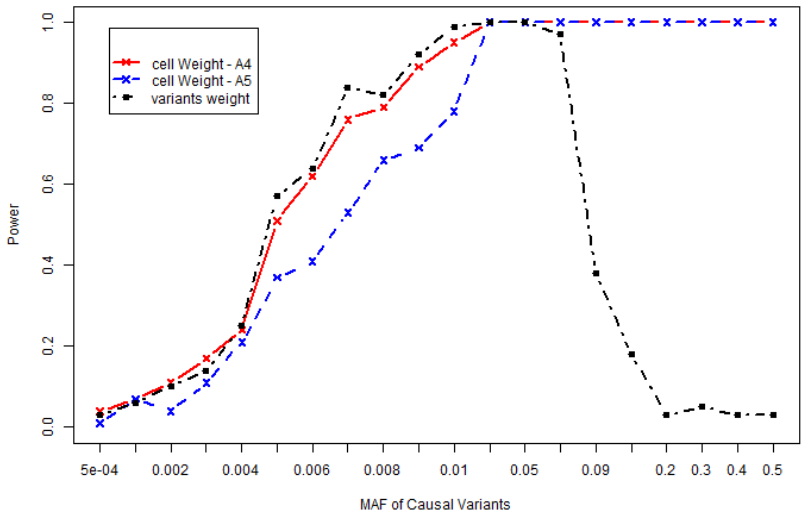


Figure 9.8: This figure shows the differences between causal variants with low and large cell-quality. In this figure, we associate all the causal and non causal variants with high quality in Scenario $A4$, and we associate them with low quality in Scenario $A5$. There are 200 variants used in this analysis 30% for each category (i.e. ERV, MRV,and large MRV), and 10% are common variants. The effect size for causal variants is fixed at $OR = 2$, and the horizontal axis represents the MAF for causal variants $(0.0005 - 0.5)$.

	Causal	Non-Causal
ERV	OR=2 [0.0005-0.005]	✓30%
MRV	OR=2 [0.005-0.01]	✓30%
large MRV	OR=2 [0.01-0.05]	✓30%
CV	OR=2 [0.05-0.5]	✓10%

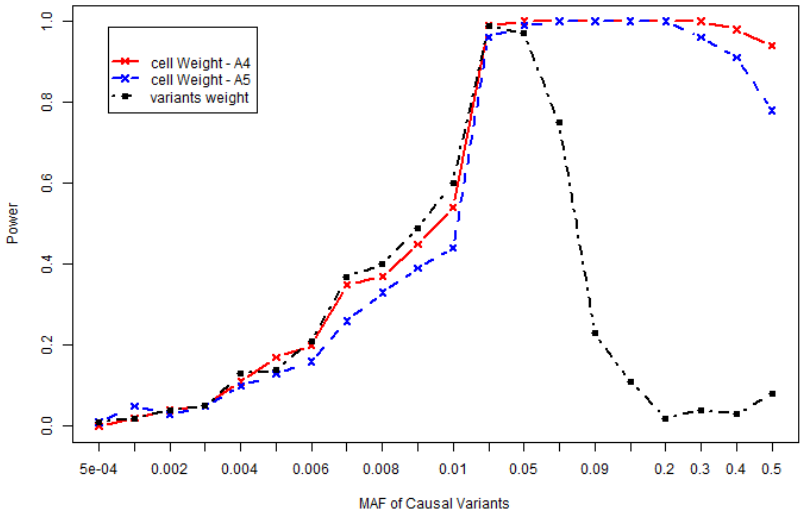


Figure 9.9: This figure shows the differences between causal variants with low and large cell-quality. In this figure, we associate all the causal and non causal variants with high quality in Scenario $A4$, and we associate them with low quality in Scenario $A5$. There are 200 variants used in this analysis 30% for each category (i.e. ERV, MRV, and large MRV), and 10% are common variants. The effect size for causal variants is fixed at $OR = 1.5$, and the horizontal axis represents the MAF for causal variants $(0.005 - 0.5)$.

	Causal	Non-Causal
ERV	OR=1.5 [0.0005-0.005]	✓30%
MRV	OR=1.5 [0.005-0.01]	✓30%
large MRV	OR=1.5 [0.01-0.05]	✓30%
CV	OR=1.5 [0.05-0.5]	✓10%

Figures in 9.10 show the result of the B scenario simulation. As we can see, the high qualities are associated with the cell that is considered rare, and low qualities are at the common variants.

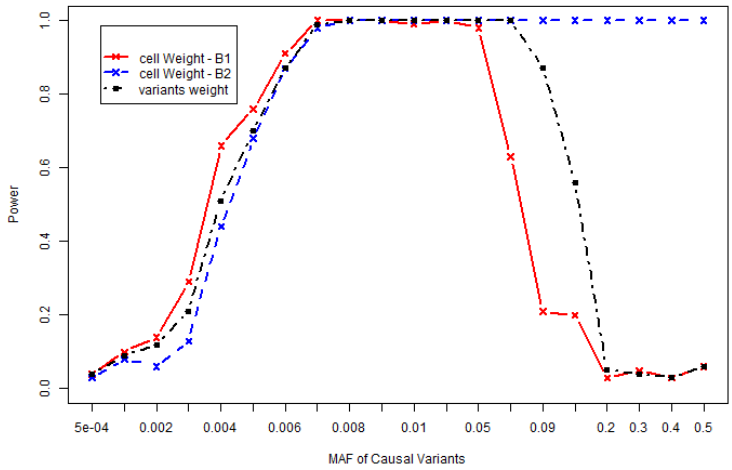


Figure 9.10: In this figure, we show the impact of low quality on the signal of association. There are 200 variants used in this analysis 80% are rare (40% ERV and 40% MRV), 20% are common. We ranged the causal variants between $0.0005 - 0.5$. The odds ratio is fixed at 2, and the quality is dependent on the MAF (scenarios $B1$ and $B2$); when MAF gets small, the quality increases.

	Causal	Non-Causal
ERV	OR=2 [0.0005-0.005]	✓40%
MRV	OR=2 [0.005-0.05]	✓40%
CV	OR=2 [0.05-0.5]	✓20%

Figure 9.11 shows the results of scenario $C1$ and $C2$. We can see that filtering or removing information (e.g., genotypes) from the data may cause the loss of important information. In these two figures, we randomly assign the causal to be in the ERV with random quality, and then we remove the genotype with low quality and keep the genotype with high quality (greater than 20%). We reduce the OR from 3 in 9.11 (top figure) to 2 in the bottom figure to illustrate the difference between all the scenarios.

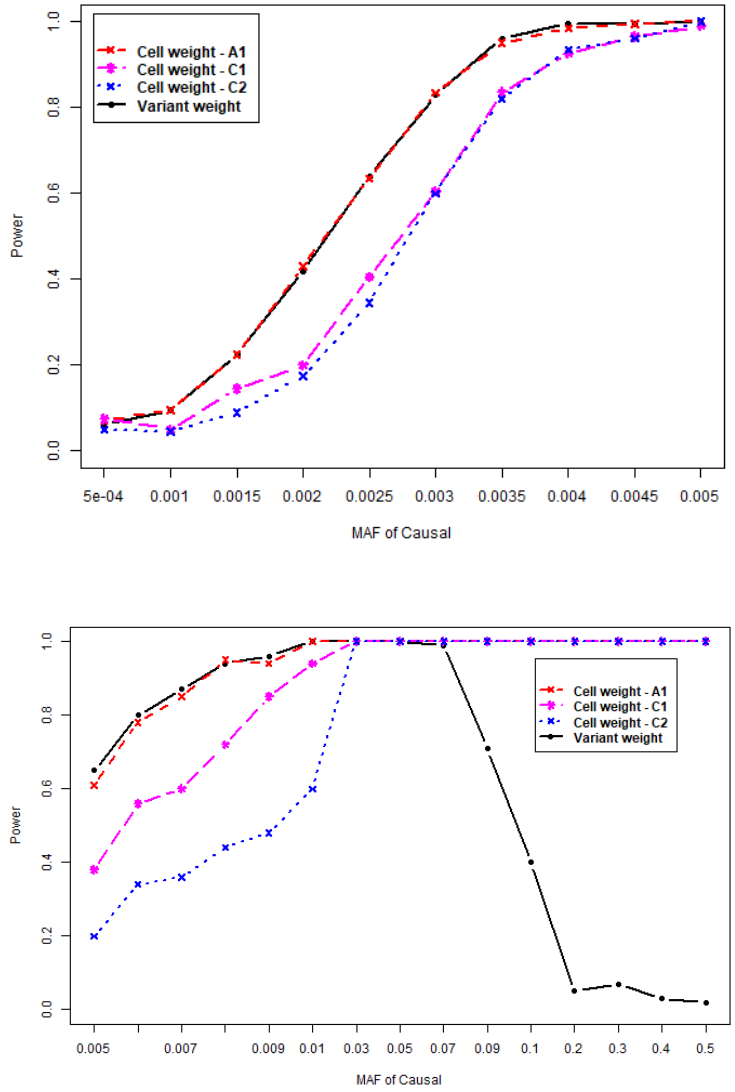


Figure 9.11: We generated 200 variants: 30% for each category (extremely and moderately rare and large moderate) and 10% common variants. The causal variants are 7% and for top figure $OR = 3$ and bottom one $OR = 2$. We vary the MAF of causal variants along the horizontal axis; it is between $0.0005 - 0.005$ in the top figure and $0.005 - 0.5$ in the bottom one. In this Figure is the result of the scenarios ($C1$ and $C2$) compared to $A1$.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005]	✓30%
MRV	.	✓30%
large MRV	.	✓30%
CV	.	✓10%

	Causal	Non-Causal
ERV	.	✓30%
MRV	OR=2 [0.005-0.01]	✓30%
large MRV	OR=2 [0.01-0.05]	✓30%
CV	OR=2 [0.05-0.5]	✓10%

Figure 9.12 shows the comparison between simulated data from scenario $A3$ and sampled quality from real genotype quality. In both scenarios (simulated quality and sampled quality), we see that the genotype quality can affect the signal of association. When the causal variants have high quality so that the likelihood of mistyping is low, it can help boost the power. We zoom in on Figure 9.13, so the causal variants is between (0.0005–0.005), and we can see the drop in in SKAT weight.

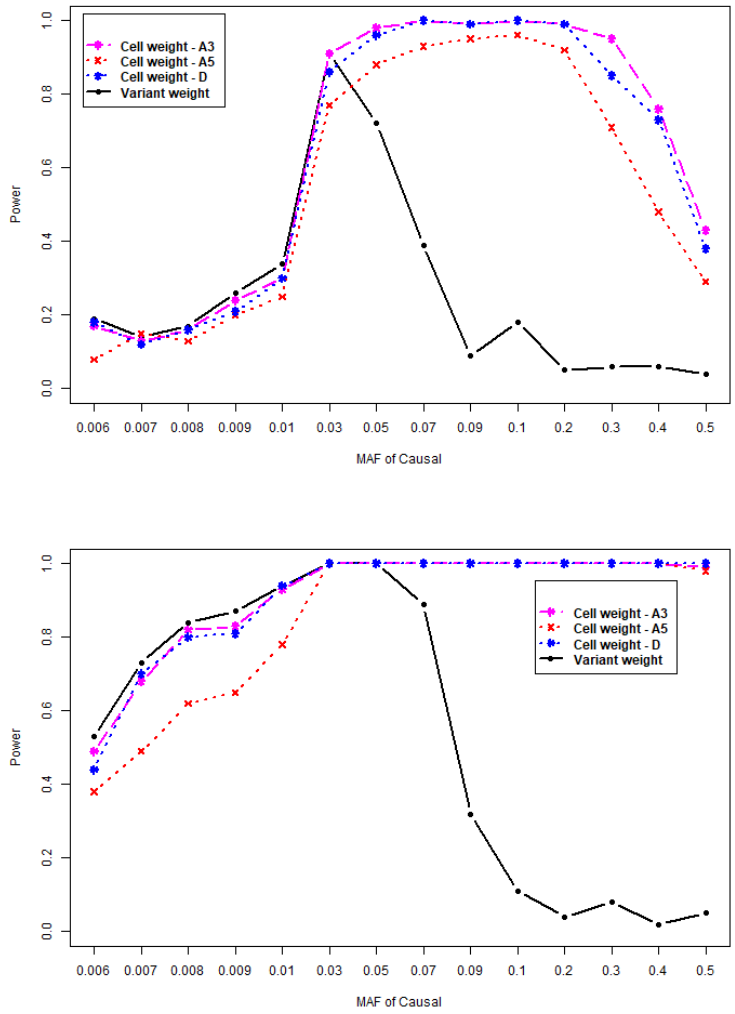


Figure 9.12: We generated 200 variants, 30% for each category (extremely and moderately rare and large moderate) and 10% common variants. The causal variants are 7%, and $OR = 1.5$ for the top figure, and $OR = 2$ for the bottom one. We vary the MAF of causal variants along the horizontal axis. In this Figure is the result of the scenarios (D) compared to $A3$ and $A5$.

	Causal	Non-Causal
ERV	.	✓30%
MRV	OR=1.5 [0.005-0.01]	✓30%
large MRV	OR=1.5 [0.01-0.05]	✓30%
CV	OR=1.5 [0.05-0.5]	✓10%

	Causal	Non-Causal
ERV	.	✓30%
MRV	OR=2 [0.005-0.01]	✓30%
large MRV	OR=2 [0.01-0.05]	✓30%
CV	OR=2 [0.05-0.5]	✓10%

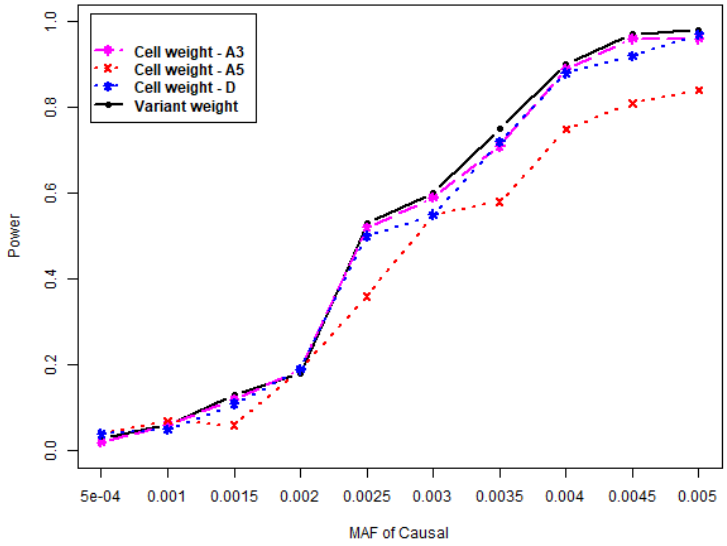


Figure 9.13: We generated 200 variants, 30% for each category (extremely and moderately rare and large moderate) and 10% common variants. The causal variants are 7% and $OR = 3$. We vary the MAF of causal variants along the horizontal axis. In this Figure is the result of the scenarios (D) compared to $A3$ and $A5$.

	Causal	Non-Causal
ERV	OR=3 [0.0005-0.005]	✓30%
MRV	.	✓30%
large MRV	.	✓30%
CV	.	✓10%

9.4 Conclusion

In this chapter, we introduced a new weighting scheme that can incorporate information based on individual i and variant j. The weighting scheme in this chapter can incorporate information on the cell level as well as up-weight the rare variants and down-weight common variants. We use different simulation settings to show the impact of including the quality measure of genotypes in the test, which will help to keep information and avoid removing information based on

pre-specified thresholds such as thresholds that we apply to filter out genotypes with low quality.

As we can see from the results of the simulation study, the variant weight which is based on the SKAT weight drops when the causal variants are in common regions. Using the Pareto function with its specified parameters will help detect causal variants even in common regions, especially when the effect size is large. Using SKAT-beta and Pareto assumes high quality variants will be close to each other when the causal variant is rare, and when the causal is common, Pareto appears more powerful than SKAT-beta.

Chapter 10

Conclusion and Future Research

10.1 Summary

In Chapters 2 and 3 of this thesis, we have described VCF data files and reviewed existing methods for rare variant analysis. A score test with different weight schemes was derived, as well as its distribution. Then, we proposed different variant-, individual-, and cell-level weight schemes. We have investigated the effect of different variant weight schemes and the differences between them. The empirical investigation focused on rare variants; however, we also proposed variant weights that are applicable to the joint analysis of multiple rare and common variants. Additionally, we proposed a weighting scheme that can incorporate different types of information, such as quality, in two scenarios: incorporating quality scores (QUALs) based on variant levels, which is explained in Chapter 7, and incorporating genotype quality (GQ) based on the cell level, which is explained in Chapter 9. In Chapter 8, we incorporated individual weights and combined individual variant weights.

In chapter 5, we introduced different weighting schemes based on variant weight. These weighting schemes focus on rare variants. We also introduced an adaptive weighting scheme that can be adjusted according to the data.

Previous studies of rare variants using weighting schemes have included both rare and common variants. Most studies in this field have only focused on rare variants. In Chapter 6, we extended the weighting scheme to cover common variants and present the current investigation's principal findings regarding weighting

schemes that cover the whole MAF range. When the data has more ERVs (compared to moderately rare variants), we recommend using the Cauchy adjusted or Levy weight schemes since these schemes up-weight ERVs more than others. However, if the MAFs of the data are uniformly distributed, or there is a large number of moderately rare variants (compared to ERVs), then we recommend using the Burr or beta functions.

In Chapters 7, 8, and 9, we explored an exciting opportunity to advance our knowledge of incorporating external information on the score test. In Chapter 7, we incorporated quality based on the variant level (quality call) and concluded that incorporating such information could help eliminate errors by down-weighting low-quality variants instead of removing them or up-weighting them because they are rare. We also incorporated an individual weight in Chapter 8 and a cell weight in Chapter 9 for the same reasons explained above.

This research extends our knowledge of weighting schemes in rare variant association studies to incorporate information that can help detect real variants and shed new light on weighting schemes that can eliminate possible errors and may boost the power of the tests. The present study makes several noteworthy contributions to rare variant association studies and will serve as a basis for future studies in this field.

10.2 Future Research

To develop a full picture of rare variant association studies, additional studies will be needed. Three areas regarding rare variant association are in need of further research. One area involves combining two different weights based on the variant level or combining an individual weight and variants, as well as a cell weight. For example, a grid search could be used as the weight value in the test to find an optimal value that will maximize the power.

Additionally, in many types of medical research, the main interest lies in testing whether a random effect variance component is equal to zero within a mixed-effect model framework. The variance component test has been a statistical challenge for a long time and has received considerable attention in the literature (see, for example, Self & Liang (1987) and Stram & Lee (1994)). However,

little work has been conducted on a likelihood-ratio based variance component test in generalized linear mixed models (GLMM) where the response is discrete, and the log-likelihood cannot be precisely computed. The motivation to propose the use of LRT in rare variant association to determine whether type I errors can be controlled by using a permutation test arose when Fitzmaurice *et al.* (2007) examined the variance component test using permutation in GLMM to detect the cluster effect in binary datasets and demonstrated that type I errors were controlled effectively and had a higher sensitivity than the variance component test based on scaled chi-square distributions. The hypothesis is that the LRT might produce better results than the permutation variance component test. Before applying the LRT to the variance component in GLMM, several difficulties need to be overcome, including the computation of the log-likelihood, parameter estimation, and the derivation of the null distribution of the LRT statistic. To overcome these problems, we can make use of the penalized quasi-likelihood (PQL), which is the most common estimation procedure for GLMM, and calculate the LRT statistic based on the resulting working response and quasi-likelihood so that the LRT in GLMM will be computationally feasible. The permutation procedure could possibly be used to obtain the null distribution of the LRT statistic, or we could have the opportunity to propose a mixed chi-square.

Furthermore, incorporating additional functional information in the weighting scheme is still an open area of research, and finding an optimal weight still needs more attention. Using Bayesian statistics, such as the Bayes factor, is another method that we can consider in the near future. Therefore, the weighting scheme could be incorporated in a similar Bayesian manner as prior.

Finally, the simulation conducted in chapter 7 and 9 can be extended to generate data with some errors and compare the performance of the proposed down-weighting of the low-quality calls or variants with the usual practice of excluding calls with quality levels below an arbitrary threshold. In addition, randomness in the data could be investigated in the near future since we incorporate the weighting scheme, especially in the cell weight. The study would have been more interesting if it had included this kind of simulation. The inclusion of the weighting scheme in rare variant association analysis is important, especially for incorporating sequencing information. If the objective of the future research

above can be tamed and understood it will offer an entirely new way of thinking in rare variant association studies.

Bibliography

Asimit, J.L., Day-Williams, A.G., Morris, A.P. & Zeggini, E. (2012). ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Human Heredity*, **73**, 84–94. 36

Basu, S. & Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, **35**, 606–619. 41, 57

Bodmer, W. & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics*, **40**, 695–701. 6

Bonnefond, A., Clément, N., Fawcett, K., Yengo, L., Vaillant, E., Guillaume, J.L., Dechaume, A., Payne, F., Roussel, R., Czernichow, S. *et al.* (2012). Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nature Genetics*, **44**, 297–301. 6

Box, G.E. *et al.* (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, **25**, 290–302. 66

Breslow, N.E. & Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25. 84, 85

Burdett, N.T. (2017). GWAS Catalog.Ebi.ac.uk. 2

Cirulli, E.T. & Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, **11**, 415–425. 6

Cochran, W.G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 30, 178–191, Cambridge University Press. 66

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R. & Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, **305**, 869–872. 6

Cox, D.R. (1983). Some remarks on overdispersion. *Biometrika*, **70**, 269–274. 85, 89

Davies, R.B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **29**, 323–333. 67

Daye, Z.J., Li, H. & Wei, Z. (2012). A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Research*, **40**, e60–e60. 221

Duncan, E.L. & Brown, M.A. (2013). Genome-wide association studies. 2

Fitzmaurice, G.M., Lipsitz, S.R. & Ibrahim, J.G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics*, **63**, 942–946. 293

Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R. *et al.* (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, **42**, 1118–1125. 3

Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, **10**, 241–251. 6

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, **13**, 135–145. 4

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43–56. 84

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 45–51. 84

Gorlov, I.P., Gorlova, O.Y., Frazier, M.L., Spitz, M.R. & Amos, C.I. (2011). Evolutionary evidence of the effect of rare variants on disease etiology. *Clinical genetics*, **79**, 199–206. 6

Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 245–259. 84

Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Masson, G., Agnarsson, B.A., Benediktsdottir, K.R., Sigurdsson, A., Magnusson, O.T., Gudjonsson, S.A., Magnusdottir, D.N. *et al.* (2012). A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature Genetics*, **44**, 1326–1329. 5

Imhof, J. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, **48**, 419–426. 67

Ionita-Laza, I., 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**, e1001289. 40, 41

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D. & Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, **92**, 841–853. 106, 170

Jeng, X.J., Cai, T.T. & Li, H. (2012). Simultaneous discovery of rare and common segment variants. *Biometrika*, **100**, 157–172. 106, 170

Johnston, H.R., Hu, Y. & Cutler, D.J. (2015). Population genetics identifies challenges in analyzing rare variants. *Genetic Epidemiology*, **39**, 145–148. 115, 243

Kendall, M. & Stuart, A. (1977). The advanced theory of statistics. Vol. 1: Distribution theory. *London: Griffin, 1977, 4th ed.*. 87

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T. *et al.* (2005). Complement factorH polymorphism in age-related macular degeneration. *Science*, **308**, 385–389. 3

Knaus, B.J. & Grünwald, N.J. (2017). VCFR: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, **17**, 44–53. 223

Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, **80**, 727–739. 4

Lee, S., Wu, M.C. & Lin, X. (2012a). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, **13**, 762–775. 34

Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, **95**, 5–23. 6, 34, 41, 96

Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., Wray, N.R., Consortium, S.P.G.W.A.S. *et al.* (2012b). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics*, **44**, 247–250. 3

Li, B. & Leal, S.M. (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. 33, 38, 57

Lin, D.Y. & Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, **89**, 354–367. 65

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84**, 309–326. 87, 89

Liu, D., Ghosh, D. & Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, **9**, 292. 67, 68

Liu, H., Tang, Y. & Zhang, H.H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, **53**, 853–856. 67, 68

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B. *et al.* (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823–828. 5

Madsen, B.E. & Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, **5**, e1000384. 33, 38, 39, 40, 58, 180

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A. *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, **461**, 747. 4

McCrone, J.T. & Lauring, A.S. (2016). Measurements of intrahost viral diversity are extremely sensitive to systematic errors in variant calling. *Journal of Virology*, **90**, 6884–6895. 243

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models.*. London England Chapman and Hall 1983. 61

Morgenthaler, S. & Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **615**, 28–56. 33, 36, 57

Morris, A.P. & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, **34**, 188–193. 33, 36

Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K. & Daly, M.J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, **7**, e1001322. 34, 42, 43, 58

Newton-Cheh, C. & Hirschhorn, J.N. (2005). Genetic association studies of complex traits: design and analysis issues. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **573**, 54–69. 80

Neyman, J. & Scott, E. (1965). On the use of c (alpha) optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute*, **41**, 477–497. 42

Pan, W., Kim, J., Zhang, Y., Shen, X. & Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics*, **197**, 1081–1095. 93

Patel, Z.H., Kottyan, L.C., Lazaro, S., Williams, M.S., Ledbetter, D.H. *et al.* (2014). The struggle to find reliable results in exome sequencing data: filtering out mendelian errors. *Frontiers in Genetics*, **5**. 250, 276

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press. 63, 246

Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J. & Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, **86**, 832–838. 33, 152

Rivas, M.A., Voight, B.F., Devlin, B., Orho-Melander, M., Kathiresan, S., Roeder, K., Neale, B.M., Altshuler, D.M., Purcell, S. & Daly, M.J. (2011). Testing for an unusual distribution of rare variants. 41

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 15–32. 84

Rodriguez, G. & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 73–89. 85

Saint Pierre, A. & Génin, E. (2014). How important are rare variants in common disease? *Briefings in Functional Genomics*, **13**, 353–361. 6, 8

Self, S.G. & Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610. 292

Stiratelli, R., Laird, N. & Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 961–971. 83

Stram, D.O. & Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 1171–1177. 292

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J. *et al.* (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713. 3

Tony Cai, T., Jessie Jeng, X. & Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 629–662. 106, 170

Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, **90**, 7–24. 3

Wall, J.D., Tang, L.F., Zerbe, B., Kvale, M.N., Kwok, P.Y., Schaefer, C. & Risch, N. (2014). Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research*, **24**, 1734–1739. 271

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, **89**, 82–93. 34, 42, 44, 45, 58, 72, 96, 97, 152, 157, 158, 179, 180, 181, 195, 196, 279

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G. *et al.* (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, **43**, 519–525. 3

Zeger, S.L., Liang, K.Y. & Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049–1060. 83

Zelterman, D. & Chen, C.F. (1988). Homogeneity tests against central-mixture alternatives. *Journal of the American Statistical Association*, **83**, 179–182. 42

Zhang, D. & Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In *Random effect and latent variable model selection*, 19–36, Springer. 91, 93

Zhang, Q. (2015). Associating rare genetic variants with human diseases. *Frontiers in Genetics*, **6**. 57

Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R. & Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, **111**, E455–E464. 4, 5, 7