Statistical Methods for Next Generation Sequencing Data

by

Miao Zhang

A Dissertation Submitted to the Faculty of the

Graduate Interdisciplinary Program in Statistics

In Partial Fulfillment of the Requirements
For the Degree of

Doctor of Philosophy

In the Graduate College

The University of Arizona

2 0 1 8

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Miao Zhang entitled "Statistical Methods for Next Generation Sequencing Data" and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

____________________ Date: May 11, 2018
Jin Zhou

Joseph C Watkins ____________________ Date: May 11, 2018
Joseph C. Watkins

____________________ Date: May 11, 2018
Lingling An

____________________ Date: May 11, 2018
Ryan Gutenkunst

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.
I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

____________________ Date: May 11, 2018
Dissertation Director: Jin Zhou

Joseph C Watkins ____________________ Date: May 11, 2018
Dissertation Director: Joseph C. Watkins

Statement by Author

Signed: Miao Zhang

Acknowledgments

First and foremost, my heartfelt thanks go to my co-advisors: Dr. Jin Zhou and Dr. Joseph C. Watkins. They both provided me so much support over the course of my time at the University of Arizona. Joe has invested countless hours to train me in being productive as an academic, scientific writing, critical thinking, and much more. From Jin, I have learned how to be a statistician, especially how to think as a statistician. Her excellent guidance and her creative ideas were essential to the completion of my dissertation; her encouragement and our discussions are invaluable for my professional career. Interacting with Yves and Helen has not only shaped the way I think about single-subject analysis in precision medicine but has also improved the way I reason and communicate more generally.

I greatly appreciate many people in the program of Statistics GIDP. Thanks to Dr. Joe Watkins as Chair of the Statistics GIDP, for financial support and friendly discussions about many topics. A special thank you to Dr. Ryan Gutenkunst for serving on my committee and providing the guidance for my third project. Dr. Lingling An has provided me many help, guidance, and encouragement throughout my time in Statistics GIDP. And thank you my fellow students for their support and friendship.

I would like to show my gratitude to Dr. Joe Watkins again again for helping me in my most difficult time when my former advisor quieted and left.

Finally, I would like to thank my parents for their love and unconditional support over the years. Thank you to my aunt, grandmother and cousin for their support and encouragement. Their love helps me to survive the 6-year PhD study. Without them, it is very hard for me to have this achievement.

Contents

Contents—*Continued*

List of Figures

List of Tables

Abstract

Statistical genetics is a scientific field concerned with the development of statistical methods for drawing inferences from genetic data. Research in statistical genetics generally involves developing theory or methodology to support research in one of three related areas: population genetics, genetic epidemiology and quantitative genetics. This dissertation is an ensemble of my research work in statistical genetics, including three projects with varying focuses. The first project applies a rare variant region-based test to identify sets of common or rare variants aggregated in and around genes associated with Dravet Syndrome. The second project proposes a score-based test to investigate the association for a set of rare variants and ordinal traits. The third project describes an implement of dimensionality reduction method in genotype data for population inference.

Advisor: Jin Zhou Joseph C. Watkins

Chapter 1
Introduction

1.1 Next Generation Sequencing Data

Next generation sequencing (NGS) technologies (Mardis, 2008) (DePristo et al., 2011) are high-throughput parallel-sequencing approaches, which can generate billions of short sequence reads for a modest cost. The advances in NGS technologies provide the new approach to large-scale resequencing of human samples for medical and population genetics. It has led to new approaches to enable exome and whole-genome sequencing (WGS) (Bick and Dimmock, 2011) studies of complex diseases. In recent years, the price of sequencing has fallen dramatically, enabling whole-exome sequencing (WES) and whole-genome sequencing (WGS) studies to have sufficient statistical power to tackle association studies of complex diseases.

1.2 Association Studies

1.2.1 Genome-wide Association Studies (GWAS)

In statistical genetics, GWAS (Hirschhorn and Daly, 2005) is an examination of a set of genetic variants in different individuals to assess whether any variant is associated with a given disease or trait. In the last decade, GWAS have been extensively used (Klein et al., 2005) (MacArthur et al., 2017) (Burton et al., 2007) (Clarke et al., 2011) to dissect the genetic architecture of complex diseases and quantitative traits. These studies systematically evaluate common variants, typically defined as having a minor allele frequency (MAF) $> 5\%$. More than 2,000 disease-associated common variants have been identified through GWAS, which provided many new clues about disease biology (Hindorff et al., 2011). Despite these discoveries, GWAS was designed to identify common variants contributing to common disease and these variants are usually located in intronic regions of the chromosome, which do not code for specific functional proteins. Instead, to claim an association between the gene and a disease, these single-nucleotide polymorphisms (SNPs) either have to lie hundreds of kilobases (kb) up or downstream from a known causative gene, or the SNPs have to be in linkage disequilibrium with other SNPs of potential interest, meaning that they would have to be associated more frequently than would be expected randomly. Thus, a SNP's involvement is inferred due to its being in a place or associated with a gene known to have an effect on the disease or pathology of interest.

1.2.2 Rare-variant Association Tests

Due to the fact SNPs identified in GWAS are usually not casual, much of the genetic contribution to complex traits remains unexplained. Conceding that that single genetic variations cannot account for much of the heritability of diseases, behaviors, and other phenotypes, we denote the issues of as-yet unexplained traits as the problem of missing heritability (Manolio et al., 2009) (Eichler et al., 2010). By contrast, WES looks at exonic regions of the chromosome, which code for functional proteins. This approach is more likely to identify mutations, which are often rare. These variants can be broadly classified as low-frequency ($0.5\% < \text{MAF} < 5\%$) and rare ($\text{MAF} < 0.5\%$) variants and it could explain additional disease risk or trait variability. However, detecting the association of the rare variants requires either a large effect size or a large sample size if we are able to use GWAS methods to analyze a single SNP at a time. To address these issues, region-based methods are designed for rare-variant association analysis to boost power. These methods evaluate association for multiple variants in a biologically relevant region, such as a gene. Among the proposed region-based tests (Lee et al., 2014), burden tests and variance-component tests (Li and Leal, 2008) (Morgenthaler and Thilly, 2007) (Madsen and Browning, 2009) (Pan, 2009) (Neale et al., 2011) are two commonly-used methods and implemented in various packages. Burden tests collapse rare variants into genetic scores by assuming that a large proportion of variants are causal and their effects are in the same direction. In the contrast, variance-component tests assume that a small fraction of causal variants and effects might be in different directions.

1.2.3 Ordinal Trait

In genetic epidemiology studies, many human conditions (e.g., cancer and most behavioral and psychiatric disorders) are measured in a discrete ordinal scale, and their levels represent the degree of severity of a symptom. For example, doctors may record intellectual or motor disability as normal, mild, moderate, or severe. Moreover, critical information is lost if we reduce our data to a simply binary trait, e.g., mild vs severe. For example, epilepsy is a group of neurological disorders characterized by epileptic seizures, and commonly found in people with intellectual disability (ID). Physicians may record intellectual or motor disability/ID as normal, mild, moderate, or severe to present the severity of the patients. However, most statistical procedures are designed for either continuous or binary outcomes. Because no closed form of the likelihood function is known, association tests of rare variants and ordinal traits have two potential challenges: (1) deriving an effective test statistic and (2) bias due to the sparsity of genotype data.

1.3 Low Coverage Data and Genotype Dosage Value

Sequencing depth refers to the average number of reads that cover each base. Deep WGS can identify nearly all variants in genome with high confidence but is usually impractical because of the high cost. By contrast, low-depth is a cost-effective alternative and allows to include more individuals. In comparison to high coverage sequencing, low coverage sequencing relies on linkage-disequilibrium (LD)-based methods and result in higher genotyping error rates. Researchers developed methods to take genotype uncertainty into account. Instead of basing the analysis on called genotypes, it works directly on the raw sequencing data or by using genotype likelihoods/posterior probability of genotype. Based on posterior probability of genotype, we can calculate the genotype dosage value, which is a continuous variable between 0 and 2 and represents the predicted dosage of the non reference allele given the data available. It indicates how well the genotype is supported by the imputation engine.

1.4 Principal Component Analysis

Principal component analysis (PCA) is a very powerful technique that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations. PCA projects the original variables on a low dimensional space embedding that maximally explain variance of the data. PCA has become a standard tool in genetics. It is the most widely and popular used approach in Genome-Wide Association Studies (GWAS) for identifying and adjusting for ancestry difference among sample individuals. (Price et al., 2006) proposed a method (EIGENSTRAT) to correct for population stratification in GWAS to avoid the false positive due to differences in allele frequency among groups. It uses the top principal components from PCA as covariates in a multilinear regression model to correct for sample structure. PCA is also used to infer population structure in genetic data (Patterson et al., 2006). A Tracy-Widom-like statistic is developed to compute the p-value corresponding to the significance in population.

1.5 Proposed Work

1.5.1 Region-based Association Test for Ordinal Traits

To address the challenges raised in ordinal traits, we develop a computationally very efficient score-based test to investigate the association for a set of rare variants and ordinal traits. Through simulation studies, we demonstrate that our method is more powerful when the trait is ordinal in comparison to other methods. In addition, we showed our method is able to control the type I error very well even when the model is misspecified.

1.5.2 Non-linear PCA for Population Inference and Stratification

The sequencer can not identify a certain number of bases with sufficient confidence at low coverage. To handle the genotype uncertainty in low-depth sequencing date, we applied a data-adaptive kernel PCA to do data dimensionality reduction by using the genotype dosage value. It considers possibly nonlinear correlations among markers and uses different mapping patterns for each feature.

1.6 Organization of This Dissertation

My dissertation is composed of the three projects. The overarching goal is to develop innovative statistical methods to solve the genetic problems raised from NGS data. Each of Chapters 2-4 summarizes one project by stating the objective, the proposed methods, and the main results. The concluding chapter provides an overall discussion and projects future directions.

Chapter 2

Region-Based Association Test and Its Application on Dravet Syndrome

2.1 Abstract

Dravet syndrome (DS) is a rare, devastating form of childhood epilepsy. It begins in the first year of life in a healthy infant with frequent seizures. As seizures continue, most children develop some level of disability and other conditions associated with the syndrome. DS is often associated with mutations in the voltage-gated sodium channel gene, SCN1A. Haploinsufficiency of SCN1A, either through mutations or deletions, is the major cause of Dravet Syndrome. There is considerable variability in expressivity within families, as well as among individuals carrying the same primary mutation, and it suggests that clinical outcome might be modulated by variants at other genes, which we call the modifier genes. In this project, we identified 22 patients from a Japanese cohort with either mild or severe clinical expression as determined by cognitive and motor skills to search for modifier genes. Then we sequenced the exomes of these patients and performed rare variant region-based test, SNP-set kernel association test (SKAT) to identify sets of both common and rare variants aggregated in and around genes associated with the phenotype. A gene, FKBP5 on chromosome 6 was found statistically significant. Interestingly, this finding suggests that sex-associated factors may play a role in this association. Furthermore, we investigated whether this finding is a false positive reason due to a bias in the adjustment of motor delay, a possible heritable covariate.

2.2 Introduction

In statistical genetics, genome-wide association studies (GWAS) (Hirschhorn and Daly, 2005) is an examination of a set of genetic variants in different individuals to a given disease or trait. In the last decade, GWAS have been extensively used (Klein et al., 2005) (MacArthur et al., 2017) (Burton et al., 2007) (Clarke et al., 2011) to dissect the genetic architecture of complex diseases and quantitative traits. These studies systematically evaluate common variants, typically defined as having a minor allele frequency (MAF) $> 5\%$. More than 2,000 disease-associated common variants have been identified through GWAS, which provided many new clues about disease biology (Hindorff et al., 2011). Despite these discoveries, much of the genetic contribution to complex traits remains unexplained.

Conceding that that single genetic variations cannot account for much of the heritability of diseases, behaviors, and other phenotypes, the issues of as-yet unexplained traits is commonly referred to as the problem of missing heritability (Manolio et al., 2009) (Eichler et al., 2010).

Several explanations have been proposed to account for this missing heritability (Zuk et al., 2012). Since GWAS was predicated on the common disease common variant (CDCV) hypothesis (Cirulli and Goldstein, 2010), which assumes that common diseases are caused by common variants of small to modest effect. A plausible alternative explanation, the common disease rare variant (CDRV) hypothesis suggests low-frequency and rare variants could explain additional trait variability. CDRV requires substantially different statistical approaches. Because the CDRV hypothesis postulates that multiple rare variants work collectively to influence common diseases with moderate to large effect, substantially different, new and innovative statistical methods were necessary.

The advent of next-generation sequencing technology (Mardis, 2008) has led to new approaches to enable exome and whole-genome sequencing (WGS) (Bick and Dimmock, 2011) studies of complex diseases. Next generation sequencing (NGS) technologies are high-throughput parallel-sequencing approaches, which can generate billions of short sequence reads for modest cost. In recent years, the price of sequencing has fallen dramatically, enabling whole-exome sequencing (WES) and whole-genome sequencing (WGS) studies to have sufficient statistical power to tackle association studies of complex diseases. Detecting the association of the rare variants requires either a large effect size or a large sample size if we are able to use GWAS methods to analyze a single SNP at a time. In its place, region-based methods are designed for evaluating association for multiple variants in a biologically relevant region, such as a gene. Lee et al. (Lee et al., 2014) summarized various region-based association tests.

Epilepsy disorders caused by sodium channel mutations represent a disease phenotype, which may be strongly influenced by modifier genes. Truncation mutations in the voltage-gated sodium channel gene, SCN1A, usually result in Dravet Syndrome (DS) (Mulley et al., 2005) (Depienne et al., 2008) (Ishii et al., 2017). Most affected individuals are sporadic cases whose causal variant is likey a *de novo* mutation; however, there are known examples of families in which an SCN1A truncation mutation is transmitted through multiple generations, with some members exhibiting typical DS, and others with mild forms of epilepsy. Similarly, SCN1A missense variants often segregate in families in which there are members with DS and milder forms of epilepsy, along with unaffected carriers. Studies in mice have mapped several modifiers of SCN1A, many of which turn out to be genes coding for other ion channels. In this project, we employ an extreme phenotype approach to search for genetic modifiers in a cohort of Japanese patients with SCN1A truncation mutation. We sequenced the exomes of 22 individuals (Hammer et al., 2017) diagnosed either mild or severe. In order to control the variation associated with different mutation types, we limited inclusion to individuals with a *de novo* truncation mutation. We performed region-based test aimed at identifying sets of common or rare variants aggregated in and around genes associated with the clinical outcome.

This chapter is developed as follows. First, we introduce statistical methods for region-based association. Second, we implement these methods on DS data and the results from the association study (Hammer et al., 2017). Finally, we provided the biological interpretation behind the statistical result.

2.3 Statistical Model

Among all kinds of region-based tests, burden test and variance-component test (Li and Leal, 2008) (Morgenthaler and Thilly, 2007) (Madsen and Browning, 2009) (Pan, 2009) (Neale et al., 2011) are two commonly-used methods and implemented in various packages. Burden test collapse rare variants into genetic scores by assuming a large proportion of variants are causal and effects are in the same direction. In the contrast, variance-component test assumes a small fraction of causal variants and allows the effects to be in the different directions, to test variance of genetic effects.

Assume n subjects are sequenced in a region with p genetic variants. For subject i, let Y_i denote the phenotype, $X_i = (X_{i1}, \cdots, X_{im})'$ are m covariates, and $G_i = (G_{i1}, \cdots, G_{ip})'$ are genotypes for p genetic variants with the region. Here we consider the the phenotype to be a binary trait (e.g. $y = 0$ or 1 for control or case), and assume an additive genetic model, which $G_{ij} = 0, 1$ or 2 representing the number of minor alleles carried by the individual. Thus, we can relate the sequence variants in the region to the binary outcome in a logistic model

$$\text{logitPr}(Y_i = 1) = \alpha_0 + \alpha' X_i + \beta' G_i$$

where α_0 is an intercept, $\alpha = [\alpha_1, \cdots, \alpha_m]'$ is a vector of coefficients for covariates and $\beta = [\beta_1, \cdots, \beta_p]'$ is a vector of coefficients for genetic variants within a region, and the null hypothesis is that there is no genetic association,

$$H_0 : \beta = 0$$

By taking the first derivative of the log-likelihood function, $\sum_{i=1}^{n} \{y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)\}$ with respect to β, we can obtain the marginal score statistic for variant j under the null as

$$S_j = \sum_{i=1}^{n} G_{ij}(y_i - \hat{\mu_i})$$

where $\hat{\mu_i}$ is the estimated mean of y_i under the null hypothesis. In the next three subsections, I will introduce the burden test, variance-component (Sequence Kernel Association Test, SKAT (Wu et al., 2011)) and the combined burden and SKAT.

2.3.1 Burden Test

Burden test collapses information for multiple genetic variants into a single genetic score and tests for the association between the score and the phenotype. The summary combined score for subject i is

$$C_i = \sum_{j=1}^{p} w_j G_{ij}$$

where w_j is a weight for variant j. Modeling assumption of weights are often based on biological mechanism and empirical considerations. The first and simplest case is to have C_i count the total number of minor alleles carried by individual i. Thus, set $w_j = 1$ for all j. If we have strong evidence that only rare variants are associated with the outcome, a reasonable choice is to set $w_j = 1$ if minor allele frequency at the jth site, $MAF_j \leq c$ and c is a threshold. A third choice, used in SKAT is to replace the threshold by the square of a beta distribution, $w_j \sim Beta(\alpha_1, \alpha_2)$, with prespecified parameters α_1 and α_2. This choice accommodates the desire to increase the weight of rare variants (if α_1 is much less than α_2) while still putting some weight for common variants. In addition, Price (Price et al., 2010a) developed a Variable-Threshold (VT) approach by taking the maximum of the test statistics (i.e., Z-scores) over all allele-frequency thresholds and assessing statistical significance using permutation methods. In the SKAT R package, the default weight is beta distribution density function of the MAF with parameters 1 and 25, also called Wu's weights (Lee et al., 2012). The corresponding score statistic to test the null hypothesis is

$$Q_{\text{burden}} = (\sum_{j=1}^{p} w_j S_j)^2$$

and a $p - value$ can be obtained by comparison to χ^2 distribution with 1 degree of freedom.

2.3.2 SKAT

SKAT tests the association by evaluating the distribution of the genetic effects for a group of variants by assuming each β_j follows an arbitrary distribution with a mean of zero and variance of $w_j\tau$, where τ

is a variance component. Under this assumption, the original null hypothesis $H_0 : \beta = 0$ is equivalent to testing $H_0 : \tau = 0$. The test statistic of SKAT can be written as

$$Q_{\text{SKAT}} = \sum_{j=1}^{p} w_j^2 S_j^2$$

SKAT has been shown to be more powerful than burden test whenever a region contains many non-causal SNPs that have different directions of association. In contrast, if the percentage of causal SNPs in a region is high and they work in the same direction, the burden test is more powerful than SKAT.

2.3.3 Combine burden and SKAT

Because either use burden test or variance-component test depends on our understanding of the genetic scenarios, it may be desirable to combine these two approaches to boost the power to detect the association. One approach is to combine the burden and SKAT test statistics. Lee (Lee et al., 2012) proposed a linear combination of burden and SKAT test statistics:

$$Q_\rho = (1-\rho)Q_{SKAT} + \rho Q_{burden}, 0 \leq \rho \leq 1$$

where the parameter ρ is interpreted as a pairwise correlation among the genetic-effect coefficients β_j. Since ρ is unknown in practice, they developed an adaptive procedure, which approximates the test by using an optimal value of ρ estimated corresponding to the minimum p-value over a grid search of ρ's. This method is implemented in the SKAT R package, called SKAT-O.

2.4 Dravet Syndrome Dataset

In this project, we employ an extreme phenotype approach to search for genetic modifiers in a cohort of Japanese patients with SCN1A truncation mutation. In order to control the variation associated with different mutation types, we limited inclusion to individuals with a *de novo* truncation mutation resulting in SCN1A haploinsufficiency. We sequenced the exomes of 22 individuals diagnosed either mild or severe. Details of patients' information, the exome sequencing procedure, and variant filtering are provided in Hammer et al. (Hammer et al., 2017). To fit our data into existing methods, we take severity as our outcome and denote mild=0 and severe=1. Among those 22 patients, 12 are chosen of the most severely affected patients – those that were classified as severe in the shortest times after seizure onset (age range 57-150 months), and 10 are chosen of the most mildly affected patients –

those that remained mild (i.e., normal, border or mild) for the longest period of time after seizure onset (age range 68-267 months). When we resurveyed physicians in 2016, one mild patient (SMEI 458) progressed to moderate phenotype and another mild patient (SMEI 584) progressed to severe phenotype. In addition, one subject (SMEI 537) is diagnosed as Korean offspring. Thus, we consider 5 possible sample settings (Table 1.1) in our association studies.

Model	Number of Subjects	Subject Removed
1	10 mild: 12 severe	None
2	10 mild: 11 severe	SMEI 537 (Korean offspring)
3	9 mild: 11 severe	SMEI 537 (Korean offspring) and SMEI 458 (mild to moderate)
4	9 mild: 12 severe	SMEI 458 (mild to moderate)
5	9 mild: 11 severe	SMEI 458 (mild to moderate) and SMEI 584 (mild to severe)

Table 2.1: 5 different sample settings from 22 Japanese DS patients

To search for modifier genes, we consider gene-based association tests on the WES data to assess the joint effects of multiple SNPs in a region on a binary outcome phenotype (mild or severe). There are a total of 20,656 genes included in the association tests. To maximum the power to find possible modifier genes, we used an adaptively combined burden test and variance-component test (SKAT) to search for common and/or rare variants aggregated in and around genes throughout the genome in our exome sequence data. We also adjust for four covariates, age of seizure onset, time since onset (to present for mild, to first diagnosis of severe for severe), gender and motor delay. We obtain parameter estimates and residuals under the null hypothesis, which assumes there is no genetic association between multiple SNPs and phenotypic outcomes. In addition, for each gene, SKAT package analytically calculates a p-value for association.

2.5 Results

2.5.1 Association test

For all 5 models described above, we find that exactly one gene region, FKBP5 on chromosome 6 shows a highly statistically significant association with phenotypic class for all models (Table 1.2). There are 13 variants in this region, 5 of which are common SNPs (MAF$\geq$ 5%). The MAFs for those 13 SNPs shows in Table 1.3. Here, we focus on Model 5, which two patients were removed due to the fact they progressed to more severe conditions. Table 1.4 shows the top 20 most significant genes in Model 5. In addition, Figure 1.1 shows the Manhattan plot in Model 5. Importantly, statistical significance was greatly reduced when we did not include gender as a covariate ($p = 1.7 \times 10^{-3}$), which suggests that sex might be an important driver of this association. Next, we checked a few possibilities of false positive.

Model	p-value
1	9.1×10^{-7}
2	2.1×10^{-6}
3	1.2×10^{-6}
4	6.2×10^{-7}
5	5.7×10^{-8}

Table 2.2: p-values of FKBP5 in 5 different models

Func.refGene	POS	ID	MAF(DS cohort)	MAF(All Human:UCSC)
exonic	35544987	rs78178707	0.02273	0.003195
intronic	35548180	rs77550012	0.04545	0.002796
intronic	35548236	rs7757037	0.4318	0.527356
UTR3	35554624	rs201661401	0.1364	-
intronic	35555110	rs2294807	0.04545	0.055911
intronic	35564917	rs112917942	0.04545	0.016973
intronic	35610215	rs9348979	0.02273	0.197484
intronic	35610308	rs6902124	0.5	0.352835
exonic	35610590	rs146032538	0.1364	0.007897
intronic	35610681	rs2143404	0.2273	0.127596
intronic	35610751	rs12527329	0.04545	0.053115
intronic	35632584	rs79540155	0.02273	0.001997
intronic	35632727	rs144270670	0.04545	0.007788

Table 2.3: 13 SNPs in FKBP5

2.5.2 Bias from adjusting heritable covariates

Given the correlation between severity (mild and severe) and motor delay is 0.612 (Pearson and Spearman), we first wanted to ask whether the signal was from one or both of these covariates. H Aschar et al. (Aschard et al., 2015) argued that adjusting for heritable covariates can introduce bias to estimates in GWAS. The main goal to adjust covariates is to account for potential confounding factors that can bias SNP-effect estimates and improve statistical power by reducing residual variance. Based on this notion, recently researchers have also begun to adjust for other heritable covariates. They were motivated by the goal of identifying genetic variants associated only with the primary outcome. However, an important difference between environmental/demographic factors and heritable human

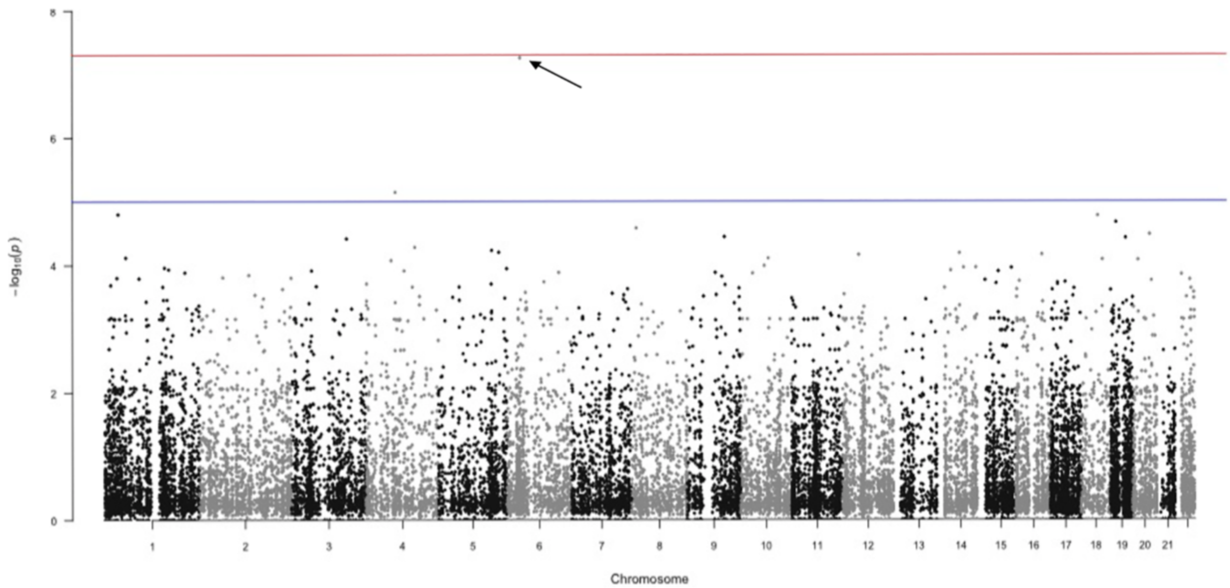


Figure 2.1: Manhattan plot in Model 5 (the arrow points KFBP5)

Gene Name	Chromosome	Set Size	p-value
FKBP5	6	16	5.7×10^{-8}
CMAHP	6	20	7.7×10^{-6}
NDUFB7	19	3	2.1×10^{-5}
RPL11	1	9	2.9×10^{-5}
ECI1	16	6	3.3×10^{-5}
SH2B3	12	7	3.2×10^{-5}
ZSWIM3	20	9	3.5×10^{-5}
LIPM	10	19	3.8×10^{-5}
ST8SIA5	18	10	4.1×10^{-5}
LINC01269	14	4	4.5×10^{-5}
DPT	1	8	5.2×10^{-5}
SLC25A31	4	7	5.2×10^{-5}
CCPG1	15	3	5.2×10^{-5}
GPR150	5	3	5.5×10^{-5}
CECR3	22	2	5.9×10^{-5}
RTN1	14	12	6.2×10^{-5}
PSMC6	14	8	6.3×10^{-5}
LINC01541	18	4	6.6×10^{-5}
PRICKLE1	12	20	6.8×10^{-5}
GALR3	22	6	7.1×10^{-5}

Table 2.4: p-values of FKBP5 in 5 different models

traits is that the latter have genetic associations in that the adjusted and unadjusted estimated effects of the genetic variant on the outcome will differ. Therefore, our questions are: is the signal from both the motor delay and severity? Is the signal from the motor delay only? In order to answer these questions, we switched primary outcome to motor delay and severity to a covariate. Under this model assumption, FKBP5 no longer shows a statistically significant association ($p = 0.022$). In addition, we also excluded severity from the covariates in regression model and FKBP5, again, does not show a statistically significant association ($p = 0.381$). Therefore we exclude the possibility that the signal from FKBP5 is due to bias from adjusting covariates.

2.5.3 Redefine motor delay variable

Because the variable motor delay is relatively high correlated to the trait, we also consider the model, which motor delay is treated as categorical variable (0=bedding, 1=standing, 2=walking and 3=running). Then we set bedding as reference group with other 3 binary variables where standing, walking and running. However, no signal found on chromosome 6. With fewer sample size, power is likely to be very low.

2.5.4 Variable selection

Then we attempted to detect the causal SNPs within FKBP5. We implemented a forward algorithm by including one SNP in the null logistic model at a time to calculate the p-value. We dropped one SNP out of 13 at a time by including the SNP as covariate and Table 1.5 shows the change of p-value after modification. Based on p-value, 5 SNPs lead to power loss after subtracting the SNP effect. The p-value increases dramatically when we adjust 5 SNPs effects in the regression model. We only use these 5 SNPs for association test, the p-value is 0.0094. This leads to the conclusion that these 5 SNPs are possibly causal SNPs within gene region FKBP5. A stronger statement would require a larger data set and an investigation into the genetic background of these 5 SNPs.

2.5.5 Haplotype analysis

We used SHAPEIT to phase the genotype data by considering 12 SNPs, which pass the filtering in FKBP5. One SNP has missing SNP ID and one SNP, rs201661401, is not qualified for the association study due to fail the filtering criterion. Table 1.6 summarizes the haplotype data. FKBP5 data set has total of 13 haplotypes among the 22 subjects. Four haplotypes (C', C", E' and F') appear only one time. Haplotype A and C have the highest frequencies. and we observe there is a pattern between mild and severe, male and female in those two haplotypes. We plot the histogram of severity and gender for all haplotypes (Figure 1.2 and 1.3). Based on the histograms, haplotype A and C appear more frequently (Table 1.7) in severe group than mild group, and haplotype C appear more

SNP ID	p-value
rs112917942	5.8×10^{-8}
rs12527329	1.5×10^{-7}
rs144270670	4.1×10^{-8}
rs146032538	1.2×10^{-6}
rs201661401***	5.0×10^{-1}
rs2143404	5.8×10^{-8}
rs2294807	6.1×10^{-8}
rs6902124***	1.0×10^{-3}
rs77550012	3.9×10^{-8}
rs7757037	4.0×10^{-6}
rs78178707***	5.0×10^{-1}
rs79540155***	1.9×10^{-3}
rs934879***	5.0×10^{-1}

Table 2.5: Detecting causal SNPs within FKBP5 by using forward algorithm

frequently in female patients than male patients. For male subjects, the haplotype A has equal frequency between mild and severe groups. However haplotype A is only found in female, who is diagnosed severe. It suggests that haplotype A may have negative effects on female. For male subjects, the haplotype C only is found in the severe group, and it is observed both in mild and severe but the percentage in severe is a little higher than mild in female. It suggests that haplotype C probably has a negative effect on both male and female.

Haplotype ID	Count	Freq
A	8	0.182
B	2	0.045
C	13	0.295
C'	1	0.023
C"	1	0.023
D	3	0.068
E	5	0.114
E'	1	0.023
F	3	0.068
F'	1	0.023
G	2	0.045
H	2	0.045
I	2	0.045
Total	44	1

Table 2.6: Haplotye information of FKBP5

	Mild	Severe
A	2	6
C	3	7

Table 2.7: A 2 by 2 table of total number of haplotype A and C between two groups

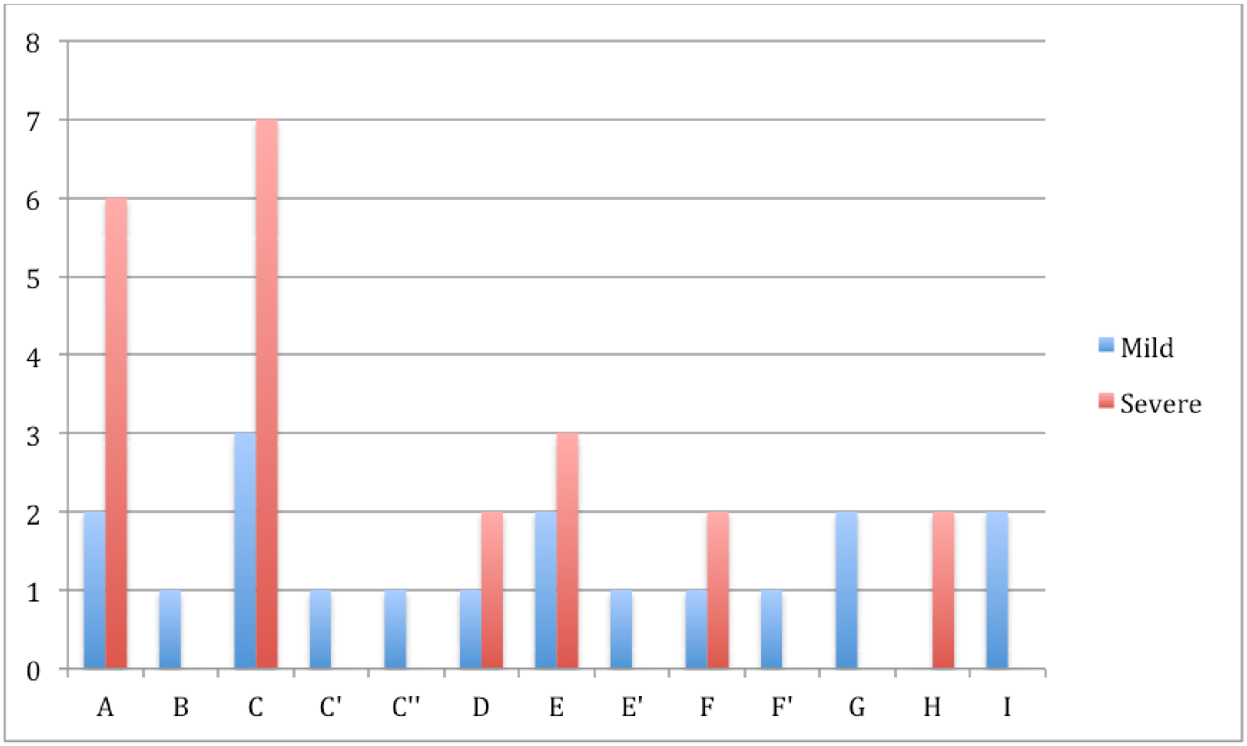


Figure 2.2: Histogram of haplotypes between mild and severe

2.5.6 Investigating the score test statistic

In an effort to detect the causal SNP, we performed a logistic regression anslysis using the same four covariates and computed the contribution of each SNP to the test statistic following the methods of Lee et al. The total score statistic using equal weights is 336.7 (mean = 25.9 sd= 32.5). The partial scores for each SNP showed that nearly half of the signal comes from two SNPS: rs9348979 and rs79549155. For these two SNPs, only a single subject in our study carried the minor allele. Moreover, this particular individual in the mild group (SMEI 733) has a large residual (-8.84). When we remove SMEI 733 from the analysis, the total score statistic is reduced to 148.4 (mean = 12.4, sd=15.6). This result suggests that our finding of an association between FKBP5 and clinical severity may be biased by inclusion of a single female individual in the mild group.

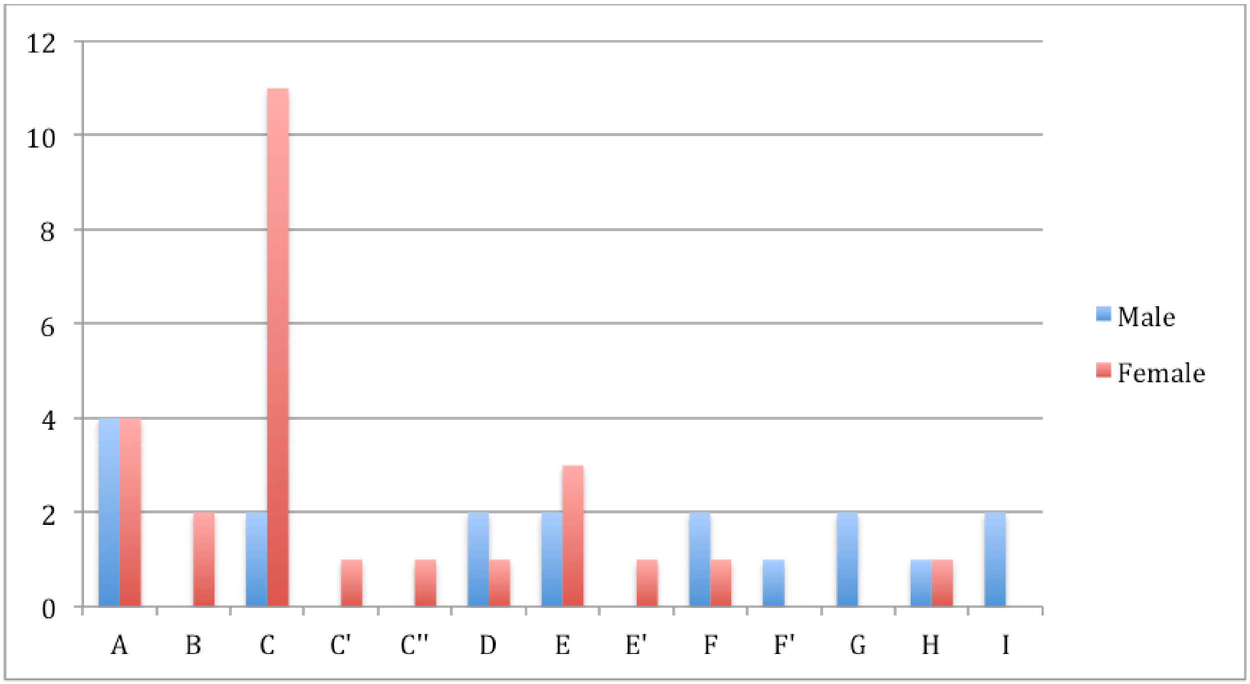


Figure 2.3: Histogram of haplotypes between male and female

2.6 Discussion

An important goal in human genetics is to identify genetic factors that modulate the phenotype. Characterization of modifier genes not only leads to a better understanding of syndrome-specific pathophysiology, it also helps to facilitate novel therapeutic approaches. In order to search for genetic variants affecting the severity of DS, we sequenced the exomes of 22 Japanese patients with SCN1A *de novo* truncation mutation. We controlled for the effect of mutation type on clinical outcome. By restricting our study to Japanese patients, we had hoped to minimize the power loss due to the possibility of more than a single common modifier. A liability of this approach is that our sample size was limited by the rarity of patients with both haploinsufficiency and mild clinical symptoms. By using the gene-based association test, SKAT, we found 13 SNPs positioned across the FKBP5 locus produced a statistically significant association. Interestingly, this association appears when clinical severity was selected as the outcome and gender as a covariate. This association disappeared when gender was ignored, suggesting that sex-associated factors played a role in this association. However, we also found that statistical significance was lost when gene-based association analysis was performed after removing a single female in our cohort, suggesting that that this association may be a false positive. Further work with a much greater sample size is needed to confirm an association between variation at FKBP5 and clinical outcome in DS.

Clinical assessments are more often recorded in a discrete ordinal scale, and their levels stands the degree of severity of a symptom. In my second chapter, I extend the region-based association tests from a binary response to a ordinal categorical outcome. For the additive regression model, the genetic coefficient gives the population increase in the mean phenotype as one alternative allele added to the genotype. For the additive genotype coding, the genotype value represents the number of alternative alleles carried by the subject at a certain locus. However, we only can observe discrete additive coding (0, 1 and 2). Instead of using discrete additive genotype, it is plausible to use a continuous value between 0 and 2, called the genotype dosage value to represent the predicted dosage of the non-reference allele given the data. In my third chapter, I discuss the implement of a data-adaptive kernel PCA on the dosage value for population inference.

Chapter 3

Association Test for Ordinal Categorical Outcome

3.1 Abstract

In Chapter 1, we reviewed the statistical methods for rare variants association tests and applied on Dravet Syndrome cohort to search for modifier genes. However, clinical assessments are often recorded using categorical ordinal data in genetic epidemiology studies. For example, doctors may record intellectual or motor disability as normal, mild, moderate, or severe. Moreover, critical information is lost if we reduce our data to a simply binary trait, e.g., mild vs severe. Due to the fact GWAS findings only explain a small proportion of heritability, group tests have been widely developed to detect the association of rare genetic variants in sequencing studies. However, most statistical procedures are designed for either continuous or binary outcomes. In this chatper, we present a computationally very efficient score-based test to investigate the association for a set of rare variants and ordinal traits. Through simulation, we evaluated the performance of our method. The proposed statistic obtains an accurate type I error rate, and it has higher power than many comparable binary trait methods. As a score-based test, our method can quickly calculate p-values, and so can easily be applied to genome-wide data sets of many hundreds of individuals. We apply this methodology to an epilepsy exon sequencing study to show the practical relevance of this approach.

3.2 Introduction

In genetic epidemiology studies, many human conditions (e.g., cancer and many behavioral and psychiatric disorders) are measured in a discrete ordinal scale, and their levels represent the degree of severity of a symptom. For example, epilepsy is a group of neurological disorders characterized by epileptic seizures, and commonly found in people with intellectual disability (ID). Physicians may record intellectual or motor disability/ID as normal, mild, moderate, or severe to present the severity of the patients. The prevalence of epilepsy increases with the severity of ID (Kumari et al., 2013). showed lower IQ score in children with ID has association with occurrence of epilepsy. A major challenge in epilepsy research is to unravel the complex genetic mechanisms underlying both common and rare forms of epilepsy. In the past two decades, great progress has been made in exploring the genetic bases of epilepsy.

GWAS have been successful in identifying risk factors for common diseases, and have been applied to epilepsy (Heinzen et al., 2010) (Guo et al., 2011). As mentioned in Chapter 1, there is growing interest in the possibility that variants too rare to be represented well on GWAS chips may be important contributors to common diseases, including various types of epilepsy. Indeed, (Veeramah et al., 2013) has suggested that the impact individual single-nucleotide variants in this disease is small, and indicates possibly that gene-based approaches might be more successful for future sequencing studies of epilepsy predisposition.

The advent of next-generation sequencing technology has led to new approaches for comprehensive characterization of variants across the genome. The rare variant hypothesis has been supported by the recent observation that rare copy number variants (CNVs) can have large effects on neuropsychiatric disease risk (Need and Goldstein, 2010). The corresponding gene-based tests have been developed for both binary and quantitative traits (Lee et al., 2012) (Lee et al., 2014). Because clinical assessments are frequently given by an ordinal scale (Zhang et al., 2003) (Feng et al., 2004), an unnecessary collapse of trait levels could reduce the power in association studies. Therefore, methods need to be developed to investigate the association of genes and ordinal traits. (Zhang et al., 2006) proposed an ordinal transmission/disequilibrium test (O-TDT), and (Wang et al., 2006) presented a class of family-based association tests (FBATs) for ordinal traits. Both required nuclear family data and designed for testing for association between genetic a single variant. Due to the unusually large number of zeros in rare variants, the genetic coefficients can be hardly estimated. These approaches have limited power unless sample sizes are very large, the effects are substantial, or the variants are not too rare.

Epilepsy associated genes mostly has been identified in *voltage-gated* or *ligand-gated* ion channels. Some individuals do not inherit epilepsy but rather the epilepsy is the result of a mutation occurring in the germ cell line of the parents. Such *de novo* mutations are extremely important in sever myoclonic epilepsy of infancy (SMEI) (Özmen et al., 2011). One example is Dravet syndrome (DS). Dravet syndrome is a rare, devastating form of childhood epilepsy, which is often associated with a truncation mutation in the voltage-gated sodium channel gene, SCN1A. There is considerable variability in expressivity within families, as well as among individuals carrying the same primary mutation, which suggests that clinical outcome is modulated by variants at other genes. Our interest is to search for the possible modifier genes. To identify modifier gene variants that contribute to clinical outcome of DS we sequenced the exomes of 22 individuals diagnosed as either mild or severe cognitive condition. We controlled for variation associated with different mutation types by limiting inclusion to individuals with a *de novo* truncation mutation resulting in SCN1A haploinsufficiency. In psychiatry and schizophrenia research, the clinical global impression - severity scale (CGI-S) (Suzuki et al., 2012) is measure of symptom severity. CGI-S has 7 categories and each represents the severity of the patient's illness at the time of assessment, relative to the clinician's past experience with patients who have the same diagnosis. Physicians classified the patients in different categories (normal, border, mild, moderate and severe) on the basis of IQ and DQ assessment tests. The trait is characterized in a discrete ordinal scale, and their levels represent the degree of severity of DS.

To our knowledge, statistical procedures of rare variants are designed for either continuous or binary outcomes. Critical information and power may be lost if we reduce our data to a simply binary trait, especially when sample size is small. Because no closed form of the likelihood function is known, association test of rare variants and ordinal traits has two potential challenges: (1) deriving an effective test statistic and (2) the bias due to the sparsity of genotype data. In this chapter, we present a computational efficient score-based test to investigate the association for a set of rare variants and ordinal traits while easily adjusting for covariates. We apply this methodology to a whole-exome sequencing data of DS to show the practical relevance of this approach.

3.3 Statistical Model

3.3.1 Proportional odds model

The proportional odds model (Agresti, 2010) (McCullagh, 1980) is a common method to analyze ordinal categorical outcome. Assume the data is taken from n subjects. We consider an ordinal trait consisting of J categories. Let y_i be the trait value for the ith individual, $\mathbf{X_i} = (X_{i1}, \cdots, X_{ip})$ be the covariates (e.g gender, age, etc), and $\mathbf{G_i} = (G_{i1}, \cdots, G_{iq})$ be the number $\{0, 1, 2\}$ of variant alleles of the q polymorphic sites. We propose the following cumulative logit model:

$$\text{logit}(\Pr(y_i \leq j | \mathbf{X_i}, \mathbf{G_i})) = \theta_j - \beta^T \mathbf{X_i} - \alpha^T \mathbf{G_i} \tag{3.3.1}$$

where θ_j is the intercept parameter corresponding to category j, and $-\infty = \theta_0 \leq \theta_1 \leq \cdots \leq \theta_J = \infty$. $\beta = (\beta_1, \cdots, \beta_p)^T$ are coefficients of environmental effects. $\alpha = (\alpha_1, \cdots, \alpha_q)^T$ are coefficients of genetic effects.

We denote $\pi_j(y_i)$ the probability for ith subject falls jth category ($j = 1, \cdots, J$), where $\pi_j(y_i) = P(y_i = j | \mathbf{X_i}, \mathbf{G_i})$. We can express $\pi_j(y_i)$ as

$$\pi_j(y_i) = \Pr(y_i \le j|\mathbf{X_i}, \mathbf{G_i}) - \Pr(y_i \le j-1|\mathbf{X_i}, \mathbf{G_i})$$
$$= \frac{1}{\exp(\beta^T\mathbf{X_i} + \alpha^T\mathbf{G_i} - \theta_j) + 1} - \frac{1}{\exp(\beta^T\mathbf{X_i} + \alpha^T\mathbf{G_i} - \theta_{j-1}) + 1} \tag{3.3.2}$$

For jth category, we have n_j observations $n = \sum_{j=1}^{J} n_j$. Set $\psi(t) = \frac{1}{1+\exp(-t)}$, we have

$$\pi_1(y_i) = \psi(\theta_1 - \beta^T\mathbf{X_i} - \alpha^T\mathbf{G_i})$$

$$\pi_j(y_i) = \psi(\theta_j - \beta^T\mathbf{X_i} - \alpha^T\mathbf{G_i}) - \psi(\theta_{j-1} - \beta^T\mathbf{X_i} - \alpha^T\mathbf{G_i})$$
$$\pi_J(y_i) = 1 - \psi(\theta_{J-1} - \beta^T\mathbf{X_i} - \alpha^T\mathbf{G_i})$$

Our interest is to investigate association between the gene-based region and the ordinal trait. Thus, the null hypothesis $H_0 : \alpha = 0$. We can write $\alpha = \tau\xi$, where τ is a scalar constant, and $\xi = \alpha/\tau$. Set $S_i = \xi^T\mathbf{G_i}$, and $\xi = (\xi_1, \cdots, \xi_q)^T$ is a q vector of weights. Thus, S_i is a weighted linear combination of $G_{i1}, \cdots, G_{i1}$ with G_{ik} having the weight ξ_k. These weights can assume one of several modeling assumptions described in Chapter 1. The null hypothesis becomes H_0: $\tau = 0$ and the Equation 1 can be rewritten as

$$\Pr(y_i \le j|\mathbf{X_i}, S_i) = \frac{1}{\exp(\beta^T\mathbf{X_i} + \tau S_i - \theta_j) + 1}$$

and the cumulative probability in Equation 2 can be rewritten as

$$\pi_j(y_i) = \frac{1}{\exp(\beta^T\mathbf{X_i} + \tau S_i - \theta_j) + 1} - \frac{1}{\exp(\beta^T\mathbf{X_i} + \tau S_i - \theta_{j-1}) + 1}$$

The likelihood function based on cumulative probability is

$$L(\beta, \theta, \tau|\mathbf{X_i}, \mathbf{G_i}) = \prod_{i=1}^{n}[\prod_{j=1}^{J}\pi_j(y_i)^{\mathbf{I}\{y_i=j\}}] = \prod_{i=1}^{n}\{\prod_{j=1}^{J}[\Pr(y_i \le j|\mathbf{X_i}, \mathbf{G_i}) - \Pr(y_i \le j-1|\mathbf{X_i}, \mathbf{G_i})]^{\mathbf{I}\{y_i=j\}}\}$$

$$= \prod_{i=1}^{n}\{\prod_{j=1}^{J}[\psi(\theta_j - \beta^T\mathbf{X_i} - \tau S_i) - \psi(\theta_{j-1} - \beta^T\mathbf{X_i} - \tau S_i)]^{\mathbf{I}\{y_i=j\}}\}$$
$$= \prod_{j=1}^{J}\prod_{i=1}^{n_j}[\psi(\theta_j - \beta^T\mathbf{X_i} - \tau S_i) - \psi(\theta_{j-1} - \beta^T\mathbf{X_i} - \tau S_i)]$$
$$= \prod_{j=1}^{J}\prod_{i=1}^{n_j}[\frac{1}{\exp(\beta^T\mathbf{X_i} + \tau S_i - \theta_j) + 1} - \frac{1}{\exp(\beta^T\mathbf{X_i} + \tau S_i - \theta_{j-1}) + 1}]$$

where $\mathbf{I}\{y_i = j\}$ is an indicator function and the corresponding log-likelihood is

$$l(\beta,\theta,\tau|\mathbf{X_i},S_i) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\ln[\psi(\theta_j-\beta^T\mathbf{X_i}-\tau S_i)-\psi(\theta_{j-1}-\beta^T\mathbf{X_i}-\tau S_i)]$$

$$= \sum_{j=1}^{J}\sum_{i=1}^{n_j}\ln[\frac{1}{\exp(\beta^T\mathbf{X_i}+\tau S_i-\theta_j)+1}-\frac{1}{\exp(\beta^T\mathbf{X_i}+\tau S_i-\theta_{j-1})+1}]$$

since $\frac{dln(\psi(t))}{t} = 1-\psi(t)$, the score function with respect to τ is

$$\frac{\partial l(\beta,\theta,\tau|\mathbf{X_i},S_i)}{\partial\tau} = -\sum_{j=1}^{J}\sum_{i=1}^{n_j}S_i[1-\psi(\theta_j-\beta^T\mathbf{X_i}-\tau S_i)-\psi(\theta_{j-1}-\beta^T\mathbf{X_i}-\tau S_i)]$$

the MLE of θ under H_0 can be obtained by solving by using Fisher scoring or Newton-Raphson algorithms. R package VGAM (Yee, 2008) can be used to fit null model of Equation (1). Substitute $\hat{\beta}$ and $\hat{\theta}$ for β and θ, the score function under the null hypothesis is

$$U = -\{\sum_{i=1}^{n_1}S_i[1-\psi(\hat{\theta}_1-\hat{\beta}^T\mathbf{X_i})]+\sum_{j=2}^{J-1}\sum_{i=1}^{n_j}S_i[1-\psi(\hat{\theta}_j-\hat{\beta}^T\mathbf{X_i})-\psi(\hat{\theta}_{j-1}-\hat{\beta}^T\mathbf{X_i})]+\sum_{i=1}^{n_J}S_i[-\psi(\hat{\theta}_{J-1}-\hat{\beta}^T\mathbf{X_i})]\}$$

The Fisher information is $I(\Theta) = -E[\frac{\partial^2}{\partial\Theta^2}l(\Theta)|\Theta]$, where $\Theta = (\theta,\alpha,\beta)$. Our test, $\frac{U(\Theta_0)^2}{I(\Theta_0)}$ follows asymptotic distribution of χ^2_1. Because the asymptotic approximation depends on the large sample theory it might not hold in small samples. Consequently, we shall provide a permutation test for small sample size. We permute the combined genotype values S_i's and calculate the test statistic for each permutation.

3.4 Simulations

Our simulation have two purposes. First, we assess the type I error rate of our test statistic with respect to four nominal levels (0.05, 0.01, 0.001 and 0.0001). Second, we evaluate the power in comparison with other established test statistics. Because there is no existing method of analyzing ordinal outcome in gene-based association study, we compare with binary methods by merging categories into two. We compared six approaches: our method (Ordinal) , T1 test (Set the weight to 1 if MAF < 1% and to 0 otherwise), T5 test (Set the weight to 1 if MAF < 5% and to 0 otherwise), Madsen and Browning (MB) test (Madsen and Browning, 2009), VT (Price et al., 2010a) and SKAT. We merged later two categories into one to compute the p-values for T1, T5, MB, VT and SKAT. The package we used to implement other methods is SCORE-Seq (Lin and Tang, 2011). The genotype data are generated by the calibration coalescent model (COSI) (Schaffner et al., 2005) with mimicking linkage disequilibrium (LD) structure of European ancestry. In their dataset, they generated 10,000 haplotypes over 200k BP region mimicking the linkage disequilibrium (LD) structure of European ancestry. We randomly draw n-pair haplotypes within each 2kb subregion and add each pair together to get the additive genotype coding. Then, we calculate sample MAF and keep SNPs with observed MAF between 0 and 0.05. To assess the type I error rate, we generate the ordinal outcome under the null hypothesis using an underlying latent quantitative variable. Then we discretize the latent variable. To evaluate the robustness of our method, we also simulate the ordinal trait under a multinomial logit model (Rodrguez, G. 2007), and thus violating the proportionality assumption is violated. To evaluate the power, we considered effect sizes, ln(2), ln(3), ln(4) and ln(5) with the ratio of causal SNPs, 30%, 40%, 50%. For both type I error and power simulation, we considered two covariates: one continuous and one binary for all scenarios and simply count the total number of rare mutations to aggregate multiple SNPs into a score.

3.4.1 Type I error rate comparison

To fit the proportional odds model, we introduced an underlying latent variable (Agresti, 2010) and generate the outcome following (Chen et al., 2013) parameter setting (except the random term)

$$y^* = 0.05age + 0.5gender + \epsilon$$

where y^* is the latent variable, a vector of continuous trait, age is a vector of continuous covariate from a normal distribution with mean 50 and standard deviation 5, gender is a vector of dichotomous covariate from a Bernoulli distribution with probability 0.5, ϵ follows a standard logistic distribution. After we generated a quantitative trait, 2 cutpoints were used to separate the continuous trait into 3 discrete scales.

In Table 3.1, we compared the nominal levels of type I error rate with other five methods (T1, T5, MB, VT and SKAT) by the simulation in 10^6 replications when ordinal traits were generated from proportional odds models, and genotype drawn from 2k BP. Because we derived our score test from a proportional odds model, we deliberately generated data from non-proportional odds to assess the robustness of our score test with respect to the proportionality assumption. The type I error simulation result is shown in Table 3.2.

The empirical type I error rates estimated from all methods were numerically close to the nominal levels (0.05 and 0.01). On a relative scale, we observed a proportionally larger deviation between the empirical and nominal levels of type I error rates among binary methods for $\alpha = 0.001$ and 0.0001.

For small sample sizes ($n = 500$), T1 and T5 is inflated under the nominal level 0.0001. MB test is anti-conservative when the sample size is large. VT is dramatically inflated under the nominal level 0.001 and 0.0001 across all sample sizes. However, our method controlled the type I error very well even when the model is misspecified.

Table 3.1: Type I Error for Proportional Odds Model (2kb)

α	Ordinal	T1	T5	MB	VT	SKAT
Sample Size = 500						
0.05	0.050	0.049	0.050	0.050	0.046	0.046
0.01	0.010	0.009	0.010	0.009	0.009	0.009
10^{-3}	1.02×10^{-3}	0.85×10^{-3}	0.94×10^{-3}	0.86×10^{-3}	0.87×10^{-3}	0.75×10^{-3}
10^{-4}	1.25×10^{-4}	0.93×10^{-4}	1.03×10^{-4}	0.85×10^{-4}	0.86×10^{-4}	0.75×10^{-4}
Sample Size = 1000						
0.05	0.050	0.050	0.050	0.050	0.046	0.048
0.01	0.010	0.009	0.010	0.009	0.009	0.009
10^{-3}	0.95×10^{-3}	0.90×10^{-3}	0.97×10^{-3}	0.94×10^{-3}	0.92×10^{-3}	0.85×10^{-3}
10^{-4}	0.85×10^{-4}	1.10×10^{-4}	0.82×10^{-4}	0.93×10^{-4}	0.90×10^{-4}	0.88×10^{-4}
Sample Size = 1500						
0.05	0.049	0.049	0.050	0.050	0.047	0.049
0.01	0.010	0.010	0.010	0.010	0.009	0.009
10^{-3}	0.96×10^{-3}	0.90×10^{-3}	1.01×10^{-3}	0.93×10^{-3}	0.93×10^{-3}	0.92×10^{-3}
10^{-4}	1.02×10^{-4}	0.92×10^{-4}	0.95×10^{-4}	0.72×10^{-4}	0.90×10^{-4}	0.89×10^{-4}
Sample Size = 2000						
0.05	0.049	0.050	0.051	0.050	0.047	0.049
0.01	0.010	0.010	0.010	0.010	0.009	0.010
10^{-3}	0.90×10^{-3}	0.93×10^{-3}	1.00×10^{-3}	0.93×10^{-3}	0.95×10^{-3}	0.97×10^{-3}
10^{-4}	0.81×10^{-4}	0.98×10^{-4}	0.96×10^{-4}	0.92×10^{-4}	1.01×10^{-4}	0.95×10^{-4}

Table 3.2: Type I Error for Non-Proportional Odds Model (2kb)

α	Ordinal	T1	T5	MB	VT	SKAT
Sample Size = 500						
0.05	0.050	0.048	0.050	0.049	0.046	0.047
0.01	0.010	0.009	0.010	0.009	0.010	0.009
10^{-3}	0.89×10^{-3}	1.05×10^{-3}	0.92×10^{-3}	0.95×10^{-3}	1.21×10^{-3}	0.91×10^{-3}
10^{-4}	0.90×10^{-4}	1.47×10^{-4}	0.99×10^{-4}	0.97×10^{-4}	1.40×10^{-4}	1.07×10^{-4}
Sample Size = 1000						
0.05	0.049	0.049	0.050	0.050	0.047	0.048
0.01	0.010	0.010	0.01	0.010	0.010	0.009
10^{-3}	1.01×10^{-3}	0.99×10^{-3}	0.99×10^{-3}	0.98×10^{-3}	1.18×10^{-3}	0.99×10^{-3}
10^{-4}	1.14×10^{-4}	1.17×10^{-4}	1.22×10^{-4}	1.00×10^{-4}	1.48×10^{-4}	1.12×10^{-4}
Sample Size = 1500						
0.05	0.048	0.050	0.050	0.050	0.047	0.049
0.01	0.009	0.010	0.010	0.010	0.010	0.010
10^{-3}	0.92×10^{-3}	0.98×10^{-3}	0.99×10^{-3}	1.00×10^{-3}	1.22×10^{-3}	1.02×10^{-3}
10^{-4}	0.87×10^{-4}	1.31×10^{-4}	1.10×10^{-4}	1.01×10^{-4}	1.67×10^{-4}	1.06×10^{-4}
Sample Size = 2000						
0.05	0.048	0.049	0.05	0.050	0.047	0.049
0.01	0.009	0.010	0.010	0.010	0.010	0.010
10^{-3}	0.99×10^{-3}	0.95×10^{-3}	0.98×10^{-3}	1.01×10^{-3}	1.22×10^{-3}	1.04×10^{-3}
10^{-4}	0.89×10^{-4}	1.09×10^{-4}	1.09×10^{-4}	1.24×10^{-4}	1.52×10^{-4}	0.93×10^{-4}

3.4.2 Power comparison

To evaluate power, we compared six approaches when generate the trait under the alternative hypothesis. Because we showed the type I error was controlled well when the model was misspecified, we only focused on the proportional odds model for power simulation. We simulated 10^4 replicates under the alternative hypothesis. To fit the proportional odds model, we included the genotype by using the model

$$y^* = 0.05age + 0.5gender + G\alpha + \epsilon$$

where y^*, age, gender and ϵ are generated in the same way as in the type I error simulations. α is a vector consisting of the genetic effect sizes of the causal SNPs. We varied the effect sizes of causal SNPs from ln(2), ln(3), ln(4) and ln(5), and the proportion of causal SNPs or causal ratio from 30%, 40%, 50%. Causal SNPs were randomly drawn from 2kb subregion for each replicate. Figure 3.1-3.3 display the power comparison at four nominal levels when the effect size is ln(3) varying different

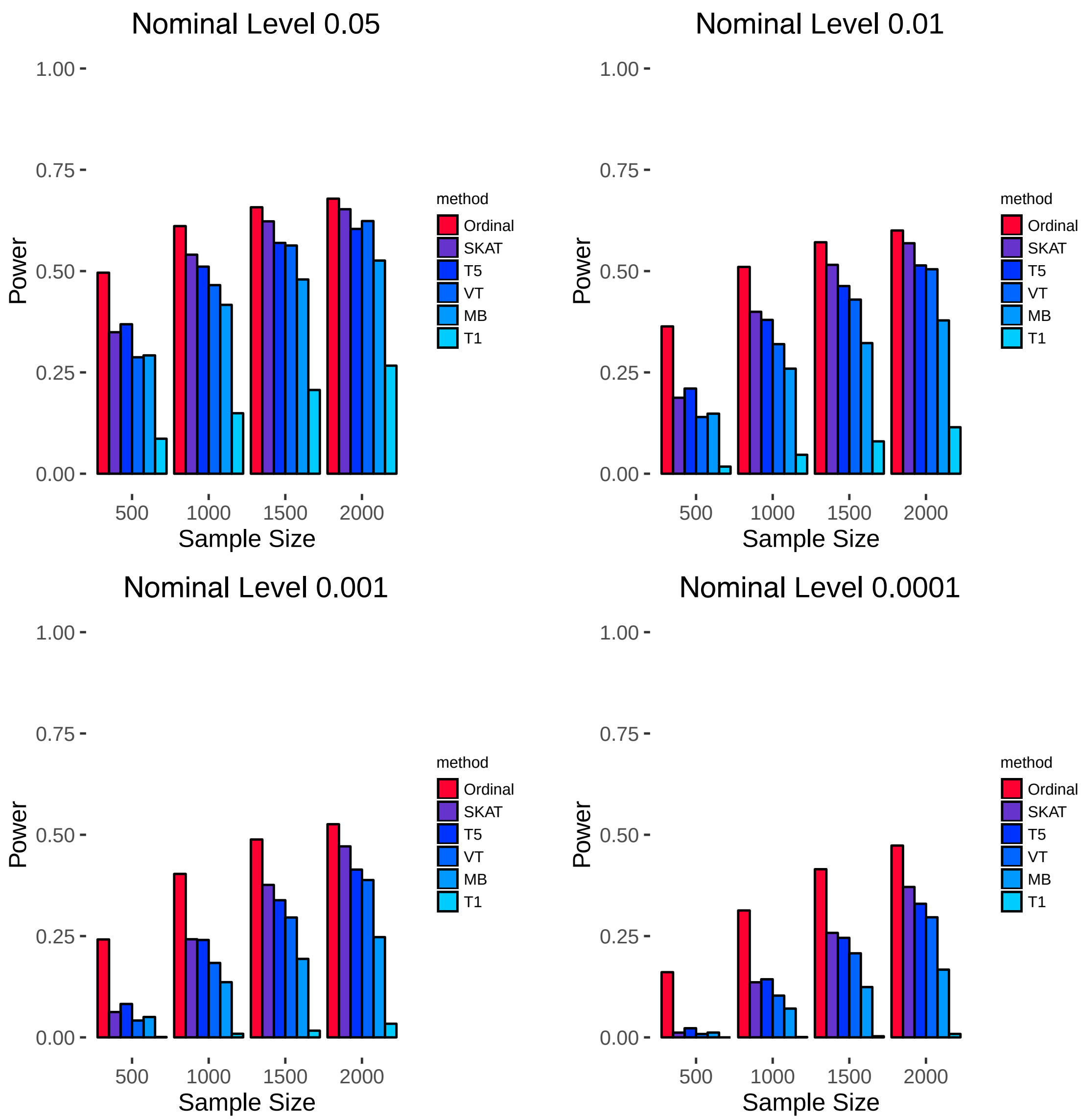


Figure 3.1: Power Comparison causal ratio 0.3 and effect size log(3)

sample size. Figure 3.4-3.6 show the power under different causal ratios when the effect size is ln(5) for sample 1000, 1500 and 2000 individuals, respectively. Figure 3.7-3.9 show the power over different effect sizes when the causal ratio is 50% for sample 1000, 1500 and 2000 individuals, respectively. As expected, we can see that there is a substantial loss of power when dichotomizing an ordinal trait.

Since the methods, T1 and T5 filtered out the variants beyond the thresholds, 1% and 5%, they suffer from substantial power loss if a large proportion of casual SNPs or all casual SNPs were beyond the thresholds. In those scenarios, the results were incomparable. Thus, we simulated the outcome, in which all casual SNPs were either below 1% or 5% and compared the power. The results were shown in Table 3.3 and Table 3.4.

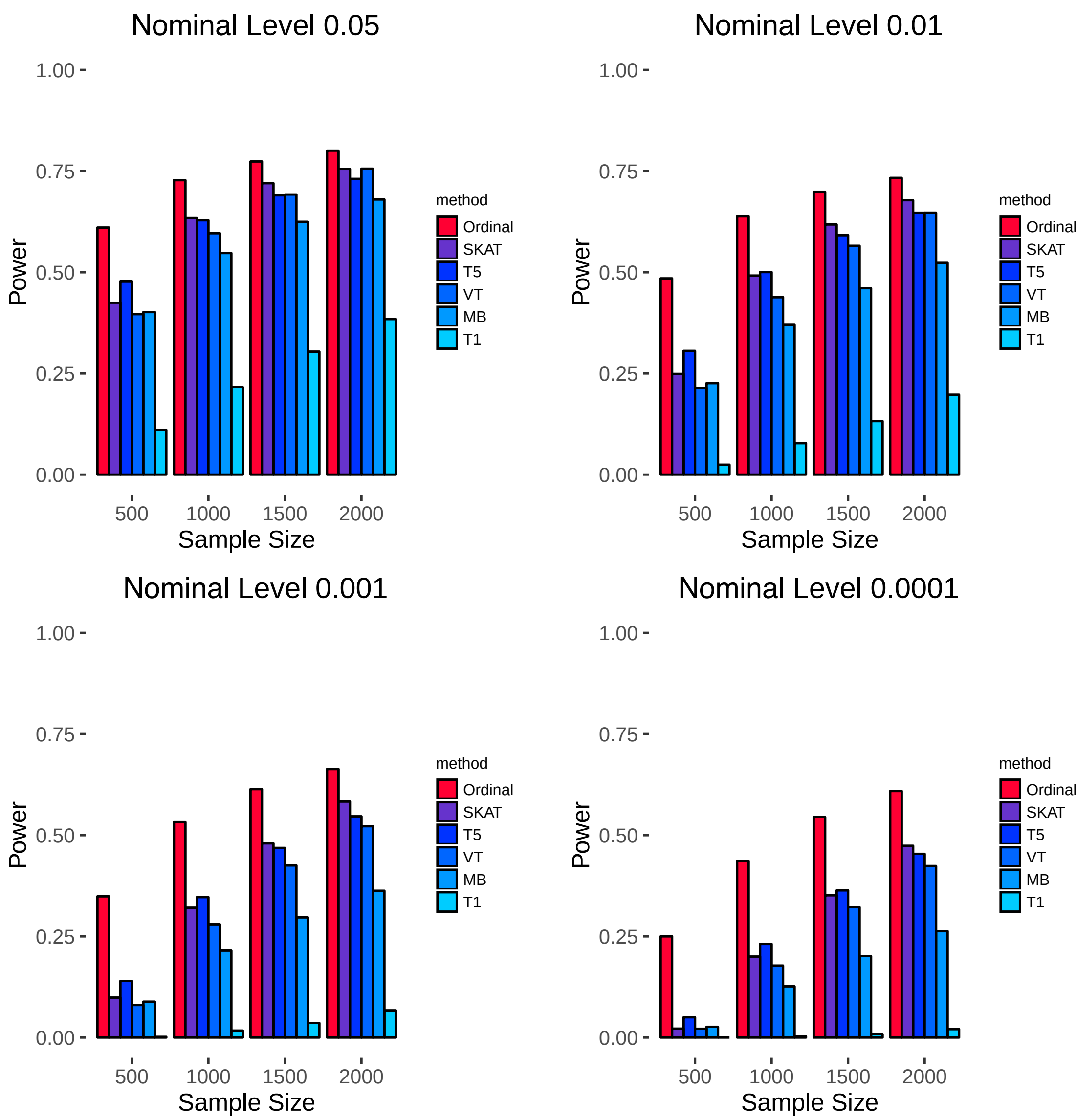


Figure 3.2: Power Comparison causal ratio 0.4 and effect size log(3)

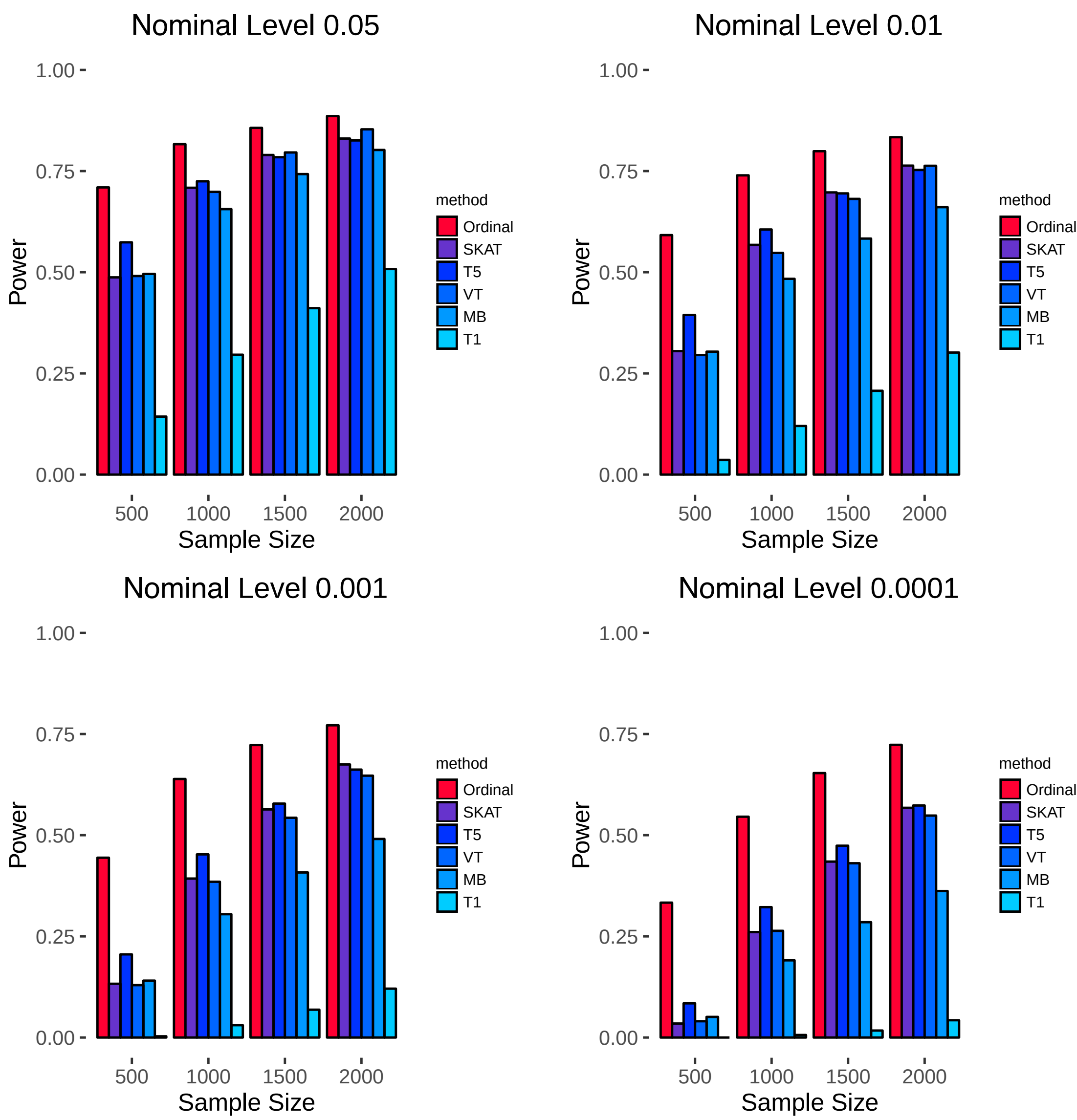


Figure 3.3: Power Comparison causal ratio 0.5 and effect size log(3)

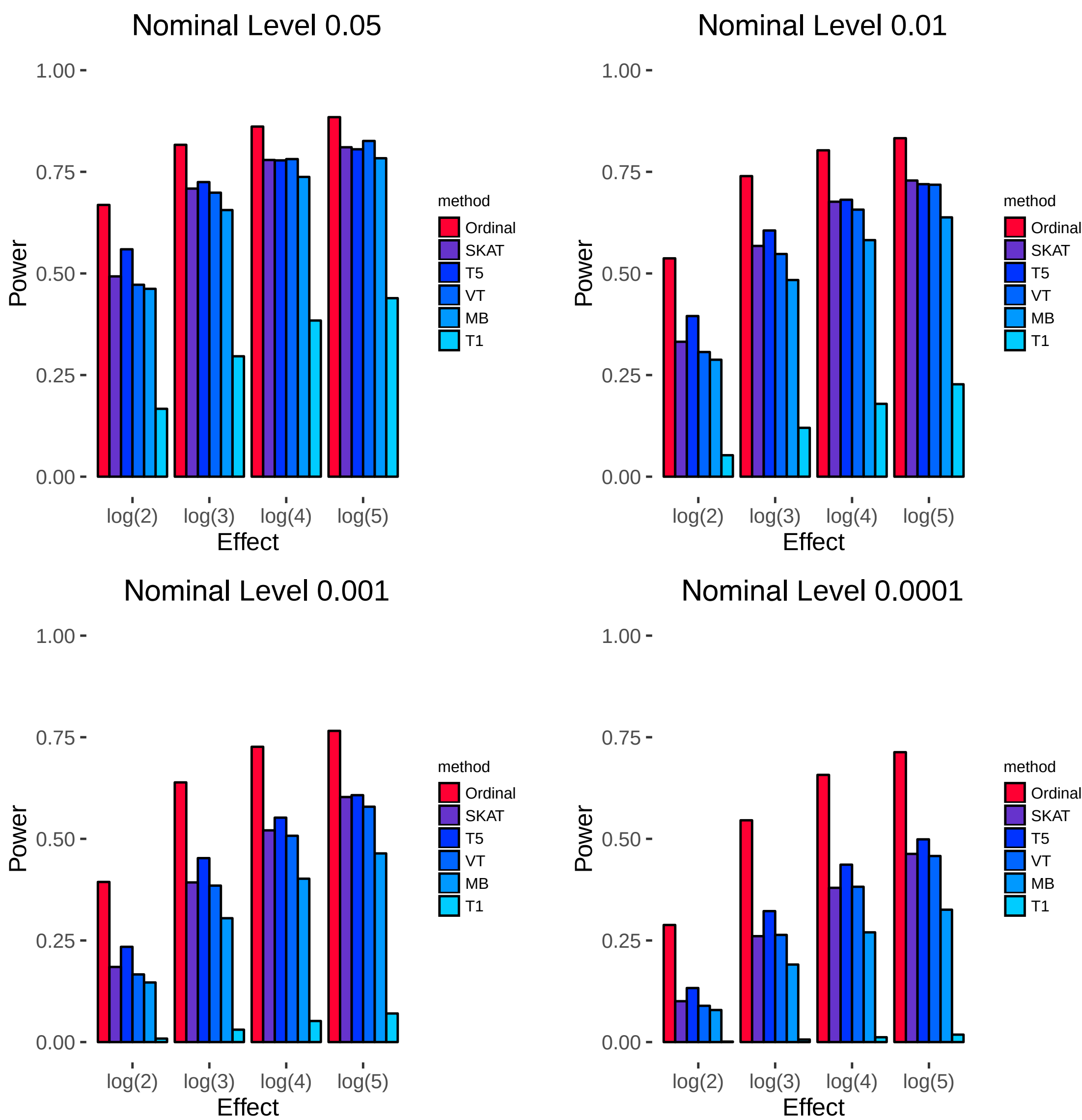


Figure 3.4: Power Comparison sample size 1000 and effect size log(5)

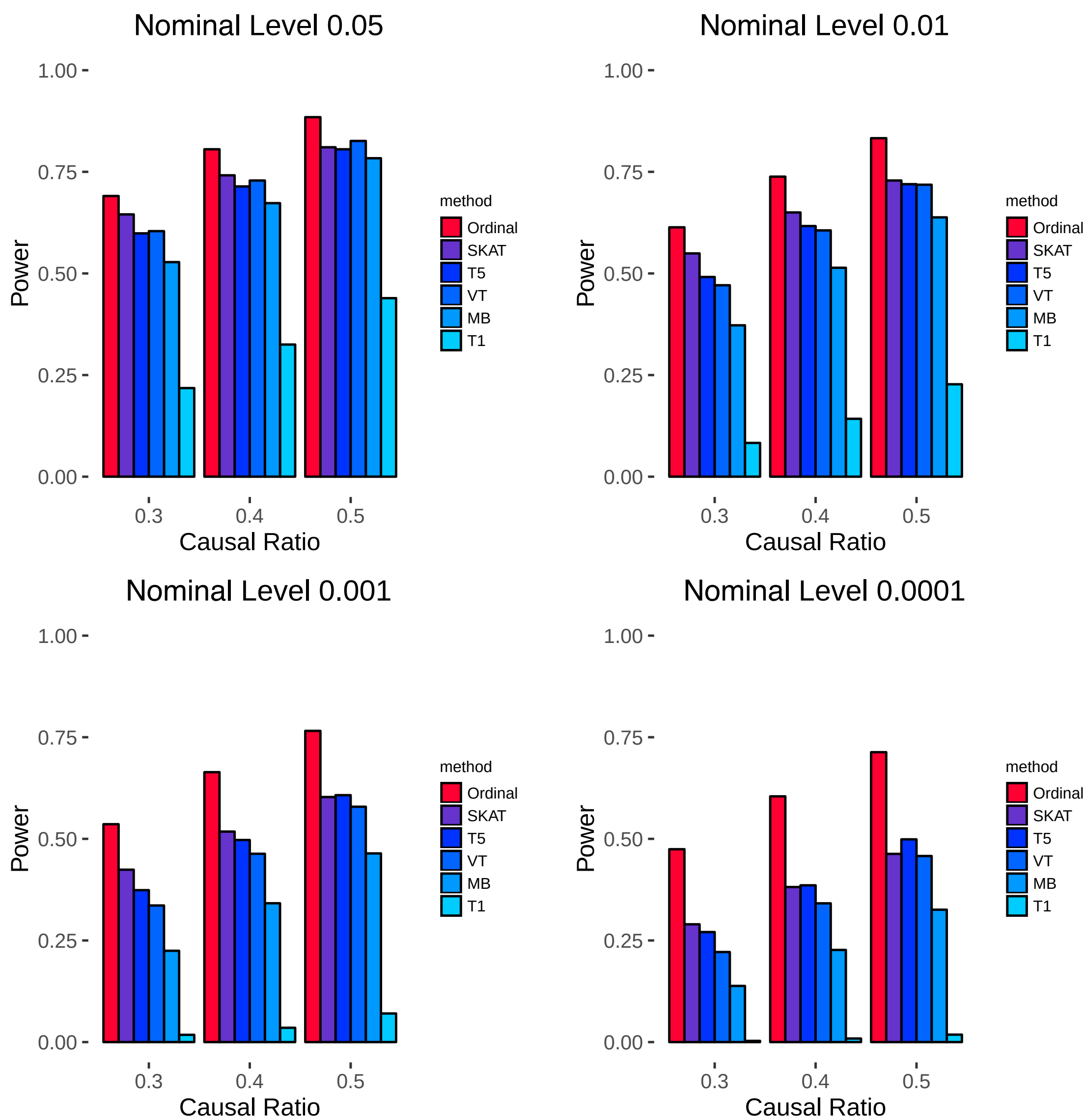


Figure 3.5: Power Comparison sample size 1500 and effect size log(5)

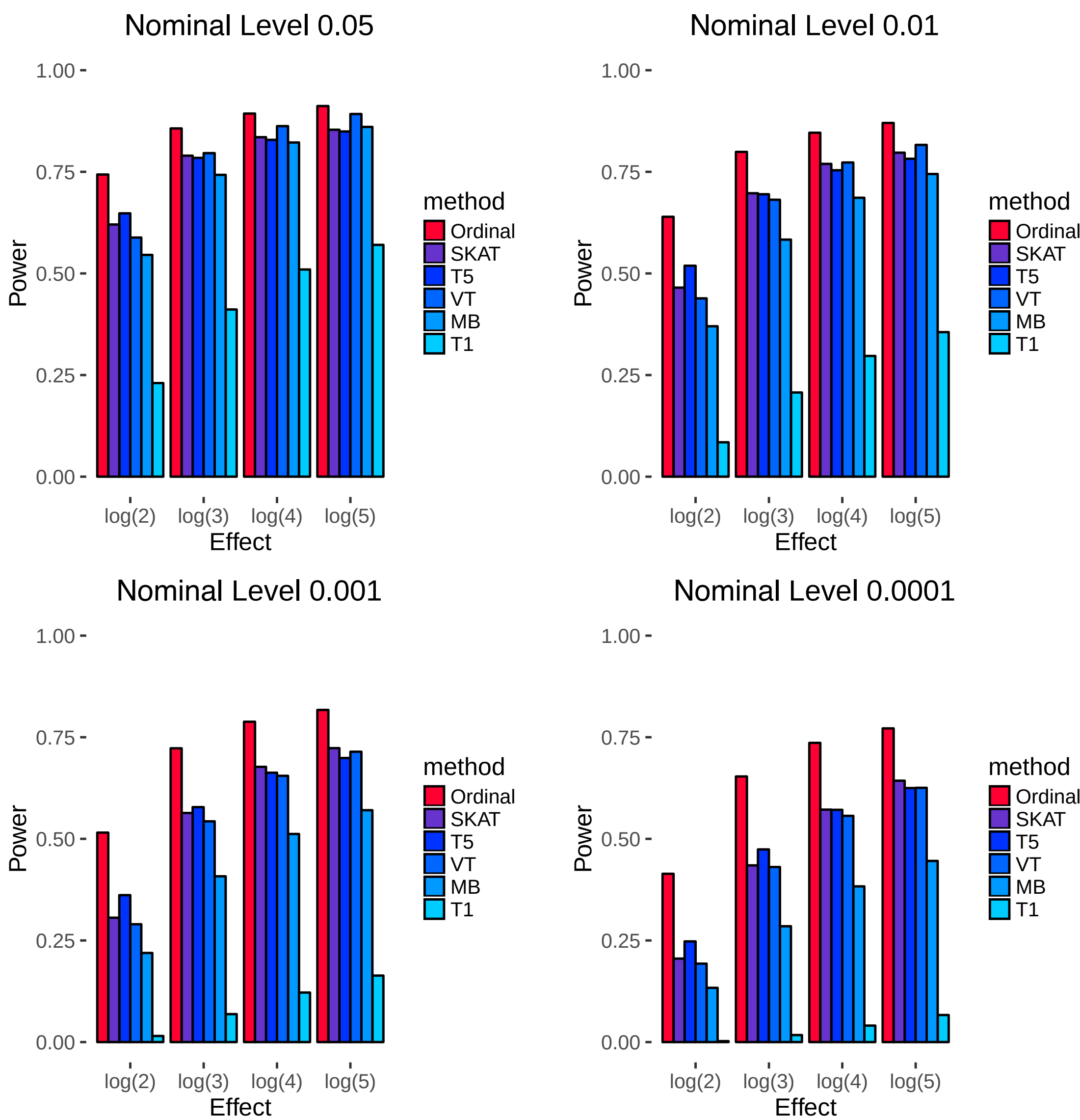


Figure 3.6: Power Comparison sample size 2000 and effect size log(5)

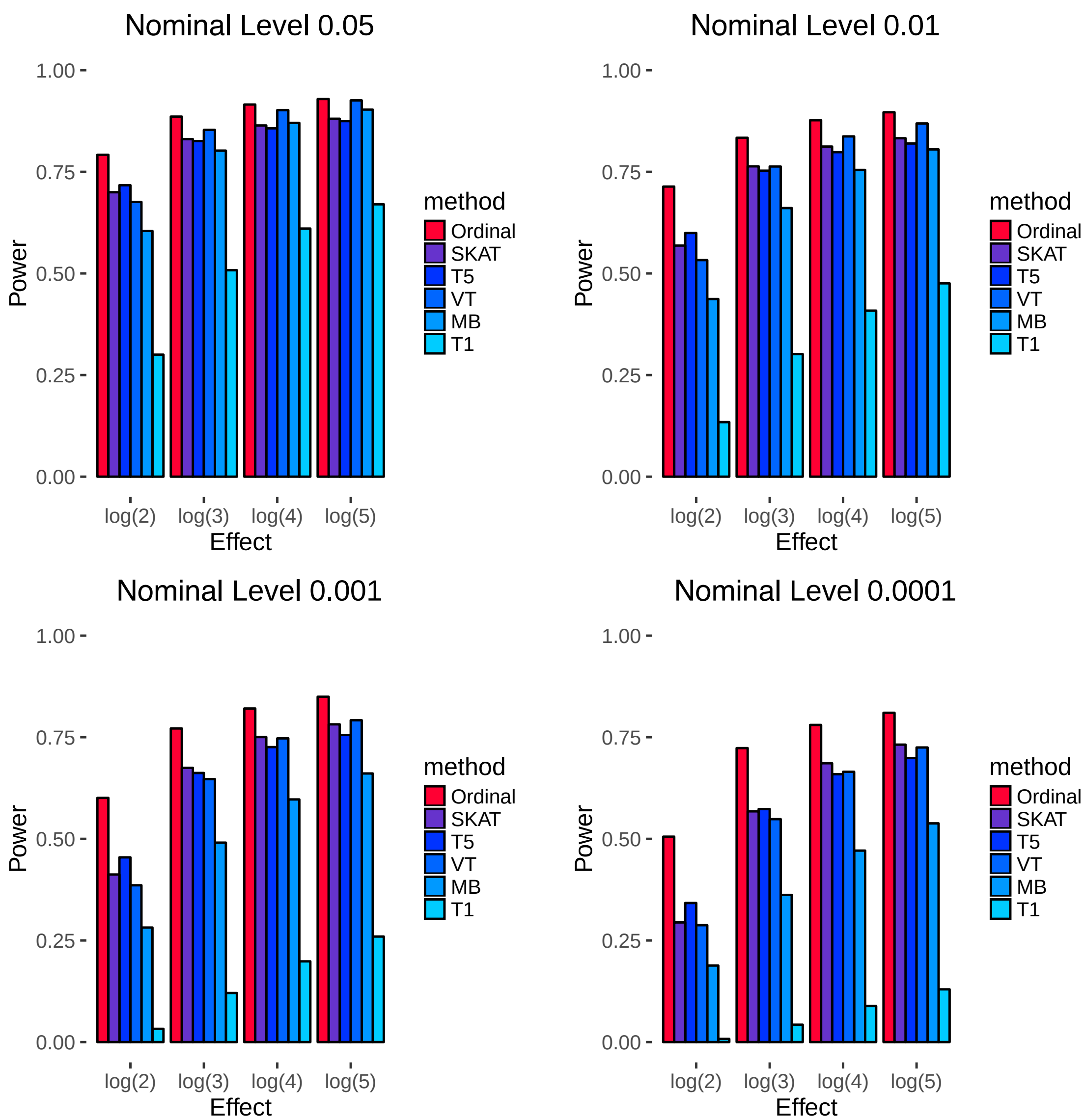


Figure 3.7: Power Comparison sample size 1000 and causal ratio 0.5

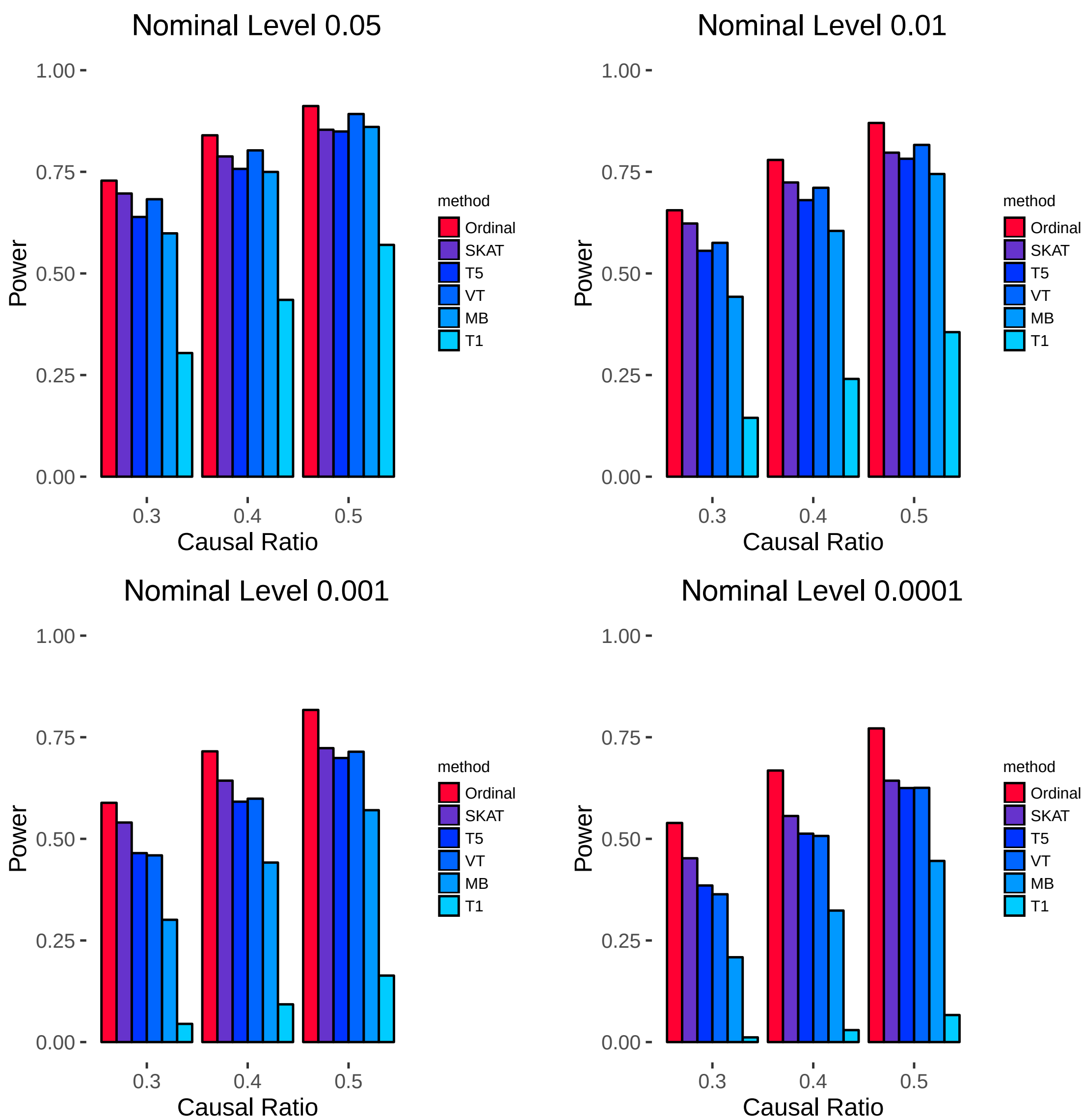


Figure 3.8: Power Comparison sample size 1500 and causal ratio 0.5

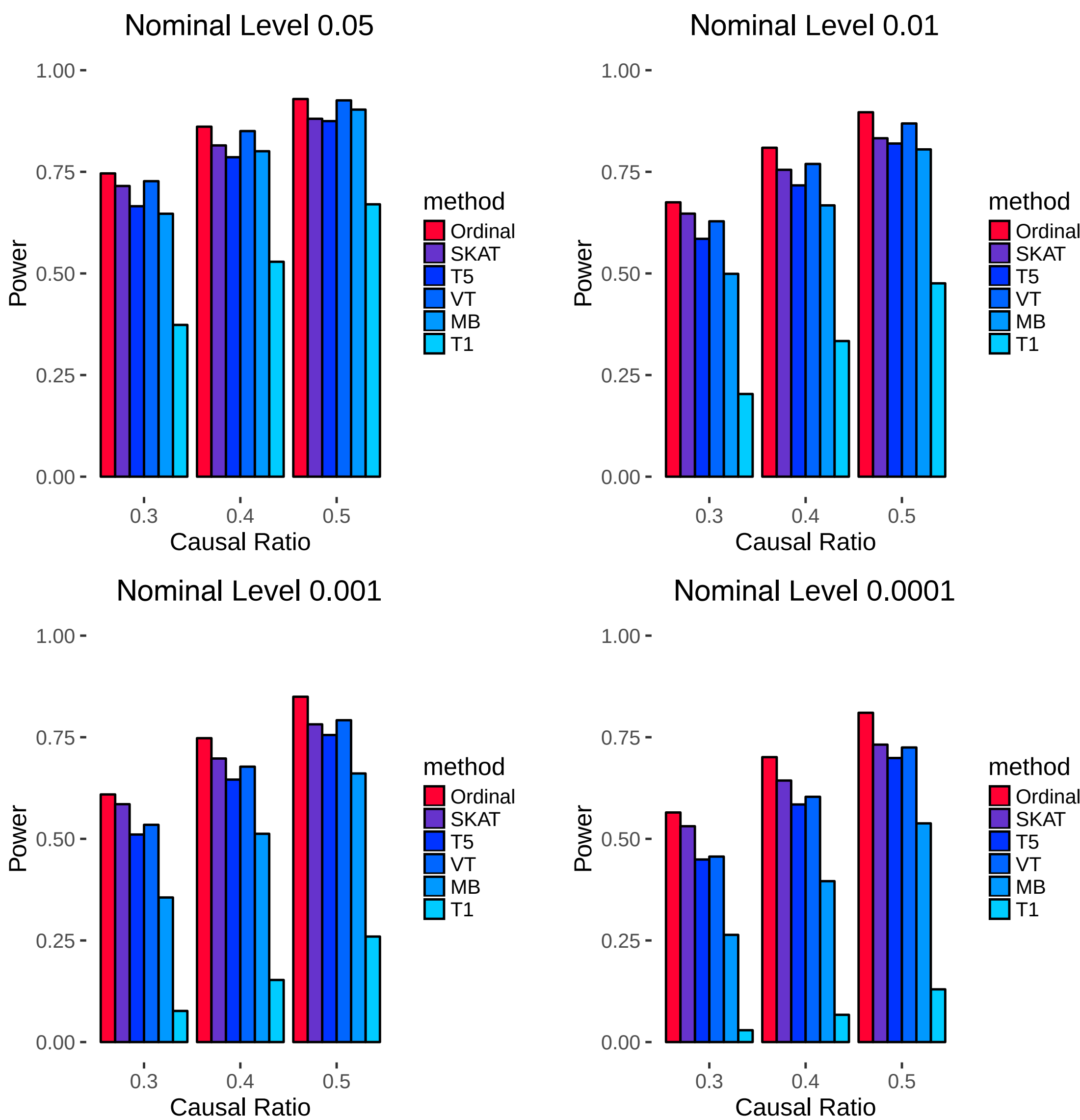


Figure 3.9: Power Comparison sample size 2000 and causal ratio 0.5

Table 3.3: Power For MAF 1%

Causal Ratio	Ordinal	T1	T5	MB	VT	SKAT
Effect Size = log(2)						
30%	0.080	0.033	0.033	0.025	0.016	0.011
40%	0.138	0.058	0.058	0.045	0.028	0.012
50%	0.218	0.094	0.094	0.074	0.047	0.016
Effect Size = log(3)						
30%	0.191	0.071	0.071	0.054	0.037	0.026
40%	0.342	0.138	0.138	0.107	0.080	0.036
50%	0.514	0.233	0.233	0.185	0.144	0.048
Effect Size = log(4)						
30%	0.295	0.106	0.106	0.076	0.056	0.044
40%	0.485	0.203	0.203	0.155	0.126	0.062
50%	0.689	0.339	0.339	0.274	0.233	0.086
Effect Size = log(5)						
30%	0.370	0.134	0.134	0.094	0.074	0.063
40%	0.584	0.256	0.256	0.195	0.167	0.091
50%	0.779	0.418	0.418	0.345	0.307	0.127

This is the power simulation result, where sample size = 1000, SNP number = 30 and MAF < 1% under the nominal level 0.01

Table 3.4: Power For MAF 5%

CausalRatio	Ordinal	T1	T5	MB	VT	SKAT
Effect Size = log(2)						
30%	0.332	0.029	0.222	0.141	0.157	0.194
40%	0.470	0.049	0.330	0.227	0.251	0.268
50%	0.608	0.081	0.448	0.320	0.357	0.339
Effect Size = log(3)						
30%	0.521	0.064	0.380	0.262	0.329	0.394
40%	0.688	0.119	0.532	0.406	0.483	0.509
50%	0.816	0.199	0.673	0.563	0.636	0.608
Effect Size = log(4)						
30%	0.603	0.093	0.453	0.333	0.419	0.502
40%	0.769	0.180	0.616	0.508	0.596	0.627
50%	0.877	0.296	0.752	0.681	0.748	0.723
Effect Size = log(5)						
30%	0.648	0.117	0.492	0.380	0.479	0.557
40%	0.808	0.226	0.659	0.575	0.660	0.6812
50%	0.905	0.361	0.790	0.750	0.804	0.775

This is the power simulation result, where sample size = 1000, SNP number = 30 and MAF $< 5\%$ under the nominal level 0.01

3.5 Real Data analyses

In this section, we test the association between candidate genes and IQ to identify modifier gene variants that contribute to clinical outcome of DS. We sequenced the exomes of 22 individuals (Hammer et al., 2017) diagnosed as either mild or severe cognitive condition. Those 22 individuals are from a Japanese cohort. 11 subjects diagnosed as mild, and another 11 subjects are severe when they enrolled. In our study, we analyzed IQ as a discrete ordinal scale, and adjusted for age (or age of onset) and gender. We applied permutation-based test instead of asymptotic approximation in our method because the sample size is small and the large sample theory may not hold. Table 3.5 and Table 3.6 show the 10 genes with the smallest p-values. We also used false discovery rate (FDR) to adjust the p-values but none shows significant after correction.

Table 3.5: Application on DS data

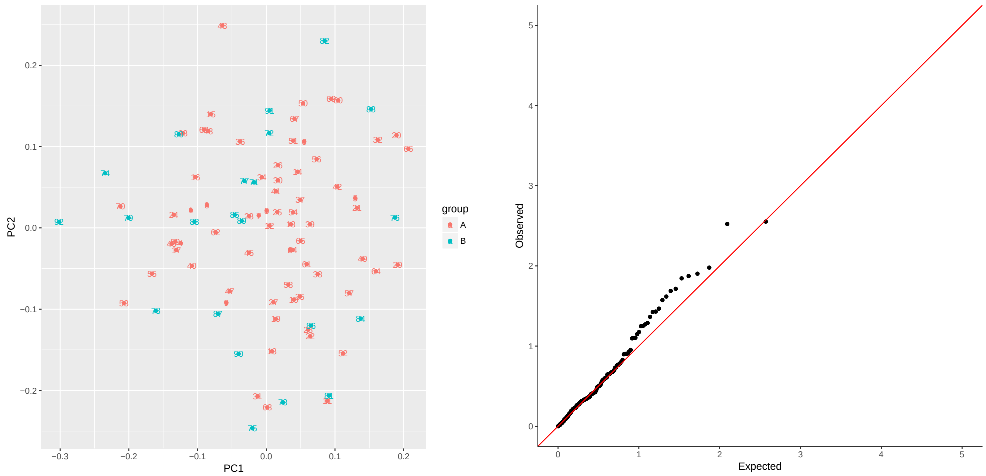
Gene	Chr	Start	Stop	N SNPs	p-value	adjusted p-value (FDR)
CIDEC	3	9908761	9908761	2	4.3×10^{-5}	0.5668025
SERPINB11	18	61377320	61391738	26	5.5×10^{-5}	0.5668025
C12orf5	12	4430317	4462040	18	1.0×10^{-4}	0.6870333
MIR3163	6	152476952	152545865	40	2.0×10^{-4}	0.9571335
LMOD1	1	201865763	201865763	6	3.0×10^{-4}	0.9571335
MAP2K2	19	4090572	4090572	21	5.0×10^{-4}	0.9571335
C12orf66	12	64587601	64609890	6	6.0×10^{-4}	0.9571335
MIR4455	4	185859505	185859505	2	7.0×10^{-4}	0.9571335
SLC4A9	5	139739974	139739974	9	7.0×10^{-4}	0.9571335
MAP3K13	3	185046756	185046756	25	9.0×10^{-4}	0.9571335

Covariate adjustment for age of onset and gender

Table 3.6: Application on DS data

Gene	Chr	Start	Stop	N SNPs	p-value	adjusted p-value (FDR)
PGS1	17	76374978	76420373	41	1.3×10^{-5}	0.2679430
CHERP	19	16630994	16653249	14	4.8×10^{-5}	0.4946640
SLC4A9	5	139739974	139748283	9	8.5×10^{-5}	0.5839783
GOLGA5	14	93263982	93303855	27	2.0×10^{-4}	0.9739546
C1orf177	1	55271730	55307693	32	2.8×10^{-4}	0.9739546
TRIM8	10	104403391	104416011	11	4.0×10^{-4}	0.9739546
ADGRE2	19	14846692	14887349	61	5.0×10^{-4}	0.9739546
TF	3	133465201	133465201	39	5.0×10^{-4}	0.9739546
MAP3K13	3	185046756	185046756	25	5.0×10^{-4}	0.9739546
BRAT1	7	2577781	2577781	31	6.0×10^{-4}	0.9739546

Covariate adjustment for age and gender



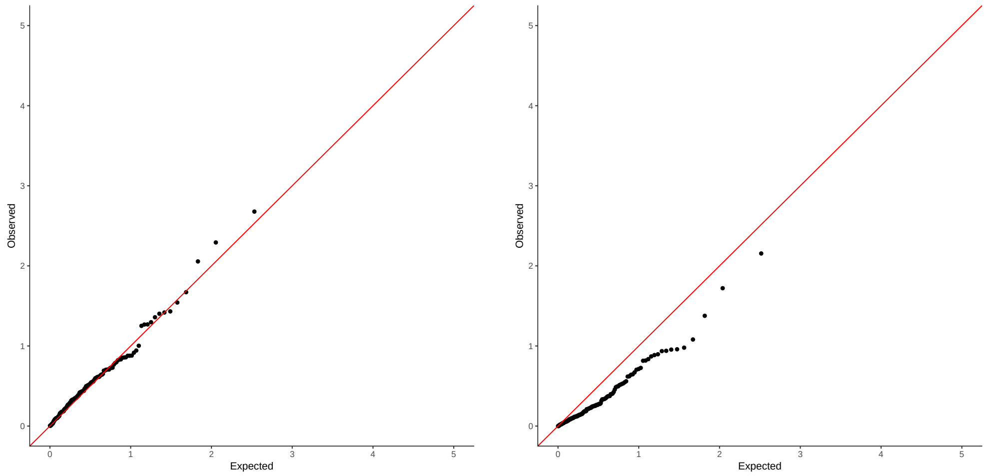
(a) PCA (b) Adjust for gender

Figure 3.10: PCA plot (left) and Quantile-Quantile Plot (right) with 92 subjects (70 Japanese exomes and 22 samples)

In addition, we accessed the genotype data for 70 East Asian exomes made public by the 1000 Genomes project and investigated the association jointly with our DS cohort. The data is based on exome sequencing and focus on epilepsy-related genes (about 600). Because the VCF format does not report sites that match the reference for all samples in the set, some sites called as variant in DS cohort do not appear in 1000 Genomes data set. Due to the fact that the sequencing data were from two different platforms, there might exist batch effect. Therefore, we first do the PCA to see if the two group separate from each other and noticed no obvious clusters (See Figure 3.10, left). Then, we conducted the association test by using healthy, mild and severe as our ordinal trait, and adjusted for gender. Figure 3.10 (right) shows the normal QQ-plot but we did not find any statistical significant candidate genes.

3.5.1 Pathway-Based Analysis

There are two ways to extend the gene-based rare variant association methods to a pathway level analysis (Wu and Zhi, 2013). The current rare variant association methods are designed for gene-based tests. However, the same methodology can be applied to pathway level, and test the effect of multiple SNPs nested within whole genes in one pathway. The second one is to apply these methods at a gene-level and then the rank gene-level test statistic to form a weighted Kolmogorov Smirnov (WKS)-based enrichment score for pathway level analysis (Wang et al., 2007) (Subramanian et al., 2005). The first is called self-contained null hypotheses focus on the target gene set and assumes no association between the target gene set and the trait; while the later, called a competitive null



(a) Self-contained pathway analysis (b) Competitive pathway analysis

Figure 3.11: Quantile-Quantile Plots of p Values on the $-log_{10}$ Scale for Pathway Analysis

hypotheses assuming that the target gene set has the same magnitude of association as that of complement genes in the genome (Tian et al., 2005). Here, we applied both methods. For self-contained null hypotheses, we simply extended the region from gene level to pathway level. For competitive method, we modified Kai Wang's Gene Set Enrichment Analysis (GSEA) method (Wang, et al. 2007). For each gene, we calculate the gene-based test statistic introduced in method part. For all N genes there are N corresponding test statistics, denoted by $U_1, U_2, \cdots, U_N$. We sort these test statistic from the largest to the smallest, denoted by $U_{(1)}, U_{(2)}, \cdots, U_{(N)}$. For a given pathway S, composed of N_H genes, we calculate a WKS-like running-sum statistic that reflects the overrepresentation of genes within the gene set at the top of the ranked list of genes:

$$ES(S) = \max_{1 \leq j \leq N}\{\sum_{G_{j*} \in S, j^* \leq j} \frac{|U_{(j^*)}|}{N_R} - \sum_{G_{j*} \notin S, j^* \leq j} \frac{1}{N - N_H}\}$$

where $N_R = \sum_{G_{j^*} \in S} |U_{(j^*)}|$. We also normalize the enrichment score. $ES(S)$ by permuting our phenotype or genotype to recalculate the gene-level test statistic. We used KEGG to annotate genes into pathways. We have 4323 genes and 170 pathways for analysis. Figure 3.11 shows the QQ-plots for self-contained pathway analysis (left) and competitive pathway analysis (right). Unfortunately, we did not find any significant pathway from either approach. This in part can ascribed to the small number of observations in this dataset.

In genetic epidemiology studies, many human conditions (e.g., cancer and many behavioral and psychiatric disorders) are measured in a discrete ordinal scale, and their levels represent the degree of severity of a symptom. For example, epilepsy is a group of neurological disorders characterized by epileptic seizures, and commonly found in people with intellectual disability (ID). Physicians may record intellectual or motor disability/ID as normal, mild, moderate, or severe to present the severity of the patients. The prevalence of epilepsy increases with the severity of ID (Kumari et al., 2013). showed lower IQ score in children with ID has association with occurrence of epilepsy. A major challenge in epilepsy research is to unravel the complex genetic mechanisms underlying both common and rare forms of epilepsy. In the past two decades, great progress has been made in exploring the genetic bases of epilepsy.

GWAS have been successful in identifying risk factors for common diseases, and have been applied to epilepsy (Heinzen et al., 2010) (Guo et al., 2011). As mentioned in Chapter 1, there is growing interest in the possibility that variants too rare to be represented well on GWAS chips may be important contributors to common diseases, including various types of epilepsy. Indeed, (Veeramah et al., 2013) has suggested that the impact individual single-nucleotide variants in this disease is small, and indicates possibly that gene-based approaches might be more successful for future sequencing studies of epilepsy predisposition.

The advent of next-generation sequencing technology has led to new approaches for comprehensive characterization of variants across the genome. The rare variant hypothesis has been supported by the recent observation that rare copy number variants (CNVs) can have large effects on neuropsychiatric disease risk (Need and Goldstein, 2010). The corresponding gene-based tests have been developed for both binary and quantitative traits (Lee et al., 2012) (Lee et al., 2014). Because clinical assessments are frequently given by an ordinal scale (Zhang et al., 2003) (Feng et al., 2004), an unnecessary collapse of trait levels could reduce the power in association studies. Therefore, methods need to be developed to investigate the association of genes and ordinal traits. (Zhang et al., 2006) proposed an ordinal transmission/disequilibrium test (O-TDT), and (Wang et al., 2006) presented a class of family-based association tests (FBATs) for ordinal traits. Both required nuclear family data and designed for testing for association between genetic a single variant. Due to the unusually large number of zeros in rare variants, the genetic coefficients can be hardly estimated. These approaches have limited power unless sample sizes are very large, the effects are substantial, or the variants are not too rare.

Epilepsy associated genes mostly has been identified in *voltage-gated* or *ligand-gated* ion channels.

Some individuals do not inherit epilepsy but rather the epilepsy is the result of a mutation occurring in the germ cell line of the parents. Such *de novo* mutations are extremely important in sever myoclonic epilepsy of infancy (SMEI) (Özmen et al., 2011). One example is Dravet syndrome (DS). Dravet syndrome is a rare, devastating form of childhood epilepsy, which is often associated with a truncation mutation in the voltage-gated sodium channel gene, SCN1A. There is considerable variability in expressivity within families, as well as among individuals carrying the same primary mutation, which suggests that clinical outcome is modulated by variants at other genes. Our interest is to search for the possible modifier genes. To identify modifier gene variants that contribute to clinical outcome of DS we sequenced the exomes of 22 individuals diagnosed as either mild or severe cognitive condition. We controlled for variation associated with different mutation types by limiting inclusion to individuals with a *de novo* truncation mutation resulting in SCN1A haploinsufficiency. In psychiatry and schizophrenia research, the clinical global impression - severity scale (CGI-S) (Suzuki et al., 2012) is measure of symptom severity. CGI-S has 7 categories and each represents the severity of the patient's illness at the time of assessment, relative to the clinician's past experience with patients who have the same diagnosis. Physicians classified the patients in different categories (normal, border, mild, moderate and severe) on the basis of IQ and DQ assessment tests. The trait is characterized in a discrete ordinal scale, and their levels represent the degree of severity of DS.

To our knowledge, statistical procedures of rare variants are designed for either continuous or binary outcomes. Critical information and power may be lost if we reduce our data to a simply binary trait, especially when sample size is small. Because no closed form of the likelihood function is known, association test of rare variants and ordinal traits has two potential challenges: (1) deriving an effective test statistic and (2) the bias due to the sparsity of genotype data. In this chapter, we present a computational efficient score-based test to investigate the association for a set of rare variants and ordinal traits while easily adjusting for covariates. We apply this methodology to a whole-exome sequencing data of DS to show the practical relevance of this approach.

3.6 Discussion

In this chapter, we discuss the importance of bringing in the ordinal scale of a trait for association studies. In genetic epidemiology studies, the clinical assessments are often recorded in a discrete ordinal scale, and the levels represent the severity degree. Simply dichotomizing an ordinal trait might lead substantial lose of information and consequently reduce the statistical power. To our knowledge, there does not exist statistical procedures of rare variants designed for ordinal outcomes.

With this motivation, we present a score-based statistic that tests for association between region-based (gene-based) and ordinal phenotypic traits in sequencing-based association studies. First, we build our method upon the proportional odds model and collapse the multiple variants within a region into genetic scores (Burden tests). Then, we derive the score-based test statistic from the likelihood function based on cumulative probability. Due to the sparsity of the genotype, we use Davis's method (Davis, 1982) to adjust the distribution of our test statistic. In simulation section, we evaluate our method from two aspects, type I error rate and power with compared to other five methods. Through simulation studies, we demonstrate that our method is more powerful when the trait is ordinal in comparison to other methods. Since many human conditions and diseases are categorized according to an ordinal scale, an unnecessary collapse of trait levels could lead a substantial power loss in association analyses. In addition, our method is able to control the type I error very well even when the model is misspecified. However, other competing methods designed for the binary outcome show weaker control with a tendency towards an anti-conservative type I error. Moreover, resampling method is easily implemented in our method to calculate the p-value if the sample size is small and the large sample theory might not hold.

In Dravet Syndrome data analysis section, we implement our method to test the association signals between candidate genes and IQ to identify modifier gene variants that contribute to clinical outcome of DS. Due to the small sample size, we apply permutation-based test instead of asymptotic approximation. However, we do not find any significant genes after multiple comparison (FDR) to adjust the p-value. In addition, we demonstrate our method can be easily extend the gene-based association test to a pathway level analysis by using a weighted Kolmogorov Smirnov (WKS)-based enrichment score, which provide a possible to elucidate the complete genetic basis of human complex diseases.

While the advent of next-generation sequencing technology have made large-scale sequencing studies for large cohorts feasible, there are still many sites where there is not enough data to call genotypes accurately even when deep sequencing is performed. A substantial power will loss and lead to spurious association signals if we only based on called genotypes. Some methods taking the uncertainty of genotype calls into account have been proposed. In addition, association studies can be confounded by population stratification, in which genotype is not causally related to phenotype but genotype and phenotype are both correlated to ancestry. Principal component analysis (PCA) has been used to control for population stratification in GWAS. PCA is also used to infer population structure in genetic data. In Chapter 3, I followed the idea of considering the uncertainty of genotype calls and combined it with a data-adaptive kernel PCA for population inference. I showed my method can achieve higher fraction of variance explained using top principal components by considering the uncertainty of genotype calls.

Chapter 4

Maximally Correlated Principal Component Analysis for low-coverage sequencing data

4.1 Abstract

With rapid advances in DNA sequencing technologies, different sequencing designs and platforms appear and provide options balancing between accuracy and cost. As a cost-effective choice, low to medium depth sequencing allows researchers afford to sequence more samples. But this sequencing relies on linkage-disequilibrium(LD)-based methods and has lower accuracy for genotype calling, especially rare-variant identification. Downstream association analyses are advised to incorporate sequencing uncertainties in order to increase the statistical power. As a fundamental dimension reduction technique, principal component analysis (PCA) is widely adopted in genetic studies to infer population structure and to adjust the bias caused by population differences in the association testing procedures. However, PCA is not designed to reveal non-linear patterns that arise in genetics data, for example, in low-depth genomic data. We propose an advancement of a data-adaptive approach, `MCPCA_PopGen`. Our method uses genotype likelihoods rather than genotypes and, thus, takes into account the uncertainty of raw sequencing reads to discover the non-linear patterns. `MCPCA_PopGen` is defined via an optimization that infers transformations of genotype likelihoods that maximizes Ky Fan norm of the transformed covariance matrix. We propose an efficient block coordinate descent algorithm determine transformation. Moreover, we propose a distributed algorithm to compute an approximation of MCPCA for dense sequencing data by chromosomal partitions. In the simulation, we reveal two major properties of our method for low-coverage sequencing data. For a locus with a low MAF, the transformation emphasizes the uncertainty in calling between heterozygous and the major homozygous loci. On the other hand, the transformation is more linear when both variants are common. We apply `MCPCA_PopGen` to a human sequencing sample from two indigenous Siberian populations, the Nganasan (nomadic hunters, NGA) from the Taymyr Peninsula in the Arctic Ocean, and the Yakut (herders, YAK) of North-Central Siberia with ?6x and ?4x coverage on average for the NGA and YAK samples, respectively. Our method can separate two groups with much higher confidence in comparison with many existing methods, including PCA.

4.2 Introduction

PCA is a fundamental dimension reduction technique that leads to low-dimensional representations of high-dimensional datasets. In the case of multivariate normally distributed data, PCA projects the original variables on a low-dimensional space that maximally explains the variance of the data. Among many its applications, PCA is a widely adopted tool in genetics studies to infer population structure and to adjust for the bias causing by population differences in the association testing procedures (Price et al., 2006; Patterson et al., 2006). Our method extends PCA approach in genetics studies with the application to low coverage NGS datasets.

With recent rapid advances in DNA sequencing technologies, high-throughput parallel-sequencing approaches has generated billions of short sequence reads on scale (Cirulli and Goldstein, 2010). Different sequencing designs and platforms provide options balancing between accuracy and cost. High-depth whole-genome sequencing identifies nearly all variants along genome with high confidence, but at high cost (Bentley et al., 2008). As a cost-effective alternative, low to medium depth NGS, which relies on linkage-disequilibrium (LD)-based methods, has lower accuracy. especially for rare-variant identification and genotype calling, but atmuch lower cost (Li et al., 2011; Pasaniuc et al., 2012). For low to medium depth NGS data, downstream analyses based on the raw sequencing data incorporating uncertainties is advantageous and comparable to high depth NGS (Korneliussen et al., 2014). Therefore researchers can afford to sequence more samples without sacrificing accuracy.

Using sequencing data, PCA has allowed us is to visualize differences between subpopulations from both case-control studies and structured populations. However, PCA is not designed to reveal non-linear associations that may arise, for example, from the uncertainties in low-depth genomic data. Several methods, including IsoMap (Tenenbaum et al., 2000), locally linear embedding (LLE) (Roweis and Saul, 2000) and Kernel PCA (KPCA) (Schölkopf et al., 1998) have been developed to capture non-linear patterns. KPCA, which has been has been widely adopted in population genetic studies, enables us easily to construct nonlinear versions of any algorithm (Schölkopf et al., 1998). In addition, KPCA has been successfully applied to gene expression data (Liu et al., 2005) and for the classification of samples, such as the types of tumor (Reverter et al., 2014). However, KPCA suffers from three major limitations: i) the kernel must be prespecified; ii) the corresponding transformation is identical at each loci for each feature; iii) the transformation has little effect for high coverage data where we are essentially limited to only 3 discrete categories, $(0, 1, 2)$, corresponding to the number of minor alleles.

With this motivation, we propose an advancement of a data-adaptive approach, Maximally Correlated Principal Component Analysis (MCPCA) introduced by Feizi and Tse (2017). As we shall see, this approach naturally addresses the first two limitations. To address the third, our data will use genotype likelihoods rather than a hard decision of any single genotype. Taking into account the uncertainty of raw sequencing reads provides an opportunity to model the nonlinear patterns in population genetic data. In particular, we will employ a continuous value, i.e, SNP dosage, to summarize the uncertainty in genotype calling. MCPCA is designed to transform dosage values at each loci to maximize the sum of a prespecified number of eigenvalues of the transformed covariance matrix. Our simulation reveals two major properties for MCPCA for low-coverage sequencing data. For a locus with a low MAF, the transformation emphasizes the uncertainly in calling between heterozygous and the major homozygous loci. On the other hand, the transformation is more linear when both variants are common. In the real data analysis, our method more clearly distinguishes the two populations even if using the genetic information from a single chromosome. In association studies, the method controls the inflation caused by the population stratification better than PCA.

We organize the paper as follows, first we introduce our optimization algorithm for dosage data. Second, we undergo an extensive simulation-based investigation to determine the type of data that will best benefit from an analysis based on MCPCA in comparison to the performance of PCA through various simulations in population stratification adjustments and a range of coverage depths.

4.3 Methods

In this section, we introduce the data type used for our method and the approaches to account for the genotype uncertainty. Then, we will lead to our statistical methodology, which we shall call `MCPCA_PopGen`.

4.3.1 Genotype Dosage Value

We use dosage as our measure of genotype uncertainty (Zheng et al., 2011). Dosage is the posterior mean of the genotype under additive coding. Our method can also be applied to dosage data imputed by Mach/Thunder (Li et al., 2011). With values 0, 1 and 2 assigned to the genotypes (major, major), (major, minor) and (minor, minor), respectively, the dosage, DS = Pr(1 | Data) + 2Pr(2 | Data).

4.3.2 MCPCA for Genotype Dosage Value

Let $\boldsymbol{X}$ be a $n \times p$ genotype matrix, $\boldsymbol{x}_i$ represent a vector of dosages for ith individual across p SNPs, and $\boldsymbol{x}^j$ represent dosage vector of jth SNP across n individuals. Thus, $X_{i,j}$, or $X(i,j)$ denotes the dosage value for the ith person at the jth SNP. Consequently, $0 \leq X(i,j) \leq 2$.Let $\phi(\boldsymbol{X}) = (\phi_1(\boldsymbol{x}^1), \cdots, \phi_p(\boldsymbol{x}^p))$ be any transformed scaled dosage matrix having zero means, $E[\phi_i(\boldsymbol{x}^i)] = 0$, and unit variances, $E[\phi_i(\boldsymbol{x}^i)^2] = 1$, for all $1 \leq i \leq p$. In addition, let $\boldsymbol{K}(i,i') = E[\phi_i(\boldsymbol{x}^i)\phi_{i'}(\boldsymbol{x}^{i'})]$ be the covariance matrix of transformed data. Define

$$\kappa_X = \{\boldsymbol{K} \in R^{p \times p} : E[\phi_i(\boldsymbol{x}^i)] = 0, E[\phi_i(\boldsymbol{x}^i)^2] = 1\}.$$

Our goal is to select $\boldsymbol{K} \in \kappa_X$ to maximize the sum of the top q eigenvalues $\lambda_1, \cdots, \lambda_q$,

$$\max_{\boldsymbol{K}} \sum_{r=1}^{q} \lambda_r(\boldsymbol{K}).$$

This maximum is commonly know at the Ky Fan q-norm.

To solve this optimization problem, we implement a block coordinate descent algorithm (Feizi and Tse, 2017). Figure 4.1 shows the pipeline of our method for the dosage data. In order to create a computationally feasible algorithm, we discretize dosage values. Care must be taken in creating this discretization. As we shall see, too many categories results in overfitting and too few sacrifices power. Here we consider three candidate methods: equal interval width, equal frequency and k-means clustering. We call our discretized variable, $X_d(i,j) \in \{1, 2, 3, \ldots m\}$, the scaled dosage. When genotype uncertainty is neglected, $m = 3$ and our method performs essentially the same as PCA. In the next section, we evaluate the performance of our method using both synthetic and real datasets.

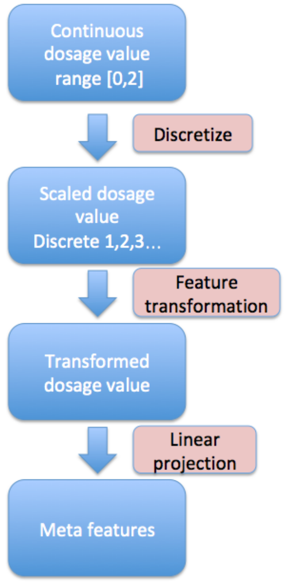


Figure 4.1: The pipeline of our method

4.4 Model Assessment

In this section, we describe how we generate the dosage data and then assess our method under different scenarios:

- Step 1: Establishing discretization schemes.
- Step 2: Choosing m, the number of categories.
- Step 3: Comparing variance explained under dosage discretization schemes, the true genotype, and PCA from observed genotypes.
- Step 4: Repeat step 3 with the data set reduced to highest quality SNPs.

4.4.1 Synthetic Data

To simulate the genotype data under a variety of assumptions concerning migration, recombination rate and population size under neutral models, we used the command line-based coalescent simulator *ms* (Hudson, 2002) to first simulate haplotype data for each population (African, Caucasian and Asian) and then use the *ms* output to generate latent genotypes of admixed individuals (see Appendix for exact *ms* commands used). We consider 50 subjects for each population. After obtaining latent genotypes, we filtered out rare variants for each population by removing those SNPs with MAF below 0.05, leaving us with a total of 1,081 latent genotype, G_{ij}. Using the method of Daye et al. (2012), we generated observed genotypes, $\tilde{G}_{ij}$ by perturbing G_{ij} with sequencing qualities sampled from the 1000 Genomes project. The method returns us the observed genotypes and the corresponding base-calling error probabilities. Then, we multiply the perturbed genotypes with the probabilities for correct base-calling to obtain the continuous dosage value data.

4.4.2 Step 1

Since if we treat each unique dosage as a category, we will end up with over 100 categories for each SNP. However, most of them are redundant and causes over-fitting problem. To address this issue, we first convert continuous variables into a categorical variables based on three basic unsupervised methods, equal-interval, equal-frequency and k-means. For all three methods, we need to specify the number of the finite categories (e.g. m categories) and each method separates the dosage into a m subset. Equal-interval method first records the minimum and the maximum of the dosage values observed from a SNP and splits the observed range into m bins. Equal-frequency method uses the similar strategy by replacing the observed dosage range with the observed dosage frequency. But note that equal frequency does not achieve perfect equally sized groups if the data contains duplicated values. K-means method assigns each dosage to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean and calculates the new means to be the centroids of the dosage in the new clusters.

Figure 4.2 shows the histograms of dosage values as well as the discretized dosage values with equal interval, equal frequency and k-means under low-coverage 5x. Not surprisingly, most values are clustered around 0, 1 and 2. We see using equal interval method just simply relabels the values and the data still maintains the same shape. The equal frequency and k-means options more strongly equalize the numbers of observations for each category. As expected, lower coverage depth results in higher uncertainty in the genotype data. `MCPCA_PopGen` aims at taking into account those low-qualities SNPs because they contains significant information on population differences. Table 4.1 shows the average Euclidean distance and percentage error. We notice that there is less uncertainty in the observed genotype when the coverage is high. Figure 4.3 shows the dosage value box-plot across different coverage depths, we can see the dosage values are highly sparse under coverage 30x.

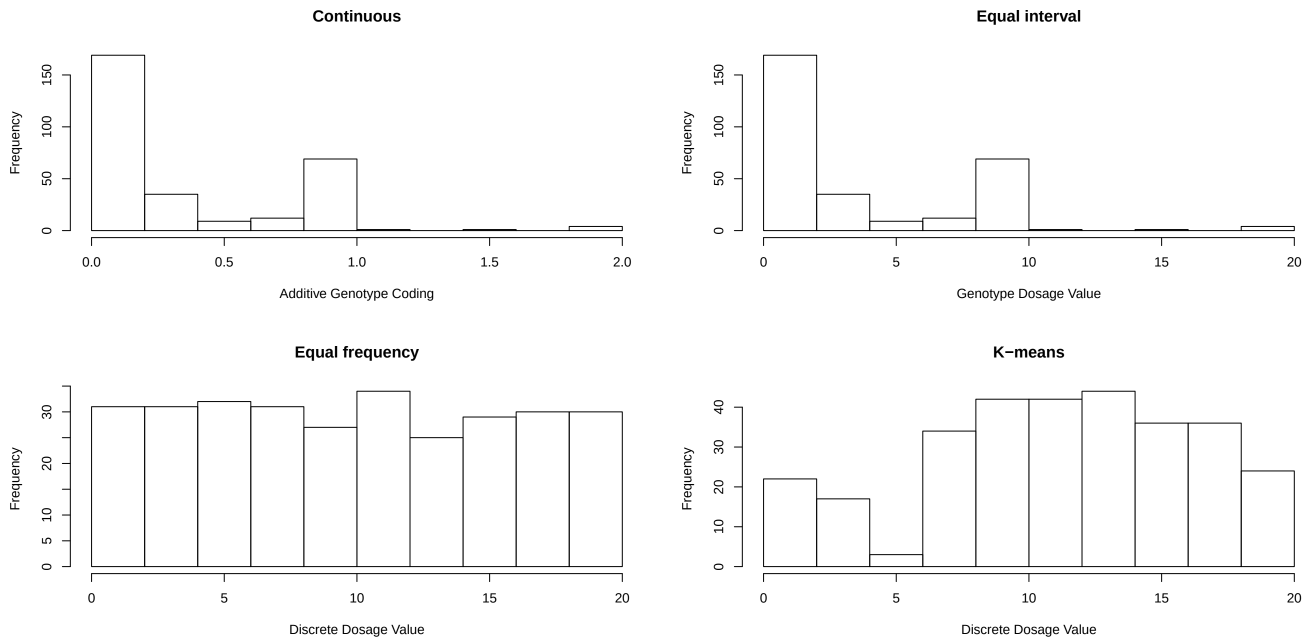


Figure 4.2: The histogram of genotype dosage values and discrete dosage values with 3 discretization methods: equal interval, equal frequency and k-means. The observed MAF = 0.16

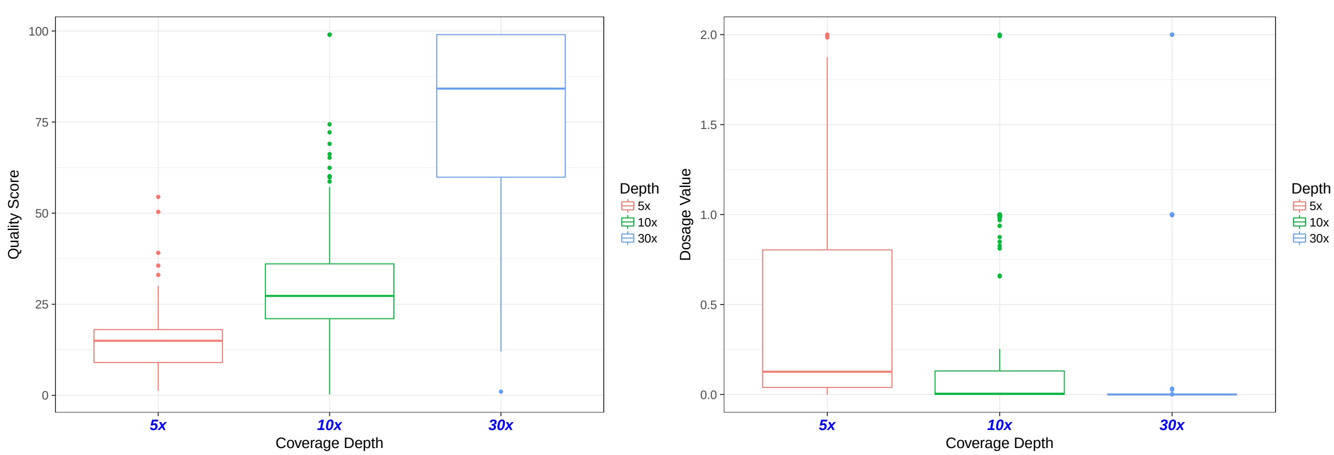


Figure 4.3: A: Box-plots of phred quality scores under three coverage depths. Notice that the quality is dramatically lower for 5x coverage. B: Box-plots of dosage values under three coverage depths. The width of the interquartile range shows the uncertainty in calling. This highlights the uncertainty in calling with 5x coverage that all but disappears with 30x coverage.

Table 4.1

Coverage depth	Average Euclidean distance	Percentage of error calling
5x	6.843	13.49%
10x	3.583	3.55%
30x	1.010	0.42%

This table shows the average Euclidean distance between true genotype and disturbed genotype at different coverage depths across all SNPs and the percentage of error calling for common variants

4.4.3 Step 2

Next, we assess the performance of `MCPCA_PopGen` using different data types. The green dash line in Figure 4.4 corresponds to the top 20 and the top 50 eigenvalues, respectively. The light blue line is the fraction curve of `MCPCA_PopGen` using unscaled dosage value data. We can see over-fitting due to an overdetermined number of categories. The blue solid line represents true genotype data and the orange dash line represents PCA using observed genotype data. The middle three solid lines are using `MCPCA_PopGen` under all three discretization options for dosage value data. All three options achieved better performance than PCA in observed genotype data. The sharp bend in these curves depends on the value of q. The larger q has a smoother curve but computationally more demanding. We also include the `MCPCA_PopGen` for true genotype, `MCPCA_PopGen` for noisy genotype and PCA for unscaled dosage values (Figure **??** Appendix). We use a nonparametric method, local polynomial regression (Jacoby, 2000) for fitting smooth curves to optimized nonlinear transformation, $\phi(.)$. Figure **??** (Appendix) shows the scatterplots and fitted curves across three discretization methods. We can see `MCPCA_PopGen` algorithm has little impact for the case of discrete genotype data. Consequently, `MCPCA_PopGen` on genotype data is very close to PCA.

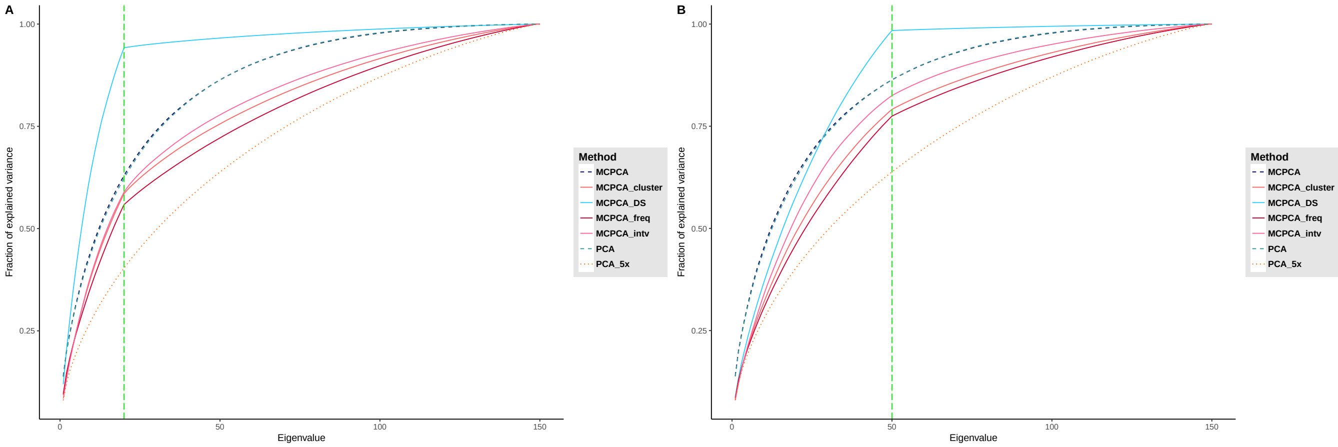


Figure 4.4: The figure shows the fraction of variance explained using q meta features computed by MCPCA with 3 different discretization methods at the low-coverage 5x; A uses $q = 20$ for all MCPCAs and B uses $q = 50$ for all MCPCAs; MCPCA–ds: MCPCA on dosage value; MCPCA–intv: MCPCA on discretized dosage value using equal interval; MCPCA–freq: MCPCA on discretized dosage value using equal frequency; MCPCA–cluster: MCPCA on discretized dosage value using k-means

4.4.4 Step 3

In the step, we evaluate the fractions of the variance explained by top 20 meta features under two coverage depths, 5x and 10x using 100 simulation replicates. Table 4.2 shows the average fractions and the standard errors across all methods. Each discretization version achieves a higher explained fraction than PCA, and the equal interval option is very close to true genotype data. Figure 4.5 shows the corresponding box-plots of the fractions of explained variation.

Table 4.2: Fraction of Variance Explained by Top 20 Meta Features (100 replicates)

Method	Coverage Depth 5x	Coverage Depth 10x
True genotype	61.94% (±0.66%)	61.94% (±0.66%)
PCA	40.98% (±0.44%)	52.58% (±0.56%)
MCPCA(equal interval)	59.36% (±1.63%)	61.63% (±0.99%)
MCPCA(equal frequency)	56.02% (±0.29%)	60.39% (±0.46%)
MCPCA(k-means)	56.63% (±0.73%)	61.57% (±0.49%)

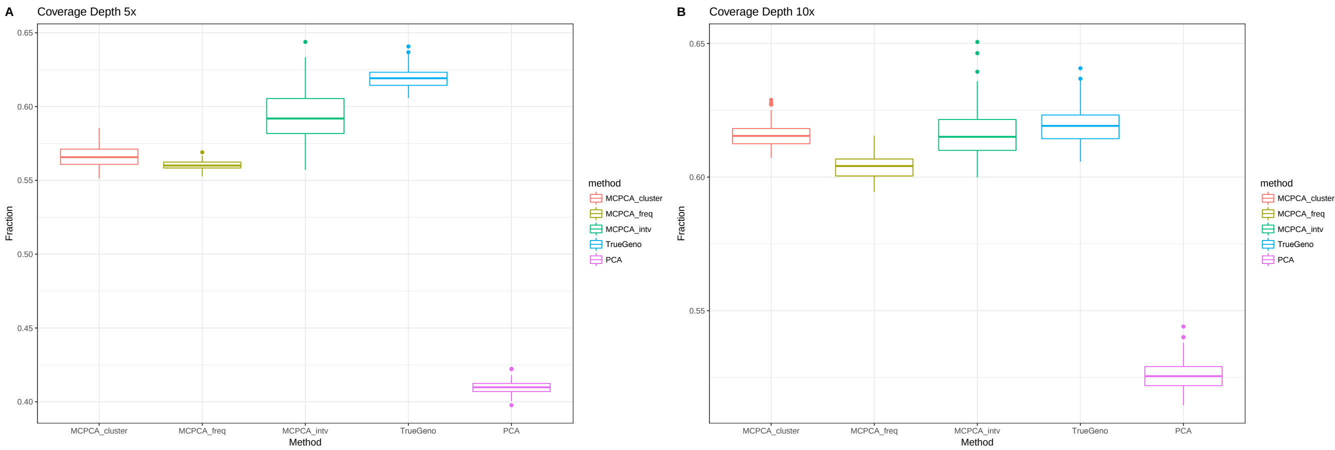


Figure 4.5: The figure shows the fraction of variance explained using top 20 meta features computed by PCA/MCPCA at different coverage depths; we set top 20 eigenvalues ($q = 20$ green dash line) as MCPCA optimization parameter

4.4.5 Step 4

As a quality control step, researchers frequently filter out reads with low quality scores and perform PCA analysis using only high quality SNPs. For this step, we compare how the fraction explained is impacted by this choice. At 5x depth and 10x depth, we filter those SNPs with mean quality scores below 15 and 29, respectively. We see no significant improvement for PCA between using all SNPs and high-quality SNPs, and the results remain well below our method using scaled dosage values. Therefore, `MCPCA_PopGen` can dramatically increase the percentage of variation explained under the low coverage both with and without filtering.

Table 4.3: Quality Comparison(100 replicates)

Coverage Depth	True Genotype	PCA (all SNPs)	PCA (High Quality SNPs)	MCPCA(k-means)
5x	61.94% (±0.66%)	40.98% (±0.44%)	41.15% (±0.44%)	56.63% (±0.73%)
10x	61.94% (±0.66%)	52.58% (±0.56%)	52.65% (±0.62%)	61.57% (±0.49%)

4.5 Guidance of Choosing Discretized Categories

In Step 2, we show that our method using 3 different discretization versions can achieve higher percentage of variance explained with compared to PCA. The question is that how we choose the discretized categories. We discretize the continuous dosage values into a smaller set to remove the redundant categories, but it is hard for us to find a uniform way to apply to all sequencing data. Figure 4.6 shows the distributions of MAF under different sequencing depths. We observe that with lower coverage depth the distribution is more far away from the true genotype distribution, and it is expected to contain more noise. It is very difficult to capture the low-frequency alleles when the coverage depth is 1x and most low-frequency alleles are treated as common variants. In addition, we characterize alleles into three categories: low-allele (MAF between 0 and 0.1), medium-allele (MAF between 0.1 and 0.4) and high-allele (MAF between 0.4 and 0.5). Figure 4.7 4.8 4.9 show the distribution of true genotype and dosage values under different coverage depths. We see the dosage is more close to true genotype under coverage 10x and the distribution is more continuous and far away from true genotype under extremely low coverage 1x. Therefore, we suggest that we should assign more categories for lower sequencing depth and less categories for relatively higher categories.

Here we choose 3 categories (5, 20 and 50) to discretize dosage value and assess how each category work on different sequencing coverages (Figure 4.10 4.11 4.12). If we use more categories the fraction curves are improved more but we are under a risk of overfitting. We see the curve of 5x using 50 categories and the curve of 10x using both 20 and 50 categories are overfitted (higher than the curve of true genotype). Therefore, we suggest that 5 categories for 10x and above, 10 to 20 categories for 5x to 10x and 50 categories for below 5x. However the argument of discretized categories can be extended to consider other nonuniform and data-dependent discretization as well. For example, a bayesian approach could be used based on MAF, LD and coverage depth information.

An easy way to give different categories to each locus is based on the observed MAF. Figure 4.6 shows that it is very common to detect the low-frequency alleles as common alleles under the extremely low coverage, 1x. Therefore, we will observe more heterozygous and major homozygous and transform into genotype dosage we will observe more dosages cluster around 0.5 and 1.5. In that case, the dosage will spread wider compared those with low observed MAF and it suggests that we should give larger categories to handle the uncertainty. Thus, we assign different categories to each locus according to their observed MAFs. For those MAFs between 0 and 0.1, we give 5 categories; 20 categories for loci whose MAFs are between 0.1 and 0.4; 50 categories if the MAFs are above 0.4. Figures 4.13 and 4.14 shows the fraction curves using mixed categories for 3 coverage depths data with $q = 20$ and 50, respectively. All curves control the overfitting well and meanwhile maximize the variation explained.

In addition, we also did several experiment by varying the same size, 90, 300 and 900 total individuals. The results are shown in Appendix, we observe that all curves control the overfitting well and meanwhile maximize the variation explained. Therefore, using the mixed category assignment based on MAF range is a good option.

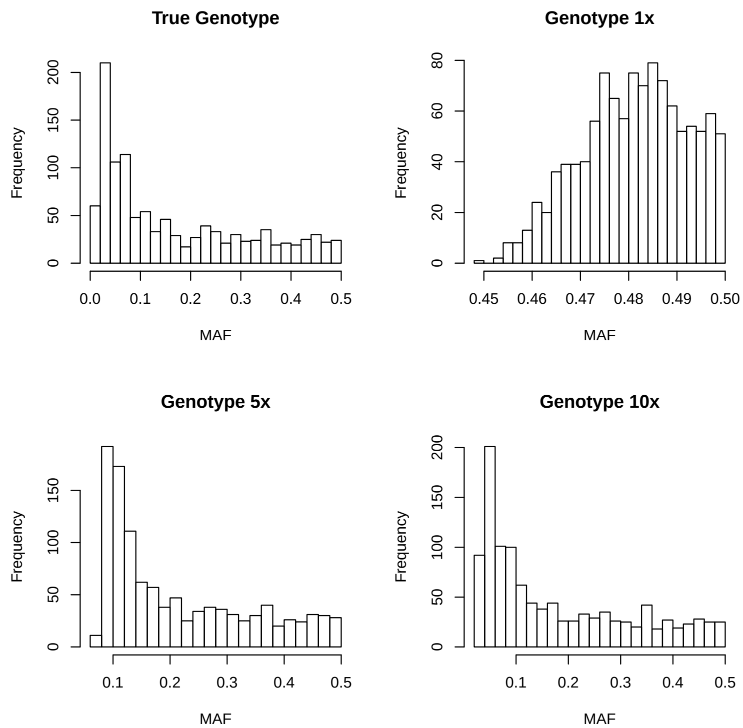


Figure 4.6: The histograms of MAFs: true genotype (top left); observed genotype 1x (top right); observed genotype 5x (top right); observed genotype 10x (top right)

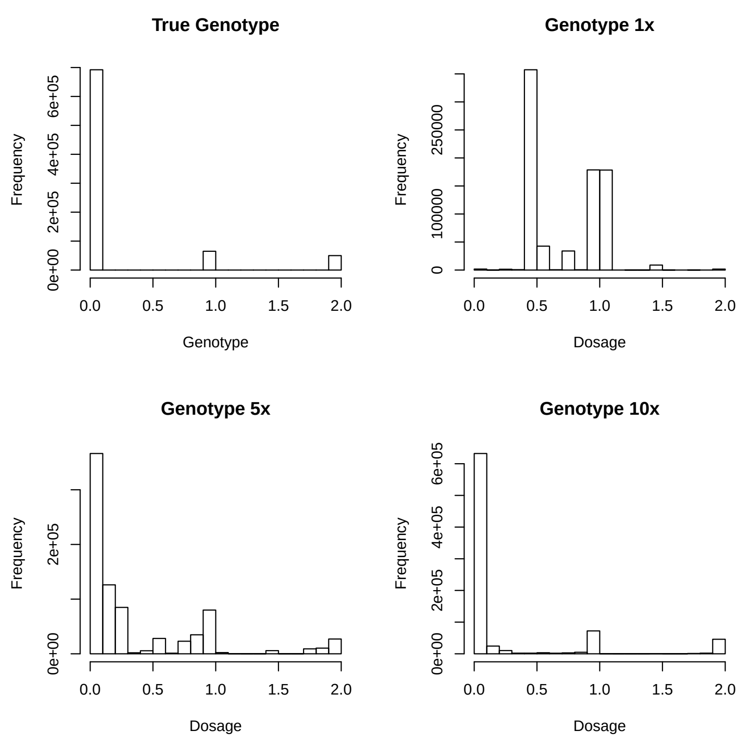


Figure 4.7: The histograms of dosages at low MAF: true genotype (top left); dosage value 1x (top right); dosage value 5x (top right); dosage value 10x (top right)

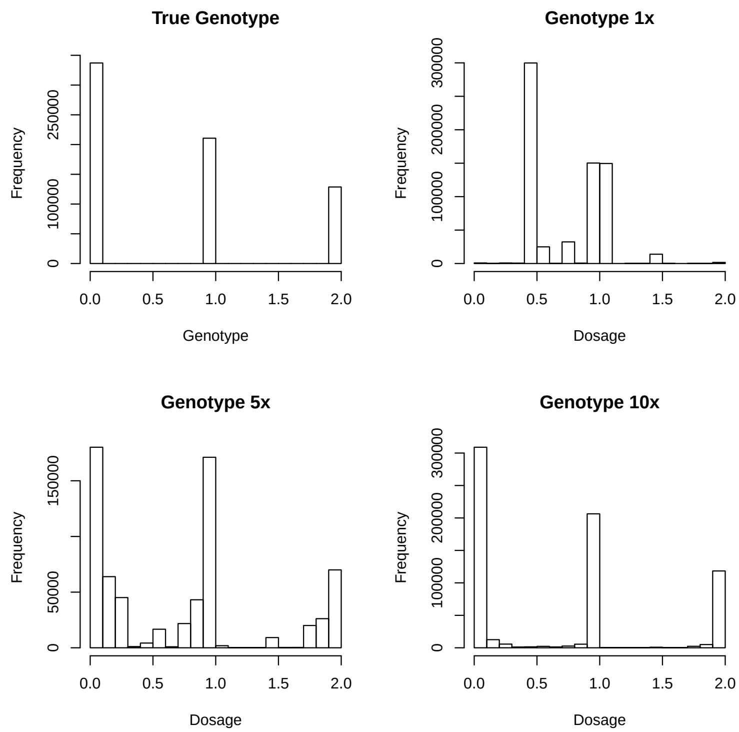


Figure 4.8: The histograms of dosages at medium MAF: true genotype (top left); dosage value 1x (top right); dosage value 5x (top right); dosage value 10x (top right)

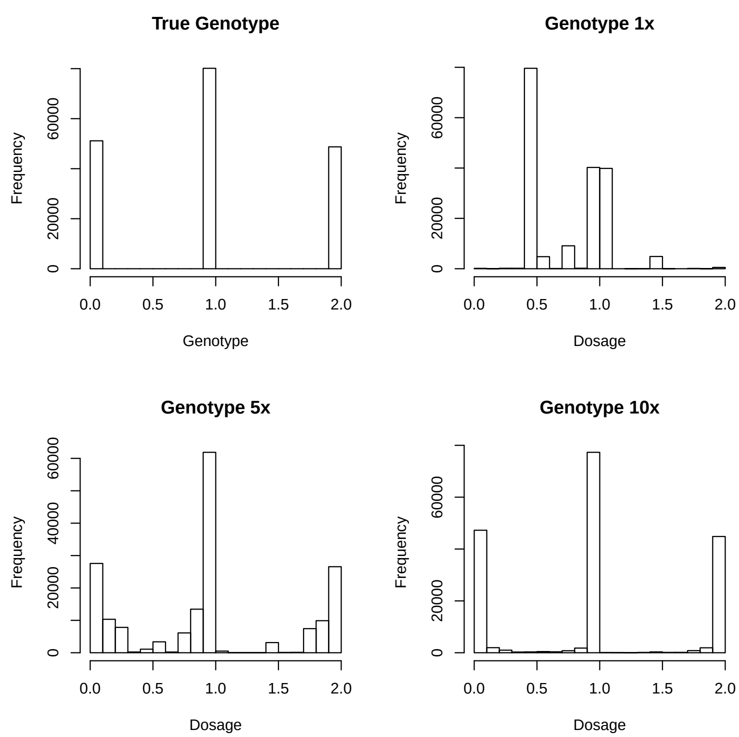


Figure 4.9: The histograms of dosages at high MAF: true genotype (top left); dosage value 1x (top right); dosage value 5x (top right); dosage value 10x (top right)

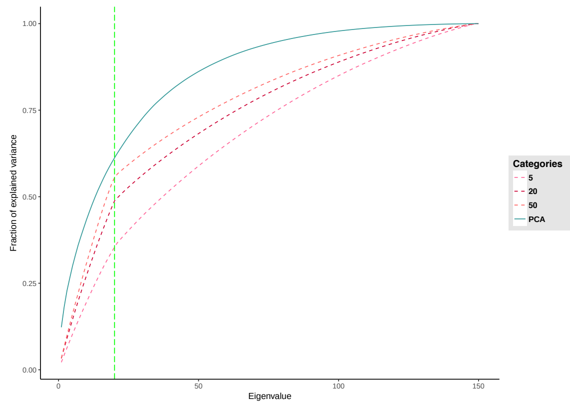


Figure 4.10: The fraction curves of variance explained at coverage depth 1x under three categories (5, 20 and 50)

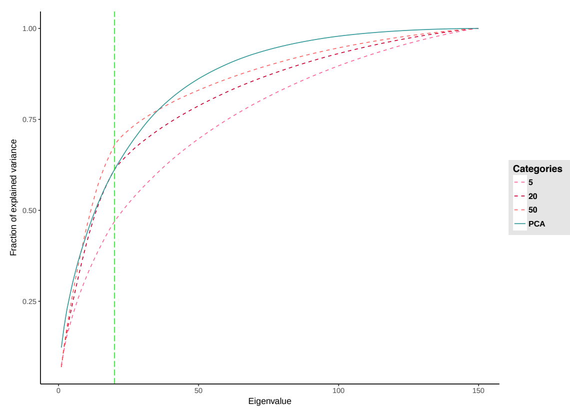


Figure 4.11: The fraction curves of variance explained at coverage depth 5x under three categories (5, 20 and 50)

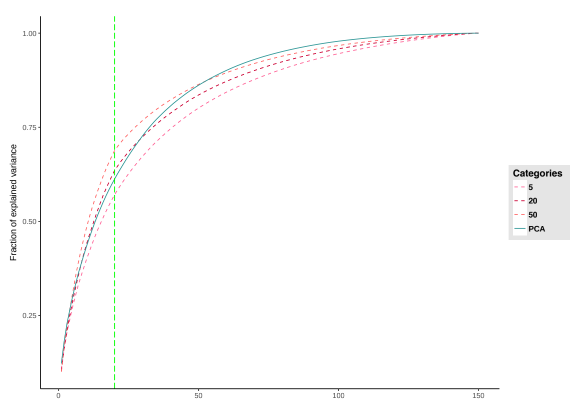


Figure 4.12: The fraction curves of variance explained at coverage depth 10x under three categories (5, 20 and 50)

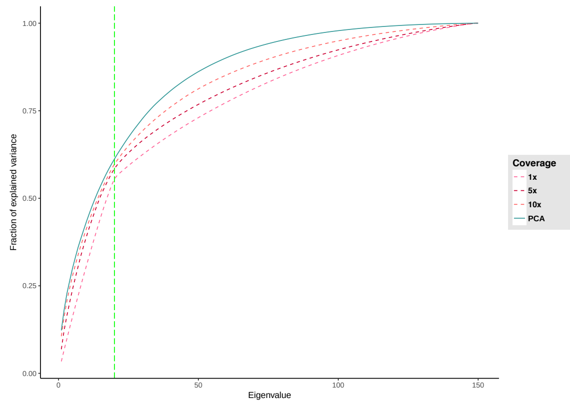


Figure 4.13: The fraction curves using mixed categories under different coverage depths

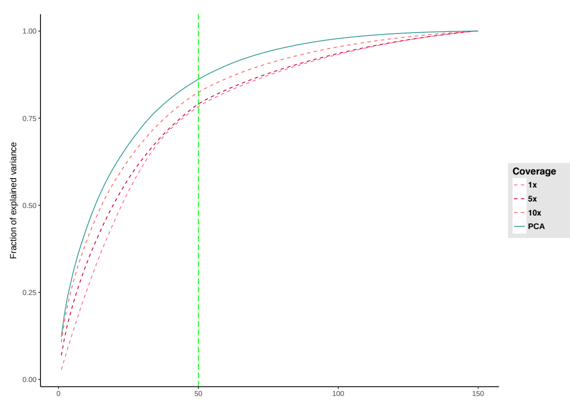


Figure 4.14: The fraction curves using mixed categories under different coverage depths

4.6 Example Using Siberian Populations

In (Hsieh et al., 2017), were collected human samples from two indigenous Siberian populations, the Nganasan (nomadic hunters, NGA) from the Taymyr Peninsula in the Arctic Ocean, and the Yakut (herders, YAK) of North-Central Siberia. The data are whole exome sequences of 21 individuals from each of the two Siberian populations. Exomes of these samples were captured using the Agilent SureSelect All Exon v5 no UTRs capture kit and pooled sequenced using Illumina HiSeq 2000. After applying the standard quality-control procedures of the University of Arizona Genetics Core, read mate pairs were aligned to the human reference genome (hg19), resulting in ∼6x and ∼4x coverage on average for the NGA and YAK samples, respectively. In their study, they developed a hypothesis-testing approach for inferring both the demographic and adaptive histories of indigenous Siberians. Because of the low coverage in the data, we called the posterior probability of the genotypes using the program ANGSD (Korneliussen et al., 2014) from exome BAM files, then used the posterior probability to calculate the genotype dosage value. We implemented our method on the dosage value data to classify the two groups. We compared out method with the program ngsPopGen (Fumagalli et al., 2014). Their idea is similar to ours, using posterior probabilities of sample allele frequencies to compute the probability of each site without calling genotype. It also accounts for the uncertainty of low quality and/or coverage sequencing data. In Figure 4.15, we computed the principal components of markers from chromosome 20, 21 and 22, respectively (set -minInd=10 in ANGSD to filter out the SNPs, which have less than 10 individuals with base calling). It shows `MCPCA_PopGen` can separate two groups better with higher percentage of explaining variance in comparison with PCA (ngsPopGen). In Figure 4.16, we called the posterior probability of the genotype likelihood across chromosome 1-22 (set -minInd=21 in ANGSD), and total 80,769 SNPs were used after filtering procedure. We draw the PCA plot using top 6 PCs from our method and ngsPopGen. Both methods can successfully identify two groups with PC1/PC2, but `MCPCA_PopGen` can also separate two groups with PC3/PC4 and PC5/PC6. In addition, `MCPCA_PopGen` can achieve bigger percentage of variance explained for each PC.

Next, we assess the performance of our method to control for population stratification in association tests. The population structure can be a confounding factor in association studies (Price et al., 2010b). Due to differences in allele frequency between two populations, we will have high rate of false positive discovery even there does not exist the genetic correlation between the genotype and the phenotype. Therefore, we need to consider this confounding factor and adjust it. By following the idea, we used 250 SNPs from the two Siberian populations that had a minor allele frequency (MAF) of at least 5%, and simulated the quantitive phenotype with $\sim N(0,1)$ for YAK and $\sim N(1,1)$ for NGA, respectively. We used top 2 principal components from our method to correct for population stratification and conducted the association test using a score statistic based method (Skotte et al., 2012) implemented in ANGSD. The method uses the joint likelihood of phenotypes and sequencing data without calling genotypes. Figure 4.17 shows the QQ-plot using chromosome 21. There is an obvious inflation if we do not adjust for the population stratification even there is no genetic association. Both adjustments lowered down the false positive rates, but our method can better control the population stratification (see the genomic inflation factor).

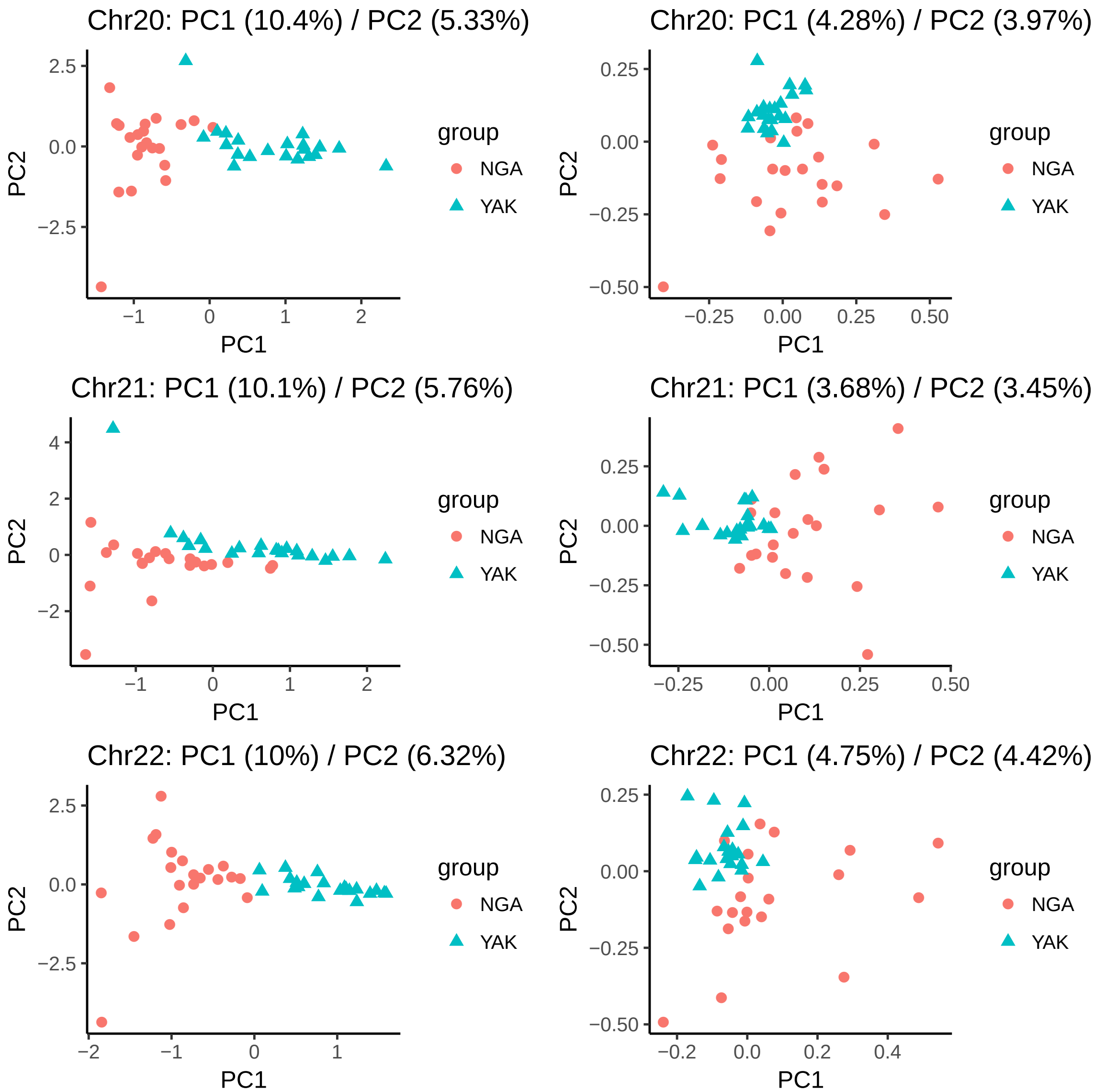


Figure 4.15: Left: MCPCA plot for chromosome 20-22; Right: PCA plot (ngsPopGen) for chromosome 20-22

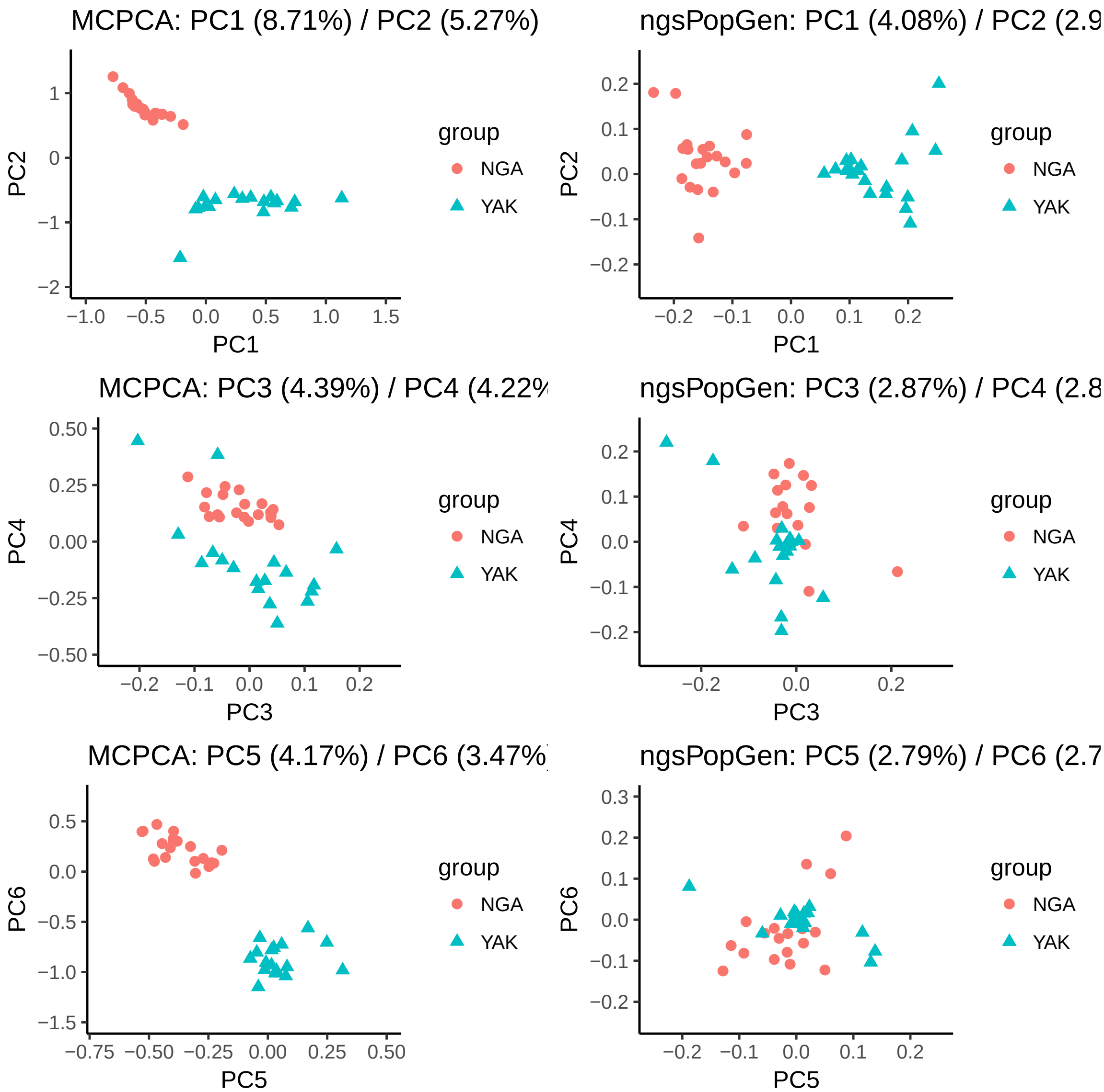


Figure 4.16: Left: MCPCA plot for top 6 PCs; Right: PCA plot (ngsPopGen) for top 6 PCs. 80769 SNPs across Chr1-22 used

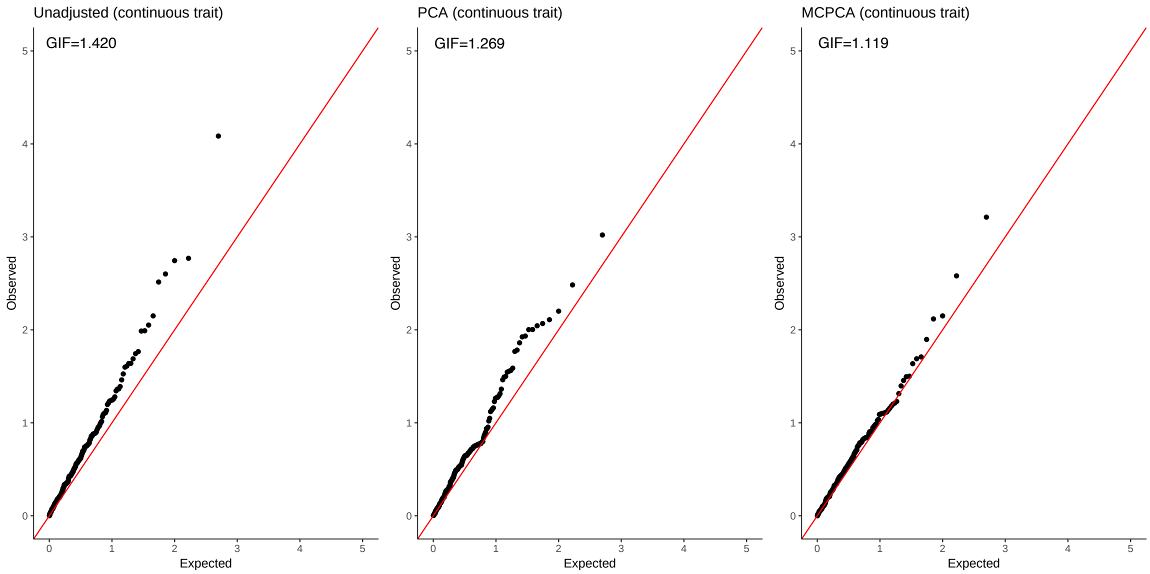


Figure 4.17: QQ-plot of p-values from association tests on chromosome 21; The genomic inflation factor (GIF) is provided

4.7 Discussion

In genetics studies, PCA is a widely adopted dimension reduction tool to infer population structure and to adjust for population stratification. Unlike high-density SNP array, new sequencing technologies allow us to model the genotype uncertainty of raw sequencing reads rather than a hard decision of any single genotype and provide options balancing between accuracy and cost. New approaches have to be developed in order to better fit this type of data.

In this article, we introduce a dimension reduction approach for low-coverage sequencing data. To account for the genotype uncertainty, we propose to use dosage values instead of the discrete genotype data. By considering both the genotype uncertainty and nonlinear correlations, our method transforms each SNP sequentially with maximizing the sum of top q eigenvalues of the transformed covariance matrix. The advantage of our method is that the transformation for each SNP do not necessarily need to be identical, which as required in KPCA. Instead, the transformation is more nonlinear with emphasizing the difference between heterozygous and the major homozygous for the SNP with low MAF and more linear when variants are common. To balance among computational feasibility, overfitting and statistical power, we suggest to discretize dosage values and provide three candidate methods. In simulation, we demonstrate that our method archives higher fractions of the variance explained by meta features with compared to PCA. In the Siberian data analysis, our method more clearly distinguishes the two populations even if using the genetic information from one chromosome, and controls the inflation caused by the population stratification better than PCA in in association studies.

Our method provides great power for low-coverage sequencing data. Our method provides a option for researchers with a limited budget to study in medical and population genetics. With the advantage in low-coverage data, we believe our it will be a potential approach to the study of non-model organisms (da Fonseca et al., 2016), which is often associated with the absence of closely related reference genomes and challenging sample material issues.

Chapter 5
Conclusion

In this dissertation, I summarizes three projects, which focus on applying and developing statistical methods to solve the problems raised from NGS data. In the first chapter, it discusses using the gene-based association test, SKAT to search for the modifier genes, which affect the clinical severity of DS. By using the technique, we found a gene, FKBP5 on chromosome 6 showing highly statistical significance and this signal might be sex-associated. However, we failed to identify any single SNPs mainly contribute to the DS phenotype and the liability of this approach is limited by the sample size and rarity of patients with both haploinsufficiency and mild clinical symptoms. Further work with a much greater sample size is needed to confirm an association between variation at FKBP5 and clinical outcome in DS.

The second chapter is motivated by the first chapter, since most clinical assessments are recorded in a discrete ordinal scale, e.g. the severity of DS, there is a need to develop a association test, which can take the ordinal outcome. With this motivation, I extend the region-based methods from a binary response to a ordinal categorical outcome. I build the method upon the proportional odds model and collapse the multiple variants within a region into genetic scores (Burden tests). Then, I derive the score-based test statistic from the likelihood function based on cumulative probability. Finally, I use Davis's method (Davis, 1982) to adjust the distribution of the test statistic to handle the issue from the sparsity of the genotype. Through the simulations, it shows simply dichotomizing an ordinal trait might lead substantial lose of information and consequently reduce the statistical power. But my method can control the type I error rate and increase power with compared to other five methods, even when the model is misspecified. Moreover, resampling method is easily implemented in our method to calculate the p-value if the sample size is small and the large sample theory might not hold.

In the last chapter, I introduce a novel dimension reduction approach for low-coverage sequencing data. By following the idea of considering the genotype uncertainty, I propose to use dosage values instead of the discrete additive genotypes into my method. By incorporating both the genotype uncertainty and non-linear correlations, I develop a data-adaptive kernel PCA to analyze the data.

By following the idea of considering the uncertainty of genotype calls and combined it with a data-adaptive kernel PCA for population inference. I showed my method can achieve higher fraction of variance explained using top principal components by considering the uncertainty of genotype calls. The method transforms each SNP sequentially with maximizing the sum of top q eigenvalues of the transformed covariance matrix. Since it is data-adaptive, it does not require the transformation for each SNP necessarily need to be the same. The advantage is that the transformation is more non-linear with emphasizing the difference between heterozygous and the major homozygous for the SNP with low MAF and more linear when variants are common. To balance among computational feasibility, overfitting and statistical power, I also provide the guidance of choosing discretized categories. In simulation, I demonstrate that the method archives higher fractions of the variance explained by meta features with compared to PCA. In the Siberian data analysis, my method more clearly distinguishes the two populations even if using the genetic information from one chromosome, and controls the inflation caused by the population stratification better than PCA in in association studies.

Bibliography

Agresti, A. (2010). *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons.

Alanis-Lobato, G., Cannistraci, C. V., Eriksson, A., Manica, A., and Ravasi, T. (2015). Highlighting nonlinear patterns in population genetics datasets. *Scientific reports*, 5:8140.

Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *The American Journal of Human Genetics*, 96(2):329–339.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218):53.

Bick, D. and Dimmock, D. (2011). Whole exome and whole genome sequencing. *Current opinion in pediatrics*, 23(6):594–600.

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Chen, H., Meigs, J. B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genetic epidemiology*, 37(2):196–204.

Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature protocols*, 6(2):121–133.

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., Zepeda-Mendoza, M. L., Campos, P. F., Heller, R., and Pereira, R. J. (2016). Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine genomics*, 30:3–13.

Davis, C. S. (1982). The distribution of a linear combination of chi square variables.

Daye, Z. J., Li, H., and Wei, Z. (2012). A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic acids research*, 40(8):e60–e60.

Depienne, C., Trouillard, O., Saint-Martin, C., Isabelle, A., Bouteiller, D., Carpentier, W., Keren, B., Abert, B., Gautier, A., Baulac, S., et al. (2008). Spectrum of scn1a gene mutations associated with dravet syndrome: analysis of 333 patients. *Journal of medical genetics*.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450.

Feizi, S. and Tse, D. (2017). Maximally correlated principal component analysis. *arXiv:1702.05471*.

Feng, R., Leckman, J. F., and Zhang, H. (2004). Linkage analysis of ordinal traits for pedigree data. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16739–16744.

Fumagalli, M., Vieira, F. G., Linderoth, T., and Nielsen, R. (2014). ngstools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, 30(10):1486–1487.

Guo, Y., Baum, L. W., Sham, P. C., Wong, V., Ng, P. W., Lui, C. H. T., Sin, N. C., Tsoi, T. H., Tang, C. S., Kwan, J. S., et al. (2011). Two-stage genome-wide association study identifies variants in camsap1l1 as susceptibility loci for epilepsy in chinese. *Human molecular genetics*, 21(5):1184–1189.

Hammer, M. F., Ishii, A., Johnstone, L., Tchourbanov, A., Lau, B., Sprissler, R., Hallmark, B., Zhang, M., Zhou, J., Watkins, J., et al. (2017). Rare variants of small effect size in neuronal excitability genes influence clinical outcome in japanese cases of scn1a truncation-positive dravet syndrome. *PloS one*, 12(7):e0180485.

Heinzen, E. L., Radtke, R. A., Urban, T. J., Cavalleri, G. L., Depondt, C., Need, A. C., Walley, N. M., Nicoletti, P., Ge, D., Catarino, C. B., et al. (2010). Rare deletions at 16p13. 11 predispose to a diverse spectrum of sporadic epilepsy syndromes. *The American Journal of Human Genetics*, 86(5):707–718.

Hindorff, L., Junkins, H., Mehta, J., and Manolio, T. (2011). A catalog of published genome-wide association studies 2010. *Ref Type: Generic*.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108.

Hsieh, P., Hallmark, B., Watkins, J., Karafet, T. M., Osipova, L. P., Gutenkunst, R. N., and Hammer, M. F. (2017). Exome sequencing provides evidence of polygenic adaptation to a fat-rich animal diet in indigenous siberian populations. *Molecular biology and evolution*, 34(11):2913–2926.

Hudson, R. R. (2002). Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338.

Ishii, A., Watkins, J. C., Chen, D., Hirose, S., and Hammer, M. F. (2017). Clinical implications of scn1a missense and truncation variants in a large japanese cohort with dravet syndrome. *Epilepsia*, 58(2):282–290.

Jacoby, W. G. (2000). Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19(4):577–613.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., San-Giovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389.

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. (2014). Angsd: analysis of next generation sequencing data. *BMC bioinformatics*, 15(1):356.

Kumari, R., Lakhan, R., Kumar, S., Garg, R., Misra, U., Kalita, J., and Mittal, B. (2013). Scn1aivs5-91g¿ a polymorphism is associated with susceptibility to epilepsy but not with drug responsiveness. *Biochimie*, 95(6):1350–1353.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

Li, B. and 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(3):311–321.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., and Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome research*.

Lin, D.-Y. and Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367.

Liu, Z., Chen, D., and Bensmail, H. (2005). Gene expression data classification with kernel principal component analysis. *BioMed Research International*, 2005(2):155–159.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384.

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(7265):747–753.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142.

Morgenthaler, S. and 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(1):28–56.

Mulley, J. C., Scheffer, I. E., Petrou, S., Dibbens, L. M., Berkovic, S. F., and Harkin, L. A. (2005). Scn1a mutations and epilepsy. *Human mutation*, 25(6):535–542.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322.

Need, A. C. and Goldstein, D. B. (2010). Whole genome association studies in complex diseases: where do we stand? *Dialogues in clinical neuroscience*, 12(1):37.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). Snp calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PloS one*, 7(7):e37558.

Özmen, M., Dilber, C., Tatlı, B., Aydınlı, N., Çalışkan, M., and Ekici, B. (2011). Severe myoclonic epilepsy of infancy (dravet syndrome): Clinical and genetic features of nine turkish patients. *Annals of Indian Academy of Neurology*, 14(3):178.

Pan, W. (2009). Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic epidemiology*, 33(6):497–507.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature genetics*, 44(6):631.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L.-J., and Sunyaev, S. R. (2010a). Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010b). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459.

Reverter, F., Vegas, E., and Oller, J. M. (2014). Kernel-pca data integration with enhanced interpretability. *BMC systems biology*, 8(2):S6.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome research*, 15(11):1576–1583.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319.

Skotte, L., Korneliussen, T. S., and Albrechtsen, A. (2012). Association testing for next-generation sequencing data using score statistics. *Genet. Epidemiol.*, 36(5):430–437.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Suzuki, T., Remington, G., Mulsant, B. H., Uchida, H., Rajji, T. K., Graff-Guerrero, A., Mimura, M., and Mamo, D. C. (2012). Defining treatment-resistant schizophrenia and response to antipsychotics: a review and recommendation. *Psychiatry research*, 197(1):1–6.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–13549.

Veeramah, K. R., Johnstone, L., Karafet, T. M., Wolf, D., Sprissler, R., Salogiannis, J., Barth-Maron, A., Greenberg, M. E., Stuhlmann, T., Weinert, S., et al. (2013). Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia*, 54(7):1270–1281.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6):1278–1283.

Wang, X., Ye, Y., and Zhang, H. (2006). Family-based association tests for ordinal traits adjusting for covariates. *Genetic epidemiology*, 30(8):728–736.

Wu, G. and Zhi, D. (2013). Pathway-based approaches for sequencing-based genome-wide association studies. *Genetic epidemiology*, 37(5):478–494.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Yee, T. W. (2008). Vgam: Vector generalized linear and additive models. *R package version 0.7-7, URL http://CRAN. R-project. org/package= VGAM.*

Zhang, H., Feng, R., and Zhu, H. (2003). A latent variable model of segregation analysis for ordinal traits. *Journal of the American Statistical Association*, 98(464):1023–1034.

Zhang, H., Wang, X., and Ye, Y. (2006). Detection of genes for ordinal traits in nuclear families and a unified approach for association studies. *Genetics*, 172(1):693–699.

Zheng, J., Li, Y., Abecasis, G. R., and Scheet, P. (2011). A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic epidemiology*, 35(2):102–110.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198.