RICE UNIVERSITY

Statistical Methods for Analyzing Rare Variant Complex Trait Associations via Sequence Data

by

Dajiang Liu

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE

Marek Kimmel, Professor, Chair, Statistics

Suzanne Leal, Professor, Molecular and Human Genetics, Baylor College of Medicine, Advisor

Dennis Cox, Professor, Statistics

Luay Nakhleh, Association Professor, Computer Science

John Belmont, Professor, Molecular and Human Genetics, Baylor College of Medicine

HOUSTON, TEXAS
NOVEMBER 2011

Abstract

Statistical Methods for Analyzing Rare Variant Complex Trait Associations via Sequence Data

by

Dajiang Liu

There is solid evidence that complex human diseases can be caused by rare variants. Next generation sequencing technology has revolutionized the study of complex human diseases, and made possible detecting associations with rare variants. Traditional statistical methods can be inefficient for analyzing sequence data and underpowered. In addition, due to high cost of sequencing, it is also necessary to explore novel cost effective studies in order to maximize power and reduce sequencing cost. In this thesis, three0 important problems for analyzing sequence data and detecting associations with rare variants are presented. In the first chapter, we presented a new method for detecting rare variants/binary trait associations in the presence of gene interactions. In the second chapter, we explored cost effective study designs for replicating sequence based association studies, combining both sequencing and customized genotyping. In the third chapter, we present a method for analyzing multiple phenotypes in selected samples, such that phenotypes that are commonly measured in different studies can be jointly analyzed to improve power. The methods and study designs presented are important for dissecting complex trait etiologies using sequence data.

Acknowledgments

Obtaining a doctorate degree is one of the most challenging things in the world. It cannot be done without the help from my teachers, families and friends. First, I want to thank my mentor Dr. Suzanne Leal, who introduced me to this field, taught me necessary things I need to know to become an independent research, encouraged me to pursue my own interest, and provided me all the resources that I need. I am especially thankful that she is very generous about her time, and make herself available whenever I need her. I want to thank Dr. Kimmel, who recruited me into statistics, and taught me a lot of valuable things in statistics and in life and Professor Belmont, Cox and Nakhleh for their time and valuable comments.

I am very blessed to be surrounded by an excellent and motivated group of classmates, Biao, Bingshan, Kwanghyuk Lee, Fremiet Lara, Gao Wang, Ian Gibson, Kristine, Merry-Lynn McDonald, Regie Santos, Rosa Banuelos, Stanley Hooker, Terrance Savitsky, Zongxiao He. Among them, I want to specially thank Bingshan, with whom I learned a lot of things and had a lot of good discussions on my research, and thank Gao for his help not only in academics but also in life.

Last but not least, I want to thank my parents, Changjiang Su and Naikuan Liu, both of whom passed away during my Ph.D. study. They set a role model for me and gave me endless encouragement and support whenever I need them. This dissertation is devoted to them.

Contents

Acknowledgments iii

Contents iv

List of Figures vii

List of Tables viii

Introduction 1

A Novel Adaptive Methods for Mapping Rare Variants in the Presence of Gene Interactions via Sequence Data 8

2.1. Background: 9

2.2. Results 13

2.2.1. Rare Variant Frequency Distributions in Generated Case-Control Samples... 13

2.2.2. Evaluation of Type I Error: 14

2.2.3. Power Comparison: 15

2.2.3.1. Main Effects Model without Misclassification: 15

2.2.3.2. Impact of Misclassification: 16

2.2.4. Interaction Models: 17

2.2.4.1. Within Gene Interaction Model: 17

2.2.4.2. Between Gene Interaction Model: 17

2.2.5. *ANGPTL* Variants and Energy Metabolism in Humans: 18

2.3. Discussion 20

2.4. Methods: 27

2.4.1. Sample Risk 27

2.4.2. Choice of Kernels 29

2.4.2.1. Hyper-geometric Kernel 30

2.4.2.2. Marginal Binomial Kernel 30

2.4.2.3. Asymptotic Normal Kernel 31

2.4.3. Test Statistics 31

2.4.4. Controlling for Confounders: 32

2.4.5. Monte Carlo Approximation 33

2.4.5.1. Monte Carlo Approximation under the Null Hypothesis 33

2.4.5.2. Monte Carlo Approximation under the Alternative Hypothesis: Power calculations 35
2.4.6. Rare Variant Analysis Methods Which Are Compared to the KBAC: 37
2.4.7. Generation of Genetic Data: 38
2.4.7.1. Simulation of Demographic Model and Selections: 38
2.4.7.2. Generation of Phenotype Data with only Main Effects: 39
2.4.7.3. Generation of Data with Gene Interactions 40
2.4.8. Analysis of Energy Metabolism Traits and Rare Variants in *ANGPTL 3, 4, 5* and *6* 41

Replication Strategies of Rare Variant Complex Trait Association via Sequencing 52

3.1. Background: 53
3.2. Material and Methods: 57
3.2.1. Probabilistic Model for Sequencing Errors: 58
3.2.2. Models of Genotyping Errors: 60
3.2.3. Power calculation for sequence-based and variant-based replication: 62
3.2.4. Simulations of Complex Demographic Models and Selections: 64
3.2.5. Generations of Phenotypic Model: 65
3.2.6. Applications to the Dallas Heart Study Sequence Data: 66
3.3. Results: 69
3.3.1. Discovery Rate of Rare Variant Sites and Frequencies: 69
3.3.2. Power Comparisons for Sequence-Based and Variant-Based Replication Strategies: 71
3.3.3. Applications to the Dallas Heart Study data: 73
3.4. Discussion: 75

A Flexible Likelihood Framework for Detecting Associations with Secondary Phenotypes in Selected Samples: Applications to Sequence Data 87

4.1. Background 88
4.2. Materials and Methods: 91
4.2.1. Locus Multi-site Genotype Coding Schemes 91
4.2.2. A General Probability Model for Multiple Phenotypes in Selected Samples . 92
4.2.3. Association Testing 93
4.2.4. Combining Different Cohorts for the Analyses of Secondary Phenotypes 94

4.2.5. Generation of Genetic and Phenotypic Data 95
4.2.6. Software Availability 96
4.3. Results 96
4.3.1. Power of Detecting Secondary Phenotype Rare Variant Associations 97
4.3.2. Applications to the *ANGPTL* Family of Genes 100
4.4. Discussion 102
References **112**

List of Figures

Figure 1: Quantile-Quantile (QQ) plot of p-values..... 42

Figure 2: Impact of misclassifications under main effects model with fixed genetic effects using simulated SFS for AA. 43

Figure 3: Impact of misclassifications under main effects model with variable genetic effects using simulated SFS for AA. 44

Figure 4: Power comparisons for within gene (left panel) and between gene interaction model (right panel) with simulated SFS for AA. 45

List of Tables

Table 1 :Rare variant summary statistics..... 46

Table 2 Rare variant summary statistics. 48

Table 3: Association analyses of the *ANGPTL 3,4,5* and *6* gene variants with human energy metabolism phenotypes. 50

Table 4 Discoveries of rare variants in small and large scale genetic studies 81

Table 5 Power comparisons of sequencing-based and variant-based replication under variable effects model 82

Table 6: Power comparisons of sequence-based and variant-based replication under fixed effects model 83

Table 7: Analyses of sequence data from the *ANGPTL* 3, 4, 5, and 6 genes 85

Table 8: Definitions of Selection Mechanisms 106

Table 9 Power to detect secondary trait *T* associations using case-control, extreme-trait, and multiple-trait study design. 107

Table 10: Power to detect secondary trait T associations for individual studies (case-control and multiple-trait) and the combined analysis..... 108

Table 11: Results for the secondary phenotype analyses using sequence data from the *ANGPTL3, ANGPTL4, ANGPTL5* and *ANGPTL6* genes. 109

Chapter 1

Introduction

Currently there is great interest in investigating the etiology of complex disease due to rare variants[1-6]. Until recently, indirect mapping of common variants has been the emphasis of complex trait association studies. It has been demonstrated that common variants tend to have modest phenotypic effects while rare variants are likely to have stronger phenotypic effects[7], although not strong enough to cause familial aggregation[8]. For mapping complex diseases due to common variants, instead of genotyping functional variants, tagSNPs are genotyped which act as a proxy for the underlying causal variants. For rare variant association studies, indirect mapping is not an optimal approach due to low correlations (r^2) between tagSNPs and rare variants. Instead, direct mapping should be used, where functional variants are analyzed. In order to implement direct mapping, variants must first be identified. Large scale sequencing efforts have begun including the 1000 Genome Project, which will provide a better understanding of the allelic architecture of the genome and a detailed catalog of human variants. Next generation sequencing technologies e.g. Roche 454, ABI SOLiD, and Illumina HiSeq, have made it feasible to carry-out rare variant association studies of candidate regions, exomes and genomes.

Ideally, when carrying out direct mapping, only causal variants should be tested for associations. When DNA samples are sequenced, both causal and non-causal variants are uncovered. Bioinformatics tools[9,10] or filters[1] can be used to predict functionality of variants, although tools such as PolyPhen[10] or SIFT[9]can have low sensitivity and specificity[6,11]. Empirical studies have shown that predictive errors can be as high as 47% and 37% for PolyPhen and SIFT respectively[6]; therefore, their usefulness in selecting

variants to be included in association analysis is limited. Even when functionality can be correctly inferred, whether the identified variants affect the phenotype of interest is still unknown. Two types of misclassifications of variant causality can frequently arise: 1.) non-causal variants are included in the analysis: a.) sequencing incorrectly identifies monomorphic sites as variant sites (false positive SNP discovery), b.) variants are falsely predicted to be functional or c.) variants are functional but non-causal; 2.) causal variants are excluded from the analysis: a.) due to locus heterogeneity, not all loci containing causal variants are included in the analysis, b.) region not sequenced, e.g. intronic variants, c.) variants not detected by sequencing assay (false negative SNP discovery) or d.) causal variants are falsely predicted to be non-functional.

Gene interactions are believed to be involved in a broad spectrum of complex disease etiologies[12]. Although a number of methods have been developed to detect gene interactions between common variants[13-16], their detection has been limited[13]. There is evidence that rare variant interaction also plays a role in disease etiology. In direct association mapping of rare variants, one or more genetic loci are commonly jointly analyzed in order to aggregate information, for example genes with similar functions or residing in the same pathway[3,4]. Therefore it is necessary to account for potential interactions between rare variants in different loci[17] and interactions between common and rare variants[18,19].

Driven by the advancement of sequencing technologies and availability of data, statistical and computational methods are needed for analyzing sequence data. It has been

demonstrated that methods used to analyze common variants are low powered when applied to the analysis of rare variants[20,21]. Methods to analyze rare variants have been proposed[20,21]; although they have clear advantages over implementing common variant analysis approaches, more powerful and robust methods need to be developed to analyze rare variant data, especially in the presence of variant misclassification or gene interactions.

In Chapter 2 of the thesis, a novel adaptive method Kernel Based Adaptive Cluster was developed, which is robust and powerful to variant misclassifications and gene interactions. Its statistical properties were explored and compared with several other rare variant analysis methods, including weighted sum statistics[21] and combined multivariate and collapsing[20]. Chapter 2 of the thesis has been published in PLoS Genetics [22].

In order to avoid spurious or false positive findings in association mapping, replicating significant associations discovered in an exploratory sample (stage 1) using an independent dataset (stage 2) is an indispensible part of every genetic association study. For mapping rare variants, gene based tests are usually performed, such as combined multivariate and collapsing[20] or weight sum statistic[21]. In a gene based test, multiple rare variants are jointly analyzed in order to aggregate signals from the gene region[3-5,20,21]. To replicate significant findings in stage 1 studies, two different strategies can be used. As a first strategy, only the variants at the nucleotide sites uncovered from the

original sample are followed up. Using this strategy novel nucleotide sites that are present only in the stage 2 sample will not be incorporated in the replication study. This constitutes a replication in a "strict" sense, i.e. both the gene region and the variants uncovered in the stage 1 sample are followed up in the replication sample. When only variants uncovered in the stage 1 sample are of interest, genotyping is sufficient. We will refer to this replication strategy as "variant-based". An alternative strategy is to follow-up the entire gene region identified in the stage 1 sample. For this design, analysis of the stage 2 sample is not restricted to the nucleotide sites uncovered in stage 1. Variants from novel sites in the replication sample are also assessed for their associations with the phenotype of interest. We will refer to this design as "sequence-based" replication. With this strategy, sequencing the target gene in the stage 2 sample is necessary. The efficiencies of the two proposed strategies are compared in Chapter 3 of the thesis. The results of Chapter 3 was published in AJHG – the American Journal of Human Genetics [23] .

In order to design powerful studies, it is necessary to deeply sequence samples from a large number of individuals[24]. However, many existing studies are small to moderate sized, due to the high cost of sequencing or limited availability of samples, and are therefore inadequately powered. It would be advantageous if different studies which measure the same phenotypes could be jointly analyzed to increase power. In particular, many clinically important traits, such as body mass index, systolic and diastolic blood pressure are often measured in different studies. When combined analysis is performed, in addition to incorporating studies that are targeted at the same primary traits, it is

desirable to also analyze data from studies for which the phenotype of interest is measured as an additional outcome. Combined analyses require modeling multiple phenotypes since different studies may sequence selected samples targeted at different primary traits. Similar to the idea of analysis of covariance (ANCOVA), jointly analyzing multiple phenotypes makes it possible to distinguish the phenotype covariance component that is due to gene pleiotropy and the component that is attributable to residual correlations.

Currently, most studies sequence selected samples, e.g. case-control samples or individuals with extreme phenotypes[3,4]. Sequencing selected samples reduces sequencing cost and improves power. Due to sample ascertainment, secondary traits can be associated with the gene region in a selected sample even though they are independent in the general population. For example, consider a gene that is associated with the primary trait, but not with the secondary trait in the general population (**Figure 1**). In a sample that consists of individuals with extreme primary trait values, the causative variant frequency will be different between individuals from the upper and lower extremes. The mean value for the secondary trait will also be different due to phenotypic correlations. Therefore, a spurious association can occur between the gene region and the secondary trait unless the sample ascertainment scheme is correctly modeled. The selection criteria for a sequencing study can be complicated and may involve multiple traits (multiple-trait study) or sub-phenotypes. For instance, it is hypothesized that the etiologies of type 2 diabetes (T2D) are different in obese and non-obese individuals[25,26]. In order to reduce phenotype heterogeneity and potentially improve

power, a study of T2D might be performed using an obese population. There have been methods developed for detecting associations with multiple phenotypes in selected samples[27,28]. However, these methods are limited to case-control studies. They are not applicable to more complicated study designs, e.g. the studies that sequence individuals with extreme primary traits (extreme-trait study), or the studies where secondary phenotypes are also involved in sample selection. In particular, extreme-trait study design is becoming increasingly popular and widely applied [29-31]. The results for detecting associations with secondary traits can be seriously biased if the secondary traits are not properly analyzed [27]. It is desirable to have a unified approach for analyzing secondary phenotypes from all available datasets. A flexible likelihood based methods MULTI-TRAIT-MAP were discussed in Chapter 4 of the thesis for this purpose. The material in Chapter 4 is currently under review with revision submitted.

Chapter 2

A Novel Adaptive Methods for Mapping Rare Variants in the Presence of Gene Interactions via Sequence Data

2.1. Background:

Gene interactions are believed to be involved in a broad spectrum of complex disease etiologies[12]. Although a number of methods have been developed to detect gene interactions between common variants[13-16], their detection has been limited[13]. There is evidence that rare variant interaction also plays a role in disease etiology. In direct association mapping of rare variants, one or more genetic loci are commonly jointly analyzed in order to aggregate information, for example genes with similar functions or residing in the same pathway[3,4]. Therefore it is necessary to account for potential interactions between rare variants in different loci[17] and interactions between common and rare variants[18,19].

The Kernel Based Adaptive Cluster (KBAC) was developed to overcome the problems of detecting rare variant associations in the presence of misclassification and gene interaction. Under the KBAC framework, data-based adaptive variant classification and testing of association are unified. The sample risk of a multi-site genotype is modeled using a mixture distribution with two components, where one component represents the distribution of sample risk of genotype if it is non-causal and the other component represents distribution of sample risks of causal genotypes. Ideally, if distributions for causal components were known, classification could first be performed and only the causal genotypes would be used in association studies. However, when searching for genotype-phenotype associations, it is usually unknown which variants are causal.

Instead of performing an unrealistic two-step procedure, variant classification and association testing are unified in the KBAC framework. Continuous adaptive weighting which is implemented in the KBAC is preferable, particularly for low frequency alleles, than classifying variants and carrying out a stratified analysis, because increasing classification and shrinking size of strata can increase both type I and II error. For the KBAC, adaptive weighting procedure is implemented using the cumulative distribution functions for the multi-site genotype counts. Distributions of multi-site genotype counts are compared between cases and controls. Those multi-site genotypes that are enriched in cases will be up-weighted. Under the null hypothesis, the assigned weights asymptotically follow a uniform distribution. While under the alternative hypothesis, disease causal multi-site genotypes tend to be more frequent in cases than in controls. Therefore they are more likely to be adaptively up-weighted. The weighted multi-site genotype frequencies are aggregated and contrasted between cases and controls. In order to evaluate whether there is an association, significance of the KBAC can be assessed using either permutation or Monte Carlo approximation (See Monte Carlo Approximation).

The performance of the KBAC was compared to the weighted sum statistic (WSS)[21] the combined multivariate and collapsing (CMC) method[20], and the comparison of rare variants found exclusively in cases to those found only in controls (RVE)[3] using simulated data sets. Forward time simulation[32] assuming infinite-site Wright-Fisher model was used to generate population genetic data. Demographic change and purifying

selection were both incorporated in the simulation, using parameters estimated from re-sequencing datasets from studies of African Americans (AA) and European Americans (EA)[33]. In addition to forward time simulation, population genetic data was also generated according to estimated site frequency spectrums (SFS) in AA and EA from the Dallas Heart Study (DHS) re-sequencing data of the *ANGPTL3, 4, 5,* and *6* genes.

For the simulated population data phenotypes were generated separately and motivated by epidemiological disease studies. Two types of main effects phenotypic model are considered: 1.) constant genetic effects for each causal variant and 2.) genetic effects inversely correlated with minor allele frequencies (MAF) of causal genetic variants. In order to evaluate the impact of variant misclassification, a variety of scenarios were examined where 1.) different proportions of non-causal variants were included in the analysis and 2.) different proportions of causal variants were excluded from the analysis.

Two disease models of gene interactions were also evaluated. The example of within gene interaction was motivated by Hirschsprung's disease[18,19], where an interaction between a common polymorphism in the promoter region and multiple rare non-synonymous (NS) mutations in exonic regions of the *RET* gene is hypothesized[18,19]. The example of between gene interaction is based on the observation that rare variants within the *CHEK2* gene increase risk of breast cancer in the absence of *BRCA1* and

BRCA2 mutations, but because of a shared pathway, the same *CHEK2* variants in the presence of high risk BRCA variants do not further increase risk[17,34,35].

Under each of the above scenarios, phenotype-genotype association testing is performed for rare NS variants. It is demonstrated that the KBAC has a clear advantage in power and robustness over other existing methods and this benefit is especially strong, when rare variant data is analyzed where there is either variant misclassification or gene interactions.

In order to further illustrate applications of the KBAC and other statistical methods, i.e., WSS, CMC, and RVE to carry-out association studies, energy metabolism traits and rare variants in *ANGPTL 3, 4, 5* and *6* genes obtained from sequence data were analyzed. In addition to identifying the originally reported association between triglyceride levels and *ANGPTL* 4, KBAC identified associations for a.) body mass index and *ANGPTL 5*, b.) diastolic blood pressure with *ANGPTL 6*, c.) high density lipoprotein with *ANGPTL 4* , d.) triglyceride levels with *ANGPTL 3* e.) very low density lipoprotein with *ANGPTL 3* and *ANGPTL 4*.

2.2. Results

The results presented focus on simulations using simulated SFS from AA sequence data. Similar results are found for simulations using simulated SFS for EA and estimated SFS for AA and EA (data not shown). Although the power varies dependent on the underlying model used to generate the data, in all cases the KBAC is the most powerful method followed by the WSS, CMC and then the RVE.

2.2.1. Rare Variant Frequency Distributions in Generated Case-Control Samples

Rare NS variants carrier information is summarized (Table 1) for replicates used in power comparisons in the presence of misclassifications. Under the phenotypic model with variable genetic effects, when all variants (both non-causal and causal variants) were analyzed, 5.5% of cases and 3.4% of controls are carriers, with carrier frequency in cases 61% higher than in controls. When only causal variants are included, the fractions of carriers in cases and in controls are 3.8% and 1.7% respectively. The case rare variant frequency is approximately 2.3 times of the controls frequency, which implicates that average ORs of uncovered rare variants lie between 2 to 3. For the phenotypic model with fixed genetic effects, the results are similar. The carrier frequency observed in cases is around 2.5 times the frequency in controls. Compared to the model with fixed effects, lower frequency rare causal variants have larger ORs for variable effects model. The probability that these low frequency rare variants are uncovered in a case-control sample is higher. Therefore, in all scenarios examined, more rare variants sites are uncovered for

the model with variable effects. When all the variants are included, 11% more rare NS variants sites are uncovered for the model with variable effects. The number of rare variants sites that are exclusive to cases or controls is also higher under the variable effect model. For example, when 100% of the variant sites are included in the analysis, 47.4% and 41.1% of the sites are found exclusively in either cases or controls for the variable and the fixed effects model, respectively. For both models, within a single gene, very few cases and controls carry more than one rare variant.

For the within gene interaction model (Table 2), similar patterns of NS variants sites and carrier frequencies are observed. When 100% of the rare variants are causal, 5.5% of the cases and 3.2% of the controls are carriers on average for a case-control sample. Due to interaction, frequency differences between cases and controls are mitigated. In the between gene interaction model (Table 2), higher case carrier frequency and more rare variants sites are observed for the high risk gene than for the low risk gene. The proportions of rare variants carriers for the two genes combined can be high, e.g. when 100% of the variants are causal, up to 12% of the cases can be rare variant carriers.

2.2.2. Evaluation of Type I Error:

When permutation was used to evaluate significance for the KBAC, type I error was well controlled, because p-values were obtained empirically. Additionally, in order to ensure that the type I error for RVE is well controlled permutation is also used to obtain

empirical p-values. For the WSS[21], CMC[20] method, it was previously demonstrated that for the analysis of rare variants, their type I errors are well controlled[20]. For moderate sample sizes e.g. 400 cases/400 controls, the distributions of p-values for the Monte Carlo approximation are very close to those obtained using permutations and theoretical expectations **(Fig. 1)** and additionally type I error is well controlled.

2.2.3. Power Comparison:

2.2.3.1. Main Effects Model without Misclassification:

For main effects model with fixed genetic effects and no misclassification **(Fig. 2)**, the power $(1-\beta)$ for KBAC, WSS CMC and RVE are respectively given by 82.5%, 77.7%, 73.9% and 14.8%. The power for RVE is much lower than the power for the other three methods. For the main effects model with variable genetic effects **(Fig. 3)**, the power for the four methods is given by 83.1%, 78.8%, 74.2% and 44.8%. The power of the RVE improves for the variable genetic effects model compared to the fixed genetics effect model; while the power for the other methods remains relatively unchanged. KBAC is consistently more powerful than WSS, CMC and RVE, e.g. for fixed effect model, KBAC is 6.1% more powerful than WSS, 11.6% more powerful than CMC, and 457.4% more powerful than RVE.

2.2.3.2. Impact of Misclassification:

Under both models **(Fig. 2, 3)**, the power of all methods is negatively impacted by exclusions of causal variants and inclusions of non-causal variants at a varying degree. When non-causal variants are included in the analysis, KBAC is consistently more powerful and more robust than the other three methods. For example, when 100% of the non-causal variants are included, under the variable effects model, KBAC $(1-\beta_{KBAC}=69.9\%)$ is 19.3% more powerful than WSS $(1-\beta_{WSS}=58.6\%)$, 27.6% more powerful than CMC $(1-\beta_{CMC}=54.8\%)$, and 91.0% more powerful than RVE $(1-\beta_{RVE}=36.6\%)$. When compared under the fixed effects model, the advantage of KBAC $(1-\beta_{KBAC}=71.2\%)$ over WSS $(1-\beta_{WSS}=61.1\%)$, CMC $(1-\beta_{CMC}=58.2\%)$ and RVE $(1-\beta_{RVE}=13.9\%)$ remains largely unchanged. For the scenarios where causal variants are missing, the relative performances of the methods remain to be in the order KBAC>WSS>CMC>RVE. For the variable effects model, the power advantage of WSS over CMC is greater than the advantage observed for the fixed effects model. For example, when 60% of the causal variants are excluded from the analysis, under the fixed effects model, the power for WSS drops 40.1% and the power of CMC drops 45.1%, while under the variable effects model, the power decreases for WSS and CMC are respectively 39.1%, 47.8%. The KBAC is more robust than the other methods: the power decreases under the fixed and variable effects models are respectively 34.1% and 35.6%, which are smaller than the decreases in power for WSS and CMC. Exclusion of causal

variants from the analysis is more detrimental to power than inclusion of non-causal variants.

2.2.4. Interaction Models:

2.2.4.1. Within Gene Interaction Model:

Under the within gene interaction model, KBAC is consistently the most powerful method for all scenarios with different proportions of causal variants (Figure 4). The advantage of KBAC in the presence of interactions is apparent and its advantage over other methods becomes greater with increasing proportion of non-causal variants. For example, when all variants are causal, the power of KBAC is 8.4% higher than WSS, which is the second most powerful method. But when only 50% of all variants are causal, KBAC is 30.7% more powerful than WSS. RVE is the least powerful methods for all scenarios compared.

2.2.4.2. Between Gene Interaction Model:

In the between gene interaction model, power comparisons between the four methods remain similar (Figure 4). KBAC is consistently the most powerful method and is robust against inclusion of non-causal rare variant sites. Comparing the scenario where all variants are causal with the scenario where only 50% of the variants are causal, the power for KBAC drops 36.3%, while the power for WSS drops 48.2%.

2.2.5. *ANGPTL* Variants and Energy Metabolism in Humans:

In order to further illustrate the application of KBAC and other rare variant analysis methods (i.e. WSS, CMC and RVE), rare variants in the *ANGTPL 3,4,5* and *6* genes were analyzed to determine whether they are associated with energy metabolism traits (Table 3). As in the original DHS study[36], the association of rare variants in the *ANGPTL3,4,5* and *6* genes with triglyceride (TG), low density lipoprotein (LDL), very low density lipoprotein (VLDL), high density lipoprotein (HDL), cholesterol, glucose, body mass index (BMI), systolic (SysBP) and diastolic blood pressure (DiasBP) were investigated. In the original DHS study, NS variants were analyzed using RVE, and significant associations were found between *ANGPTL3, ANGPTL 4* and TG as well as between *ANGPTL 6* and cholesterol[5,6]. In this study, NS variants, most of which are very rare[5,6], were analyzed. Individuals with confounding factors (lipid lowering drugs, diabetes mellitus and heavy alcohol use) were removed for all analyses. Multiple associations were identified with KBAC but not with other approaches, i.e. the novel associations between *ANGPTL* 6 and DiaBP $\left(p_{KBAC}=0.045, p_{WSS}=0.084, p_{CMC}=0.088, p_{RVE}=0.405\right)$, as well as between *ANGPTL* 3 and TG levels $\left(p_{KBAC}=0.015, p_{WSS}=0.053, p_{CMC}=0.058, p_{RVE}=0.312\right)$. Additionally multiple novel associations were observed for analyses carried out with KBAC, WSS and CMC: 1.) *ANGPTL4* and VLDL $\left(p_{KBAC}=0.001, p_{WSS}=0.006, p_{CMC}=0.010, p_{RVE}=0.141\right)$; 2.) *ANGPTL5* and BMI $\left(p_{KBAC}=0.001, p_{WSS}=0.003, p_{CMC}=0.006, p_{RVE}=0.263\right)$; 3.) *ANGPTL4* and HDL $\left(p_{KBAC}=0.021, p_{WSS}=0.041, p_{CMC}=0.045, p_{RVE}=0.681\right)$ and 4.) the previously reported

association between ANGPTL4 and TG levels

$(p_{KBAC} = 0.004, p_{WSS} = 0.005, p_{CMC} = 0.006, p_{RVE} = 0.087)$. It should be noted that HDL and TG levels are negatively correlated (-0.42) and individuals with HDL levels in the lower quartile had an excess of rare variants in *ANGPTL4* compared to those individuals with HDL levels in the upper quartile, while those individuals with TG levels in the upper quartile had an excess of rare variants in *ANGPTL4* compared to those with TG levels in the lower quartile. The association detected by KBAC between ANGPTL4 and VLDL and between ANGPTL5 and BMI remains significant after correcting for multiple testing. RVE, on the other hand, detected associations between *ANGPTL 5, 6* and glucose while the other three methods did not. We further investigated this association by applying a more stringent MAF cutoff 0.1% for the NS variants analyzed in *ANGPTL 5* and *6.* Using this new criterion both associations were detected by all methods (for *ANGPTL 5*, $(p_{KBAC} = 0.001, p_{WSS} = 0.006, p_{CMC} = 0.011, p_{RVE} = 0.011)$ and for *ANGPTL 6*, $(p_{KBAC} = 0.002, p_{WSS} = 0.008, p_{CMC} = 0.012, p_{RVE} = 0.012)$).

2.3. Discussion

The KBAC method developed for association mapping of rare variants combines genotype classification and hypothesis testing in a coherent framework. The risk of each multi-site genotype is modeled as a mixture distribution with two components, among which only the component representing a non-causal genotype is known and is used in the adaptive weighting. Each multi-site genotype is continuously weighted using the non-causal component. The power of the KBAC as well as the other methods investigated can be affected by inclusion of non-causal mutations or exclusion of causal variants in the sample, to a varying degree. When non-causal variants are included in the analysis, the difference in rare variant carrier frequencies observed between cases and controls is mitigated. On the other hand, when causal variants are excluded from the association analysis, the marginal effect size of existing variants can vary considerably depending on whether missing causal variants exist on the same multi-site genotype. As a result, treating each variant (or multi-site genotype) interchangeably will incur loss of power, the severity of which will depend on the proportion of misclassified variants in the data. The performance of the KBAC is superior to the other approaches that were examined.

Bioinformatics tools[9,10] and filters[1] can be used to determine which rare variants are potentially functional and should be included in the association analysis[1]. Their predictive accuracy, which can be low, is dependent on the amount of information

available for the gene understudy. If bioinformatics tools are used to predict variant functionality and determine which variants should be included in the analysis it is best to loosen stringency, because the exclusion of causal variants is more detrimental to power than inclusion of non-causal variants. Whether or not bioinformatics tools are used as a screening tool, misclassification will occur therefore the robustness of KBAC to misclassification is particularly beneficial. Additionally in order to avoid potentially erroneous exclusion of causal variants due to locus heterogeneity, joint analysis of multiple putative genetic loci that carry similar functions or reside in the same pathway can be valuable.

It is of great interest to evaluate gene x gene interactions in the study of complex diseases. The KBAC analyzes multi-site genotypes (or multi-locus genotype), which can be beneficial in detecting gene interactions[14]. This property is especially important when multiple genetic loci are jointly analyzed in order to aggregate rare variants. Interactions are more likely to occur between genes involved in the same pathways. In addition, it has been hypothesized that functions of rare variants can be modulated by common variants[8]. Since the KBAC uses adaptive weighting instead of a fixed model, unknown patterns of gene interaction can be automatically integrated into the analysis. Through models motivated by Hirschsprung's disease and breast cancer, it is shown that in the presence of interactions the KBAC outperforms other approaches. An additional advantage of the KBAC is that kernel weights computed for adaptive weighting provide a

measure with which the relative risk of each multi-site genotype can be assessed, for further replication studies.

The RVE method which compares the occurrence of variants which are exclusively observed in cases to those which are only observed in controls has the lowest power among all tests evaluated. The RVE method possesses undesired statistical properties by excluding those variants which are observed in both cases and controls. For all variants that are not fully penetrant, when sample size is large, they tend to appear in both case and control samples and would thus be excluded from the analysis using RVE. As a result, the RVE method is not asymptotically consistent; with increasing sample size power may be even lower than for smaller sample sizes[24].

Forward time simulations of locus genetic data incorporated both population demographic change and purifying selection. Both factors are known to impact SFS for observed rare variants (especially NS variants). Only NS variants were analyzed for comparing different methods, as it has been suggested that using NS variants will concentrate variations on functionally significant class of alleles, and increase signal to noise ratio[24]. There have been a number of studies on complex diseases which identified associations with NS variants[3,5,6]. When synonymous mutations are also considered in the analyses, higher proportions of non-causal variants may be introduced, so the adaptive property and the robustness of KBAC will be more advantageous.

Whether or not phenotypic effects of causal rare variants are inversely correlated with their MAF is unknown. Deleterious functional variants tend to have low frequencies[37], but the functional effect of a deleterious mutation may not be associated with the disease. On the other hand, for mutations involved in complex traits, they may not be at selective disadvantage due to the fact that most complex traits are late on-set and may not cause reductions in reproductive fitness. For both types of models, the advantage of KBAC is apparent. WSS and RVE perform better under the variable effects models, when only causal variants are present. This is because high risk causal variants are assigned higher weights. However, as low frequency non-causal variants also receive larger weights that negatively affect power, there are no measurable improvements of WSS compared to the model with fixed genetic effects. On the other hand, due to the adaptive nature of KBAC, the method performs consistently the best under both classes of models.

The KBAC test statistic does not have a closed form distribution; therefore it is necessary to evaluate significance either through permutation or using Monte Carlo approximation. For small sample sizes i.e. $\sim \leq 400$ cases and 400 controls, permutation is recommended, because it can be more reliable than Monte Carlo approximation. For larger sample sizes, Monte Carlo approximation not only controls type I error, but also the estimates of power do not differ from those obtained using permutations (data not shown). Permutation can be computationally intensive for large samples and/or genome-wide data where a large number of genetic regions are analyzed; therefore Monte Carlo

approximation can be particularly advantageous to evaluate significance due to its computational efficiency.

A well known problem of genetic association studies is spurious findings due to population substructure and/or population admixture. For rare variant association analysis this problem can occur when study subjects are sampled from different populations and the distribution of non-causal variant sites and/or aggregate frequencies of non-causal variants differ between the sampled populations. To control for population stratifications, KBAC can be coupled with principal components analysis (PCA) [38] approach and eigenvector(s) can be included as covariates in the analysis (see **2.4.4** Controlling for Confounders:). PCA approach has been shown to be a powerful tool to accurately infer geographical locations [39,40]. In addition, KBAC can also be used with clustering/matching based methods, such as structured association [41,42] to control for population stratification.

The application of KBAC as well as WSS, CMC and RVE were further illustrated by the analyses of genes in *ANGPTL* family. In the analyses, all individuals with potentially confounding factors i.e. diabetics, alcoholics, and individuals treated with lipid lowering drug were excluded. In the original studies individuals were excluded based upon both their quantitative trait values and the confounding factors. For example, only individuals treated with lipids lowering drugs in the lower quartile of TGs were

removed, but those in the upper quartile were included in the analysis. We believe excluding individuals based upon their quantitative trait values should not be done instead all individuals meeting the exclusion criteria should be removed from the analysis. KBAC performs consistently well, and identifies the most phenotype-genotype associations among all the approaches compared. The effects of mutant *ANGPTL* genes on lipoprotein lipase (LPL) have been studied through *in vitro* functional studies and *in vivo* mice studies. LPL has been known to affect glucose metabolism[43], cholesterol level[43-46], and blood pressure[47]. This biological evidence strengthens the support of the identified associations. Additionally, the association between variants in *ANGPTL*4 gene and triglyceride levels were successfully replicated using an independent dataset[5,6].

Although the examples given are for the analysis of single regions and interaction between two regions, the KBAC can also be used to analyze entire exomes (or genomes). In order to control for family-wise error rate (FWER), it is sufficient to use a Bonferroni correction, since there will be little or no linkage disequilibrium between rare variants in different genes. It is thus not necessary to control the FWER using permutations. If exome sequencing is carried out and analysis is implemented gene by gene, given that human genome contains ~20,000 genes, a significance level $\alpha = 0.05/20{,}000 = 2.5\times10^{-6}$ can be applied. The correction necessary for gene based association mapping of rare variants is less than the threshold currently used for genome-wide association studies [48] which is usually $\alpha = 5\times10^{-8}$.

The KBAC is a powerful tool to detect main association effects and gene interactions in large sequence data sets of candidate genes, exomes and in the future entire genomes. The KBAC is implemented in a user friendly R package and is available from the authors.

2.4. Methods:

2.4.1. Sample Risk

Total sample size is denoted as N, among which there are N^A affected (A) and $N^U = N - N^A$ unaffected (U). It is assumed that there are M sites within the candidate region where rare variants are observed. The rare variant multi-site genotype for each "individual" is contained in a vector $G = (g_1, g_2, \cdots, g_M)$, with the j^{th} entry being the number of rare variants observed at j^{th} site, i.e. g_j has value 2 if the site is homozygous for the rare allele, 1 if the site is heterozygous, 0 if the site is homozygous wild-type for the common major alleles. It is further assumed that $k+1$ distinct multi-site genotype vectors, i.e $G_0, G_1, G_2, ...G_k$ are observed, where $G_1, G_2, ...G_k$ are multi-site genotypes with at least one rare variant and G_0 represents the wild-type genotype without any rare variants (i.e. a vector of all 0's). The sample risk for multi-site genotype G_i is defined as

$$R_i = \frac{N_i^A}{N_i},$$

which is a consistent estimator of the ratio

$$\frac{N^A \times P[G_i \mid A]}{N^A \times P[G_i \mid A] + (N - N^A) \times P[G_i \mid U]}.$$

The ratio increases with disease penetrance of G_i and provides a sample based measure of the relative risk.

The sample risk R_i for multi-site genotype G_i is modeled using a mixture distribution with two components, $R_i \overset{D}{\sim} \pi_i k_i^0(R_i) + (1-\pi_i)k_i^A(R_i)$. The component $k_i^0(R_i)$ represents the distribution of the sample risk when multi-site genotype G_i is non-causal and is known, while $k_i^A(R_i)$ represents the unknown distribution of sample risk when G_i is causal. If the null hypothesis holds, all genotypes are non-causal, therefore, $\pi_i = 1$. Under the alternative hypothesis, each genotype can be either causal or non-causal and the probabilities π_i in the probabilistic mixtures are unknown.

If the mixture distribution under the alternative were known, then each genotype could be classified and only the causal genotypes would be used in the analysis. However, in disease gene mapping, the causality of variants is unknown. Instead of trying to 'estimate' π_i and k_i^A which are unknown, each multi-site genotype is adaptively weighted using only the known component, $k_i^0(\bullet)$. Each $k_i^0(\bullet)$ is called a kernel. The term kernel is borrowed from density estimation, where the density being estimated is spanned by a linear combination of kernel functions. The weight each rare genotype carries is given by the area under the curve which can be calculated as a generalized integral

$$w_i = \int_0^{\hat{R}_i} k_i^0(r)dr = K_i^0\left(\hat{R}_i\right),$$

where $\hat{R}_i$ is the estimated sample risk for multi-site genotype G_i.

Thereby, under the null hypothesis, the weights are uniformly distributed and under the alternative, greater weights can be placed on the multi-site genotypes that are enriched in cases. The genotypes with high sample risks will be given higher weights which can potentially separate causal from non-causal genotypes. Instead of classifying genotypes in a rigid manner with unknown likelihoods, this method weighs each genotype in a continuous fashion using only the known component $k_i^0(\bullet)$ from the mixture density. The adaptive weighting procedure in the KBAC attains a good balance between classification accuracy and the number of parameters which are estimated

2.4.2. Choice of Kernels

Three types of kernels can be used to assign weights to each rare genotype; they are asymptotically equivalent. For small to moderate sample sizes, binomial and hyper-geometric likelihoods tend to work best, while for large sample sizes the asymptotic normal kernel is computationally efficient. All examples shown were carried out using the hyper-geometric kernel.

2.4.2.1. Hyper-geometric Kernel

Under the null hypothesis of no disease/gene associations, conditioning on the genotype counts $\{N_i = n_i\}_{i\le k}$ and the count of cases and controls $\{N^A = n^A\}$, the number of diseased "individuals" having multi-site genotype G_i i.e. $n_i^A = n_i r_i$ follows a hyper-geometric distribution with kernel function given by

$$k_i^0(r_i) = P[R_i = r_i \mid \{N_i = n_i\}_{i\le k}, N^A = n^A] = \frac{\binom{n_i}{n_i r_i}\binom{n - n_i}{n^A - n_i r_i}}{\binom{n}{n^A}}$$

As this distribution is discrete, the integral is replaced by summations, i.e.

$$K_i^0(\hat{R}_i) = \sum_{r_i \in \{0, \cdots \hat{R}_i\}} k_i^0(r_i)$$

2.4.2.2. Marginal Binomial Kernel

Under the null hypothesis of no disease/gene association, conditioning on the genotype counts $\{N_i = n_i\}_{i\le k}$, marginally, the number of disease "individuals" with genotype G_i, $n_i^A = n_i r_i$ satisfies a binomial distribution, $n_i^A \sim Binom\left(n_i, \frac{n^A}{n}\right)$. Thus,

$$k_i^0(r_i) = P[R_i = r_i] = \binom{n_i}{n_i r_i}\left(\frac{n^A}{n}\right)^{n_i r_i}\left(1 - \frac{n^A}{n}\right)^{n_i(1-r_i)}.$$

The weight as above is obtained through summations, i.e.

$$K_i^0(\hat{R}_i) = \sum_{r_i \in \{0, \cdots \hat{R}_i\}} k_i^0(r_i).$$

2.4.2.3. Asymptotic Normal Kernel

Under the null distribution, the sample risk for genotype G_i is asymptotically normal, i.e.

$$\sqrt{n_i}\left(R_i - \frac{n^A}{n}\right) \xrightarrow{D} \mathrm{N}\left(0, \frac{n^A}{n}\left(1 - \frac{n^A}{n}\right)\right)$$

so the kernel is given by $k_i^0(r_i) = \frac{\sqrt{n_i}}{\sqrt{\frac{n^A}{n}\left(1 - \frac{n^A}{n}\right)}} \phi\left(\frac{\sqrt{n_i}\left(r_i - \frac{n^A}{n}\right)}{\sqrt{\frac{n^A}{n}\left(1 - \frac{n^A}{n}\right)}}\right)$, where $\phi(\bullet)$ is the probability density function for a standard normal random variable. The weight for genotype G_i is given by the integral

$$K_i^0(\hat{R}_i) = \int_0^{\hat{R}_i} k_i^0(r_i) dr_i.$$

2.4.3. Test Statistics

Each "individual" with multi-site genotype G_i in the sample will be assigned weight w_i. The weight is given by the kernel functions depending on the estimated sample risk $\hat{R}_i$ i.e. $w_i = K_i^0(\hat{R}_i)$. The weights assigned to rare genotypes are aggregated and contrasted between cases and controls.

The KBAC statistic is defined as $KBAC = \left(\sum_{i=1}^{k} \left(N_i^A / N^A - N_i^U / N^U \right) K_i^0 \left(\hat{R}_i \right) \right)^2$,

which compares the difference of weighted multi-site genotype frequencies between cases and controls. When a one sided alternative hypothesis is tested, e.g. the enrichment of causal variants in cases, a corresponding one sided version of KBAC can be used, i.e. $KBAC_1 = \sum_{i=1}^{k} \left(N_i^A / N^A - N_i^U / N^U \right) K_i^0 \left(\hat{R}_i \right)$. In this thesis, all power comparisons were based upon two sided tests for each method.

Standard permutation procedure is used to obtain empirical p-values for small sample sizes and for large sample sizes significance can be obtained through the Monte Carlo approximation.

2.4.4. Controlling for Confounders:

In order to control for sample heterogeneities such as population stratification/admixture, it is desirable to be able to incorporate covariates in the association analysis. The kernel weights computed for the KBAC statistic can be used with logistic regression. For an individual j with multi-site genotype G_i, we define a

variable for the kernel weight, i.e. $X_j = w_i$. The logistic regression model for association testing has the form

$$\log\left(\frac{P(Y_j = 1 | X_j, Z_{jl})}{1 - P(Y_j = 1 | X_j, Z_{jl})}\right) = \beta_0 + \beta_1 X_j + \sum_l \alpha_l Z_{jl}$$

where $\{Z_{jl}\}_{j,l}$ are the covariates such as age, sex or eigenvectors for genotypes.

A score statistic to test $H_0 : \beta_1 = 0$ can be computed in closed form. Due to the complexities involved in computing kernel weights, the score statistic does not follow a normal distribution. Standard permutation procedure can be applied to evaluate the significance. When no additional covariates are controlled, the score function U satisfies $U = \sum_j X_j (Y_j - \bar{Y})$[49]. Simple algebraic manipulations will lead to the equivalence of the score function U and the KBAC statistic (up to a constant scalar). In addition, when common variants in the gene are also hypothesized to play a role in the etiology of the phenotype of interest, their genotypes can be included as covariates and tested in a similar manner as for the CMC[20].

2.4.5. Monte Carlo Approximation

2.4.5.1. Monte Carlo Approximation under the Null Hypothesis

Although using permutation can provide an exact empirical distribution under the null hypothesis, it can be computationally prohibitive for large sample sizes and genome-wide

association studies. A Monte Carlo method was developed which enables fast computation of p-values efficiently. Under the null hypothesis, conditioning on the genotype counts, $\{n_i\}_{1\le i\le k}$ and the total number of cases and controls $n^A, n-n^A$, the number of cases n_i^A with multi-site genotype G_i follows a binomial distribution $n_i^A \sim Binom\left(n_i, \frac{n^A}{n}\right)$. Due to the low frequencies for each multi-site genotype containing rare variants, the n_i^A's are approximately independent of each other. Therefore, Monte Carlo simulation can be carried out as shown in algorithm 1:

Algorithm 1

Step 1: Simulate a k -vector of independent binomials: $(m_1, m_2, \cdots\cdots, m_k)$, with $m_i \overset{D}{\sim} Binom\left(n_i, \frac{n^A}{n}\right)$

Step 2: Compute $U = \left(\sum_{i=1}^{k}\left(m_i/n^A - (n_i - m_i)/(n-n^A)\right)K_i^0(m_i/n_i)\right)^2$

Step 3: Repeat step 1 and step 2 N times and record each KBAC statistic calculated as $\vec{U} = (U_1, \cdots\cdots, U_N)$. Through comparing the KBAC statistic calculated from

the original data with the N KBAC statistic from Monte Carlo simulation, the empirical p-value is given by $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}[U_i \geq KBAC]$.

2.4.5.2. Monte Carlo Approximation under the Alternative Hypothesis: Power calculations

In this section, power calculations were carried out empirically; haplotypes were generated using forward time simulations and case-control status was assigned via a linear log odds model. Power calculations can also be carried out using Monte Carlo approximation. Under the alternative hypothesis of disease-gene associations, it is assumed that the disease model is known (prevalence and population multi-site genotype frequencies $P = (p_1, p_2, \cdots, p_k)$ etc.) Therefore multi-site genotype frequencies for cases and controls can be assigned. The set of frequencies in cases and controls is denoted as $P^A = (p_1^A, p_2^A, \cdots, p_k^A)$, $P^U = (p_1^U, p_2^U, \cdots, p_k^U)$. Conditioning on the genotype counts, $\{n_i\}_{1 \leq i \leq k}$ and the total number of cases and controls $n^A, n - n^A$, the number of cases n_i^A with the multi-site genotype G_i follows a binomial distribution, i.e.

$$n_i^A \sim Binom\left(n_i, \frac{n^A p_i^A}{(n - n^A)p_i^U + n^A p_i^A}\right).$$

The power calculation under significance level α can be carried out in the following steps:

Algorithm 2

Step 0: Generate N_1 $k+1$-vectors $\left(n_1^1, n_2^1, \cdots, n_k^1, n_0^1\right), \cdots, \left(n_1^{N_1}, n_2^{N_1}, \cdots, n_k^{N_1}, n_0^{N_1}\right)$ satisfying multinomial distribution i.e.

$$\left(n_1^i, n_2^i, \cdots, n_k^i, n_0^i\right) \sim Multi\left(n; p_1, p_2, \cdots, p_k, 1-\sum_{l=1}^{k} p_l\right)$$

For each vector $\left(n_1, n_2, \cdots, n_k, n_0\right) = \left(n_1^i, n_2^i, \cdots, n_k^i, n_0^i\right)$, we follow step 1 to 4:

Step 1: Obtain an empirical distribution under the null by following step 1 and 2 in algorithm 1. The vector of U 's obtained is denoted by $\vec{U}^0$ and the $(1-\alpha)^{th}$ empirical quantile for $\vec{U}^0$ is denoted by U_α^0

Step 2: Simulate a k -vector with independent binomials: $\left(m_1, m_2, \cdots\cdots, m_k\right)$, with

$$m_l \sim Binom\left(n_l, \frac{n^A p_l^A}{\left(n-n^A\right)p_l^U + n^A p_l^A}\right), l = 1,2,\cdots,k$$

Step 3: Compute $U = \left(\sum_{l=1}^{k}\left(m_l/n^A - \left(n_l - m_l\right)/\left(n-n^A\right)\right)K_l^0\left(m_l/n_l\right)\right)^2$

Step 4: Repeat step 2 and step 3 N_2 times and record each KBAC statistic calculated as $\vec{U}^A = \left(U_1^A, \cdots\cdots, U_{N_2}^A\right)$. By comparing the KBAC statistic calculated from

Monte Carlo simulation with U_α^0, the empirical power conditional on

$(n_1, n_2, \cdots, n_k, n_0) = (n_1^i, \cdots, n_k^i, n_0^i)$ is given by $1-\hat{\beta}_i = \frac{1}{N_2}\sum_{j=1}^{N_2} \mathrm{I}[U_j^A \geq U_\alpha^0]$.

Step 5: The estimation of unconditional power is given by averaging $\hat{\beta}_i$'s, i.e.

$1-\hat{\beta} = \frac{1}{N_1}\sum_{i=1}^{N_1}(1-\hat{\beta}_i)$

2.4.6. Rare Variant Analysis Methods Which Are Compared to the KBAC:

The power of WSS, CMC and RVE were compared to KBAC in the thesis. A sketch of each method is provided here. More detailed descriptions can be found in the cited original reference. WSS was developed by Madsen and Browning[21]. It was designed to test for the differences of the number of mutations between cases and controls. Each mutation was weighted according to its frequency in controls, and lower frequency variants will be assigned higher weights. The statistical significance for the WSS statistic is obtained empirically through permutations.

CMC was developed by Li and Leal[20]. When applied to testing rare variant associations, multiple rare variants in the gene region are collapsed and carrier frequencies are compared between cases and control using Pearson's Chi-square test. The

RVE [3,4] was first introduced in the analysis of sequence data from Dallas Heart Study. It compares frequency of carriers of rare variants that are found exclusively in cases or controls using Fisher's exact test.

2.4.7. Generation of Genetic Data:

2.4.7.1. Simulation of Demographic Model and Selections:

To evaluate the performance of KBAC, population genetic data was generated using forward time simulation[32]. Genetic data from two populations, AA and EA were generated. The parameters for demographic changes and selection coefficients were estimated in Boyko et al[33]. For AA, a simple two-epoch model was used. Purifying selection was also simulated, with s and $2s$ being the selective disadvantage of heterozygous and homozygous new mutations. Scaled fitness effect $\gamma = 2N_{curr}s$ (where N_{curr} is the current effective population size) is assumed to follow a gamma distribution, which was shown to be parsimonious and fit the data well. A mutation rate of $\mu_S = 1.8 \times 10^{-8}$ per nucleotide per generation is assumed. On average, the coding region for human gene is 1500 base pairs (bp) long[50,51], therefore 1500 bps was used in the simulation to specify the locus scaled mutation rate. 100 haplotype pools were generated. When generating samples, one pool is randomly chosen for each replicate. The multi-site genotype of an "individual" is obtained by pairing two randomly sampled haplotypes.

2.4.7.2. Generation of Phenotype Data with only Main Effects:

The disease status of each "individual" is assigned based upon their multi-site genotypes consisting of only those rare NS variants (MAF ≤ 1%). Fifty-percent of the rare NS variant nucleotide sites were selected to be causal, where the rare mutant allele has an effect on the disease odds and the remaining rare variant sites are non-causal with no phenotypic effect. Two types of penetrance models were evaluated. In the first type of model, the genetic effects of causal variants are constant (OR=3) regardless of their allele frequencies. For the second class of models, the genetic effects are inversely correlated with the MAFs. Disease odds of individual rare variants varies in the range of 2 ~ 20. As a majority of rare variants are of extremely low frequencies, most of the uncovered rare variants in a case control sample have ORs between 2 and 4. This is compatible with surveys for multi-factorial diseases[8]. For both classes of penetrance specifications, a linear log odds model was applied to assign the affection status for each individual. Assignment of disease status continues until a sample of 1000 cases and 1000 controls is obtained for each replicate. To evaluate the effects of misclassification due to non-causal variants, scenarios were considered where 20%, 40%, 60%, 80%, and 100% of the non-causal variants with all of the causal variants were included in the sample. Additionally to evaluate the effect of misclassification due to exclusion of causal variants, 20%, 40%, and 60% of the causal variants were excluded from the analysis, while no non-causal variants are included in the analysis.

2.4.7.3. Generation of Data with Gene Interactions

To evaluate the within gene interaction and between gene interaction models, 1000 cases/1000 controls and 300 cases/300 controls were generated for each replicate, respectively. For each model, 25% to 100% of the simulated rare variant sites are causal while the remaining rare variant sites are non-causal. For the within gene interaction model, one site with a common variant [MAF >20%] is randomly selected. The disease status of each "individual" is assigned based upon their multi-site genotype using a linear log odds model. The genetic effects of causal rare variants are modulated by the alleles at the chosen common variant site. Each causal rare variant increases disease risk with an OR of 3 only if the rare variant is on the same haplotype as the minor allele from the common variant site, otherwise the OR=1. For the between gene interaction model, two unlinked genes are simulated for each "individual". The disease status of each "individual" is assigned based upon their joint multi-site genotype at high risk gene 1 and low risk gene 2 using a linear log odds model. Each causal rare variant in gene 2 increases disease risk with an OR of 2.0 if there are no causal rare variants in gene 1; however, if there are rare causal variants in gene 1, the causal variants in gene 2 do not increase risk and each causal variant in gene 1 increases disease risk with an OR of 4.0 regardless of the genotype at gene 2.

2.4.8. Analysis of Energy Metabolism Traits and Rare Variants in *ANGPTL 3, 4, 5* and *6*

The DHS dataset is a multi-ethnic population based probability sample [1830 AA, 601 Hispanics (H), 1045 EA, and 75 from other ethnicities] from Dallas County residents whose lipids and glucose metabolism have been characterized and recorded[36,52]. In order to investigate how sequence variations in *ANGTPL3, 4, 5* and *6* influence energy metabolism in humans, coding regions of the four gene were sequenced using DNA samples obtained from 3551 participants in DHS[5]. A total of 348 nucleotide sites of sequence variations were uncovered in all four genes. Most of them are rare and 86% of them have MAFs below 1%[5]. Individuals with diabetes mellitus, heavy alcohol use, or who were taking lipids lowering drugs were removed from the all the analyses because these factors could be potential confounders. Additionally individuals who do not belong to the AA, H or EA ethnic groups were removed from the analysis. Following the original study[5], and to control for potential confounders[53] we stratified the sample by race, sex, and quantitative trait level. For each quantitative trait, to test if the rare variants are enriched in the expected extremes, individuals from bottom and top quartiles are used to mimic a case-control type of design. The KBAC, WSS, CMC and RVE were applied to carry-out the association analysis.

Figure 1: Quantile-Quantile (QQ) plot of p-values

P values were obtained from Monte Carlo approximation (left panel), permutation (right panel) and theoretical expectations. P-values were estimated using 10,000 iterations and 10,000 permutations for Monte Carlo approximation and permutation, respectively. Four sample sizes were investigated: 200 cases/200 controls; 300 cases/300 controls, 400 cases/400 controls, and 500 cases/500 controls. A total of 3,000 replicates were used to generate the QQ plot for each sample size.

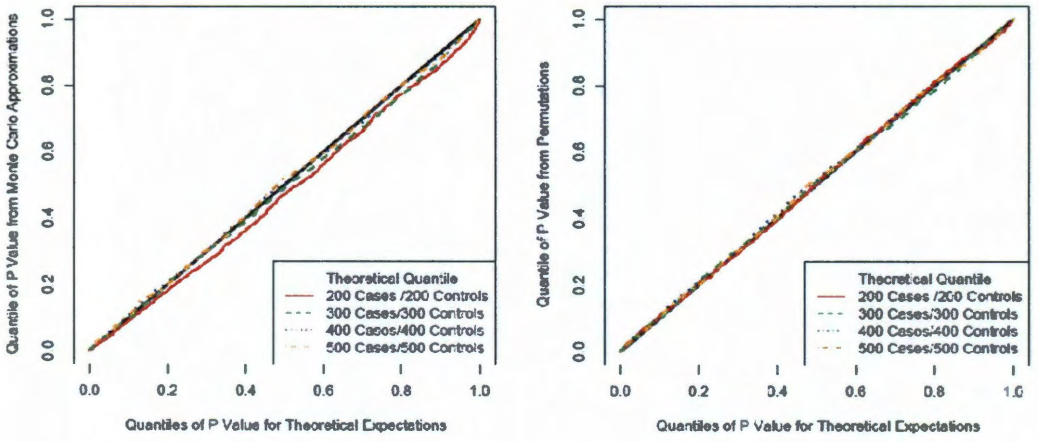

Figure 2: Impact of misclassifications under main effects model with fixed genetic effects using simulated SFS for AA.

Each causal rare variant has an OR= 3.0. Power comparisons were made for the KBAC, WSS, CMC, and RVE when 0% ~ 60% of causal rare variants are excluded from the analysis (left panel) and when 0% ~ 100% of non-causal rare variants are included (right panel). A sample size of 1000 cases and 1000 controls was used for each scenario. P-values were empirically estimated using 5,000 permutations and power was evaluated for a significance level of $\alpha = 0.05$ using 2,000 replicates for each scenario.

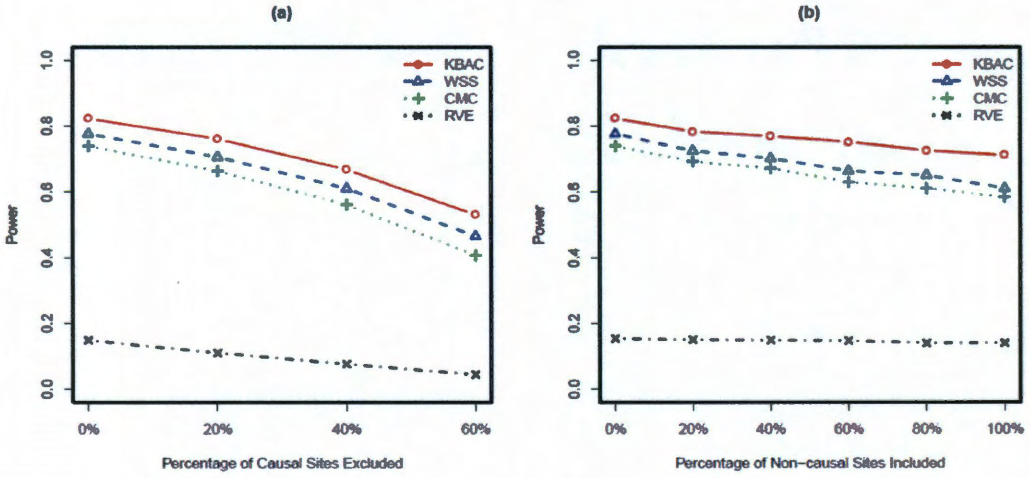

Figure 3: Impact of misclassifications under main effects model with variable genetic effects using simulated SFS for AA.

The disease odds for causal variants are inversely correlated with their MAFs and within the range of 2 ~ 20. Power comparisons were made for the KBAC, WSS, CMC, and RVE when 0% ~ 60% of causal rare variants are excluded from the analysis (left panel) and when 0% ~ 100% of non-causal rare variants are included (right panel). A sample size of 1000 cases and 1000 controls was used for each scenario. P-values were empirically estimated using 5,000 permutations and power was evaluated for a significance level of $\alpha = 0.05$ using 2,000 replicates for each scenario.

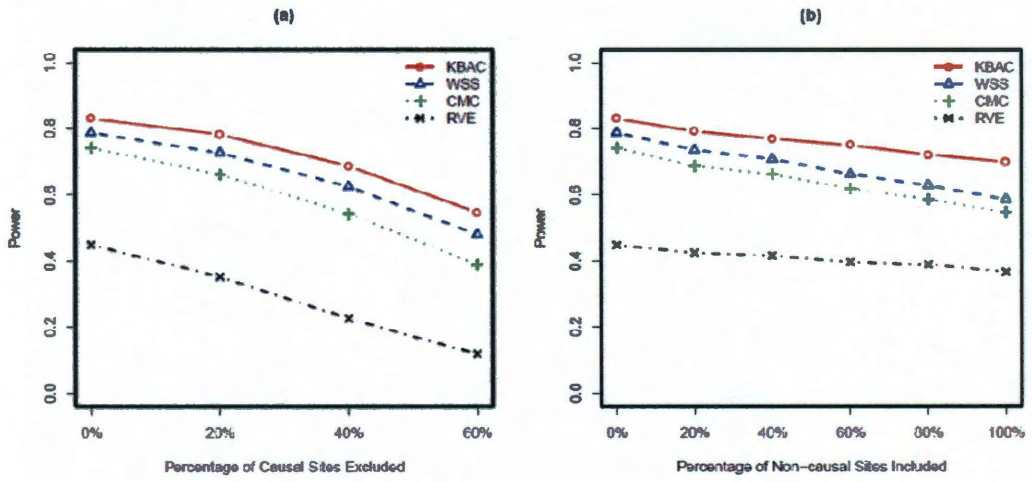

Figure 4: Power comparisons for within gene (left panel) and between gene interaction model (right panel) with simulated SFS for AA.

Power was evaluated for the KBAC, WSS, CMC and RVE. A sample size of 1000 cases and 1000 controls were used for the within interaction model, and a sample size of 300 cases and 300 controls were used for the between gene interaction model. Scenarios with different proportions of causal variants were considered. P-values were empirically estimated using 5,000 permutations and power was evaluated for a significance level of $\alpha = 0.05$ using 2,000 replicates.

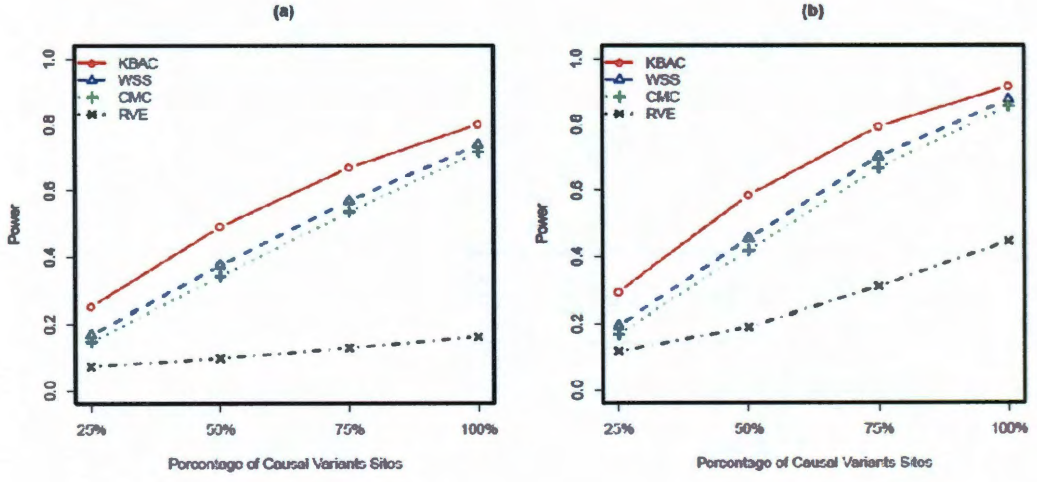

Table 1 :Rare variant summary statistics.

The summary statistics are displayed for the generated replicates under main effects model with fixed and variable genetic effects using simulated SFS from AA population. Scenarios with different proportions of causal variants excluded and scenarios with different proportions of non-causal variants included were considered. The table displays for a given sample, the information on a) the average proportion of rare NS variant carriers among cases and controls; b) the mean number of rare NS variant sites; c) the mean number of rare NS variant sites that are exclusive to cases or controls; d) the average proportion of case and control rare NS variant carriers with more than one rare variant. For each scenario, a sample size of 1,000 cases and 1,000 controls were used. 2,000 replicates were generated for each scenario.

Scenario		Rare Variant Carrier Frequencies in Cases/Controls	Mean number of Rare Variant Sites	Mean number of Rare Variant Sites Observed Exclusively in Cases/Controls	Proportions of Rare Variant Carriers with More than One Rare Variant in Case/Controls
Phenotypic Model with Variable Genetic Effects Inversely Correlated with MAFs					
Percentage of Causal Variants Excluded	20%	0.033/0.014	5.791	2.978	0.013/0.006
	40%	0.025/0.011	4.396	2.285	0.009/0.004
	60%	0.017/0.008	3.048	1.556	0.006/0.003
Perc entage of Non-causal Variants Included	0%	0.038/0.017	6.942	3.609	0.016/0.006
	20%	0.041/0.02	7.614	3.859	0.018/0.008
	40%	0.044/0.023	8.501	4.274	0.019/0.009
	60%	0.048/0.027	9.535	4.645	0.021/0.012
	80%	0.051/0.03	10.539	5.044	0.022/0.014
	100%	0.055/0.034	11.665	5.53	0.025/0.016
Phenotypic Model with Fixed Genetic Effects Unrelated to MAFs					
Perc entage of Causal Variants Excluded	20%	0.034/0.014	4.455	1.797	0.014/0.005
	40%	0.027/0.011	3.449	1.39	0.01/0.004
	60%	0.019/0.008	2.36	0.956	0.006/0.003
Perc entage of Non-causal Variants Included	0%	0.041/0.017	5.325	2.158	0.017/0.007
	20%	0.043/0.019	5.996	2.439	0.018/0.008
	40%	0.047/0.023	7.058	2.875	0.02/0.01
	60%	0.05/0.027	8.007	3.259	0.022/0.013
	80%	0.054/0.03	8.931	3.565	0.024/0.013
	100%	0.057/0.034	10.047	4.132	0.026/0.015

Table 2 Rare variant summary statistics.

The summary statistics are displayed for the generated replicates under within gene interaction model and between gene interaction model using simulated SFS from AA population. Scenarios with different proportions of causal variants were considered. The table displays for a given sample, the information on a) the average proportion of rare NS variant carriers among cases and controls; b) the mean number of rare NS variant sites; c) the mean number of rare NS variant sites that are exclusive to cases or controls; d) the average proportion of case and control rare NS variant carriers with more than one rare variant. For within gene interaction model, a sample size of 1,000 cases and 1,000 controls were used, and for the between gene interaction model, a sample size of 300 cases and 300 controls were used. 2,000 replicates were generated for each scenario.

Scenario			Rare Variant Carrier Frequencies in Cases/Controls	Mean Number of Rare Variant Sites	Mean number of Rare Variant Sites Observed Exclusively in Cases/Controls	Proportions of Rare Variant Carriers with More than One Rare Variant in Case/Controls
Between Gene Interaction Model						
Percentage of Causal Variants:	5%	Gene 1	0.049/0.035	7.348	3.612	0.022/0.015
		Gene 2	0.038/0.035	7.023	3.39	0.018/0.016
	0%	Gene 1	0.065/0.035	7.699	3.749	0.029/0.015
		Gene 2	0.042/0.034	7.174	3.475	0.019/0.016
	5%	Gene 1	0.079/0.034	8.146	4.024	0.035/0.015
		Gene 2	0.046/0.034	7.259	3.509	0.021/0.015
	00%	Gene 1	0.096/0.034	8.622	4.276	0.043/0.015
		Gene 2	0.049/0.035	7.432	3.553	0.023/0.016
Within Gene Interaction Model						
Percentage of Causal Variants	25%		0.037/0.032	9.109	2.999	0.016/0.014
	50%		0.043/0.032	9.295	3.026	0.02/0.014
	75%		0.048/0.031	9.352	3.003	0.022/0.014
	100%		0.055/0.032	9.627	3.042	0.028/0.014

Table 3: Association analyses of the *ANGPTL 3,4,5* and *6* gene variants with human energy metabolism phenotypes.

Nine phenotypes were analyzed: triglyceride (TG), high density lipoprotein (HDL), low density lipoprotein (LDL), very low density lipoprotein (VLDL), total cholesterol, glucose, body mass index (BMI), and systolic (SysBP) and diastolic (DiasBP) blood pressure. Analyses were carried-out including only NS variants. The KBAC, WSS, and CMC were used to analyze each trait and nominally significant p-values are indicated with an asterisk. The p values for KBAC, WSS and RVE were obtained empirically using 10,000 permutations, while the p-value for CMC was analytically calculated.

Phenotype	Gene Name	KBAC	WSS	CMC	RVE	Numbers of Carriers of Rare Variants Observed in Upper / Lower Quartiles	Number of Carriers of Rare Variants Observed Exclusively in either the Upper or Lower Quartiles
BMI	*ANGPTL3*	0.556	0.832	0.915	0.746	47 / 48	8 / 6
	ANGPTL4	0.999	0.331	0.412	0.104	62 / 71	2 / 7
	ANGPTL5	**0.001***	**0.003****	**0.006****	0.263	83 / 51	5 / 1
	ANGPTL6	0.128	0.189	0.217	0.410	40 / 29	9 / 5
DiasBP	*ANGPTL3*	0.237	0.805	0.759	0.950	53 / 49	6 / 6
	ANGPTL4	0.784	0.437	0.445	0.086	56 / 63	3 / 9
	ANGPTL5	0.432	0.590	0.652	0.636	71 / 65	3 / 4
	ANGPTL6	**0.045***	0.084	0.088	0.405	49 / 33	12 / 7
SysBP	*ANGPTL3*	0.455	0.965	1.000	0.919	49 / 48	7 / 6
	ANGPTL4	0.409	0.835	0.789	0.935	71 / 67	6 / 6
	ANGPTL5	0.106	0.498	0.602	0.053	77 / 71	10 / 2
	ANGPTL6	0.473	0.349	0.346	0.510	34 / 42	11 / 7
Cholesterol	*ANGPTL3*	0.950	0.299	0.326	0.906	40 / 49	7 / 7
	ANGPTL4	0.260	0.503	0.515	0.123	68 / 59	4 / 9
	ANGPTL5	0.353	0.697	0.783	0.778	68 / 63	8 / 7
	ANGPTL6	0.348	0.573	0.628	0.052	38 / 33	10 / 2
LDL	*ANGPTL3*	0.792	0.894	1.000	0.855	46 / 46	8 / 7
	ANGPTL4	0.508	0.695	0.709	0.064	66 / 60	4 / 11
	ANGPTL5	0.544	0.908	0.860	0.278	73 / 70	1 / 4
	ANGPTL6	0.307	0.745	0.813	0.388	39 / 36	9 / 5
HDL	*ANGPTL3*	0.834	0.992	1.000	0.237	50 / 51	2 / 7
	ANGPTL4	**0.021***	**0.041***	**0.045***	0.681	84 / 62	7 / 6
	ANGPTL5	0.077	0.115	0.123	0.170	85 / 67	5 / 1
	ANGPTL6	0.143	0.211	0.239	0.513	43 / 33	6 / 9
TG	*ANGPTL3*	**0.015***	0.053	0.058	0.312	34 / 52	6 / 11
	ANGPTL4	**0.004****	**0.005****	**0.006****	0.087	46 / 76	2 / 8
	ANGPTL5	0.212	0.678	0.852	0.165	62 / 64	1 / 5
	ANGPTL6	0.683	0.664	0.709	0.057	35 / 32	15 / 6
VLDL	*ANGPTL3*	**0.028***	**0.047***	0.061	0.352	35 / 53	7 / 12
	ANGPTL4	**0.001****	**0.006****	**0.010***	0.141	49 / 80	3 / 9
	ANGPTL5	0.265	0.941	1.000	0.263	67 / 68	1 / 5
	ANGPTL6	0.706	0.756	0.806	0.140	35 / 34	12 / 6
Glucose	*ANGPTL3*	0.485	0.589	0.612	0.690	49 / 55	5 / 7
	ANGPTL4	0.872	0.549	0.659	0.706	75 / 67	6 / 7
	ANGPTL5	0.407	0.896	0.862	**0.021***	76 / 72	1 / 9
	ANGPTL6	0.196	0.198	0.239	**0.026***	44 / 32	14 / 3

Chapter 3

Replication Strategies of Rare Variant Complex Trait Association via Sequencing

3.1. Background:

Currently there is worldwide interest in studying the role of rare genetic variants in the etiology of complex traits. A number of studies provide evidence that rare variants are involved in the etiology of complex diseases and quantitative phenotypes[1,3-6]. Indirect association mapping using tagSNPs is underpowered to detect associations with rare variants due to the weak correlations between higher frequency tagSNPs and rare variants[20]. Instead direct association mapping through sequencing candidate genes, exomes or entire genomes needs to be applied, where variants are discovered and tested. With the rapid development of cost effective next generation sequencing technologies such as Illumina HiSeq, ABI SOLiD, and Roche 454 as well as target enrichment methods, sequencing-based genetic association studies of complex traits have been made possible. For targeting large numbers of genetic regions, hybrid based methods such as on-array or in-solution capture with NimbleGen or Agilent products have been very beneficial[54-56]. When targeting small genetic regions is of interest, such as in candidate genes, capture methods that use molecular inversion probes are advantageous[56]. Although sequencing only captured genetic regions can be cost and time effective, high sequencing associated cost is still a concern, especially for sequencing a large number of individuals at high coverage which is necessary to accurately detect rare variants.

Another constraint of the application of next generation sequencing to association studies is error rate. Relatively high false variant discovery rates have been reported for

short reads technologies even at high coverage, e.g. Illumina Solexa (6.3%) and ABI SOLiD (7.8%)[57]. Given these concerns, there is interest in exploring alternative technologies after the variant discovery stage to extract information from targeted genetic regions, such as customized genotyping or the development of an exome genotyping chip.

In this chapter, the plausibility of applying customized genotyping and next generation sequencing in replication studies is explored from a combined genetic epidemiology and population genetics perspective. The power of the two replication strategies is dependent on the percentage of causal variants sites that were uncovered for the gene region in the stage 1 sample. If the stage 1 sample is small, there can be an advantage to sequence-based replication, since many low frequency variants may not have been observed. However, the difference between variant-based and sequence-based replication strategies will diminish if a majority of causal variants can be uncovered in stage 1. Discovery of SNPs using population-based samples have been addressed previously[58,59]. However, in these studies the population genetic models employed were overly simplistic; they did not incorporate complex human demographic history and purifying selection which are well known factors that can affect rare variant site frequency spectrums[37]. A rigorous population genetic model for Africans was used with parameters estimated using sequence data[33]. Together with realistic phenotypic models motivated by complex traits, we investigate the probability of uncovering rare variants in the context of case-control studies and demonstrate their impact on the relative performances of sequence and variants based replication.

Additionally, the relative power of the two replication strategies will also be affected by the error rates of next generation sequencing and customized genotyping technologies which are employed. To assess the impact of sequencing error on the power of rare variant association mapping, a novel sequencing error model was introduced. The parameters of the sequencing error model were estimated according to reported false positive and false negative variant discovery rates from commonly used next-generation sequencing platforms[54,55,57,60].

It is demonstrated through extensive simulations that the sequence-based replication is more powerful than variant-based replication for both small and large scale studies. In the ideal scenario where sequencing and genotyping are both of perfect quality, for small scale studies with several hundred cases and controls, a large proportion of variant nucleotide sites will not be uncovered. However, uncovered rare variants in small scale genetic studies can account for over 80% locus population attributable risk (PAR). Therefore the advantage in power can be very small. For large scale studies with thousands of cases and controls, over 90% causative variant nucleotide sites can be uncovered and nearly 100% locus PAR can be explained by the uncovered rare variants. Therefore, genotyping can be a temporal solution for replicating stage 1 studies if stage 1 and 2 samples are drawn from the same population. Resequencing based replication studies have an irreplaceable advantage in that novel variants can be discovered. This benefit can be important when the stage 1 sample is of small scale. In the presence of

sequencing errors, genotyping errors and non-converted genotyping assays, the relative performances of two replication strategies remain largely unchanged. We show that the power for sequencing based association mapping is only slightly impacted by currently attainable levels of error rates, for example, false positive rate/false negative rates of 6.3%/1%[57] or 10%/5%[61].

In order to further illustrate the relative performances of sequence-based and variant-based replication, phenotype data on energy metabolism traits and sequence data from the Dallas Heart Study on the *ANGPTL 3*, *4*, *5*, and *6* genes[5,6] were analyzed using both the CMC and WSS. The results provide solid support for the simulation experiments.

3.2. Material and Methods:

It is assumed that for a gene region of length L there are S variant sites in the study population. The major allele at site s is denoted by a^s, whereas the minor allele is labeled A^s. The underlying L-site genotype of an "individual" i is coded by a vector, i.e. $\vec{X}_i = (x_i^1, \cdots, x_i^S, x_i^{S+1}, x_i^L)$. For notational convenience, sites $1, \cdots, S$ are assumed to be variant sites in the population. Dominant genotype coding is adopted for variant nucleotide sites, i.e.

$$x_i^s = \begin{cases} 1 & \text{if the genotype at nucleotide site } s \text{ is } A^s a^s, \text{ or } A^s A^s \\ 0 & \text{otherwise} \end{cases}, s = 1, \cdots, S$$

The genotypes at monomorphic sites are identically coded as 0, i.e. $x_i^s = 0, i = S+1, \cdots, L$

Following the approach in Li and Leal[20], the collapsed genotype is introduced using an indicator function $\delta(\bullet)$, i.e.

$$X_i = \delta\left(\sum_{s=1}^{L} x_i^s > 0\right).$$

The affection status for individual i is encoded by an binary variable Y_i, which takes value 1 if the individual is affected, and 0 otherwise.

3.2.1. Probabilistic Model for Sequencing Errors:

Due to the presence of sequencing errors, the observed genotype of an "individual" i may be different from the true underlying genotype. The observed genotype from sequence data is given by $\vec{Z}_i = \left(z_i^1, z_i^2, \cdots, z_i^L\right)$, where

$$z_i^s = \begin{cases} 1 & \text{if the genotype at nucleotide site } s \text{ is called as } A^s a^s, \text{ or } A^s A^s \\ 0 & \text{otherwise} \end{cases}$$

The corresponding collapsed genotype Z_i is similarly defined as $Z_i = \delta\left(\sum_{s=1}^{L} z_i^s > 0\right)$.

Two types of sequencing error events are given probabilistically[57]. First, a false positive event is defined as $\left\{z_i^s = 1, x_i^s = 0\right\}$, where a non-variant genotype at nucleotide site s is falsely called a variant. The error rate that corresponds to the false positive event is defined as the conditional probability $e_{01}^s = P\left(z_i^s = 1 \mid x_i^s = 0\right)$. Second, a false negative event can be defined as $\left\{z_i^s = 0, x_i^s = 1\right\}$ where a variant at site s is falsely called homozygous for the reference allele. Its error probability is defined as $e_{10}^s = P\left(z_i^s = 0 \mid x_i^s = 1\right)$.

To measure and report sequencing error at a rare variant nucleotide site, it is common to use *false positive discovery rate* and *false negative error rate*[54,55,57,60,61], i.e.

$$\hat{FP}=\frac{\sum_{i,s}\delta\left(z_i^s=1,x_i^s=0\right)}{\sum_{i,s}\delta\left(z_i^s=1\right)},\ \hat{FN}=\frac{\sum_{i,s}\delta\left(z_i^s=0,x_i^s=1\right)}{\sum_{i,s}\delta\left(x_i^s=1\right)} \qquad (3.1)$$

Empirical estimates of false positive and false negative rates are usually obtained by comparing next-generation sequencing with less error prone technologies, e.g. Sanger sequencing or customized genotyping [57,61].

Using reported false positive and false negative rates, base-pair error rates can be calculated as

$$FP=\frac{\sum_{s=1}^{L}e_{01}P\left(x_i^s=0\right)}{\sum_{s=1}^{L}e_{01}P\left(x_i^s=0\right)+\left(1-e_{10}\right)P\left(x_i^s=1\right)},\ FN=\frac{\sum_{s=1}^{L}e_{10}P\left(x_i^s=1\right)}{\sum_{s=1}^{L}P\left(x_i^s=1\right)} \qquad (3.2)$$

According to formula (3.2) and simulated site frequency spectrums, a combination of $FP = 6.3\%$, $FN = 1.0\%$ corresponds to a locus average sequencing error rate of $P\left(Z_i=1\middle|X_i=0\right)=0.18\%$ and $P\left(Z_i=0\middle|X_i=1\right)=1\%$.

Based on the above sequencing error model, the distribution for the number of carriers of rare variants at each site ($\vec{M}_A = \left(m_A^1, m_A^2, \cdots, m_A^S\right), \vec{M}_U = \left(m_U^1, m_U^2, \cdots, m_U^S\right)$) can be analytically derived. For a sample with R^A cases and R^U controls, m_A^s, m_U^s follow binomial distribution i.e. $m_A^s = \sum_{i,s} \delta\left(X_i^s = 1, Y_i = 1\right) \sim Binom\left(R_A, p_A^s\right)$, $m_U^s = \sum_{i,s} \delta\left(X_i^s = 1, Y_i = 0\right) \sim Binom\left(R_U, p_U^s\right)$, where parameters $\vec{p}_A = \left(p_A^s\right)_{s=1,\cdots,L}, \vec{p}_U = \left(p_U^s\right)_{s=1,\cdots,L}$ are given by

$$\begin{cases} p_A^s = P\left(z_i^s = 1 \mid Y_i = 1\right) = \left(1 - e_{10}\right) \times P\left(X_i^s = 1 \mid Y_i = 1\right) + e_{01} \times P\left(X_i^s = 0 \mid Y_i = 1\right) \\ p_U^s = P\left(z_i^s = 1 \mid Y_i = 0\right) = \left(1 - e_{10}\right) \times P\left(X_i^s = 1 \mid Y_i = 0\right) + e_{01} \times P\left(X_i^s = 0 \mid Y_i = 0\right) \end{cases}, s = 1, \cdots, L$$

3.2.2. Models of Genotyping Errors:

It is assumed that a set K of rare variant sites are uncovered in the stage 1 sample. Rare variants from sites K are genotyped and followed up in the stage 2 replication sample. Although the accuracy for commercially available genotyping array is high, the error rate for customized genotyping is not negligible[62]. Additionally, assays on customized probes may have a low conversion rate.

The observed locus genotype from genotyping data is denoted by $\vec{W}_i = \left(w_i^1, w_i^2, \cdots, w_i^L\right)$, where

$$w_i^s = \begin{cases} 1 & \text{if the genotype at nucleotide site } s \text{ is called as } A^s a^s, \text{ or } A^s A^s \\ 0 & \text{otherwise} \end{cases}, s \in K \quad (3.3)$$

The corresponding collapsed genotype W_i is similarly defined. For a converted assay, the genotyping error is traditionally measured as sample error rate (SER)[63,64], i.e. $SER_s = P(W_i^s = 1, X_i^s = 0 \mid \mathrm{I}_C^s = 1) + P(W_i^s = 0, X_i^s = 1 \mid \mathrm{I}_C^s = 1), s \in K$, where I_C^s is an indicator for an assay being successful (e.g. converted with genotype calls generated) at nucleotide site s. Similar to sequencing errors, the genotyping error rates at converted probes are defined as:

$$f_{01}^s = P(W_i^s = 1 \mid X_i^s = 0, \mathrm{I}_C^s = 1),\ f_{10}^s = P(W_i^s = 0 \mid X_i^s = 1, \mathrm{I}_C^s = 1), s \in K. \quad (3.4)$$

To facilitate comparisons of the two replication strategies, an error ratio is introduced to measure the relative error rates for sequencing and customized genotyping, i.e. $ER = \frac{\sum_{s \in K} f_{10}^s}{\sum_{s=1}^{L} e_{10}^s} = \frac{\sum_{s \in K} f_{01}^s}{\sum_{s=1}^{L} e_{01}^s}$. When the two replication strategies are compared in the presence of imperfect technologies, two error ratios are used, i.e. $ER = 1$ or $ER = 0.5$. The rate of success for a given assay at site s, i.e $P(\mathrm{I}_C^s = 1)$ is assumed to be 90%.

For a genotyped sample with R_A cases and R_U controls, the observed counts for carriers of variants at each nucleotide site are denoted by $\vec{N}_A = (n_A^s)_{s \in K}$, $\vec{N}_U = (n_U^s)_{s \in K}$, which follow binomial distribution i.e. $n_A^s \sim Binom(R_A, q_A^s)$, and $n_U^s \sim Binom(R_U, q_U^s)$. The parameters $\vec{q}_A = (q_A^s)_{s \in K}, \vec{q}_U = (q_U^s)_{s \in K}$ are provided by

$$\begin{cases} q_A^s = P(Z_i^s = 1 \mid Y_i = 1) = (1 - f_{10}^s) \times P(I_C^s = 1) \times P(X_i^s = 1 \mid Y_i = 1) + f_{01}^s \times P(I_C^s = 1) \times P(X_i^s = 0 \mid Y_i = 1) \\ q_U^s = P(Z_i^s = 1 \mid Y_i = 0) = (1 - f_{10}^s) \times P(I_C^s = 1) \times P(X_i^s = 1 \mid Y_i = 0) + f_{01}^s \times P(I_C^s = 1) \times P(X_i^s = 0 \mid Y_i = 0) \end{cases}, s \in K \quad (3.5)$$

3.2.3. Power calculation for sequence-based and variant-based replication:

Several test statistics for mapping rare variants have been proposed, such as combined multivariate and collapsing (CMC) [20], weight sum statistic (WSS) [21]. They have been shown to be more powerful than multivariate methods such as Hotelling T^2. Here, both CMC and WSS are used in the comparisons of variant-based and sequence-based replication. The CMC has a closed form distribution, which makes it computationally efficient for candidate gene, exome- and genome-wide studies. The power of the permutation-based WSS method was also evaluated for small scale candidate gene studies. For all the scenarios evaluated, WSS is more powerful than CMC, but the comparisons for the two replication strategies are largely unaffected by the choice of the test statistics.

For both sequence and variants based replication with CMC, the association tests in both the stage 1 and stage 2 studies are implemented using Fisher exact test which compares rare variant carrier frequencies between cases and controls. When the WSS statistic is used, to guarantee that type I error is well controlled, p-values are estimated empirically based upon 2000 permutations for each replicate. Due to computation intensity of estimating small empirical p-values, the WSS was not used for the evaluation of power to replicate large scale studies.

The test statistics used for the stage 1 and sequence based stage 2 studies are denoted by T^{S1} and T^{seq} respectively. The power of successfully replicating a true significant association from the stage 1 study is investigated, i.e. $P_{H^A}\left(\left|T^{seq}\right| > z_{1-\alpha_{S2}/2} \middle| \left|T^{S1}\right| > z_{1-\alpha_{S1}/2}\right)$, where α_{S1} and α_{S2} are significance levels used for stage 1 and the replication study. Since the statistics are conditionally independent given the parameters $\vec{p}_A$ and $\vec{p}_U$, the following equation must be satisfied,

$$P_{H^A}\left(\left|T^{seq}\right| > z_{1-\alpha_{S2}/2}, \left|T^{S1}\right| > z_{1-\alpha_{S1}/2} \middle| \vec{p}^A, \vec{p}^U\right) = P_{H^A}\left(\left|T^{seq}\right| > z_{1-\alpha_{S2}/2} \middle| \vec{p}^A, \vec{p}^U\right) P_{H^A}\left(\left|T^{S1}\right| > z_{1-\alpha_{S1}/2} \middle| \vec{p}^A, \vec{p}^U\right) \quad (3.6)$$

The power for "variant-based" replication is given by $P_{H^A}\left(\left|T^{\text{var}}\right| > z_{1-\alpha_{S2}/2} \middle| \left|T^{S1}\right| > z_{1-\alpha_{S1}/2}\right)$ but unlike for sequence-based replication, the test statistics T^{var}, T^{S1} are not conditionally independent. Under the alterative hypothesis, the distribution of T^{var} depends on K which is the set of rare variant sites uncovered in stage 1. As it is impossible to enumerate the parameter space of $(\vec{p}_A, \vec{p}_U, \vec{q}_A, \vec{q}_U, K)$, an efficient Monte Carlo algorithm was developed to calculate replication power for both sequence-based and variant-based strategies.

For notational convenience, the ratio of total frequencies of uncovered rare variants to the total frequencies of all locus rare variants (including those that are not uncovered) is denoted by $f_{MAF} = \sum_{s \in K} P\left(X_i^s = 1\right) \Big/ \sum_{s=1}^{S} P\left(X_i^s = 1\right)$,

In addition, the ratio

$$f_{PAR} = \sum_{s \in K \cap C} P\left(X_i^s = 1 \middle| Y_i = 1\right) \Big/ \sum_{s \in C} P\left(X_i^s = 1 \middle| Y_i = 1\right)$$

represents the proportion of locus PAR that can be explained by the uncovered causal variants. This is asymptotically equivalent to the epidemiological definition of PAR which is the reduction of disease incidence rate that would be observed if the population were unexposed i.e. if there were no carriers of locus causative variants.

The power comparisons were performed for both the small and large scale genetic studies. In order to have enough power to detect associations, 250 cases/250 controls or 500 cases/500 controls were used for both the stage 1 and 2 samples in a small scale study. For the scenario of a large scale study, 2000 cases/2000 controls, as well as 3500 cases/3500 controls were examined. For small scale studies, examples are given using significance levels $\alpha_{S1} = \alpha_{S2} = 0.05$. The commonly accepted exome-wide significance level $\alpha_{S1} = \alpha_{S2} = 2.5 \times 10^{-6}$ is used for large scale genetic studies, which is based upon Bonferroni corrections for testing 20,000 genes. The empirical power for each scenario was estimated using 10,000 replicates.

3.2.4. Simulations of Complex Demographic Models and Selections:

To compare relative efficiencies of sequence-based and variant-based replication strategies, population genetic data was generated using forward time simulations[32]. Genetic data for the African population was generated. The parameters for demographic

changes and selections were estimated in Boyko et al.[33] The demographic change for the African population is described using a two epoch instant change model. Purifying selection was also simulated, with u and $2u$ being the selective disadvantage of heterozygous and homozygous new mutations. Scaled fitness effect $\gamma = 2N_{curr}u$ (where N_{curr} is the current effective population size) is assumed to follow a gamma distribution, i.e.

$$\gamma = -x, x \sim \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \text{ where } a = 0.184, b = 8{,}200 \quad (3.7)$$

The model was shown to be parsimonious and fit the data well. A mutation rate of $\mu_S = 1.8 \times 10^{-8}$ per nucleotide site per generation is assumed. Since the average length for human gene coding region is 1500 base pairs (bp) long[50,51], $L = 1500$bps was used in the simulation to specify the locus scaled mutation rate. Based upon the above parameter specification, 100 sets of rare variant site frequencies were generated. As suggested by Kryukov et al.[24], only non-synonymous (NS) variants were used in the analysis in order to increase the signal to noise ratio, and reduce the negative impact of non-functional variants on power.

3.2.5. Generations of Phenotypic Model:

Phenotypic effects of rare NS variants are assumed independent of their fitness[50]. 50% of the rare NS variants (with MAF ≤ 0.01) are randomly picked to be causal, and affect the binary phenotype of interest. Based upon surveys of multi-factorial

diseases[8], two types of phenotypic models were considered. For the first type of model, the genetic effects of causal variants are inversely correlated with their MAFs. It is assumed that causal variants with the smallest (or largest) MAFs (i.e. $p_{\min}$ or $p_{\max}$) have largest (or smallest) log odds ratio (log-OR) of $\beta_{\max}$ (or $\beta_{\min}$) respectively. For a causal variant with MAF p_i, the log-OR follows the interpolation relation: $\beta_i = \beta_{\max} + (\beta_{\max} - \beta_{\min})/(p_{\max} - p_{\min}) \times (p_i - p_{\min}), i \in C$. The ORs for causal variants thus satisfy an exponential relationship with their MAFs. A choice of $\beta_{\max} = \log(10), \beta_{\min} = \log(2)$ was used. For the second type of model, each causal variant has equal disease odds, which is given by $\beta_i = \log(3), i \in C$. Under both types of models, the affection status for an individual with multi-site genotype $\vec{X}$ is assigned by the following model:

$$P(Y_i = 1 | \vec{X}) = \frac{\exp(\beta_0 + \sum_{i \in C} \beta_s X_i^s)}{1 + \exp(\beta_0 + \sum_{i \in C} \beta_s X_i^s)}, \quad (3.8)$$

A baseline penetrance of 0.01 is assumed, which gives $\beta_0 = \log(0.01/(1-0.01))$.

3.2.6. Applications to the Dallas Heart Study Sequence Data:

In order to illustrate the relative efficiency of sequence-based versus variant-based replication strategies, a data set from DHS was analyzed. The dataset is a multi-ethnic population based sample [1830 African Americans (AA), 601 Hispanics (H), 1045 European Americans (EA), and 75 individuals from other ethnic groups] from Dallas County residents whose lipids and glucose metabolism have been characterized and

recorded[36,52]. In order to investigate how sequence variations in *ANGTPL3, 4, 5* and *6* influence energy metabolism in humans, coding regions of the four genes were sequenced using DNA samples obtained from 3551 participants in DHS[5]. A total of 348 nucleotide sites of sequence variations were uncovered in the four genes. Most of them are rare and 86% of them have MAFs < 1%[5]. Nine phenotypes were measured and tested for their associations with rare genetic variants, i.e. body mass index (BMI), diastolic blood pressure (DiasBP), systolic blood pressure (SysBP), total cholesterol level (TCL), low density lipoprotein (LDL), high density lipoprotein (HDL), triglyceride (TG), very low density lipoprotein (VLDL) and glucose. As a first analysis, to mimic the scenario of stage 1 study, individuals with quantitative trait values in the top and bottom 10% of the phenotypic distributions were used to form a "case-control" dataset. Individuals with intermediate quantitative trait values, i.e. in the range of 10-35% and 65%-90% were used as a replication sample. Sequence-based and variant-based replications were compared for the chosen replication dataset.

Among the identified significant results, the association between TG level and rare variants in the *ANGPTL*4 gene was supported by in-vitro functional studies and was replicated using an independent dataset[5,6]. It is highly likely to be a true association. Therefore a second experiment was performed to estimate the "power" for replicating the association between TG and rare variants in the *ANGPTL*4 gene. Individuals with TG levels in the range of top 35% and bottom 35% were used to form a case-control "cohort". 50% of the cases and 50% of the controls from the "cohort" were randomly selected as the dataset for the stage 1 study. The remaining 50% of cases and controls are used as the

stage 2 replication sample. The process was repeated 1000 times, and for each replicate, sequence-based and variant-based replication were performed. The fraction of significant stage 1 studies that were successfully replicated in stage 2 was reported as empirical replication power. Association tests with WSS and CMC were both performed.

3.3. Results:

3.3.1. Discovery Rate of Rare Variant Sites and Frequencies:

Rare variant discovery rates were compared under the assumption that sequencing data is of perfect quality (Table 4). When sequencing is not perfect, the fractions of uncovered rare variants will be lowered by false negative rate and additionally a portion of observed variants can be false positives.

When a variable effect model is assumed, relatively low proportions of variant nucleotide sites are uncovered for small scale studies. For example, in a sample of 250 cases/250 controls, 43.2% of all variant nucleotide sites and 59.9% of causal variant nucleotide sites are uncovered. On the other hand, a fairly large portion of locus PAR (91.7%) can be explained by the uncovered variants. When a fixed effect model is assumed, the results are very similar (Table 4). A slightly lower portion of variants can be uncovered but the uncovered variants explain a higher fraction of locus PAR.

When a sample of 500 cases and 500 controls was analyzed, a higher proportion of variant sites are uncovered, however, considerable fractions of rare variant sites in the "population" are still not observed in the sample for both fixed and variable effects models. For example, when the fixed genetic effects model is assumed, only 45.5% of causal variant sites are uncovered.

When an exome-wide significance level $\alpha_{S1} = 2.5 \times 10^{-6}$ is used, a much larger sample size is needed for sufficient power to detect significant associations[24]. Under the variable genetic effect model, when a sample of 2000 cases and 2000 controls was analyzed, for a gene region that attains exome-wide significance, a much larger fraction (72.8%) of rare variant nucleotide sites are present in the dataset, and nearly all (88.7%) causal nucleotide sites can be uncovered. These uncovered variants explain nearly 100% of the locus PAR. Therefore, in principle, when a large stage 1 sample is analyzed, the advantage of sequencing for novel SNP discoveries diminishes as long as the stage 2 samples are drawn from the same population. Similar results hold if a fixed effect model is assumed for the binary phenotype.

Since affected individuals are enriched in a case-control sample, nucleotide sites containing causal variants have a much higher probability of being uncovered than non-causal variant sites. For example, if a fixed effects model is assumed, 62.9% of the sites uncovered are causal variant sites for a sample of 250 cases and 250 controls. This fraction is much higher than the proportions of causal variant sites in the general "population" (50%).

3.3.2. Power Comparisons for Sequence-Based and Variant-Based Replication Strategies:

The power was compared for sequence-based and variant-based replications under different combinations of false positive/false negative variant discovery rate, genotyping assay success rate and error rate.

In the ideal scenario where both sequencing and customized genotyping qualities are perfect, the power for sequence-based and variant-based replication is jointly affected by the sample size, the proportions of rare variants uncovered and the fractions of uncovered rare variant sites that contain causal variants. For most of the examined scenarios, the power of sequence-based replication is consistently better than variant-based replication when CMC is used. For example, under the variable effects model (Table 5), for a sample size of 250 cases and 250 controls, the power for sequence-based replication is 54.2% while the power of variant-based replication is 50.7%. For large scale genetic studies, the power hardly differs between sequence and variant-based replication. This is because a large proportion of variant sites are uncovered in the stage 1 sample, and the uncovered variants account for nearly 100% of the locus PAR. For example, for a gene that attains exome-wide significance in a sample of 2000 cases and 2000 controls, the power for sequence-and variant-based replication are respectively 82.7% and 82.5%.

The power for sequence and variant-based replication is negatively impacted by sequencing and genotyping errors. The impact of sequencing error is small. If the fixed effects model is assumed (Table 6), for a sample size of 250 cases and 250 controls the power of sequenced-based replication is 44.6% in the absence of sequencing errors; when a false positive rate of 10% and a false negative rate of 5% are assumed, the power drops to 41.0% . Although a lower error rate is assumed for customized genotyping, the advantage of sequence-based replication remains. For instance, in this scenario, for a genotyping call rate of 90% and an error rate ratio of 1, the power for variant-based replication is 39.0%.

Comparisons of two replication strategies were also made when WSS is used for analysis of both the stage 1 and 2 datasets. Although the power is consistently higher for the WSS than for the CMC, the relative performances for sequence-based and variant-based replication are largely similar. One noticeable difference is that sequencing error tends to have a slightly more negative impact on power for studies using the WSS. For example, under the fixed effects model, for the scenario where the false positive rate is 10%, the false negative rate is 5%, and error rate ratio $ER = 0.5$, the power for sequence-based replication 53.3% is even lower than that for variant-based replication 54.7% if a sample of 250 case/250 control is used.

3.3.3. Applications to the Dallas Heart Study data:

For the first analysis, the stage 1 and 2 data from the *ANGPTL 3, 4, 5* and *6* genes are analyzed. Although a small sample size (individuals with trait values in the top and bottom 10%) was used for the stage 1 study, multiple (novel) associations were detected using the CMC and WSS (Table 7), i.e. a.) TCL with *ANGPTL3* $(p_{CMC} = 0.0283/p_{WSS} = 0.0218)$ b.) LDL with *ANGPTL* 4 $(p_{CMC} = 0.0208/p_{WSS} = 0.0246)$ c.) TG with *ANGPTL* 4 $(p_{CMC} = 0.0269/p_{WSS} = 0.0222)$, d.) VLDL with *ANGPTL* 4 $(p_{CMC} = 0.0373/p_{WSS} = 0.0236)$ e.) BMI with *ANGPTL* 5 $(p_{CMC} = 0.0287/p_{WSS} = 0.0207)$ f.) HDL with *ANGPTL* 5 $(p_{CMC} = 0.0252/p_{WSS} = 0.0218)$ and g.) BMI with *ANGPTL* 6 $(p_{CMC} = 0.0013/p_{WSS} = 0.0011)$. Among these, the association between BMI and *ANGPTL* 6 is significant even after Bonferonni correction for testing multiple genotypes and phenotypes. For most of the analyses, approximately 25%~40% of the nucleotide sites observed in the entire DHS sample are also observed in stage 1. The stage 2 replication sample consists of individuals with less extreme quantitative trait values. To ensure that the power of the stage 2 sample is adequate, the stage 2 samples are chosen to be larger than the stage 1 sample size. Two of the seven identified associations in the stage 1 sample were successfully replicated by both sequence and variant-based replication, i.e. associations between TG and *ANGPTL* 4 as well as between VLDL and *ANGPTL* 4. For both associations, sequence-based replication has slightly smaller p-values.

For the second analysis, the empirical power for replicating the validated association between TG and rare variants in *ANGPTL4* gene was compared. When the CMC is used, the empirically estimated "power" for sequence-based and variant-based replication is 65.3% and 62.7% respectively. The "power" for sequence based replication is only slightly better. This is very compatible with observations from simulated data. When the WSS is used, estimated power is greater but the relative performance (69.3% vs. 67.0%) is concordant.

3.4. Discussion:

In this chapter, sequence-based and variant-based replications for complex trait rare variant association studies were compared using a rigorous population genetic framework. It is demonstrated that in the ideal scenario where sequencing and genotyping are both of perfect quality, sequence-based replication is consistently more powerful. However, since the uncovered variants can account for a large proportion of locus PAR even for stage 1 studies with only a few hundred samples, the advantage in power can be very small if stage 1 and stage 2 samples are drawn from the same population. The power of sequence and variant- based replication studies is negatively impacted by sequencing and genotyping errors. For currently attainable levels of sequencing errors, the impact is minimal, and the advantage of using sequence-based replication studies remains.

It has been found previously that rare variants tend to be population specific[8]. Many studies have suggested that disease associated variants in different populations can have very different frequencies. For example, the E40K variant in *ANGPTL*4 gene was shown to be associated with TG levels. The MAF for E40K is approximately 3% in European-Americans but is very rare in African-Americans and Hispanics[5]. These differences can be observed in even more closely related populations, for example, rare variants in *CFTR, BRCA1* and *BRCA*2 genes have higher frequencies in the Ashkenazi Jewish population compared to other European Jewish populations such as Sephardic Jews and also to non-Jewish populations[65,66]. Population specific diversity of variant frequencies and sites is believed to be more pronounced for rare variants than for

common variants since rare variants tend to be younger and occur more recently in human history[8]. When stage 2 samples are drawn from a different population than the stage 1 samples, the variant-based replication studies may be at a disadvantage. Given that the demographic and selection models incorporating complex migration and admixtures are limited[33], simulation studies for variant discovery using multi-ethnic samples still remain to be explored. Evaluating the benefits and drawbacks of replication studies using samples from different populations will be very important.

One of our contributions is a model for incorporating sequencing error uncertainties into downstream association analysis. Some of the error rates discussed in this chapter (e.g. $FP = 6.3\%, FN = 1\%$) are attained when a saturated coverage depth is used. With the maturation of next-generation sequencing technologies, as well as the development of more sophisticated genotype calling algorithms, such as using pooled population samples[60], even lower rates should be attainable in the near future. For currently attainable levels of sequencing errors, their impact on the power of rare variant association mapping is minimal.

While the error model for Sanger sequencing is well known, the error model for next generation sequencing has not been extensively evaluated[61,67,68]. Due to the paucity of information on error rate estimation, our error model assumes equal error rates across different nucleotide sites. This is certainly an over simplification. In practice, for

different frequency bins, different false positive and false negative discovery rates can be expected. The proposed error model can be refined and applied to specific frequency bins when corresponding false positive and false negative rates estimates become available. Various studies have shown that non-random system errors exist, and cannot be ignored[57,60]. The system errors can be dependent on the genetic context of the variants. However, given that the main interest lies in gene-based association mapping, modeling error rate variations across different nucleotide sites may not be a necessity since only their overall impact in the gene region needs to be assessed. In particular, when the CMC method is used, the power is only affected by total number of errors, but not affected by the nucleotide sites where those errors occur. The error rates used in our model can be taken as locus averages. When comparing variant-based replication, genotyping error rates are assumed lower than sequencing error rate. It can be argued that this is sensible for two reasons. First, genotyping technology is more mature than sequencing, so it tends to have a lower error rate per base pair. Second, as customized genotyping is performed only at nucleotide sites with known polymorphisms, it is less error prone than sequencing where SNP discovery and genotype calling are performed simultaneously.

Population genetic data was generated through forward time simulations. Both demographic change and purifying selections are known to be important factors affecting rare variant site frequency spectrums. Therefore they are both modeled and incorporated in the simulation. Two types of phenotypic models were considered. According to surveys on multi-factorial disorders, most of the uncovered disease causative rare variants

have ORs between 2 and 4[8]. The choice of $OR = 3$ is therefore reasonable. Other values of causative variant odds ratios (i.e. $OR = 2$, $OR = 4$) were also experimented, and the conclusions remained the same (data not shown). On the other hand, variable effects models also have empirical support. It has been observed that lower frequency rare genetic variants tend to have larger disease odds compared to more frequent variants[7,8]. There is also evidence that highly penetrant rare genetic variants may be involved in the etiology of complex traits[69,70]. As a majority of rare variants have very low frequencies, when $OR_{\max} = 10, OR_{\min} = 2$ is used, most of the uncovered rare variants have ORs ≤ 4. The results of comparisons of replication study designs remain valid and robust under both types of phenotypic models.

In the examples discussed in this chapter, two different significance levels are used in stage 1 $\left(\alpha_{S1} = 0.05, \alpha_{S1} = 2.5 \times 10^{-6}\right)$. These significance levels are chosen for illustrative purposes. In practice, the significant levels used are dependent on the effective number of tests which can be performed. Currently for exome data where analysis is performed on a gene by gene basis, it is recommended to use an α level of 2.5×10^{-6}. This significance level is based on the Bonferroni correction for testing 20,000 genes. Since there is little linkage disequilibrium between rare variants in different genes, Bonferonni correction will not be overly conservative. If the analysis is not only performed on the gene level but pathway analysis is also performed, a more stringent α level is necessary. The choices for stage 2 significance levels are also for illustrative purposes. If gene(s) are found to be associated with a trait of interest using an

α level which adequately controls the FWER in stage 1, it is not necessary to use the same stringent α level to replicate the association. The appropriate significance level is determined by the number of tests performed in stage 2.

Sequencing based genetic studies have an irreplaceable advantage over genotyping, which is to discover novel genetic variants. Human population experienced complex patterns of demographic expansion and purifying selection[5,37]. Large numbers of very rare variant nucleotide sites exist. Based upon observations from our extensive simulations and real data, for moderate sized stage 1 studies, only a limited proportion of rare variant nucleotide sites can be uncovered. Identifying and cataloging rare variants themselves can be of great importance in genetic studies. The novel rare causal variants which are uncovered will help enhance the understanding of genetic architectures for complex traits. They can also be useful for risk prediction and personalized medicine. As a result, sequence-based replication should be eventually performed. For large scale genetic studies with thousands of cases and controls, most of the disease causative variants can be identified in stage 1. Therefore, for replicating large scale studies, customized genotyping can be a viable solution. In addition, customized genotyping can be advantageous to targeted sequencing in that multiple unlinked markers can be genotyped and used to control for population substructure/admixture. The advantage is particularly beneficial when GWAS data is not available for the replication sample.

With the rapid large-scale application of next generation sequencing, understandings of genetic etiologies of rare variants will advance to an unprecedented level. Replications of significant findings will be an indispensible part of every genetic study. Sequence-based replication for both small-scale and large-scale genetic studies is advantageous, and will eventually be affordable and widely applied. In the meantime, variant-based replication can be a temporal cost-effective solution for replications of genetic studies, and will greatly accelerate the process of identifying disease causative variants.

Table 4 Discoveries of rare variants in small and large scale genetic studies

Number of Cases/Controls in Stage 1 and 2 Samples	Proportion of Rare Variant Sites Uncovered[c]		Proportion[c]	
	All	Causal	Locus PAR Explained by Uncovered Causal Rare Variants	Causal Variant Sites among all Uncovered Rare Variant Sites
Variable Effects Phenotypic Model				
250/250[a]	0.432	0.599	0.917	0.686
500/500[a]	0.524	0.687	0.950	0.645
2000/2000[b]	0.728	0.887	0.992	0.599
3500/3500[b]	0.808	0.943	0.996	0.572
Fixed Effects Phenotypic Model				
250/250[a]	0.369	0.468	0.937	0.629
500/500[a]	0.455	0.547	0.960	0.591
2000/2000[b]	0.664	0.757	0.993	0.559
3500/3500[b]	0.751	0.827	0.995	0.538

[a]Small scale study: $\alpha_{S1} = 0.05$

[b]Large scale study: $\alpha_{S1} = 2.5 \times 10^{-6}$

[c]Rare variant data was simulated using African rare variant site frequency spectrums and case control datasets were generated using variable and fixed effects phenotypic models. A total of 10,000 replicates were generated and each reported value within the table was obtained by averaging over replicates where significant stage 1 results were obtained.

Table 5 Power comparisons of sequencing-based and variant-based replication under variable effects model

Number of Cases/Controls in Stage 1 and 2 Samples	Rates[c]				Power for Replication[d]	
	False Positive	False Negative	Assay Success	Error Ratio	Sequence-Based	Variant-Based
250/250[a]	0	0	1	1	0.542	0.507
	1%	4%	0.9	0.5	0.521	0.458
				1		0.452
	6.3%	1%	0.9	0.5	0.520	0.466
				1		0.461
	10%	5%	0.9	0.5	0.503	0.459
				1		0.447
500/500[a]	0	0	1	1	0.731	0.708
	1%	4%		0.5	0.719	0.675
				1		0.667
	6.3%	1%	0.9	0.5	0.718	0.674
				1		0.674
	10%	5%	0.9	0.5	0.701	0.672
				1		0.661
2000/2000[b]	0	0	1	1	0.827	0.825
	1%	4%	0.9	0.5	0.816	0.780
				1		0.766
	6.3%	1%	0.9	0.5	0.814	0.781
				1		0.769
	10%	5%	0.9	0.5	0.802	0.781
				1		0.755
3500/3500[b]	0	0	1	1	0.899	0.898
	1%	4%	0.9	0.5	0.893	0.870
				1		0.863
	6.3%	1%	0.9	0.5	0.893	0.868
				1		0.865
	10%	5%	0.9	0.5	0.886	0.874
				1		0.858

[a]Significance levels for small scale study: $\alpha_{S1} = 0.05$ and $\alpha_{S2} = 0.05$

[b]Significance levels for large scale study: $\alpha_{S1} = 2.5 \times 10^{-6}$ and $\alpha_{S2} = 2.5 \times 10^{-6}$

[c]The impact of different combinations of false positive/false negative rate, assay success rate and genotyping and sequencing error rate ratio on the replication power is examined.

[d]The power was empirically estimated based upon 10,000 replicates.

Table 6: Power comparisons of sequence-based and variant-based replication under fixed effects model

Number of Cases/Controls in Stage 1 and 2 Samples	Rates[c]				Power for Replication[d]	
	False Positive	False Negative	Assay Success	Error Ratio	Sequence-Based	Variant-Based
250/250[a]	0	0	1	1	0.446	0.437
	1%	4%	0.9	0.5	0.432	0.392
				1		0.386
	6.3%	1%	0.9	0.5	0.429	0.399
				1		0.390
	10%	5%	0.9	0.5	0.410	0.390
				1		0.378
500/500[a]	0	0	1	1	0.666	0.658
	1%	4%	0.9	0.5	0.650	0.619
				1		0.607
	6.3%	1%	0.9	0.5	0.652	0.623
				1		0.613
	10%	5%	0.9	0.5	0.632	0.619
				1		0.600
2000/2000[b]	0	0	1	1	0.765	0.767
	1%	4%	0.9	0.5	0.747	0.703
				1		0.689
	6.3%	1%	0.9	0.5	0.746	0.705
				1		0.694
	10%	5%	0.9	0.5	0.724	0.700
				1		0.669
3500/3500[b]	0	0	1	1	0.875	0.878
	1%	4%	0.9	0.5	0.872	0.841
				1		0.834
	6.3%	1%	0.9	0.5	0.870	0.845
				1		0.835
	10%	5%	0.9	0.5	0.856	0.843
				1		0.825

[a]Sample size for small scale study: $\alpha_{S1} = 0.05$ and $\alpha_{S2} = 0.05$

[b]Sample size for large scale study: $\alpha_{S1} = 2.5 \times 10^{-6}$ and $\alpha_{S2} = 2.5 \times 10^{-6}$

[c]The impact of different combinations of false positive/false negative rate, assay success rate and genotyping and sequencing error rate ratio on the replication power is examined.

[d]The power was empirically estimated based upon 10,000 replicates.

Table 7: Analyses of sequence data from the *ANGPTL* 3, 4, 5, and 6 genes

rait	P-values			Proportion	Ratio	Number of Rare Variants Observed	
	Stage 1 Analysis[a] (CMC/WSS)	Sequence-Based Replication[b] (CMC/WSS)	Variant-Based Replication[b] (CMC/WSS)	Nucleotide Sites Uncovered in Stage 1	Rare Variant Freq in Stage 1 Sample/Rare Variant Freq. in Entire Sample	Sequence-Based Replication	Variant Based Replication
***ANGPTL* 3**							
TCL	0.028/0.022	0.522/0.493	0.726/0.724	0.30	0.87	46/51	39/40
***ANGPTL* 4**							
LDL	0.021/0.025	0.272/0.218	0.508/0.473	0.35	0.94	78/62	70/60
TG	0.027/0.022	0.025/0.016	0.039/0.028	0.26	0.92	77/51	69/46
VLDL	0.037/0.024	0.031/0.020	0.031/0.023	0.26	0.92	75/51	69/46
***ANGPTL* 5**							
BMI	0.029/0.021	0.464/0.407	0.451/0.423	0.5	0.95	67/71	63/67
HDL	0.025/0.022	1.0/0.959	0.772/0.076	0.5	0.95	63/66	61/60
***ANGPTL* 6**							
BMI	0.001/0.001	0.909/0.874	0.794/0.774	0.21	0.78	42/40	33/30

[a]For each phenotype analyzed, individuals with QTVs from the top and bottom 10% were used as a stage 1 sample

[b]Individuals with QTVs in the range of 10%-35% and 65%-90% were used as the replication sample.

Chapter 4

A Flexible Likelihood Framework for Detecting Associations with Secondary Phenotypes in Selected Samples: Applications to Sequence Data

4.1. Background

A flexible likelihood approach MULTI-TRAIT-MAP is presented for detecting associations with multiple phenotypes in selected or randomly ascertained samples. This method can be used to detect both common and rare variant/secondary phenotype associations. MULTI-TRAIT-MAP jointly models multiple phenotypes conditional on the study subjects being ascertained. The sampling mechanisms are incorporated via a prospective likelihood approach. The MULTI-TRAIT-MAP framework is comprehensive and can be used to model multiple continuous or categorical traits. To model traits that are not continuous, a generalized linear model is used. For example, either a probit or logit link function can be applied to model binary traits. In this thesis, the discussion is focused on using the probit link function and the liability threshold model, which can be justified by the polygenic model of complex traits. It has been suggested that the liability of all complex traits can be considered as 'quantitative'[71]. For complex traits that are not measured in quantitative scale, there should exist a continuous underlying liability trait, which is due to the aggregated outcome from multiple causative variants with small effects. In this case, a multivariate liability threshold model is naturally utilized to jointly model multiple phenotypes.

The power of MULTI-TRAIT-MAP for detecting gene/secondary trait associations is examined in different selective study designs. Three study designs are considered, i.e. case-control, extreme-trait and multiple-trait. It is assumed for each of the study designs that the same continuous secondary phenotype T is measured. For comparison purposes,

study designs are also evaluated where the quantitative trait T is selected and analyzed as the primary phenotype. Simulation details for each study design can be found in (Table 8).

It is very beneficial to be able to utilize and combine selected samples from existing sequencing based genetic studies. Through extensive simulation studies, it is shown that the case-control and extreme-trait designs can be more powerful for detecting associations with secondary phenotypes than using a population based design, where individuals are randomly selected regardless of their phenotypes. The power for detecting associations with secondary phenotypes strongly depends on the aggregation of causative variants in the sample. For study designs that facilitate the enrichment of causative variants, power will be increased. In the presence of gene pleiotropy, variants that are associated with both the primary and secondary traits can be enriched through selections on the primary phenotype. When the gene region is only associated with the secondary phenotype, if the primary and secondary traits are correlated, selections on the primary phenotype can also induce selection on the secondary phenotype. In this case, for a sample of equivalent size, the power of rejecting the null hypothesis of no gene/secondary trait association in case-control or extreme-trait studies is still superior or comparable to a population based study.

The power for detecting associations with secondary phenotypes in selected samples is jointly affected by locus phenotypic effects for both the primary and secondary phenotypes, as well as residual correlations. Concordant with observations from previous studies of multiple-trait linkage/association mapping[72-74], it is demonstrated that

power is maximized when the locus induced trait correlations are in the opposite direction of the residual correlations. To further demonstrate the utility of MULTI-TRAIT-MAP in combined analysis, an example is given where samples from a case-control study and a multiple-trait study are jointly analyzed. The power for detecting associations with commonly measured phenotypes can be greatly increased when studies are combined, compared to analyzing each individual study separately.

As an application of MULTI-TRAIT-MAP, we analyzed the sequence data from the *ANGPTL*3, *ANGPTL4, ANGPTL5* and *ANGPTL6* genes generated by the Dallas Heart Study (DHS). The 3551 study participants of the DHS were phenotyped for multiple metabolism related traits, including body mass index (BMI), diastolic blood pressure (DiasBP), systolic blood pressure (SysBP), total cholesterol level (TCL), low density lipoprotein (LDL), high density lipoprotein (HDL), triglyceride (TG), and glucose (Gluc). Two primary trait analyses were first performed: 1.) analyzing all samples and 2.) analyzing selected samples whose quantitative trait values fall within the lower and upper quartiles. Next a secondary phenotype analysis was performed where within each selected sample, all other available phenotypes were analyzed as secondary traits. The results from the secondary trait analyses confirmed the primary trait analyses. These analyses established the importance of analyzing secondary phenotypes and the effectiveness of MULTI-TRAIT-MAP. They provided solid support to our simulation experiment.

4.2. Materials and Methods:

It is assumed that there are S variant nucleotide sites for a gene locus. The multi-site genotype for individual i is given by $\vec{X}_i = \left(x_i^1, x_i^2, \cdots, x_i^S\right)$, where the genotype at segregating nucleotide site s is coded by the number of minor alleles, (e.g. $x_i^s = 2$ if the individual is homozygous for the minor allele). To detect associations with rare variants, multiple rare variants in a gene locus are usually jointly analyzed[1,21,22,75,76]. The locus genotype coding for an individual i is defined as $X_i = C\left(\vec{X}_i\right)$, where $C(\bullet)$ is the coding function.

4.2.1. Locus Multi-site Genotype Coding Schemes

Many statistical methods have been developed for association studies of complex traits due to rare variants. Existing methods include combined multivariate and collapsing (CMC)[20], the test of an aggregated number of rare variants (ANRV) [76], weighted sum statistics (WSS) [21], variable threshold test (VT)[75], kernel based adaptive cluster (KBAC)[22], the data adaptive sum test (DAST)[77], C-alpha test[78] and the RARECOVER(RC) method[79], etc. Most of these methods are essentially based upon weighting or grouping variants. Among them, CMC and ANRV are regression based methods which can be incorporated into MULTI-TRAIT-MAP through the coding function $C(\bullet)$:

1.) CMC: the coding function is defined as $X_i = C^{CMC}\left(\vec{X}_i\right) = \delta\left(\sum_{s \in RV} x_i^s > 0\right)$, where $\delta(\bullet)$ is an indicator function and $\sum_{s \in RV}$ is a summation over the set of rare variant nucleotide sites RV, which can be determined by a pre-specified frequency cutoff.

2.) ANRV: the coding function belongs to a more general class of weighted sum coding (WSC), which can be defined as $X_i = C^{WSC}\left(\vec{X}_i\right) = \sum_{s \in RV} w^s x_i^s$. In the weighted sum coding scheme, the variant from nucleotide site s is assigned weight w^s. The ANRV coding assigns equal weight for all variants, i.e. $X_i = C^{ANRV}\left(\vec{X}_i\right) = \sum_{s \in RV} x_i^s$.

4.2.2. A General Probability Model for Multiple Phenotypes in Selected Samples

In order to incorporate the sample ascertainment mechanism and correct for the bias induced by phenotypic residual correlations, multiple phenotypes are jointly modeled. The primary and secondary traits are assumed to follow a multivariate generalized linear model:

$$\begin{cases} F_{Y_1}\left(\vec{\theta}_{Y_1}\right) = \beta_{01} + \beta_{11} X_i + \sum_k \alpha_{k1} W_{ki} \\ F_{Y_2}\left(\vec{\theta}_{Y_2}\right) = \beta_{02} + \beta_{12} X_i + \beta_{Y_1} Y_{1i} + \sum_k \alpha_{k2} W_{ki} \end{cases} \quad (1)$$

$F_{Y_1}(\vec{\theta}_{Y_1})$, $F_{Y_2}(\vec{\theta}_{Y_2})$ are link functions and $\vec{\theta}_{Y_1}$ and $\vec{\theta}_{Y_2}$ are the model parameters related to the primary and secondary traits. This multivariate generalized linear model can be used with any type of link functions, such as probit link function or logit link function.

For selected samples, a conditional likelihood is used, which is similar to Pearson-Aitken correction for ascertainment[80]:

$$L(\beta,\theta;X,\mathrm{Y})=\prod_{i=1}^{N}\Pr(Y_{1i},Y_{2i}|Z_i=1,X_i,\{W_{ki}\}_k) \qquad (2)$$

Z_i is an indicator of individual i being sampled, and N is the number of individuals in the sample. Each term $\Pr(Y_{1i},Y_{2i},|Z_i=1,X_i,\{W_{ki}\}_k)$ in (2) satisfies

$$\Pr(Y_{1i},Y_{2i}|X_i,Z_i=1,\{W_{ki}\}_k)=\frac{\Pr(Z_i=1|Y_{1i},Y_{2i},X_i,\{W_{ki}\}_k)\Pr(Y_{1i},Y_{2i}|X_i,\{W_{ki}\}_k)}{\int\Pr(Z_i=1|y_{1i},y_{2i})\Pr(y_{1i},y_{2i}|X_i)dy_{1i}dy_{2i}} \qquad (3)$$

The sampling mechanism is characterized by $\Pr(Z_i=1|Y_{1i},Y_{2i},X_i,\{W_{ki}\}_k)$, which can be explicitly calculated for case-control, extreme-trait and multiple-trait studies. When the probit link function is used to model binary phenotypes, the multivariate generalized linear model can be simplified.

4.2.3. Association Testing

The likelihood based score statistic can be applied to detect associations with rare variants. Using collapsing coding, p-values for the score statistics can be analytically evaluated. For the weighted sum coding, if the weights are only dependent on the multi-site genotypes, the score statistic will asymptotically follow a normal distribution and the p-values can also be analytically evaluated. Permutation procedures cannot be used to analyze secondary phenotypes in selected samples. This is because if the gene region is associated with the primary phenotype, study subjects are not interchangeable under the null hypothesis of no gene/secondary phenotype associations. The analyses were performed using the CMC coding i.e. $X_i = C^{CMC}\left(\vec{X}_i\right)$. The results remain the same when other coding schemes are used.

4.2.4. Combining Different Cohorts for the Analyses of Secondary Phenotypes

Statistical theories for combining multiple studies are well developed[81]. Since heterogeneity may exist between different cohorts, meta-analysis methods that combine test statistics should be used [29,30]. For rare variant analysis, multiple rare variants are jointly analyzed, and their phenotypic effects are not usually estimated and reported. Therefore, all the joint analyses in this study were carried-out by combining score statistics from different studies. In the joint analysis, score statistics from different studies are weighted and summed. The weight assigned for each score statistic is proportional to the square root of the sample size according to Skol et al [82].

4.2.5. Generation of Genetic and Phenotypic Data

Following Boyko et al. [33], a rigorous population genetic model incorporating demographic change and purifying selections was used to simulate the African variant data. To generate phenotypes, we assume that the phenotypic effects for causative variants are independent of their fitness. In a case-control study, the augmented phenotype $\left(A_i^*, T_i\right)$ for an individual i with multi-site genotype $\vec{X}_i = \left(x_i^1, x_i^2, \cdots, x_i^S\right)$ follows a bivariate normal distribution $\text{MVN}\left(\vec{\mu}_i^{CC}, \Sigma^{CC}\right)$, with

$$\vec{\mu}_i^{CC} = \left(\widetilde{\beta}_{A^*} \sum_{s \in CV_{A^*}} x_i^s, \widetilde{\beta}_T \sum_{s \in CV_T} x_i^s\right), \text{ and } \Sigma^{CC} = \begin{pmatrix} \sigma_{A^*}^2 & \rho_{A^*,T}\sigma_{A^*}\sigma_T \\ \rho_{A^*,T}\sigma_{A^*}\sigma_T & \sigma_T^2 \end{pmatrix} \quad (4)$$

The rare variants sites CV_{A^*} and CV_T are randomly chosen to be causative for the traits A^* and T. Either set can be empty if the gene is not associated with the corresponding trait. Variants at sites $CV_{A^*} \cap CV_T$ are pleiotropic and affect both phenotypes. The binary disorder status A_i is determined by $A_i = \delta\left(A_i^* > a^C\right)$. For each scenario, 1,000 individuals were simulated.

In order to evaluate type I errors, phenotype data was generated under the null hypothesis of no gene/secondary trait T associations, i.e. $\widetilde{\beta}_T = 0$. Scenarios were considered where 1.) the gene region is neither associated with the primary nor the secondary phenotype and 2.) the gene is associated with the primary phenotype, but not with the secondary

phenotype. Scenarios with a combination of two causative variant primary trait effects $\tilde{\beta}_{A^*} = 0.5\sigma_{A^*}$, 0 $\left(\text{or } \tilde{\beta}_B, \tilde{\beta}_{C^*}\right)$, and four residual correlations $\rho_{A^*,T} = \pm 0.3, \pm 0.6$ $\left(\text{or } \rho_{B,T}, \rho_{C^*,T}\right)$ were evaluated.

To compare the power of rejecting the null hypothesis of no gene/secondary trait associations, two causal variant secondary phenotype effects $\tilde{\beta}_T = \pm 0.5\sigma_T$ were employed. The power for the three study designs was compared under scenarios with different combinations of genetic parameter values.

4.2.6. Software Availability

An R-package implementing MULTI-TRAIT-MAP will be available at http://www.bcm.edu/genetics/leal/software

4.3. Results

Evaluation of Type I Errors

Type I errors for each study design using MULTI-TRAIT-MAP were evaluated empirically. Under the null hypothesis of no genetic/secondary phenotype associations, the quantile-quantile (Q-Q) plots of the empirical and theoretical distributions of p-values are displayed in (**Figure 2 and 3**) for the case-control study design. When the

ascertainment mechanism is correctly specified, the type I errors are controlled. Results are shown in (**Figure 2**) for the scenario where the gene region is not associated with either the primary or the secondary phenotypes and the scenario where the gene region is only associated with the primary trait. Type I errors for the extreme-trait and multiple-trait designs were also well controlled (data not shown). The impact of mis-specified sampling mechanisms was investigated. The results are shown in (**Figure 3**) when the prevalence parameter is 10%, but is incorrectly set to be 7% (**Figure 3a**) or 13% (**Figure 3b**) in the analyses. The results indicate that mis-specifying prevalence has only a very minimal impact on type I error rates as can be observed in the Q-Q plot.

4.3.1. Power of Detecting Secondary Phenotype Rare Variant Associations

The efficiency of the three selective sampling designs for detecting secondary trait associations was compared when both the primary and the secondary traits are associated with the same gene (Table 9). Scenarios were examined where 1,000 individuals are sequenced for each study design. There is considerable power for detecting secondary phenotype associations in selected samples. Analyzing secondary phenotypes in a case-control or an extreme-trait study dataset can be consistently more powerful than a randomly ascertained population dataset of equal size.

When a population based sample is used where 1,000 individuals are randomly selected regardless of their phenotypic values, the power for rejecting the null hypothesis

is only 51.7% . For a case-control sample where the secondary trait T is analyzed, the power can be higher (Table 9). For example, when the primary and secondary trait phenotypic effects and residual correlation satisfy $\tilde{\beta}_{A^*} = 0.5\sigma_{A^*}$, $\tilde{\beta}_T = 0.5\,\sigma_T$ and $\rho_{A^*,T} = -0.6$, the power is 56.5%. It is also comparable to the power (56.6%) when 200 individuals with the most extreme trait T values from a cohort of 5,000 are sequenced.

Compatible with observations from bivariate phenotype association studies [73], the power for detecting associations with secondary phenotypes is jointly determined by the sizes and directions of the locus phenotypic effects and residual correlations. The power is the highest when the correlation between the locus phenotypic effects is in the opposite direction of the trait residual correlations. For example, when the locus induced correlation is positive (i.e. $\tilde{\beta}_{A^*} = 0.5\sigma_{A^*}$ and $\tilde{\beta}_T = 0.5\,\sigma_T$), and the trait residual correlation is negative (i.e. $\rho_{A^*,T} = -0.3$), the power is 55.7%. However, if the trait residual correlation is also positive (i.e. $\rho_{A^*,T} = 0.3$), the power is 53.5% (Table 9).

Similar patterns of power comparisons are observed for detecting associations with secondary phenotypes T in extreme-trait studies. The power for an extreme-trait study can be substantially higher than that for a population based study of equivalent size. For example, if the primary and secondary trait effects and residual correlations are given by $\tilde{\beta}_{C^*} = 0.5\sigma_{C^*}$ $\tilde{\beta}_T = 0.5\,\sigma_T$ and $\rho_{C^*,T} = -0.6$, the power of rejecting the null hypothesis is 66.7% (Table 9). It is comparable to the power (70.6%) when 600

individuals with the most extreme trait T values from a cohort of 5,000 are sequenced, or the power (66.6%) when 2,000 randomly selected samples are sequenced.

When the gene region is only associated with the secondary trait T, using samples ascertained on the primary phenotype will induce selections on the secondary phenotype. For a dataset of equivalent size, the power for rejecting the null hypothesis of no gene/secondary trait associations in case-control or extreme-trait samples is still greater than (or comparable to) analyzing the same trait using a randomly ascertained population sample. For example, in an extreme-trait study which sequences 1000 individuals, when causal variants in the gene affect the secondary trait with effect $\tilde{\beta}_T = 0.5\sigma_T$ and the wtwo traits are positively correlated with correlation coefficient $\rho = 0.6$, the power is 60.2%. If the two traits are negatively correlated with $\rho = -0.6$, the power is 60.6% (Table 9). The power in these two scenarios are both superior to that of a population based study (51.6%) which sequences an equivalent number of samples.

The MULTI-TRAIT-MAP method can be applied to analyze samples ascertained on multiple phenotypes. In this example of a multiple-trait study, 500 affected individuals with trait T value above the 65th percentile are sequenced and 500 unaffected individuals are also selected regardless of their trait T values (Table 9). Compared to the extreme-trait or case-control study design, the multiple-trait study example that is given is not as powerful. This is because there is not enough phenotypic variability in the sample, since affected individuals are only sampled from the sub-population with trait T above the 65th percentile. However, in some scenarios, there can be considerable power in a multiple-

trait study, in particular when sampling on the secondary trait T increases phenotypic variability, e.g. affected or unaffected individuals are selected to have secondary T trait values from opposite extreme tails.

MULTI-TRAIT-MAP allows joint analysis of commonly measured phenotypes in different genetic studies. These studies may be targeted at different primary traits. An example is given where a multiple-trait study is implemented, and the association analysis of the secondary trait T is performed by combining a case-control study dataset (Table 10). A wide variety of scenarios were extensively evaluated, and a sizable power increase for the combined analysis is consistently observed.

4.3.2. Applications to the *ANGPTL* Family of Genes

When each of the eight phenotypes from the DHS was analyzed as primary phenotype using selected samples and the entire sample, four nominally significant associations were found for both types of analyses, i.e. *ANGPTL4* with TG (p=0.005), *ANGPTL5* with BMI (p=0.003) *ANGPTL5* with HDL (p=0.024), and *ANGPTL6* with BMI (p=0.022). All of the above significant associations were also successfully detected when TG, BMI and HDL were analyzed as secondary phenotypes. An additional association between *ANGPTL4* and HDL (p=0.018) was identified only when the entire sample was analyzed.

The association between TG and rare variants in the *ANGPTL4* gene was identified using selected samples where the primary traits are BMI (p=0.025), SysBP (p=0.012), or LDL (p=0.010) (Table 10). These traits are only weakly positively correlated with TG, i.e. $\rho_{BMI,TG} = 0.227$, $\rho_{LDL,TG} = 0.197$ and $\rho_{SysBP,TG} = 0.102$. The association between *ANGPTL4* and TG is not significant using samples with extreme DiasBP (p=0.137), TCL (p=0.065), Gluc (p=0.117) and HDL levels (p=0.107).

Although the *ANGPTL4* gene is significantly associated with HDL and the size of the correlation between HDL and TG is larger ($\rho_{HDL,TG} = -0.374$), the association of TG with *ANGPTL4* gene is not significant when TG is analyzed as a secondary trait using samples with extreme HDL levels. This could have occurred because the locus phenotypic effects for HDL and TG are negatively correlated, and the locus induced correlation lies in the same direction as the residual correlation, which is shown in our simulations to have reduced power compared to when the locus induced correlation and trait residual correlations are in opposite directions.

There is one nominally significant association that was only detected in secondary phenotype analyses, i.e. the association between Gluc and rare variants in the *ANGPTL3* gene (p=0.024). It was identified when samples with extreme LDL levels were used. But when Gluc was analyzed as primary trait, the association is not significant (p=0.64). This could either be a novel association or a false positive finding.

4.4. Discussion

In this part of the thesis, a flexible likelihood framework MULTI-TRAIT-MAP is proposed for jointly modeling multiple phenotypes in non-randomly ascertained samples, e.g. case-control samples or extreme-trait samples. By coupling multivariate generalized linear models with prospective likelihood, complicated ascertainment mechanisms can be incorporated. The approach is flexible and particularly suitable for analyzing complex traits. It can be applied to any study with known sampling mechanisms. MULTI-TRAIT-MAP allows efficient statistical inference for the genetic parameters of interest. Although the discussion in this section of the thesis is focused on analyzing sequence data, MULTI-TRAIT-MAP can also be applied to analyze genotype data.

The results presented in this section of the thesis have important implications for the design and analysis of complex traits. Most current studies, due to their limited sample size, are not adequately powered to detect associations with rare variants. It has been suggested that for an exome study, ~10,000 individuals with extreme traits from a cohort of 100,000 need to be sequenced in order to have adequate power[24]. However, the sample size well exceeds the capacity of many existing studies[24]. It is therefore particularly important that combined analysis can be performed using data from multiple studies in order to have sufficient power. Applying MULTI-TRAIT-MAP, sequencing studies that are targeted at different primary traits can be jointly analyzed for detecting associations with a variety of commonly measured secondary traits.

The power of different selective study designs was investigated. It was shown through extensive simulations that there is considerable power for detecting secondary phenotype associations in selected samples. In particular, when the secondary trait of interest is analyzed in a case-control or an extreme-trait study dataset, the power can be greater than analyzing an equivalent sized randomly ascertained sample. Utilizing data sharing platforms and protocols such as dbGaP[83], samples from existing studies can be freely obtained and analyzed. The power can be greatly increased when data from multiple studies are jointly analyzed.

Secondary phenotypes not only have their own clinical importance, but they can also be relevant for understanding the primary trait etiologies. For example, in the study of T2D, a number of studies are targeted at related quantitative traits including fasting glucose levels [54], and C-reactive protein [55]. Given that these traits are often available for individuals who participate in T2D case-control studies[84], MULTI-TRAIT-MAP can be applied to detect associations with these additional phenotypes.

MULTI-TRAIT-MAP was also applied to the analysis of sequence data from the Dallas Heart Study. Multiple associations were identified, which confirmed previous data analyses. When the traits were analyzed as secondary phenotypes, although these same set of associations was observed, they were not detected in every selected sample, e.g. the association between TG levels and *ANGTPL4* was only detected in secondary trait analyses using samples with extreme BMI, SysBP, and LDL, but not in samples with

extreme DiasBP, HDL, TCL and Gluc. This could be affected by the small sample sizes that were analyzed, the moderate effect sizes for variants involved in complex trait etiologies or the directions and magnitudes of the correlations between the primary and secondary phenotypes. Although these identified associations are only nominally significant, they all have biological support. In fact, the effects of mutant *ANGPTL* family genes on lipoprotein lipase (LPL) have been investigated through *in vitro* functional studies and *in vivo* mice studies. LPL has been known to affect glucose metabolism [43], cholesterol level [44], and blood pressure [47]. Additionally, the association between variants in the *ANGPTL*4 gene and triglyceride levels has been successfully replicated[5,6].

Sensitivity of MULTI-TRAIT-MAP to mis-specified sampling mechanisms was extensively evaluated. When the disease prevalence is reported as an interval of possible values, inferences from MULTI-TRAIT-MAP can be carried-out under different prevalence values from the interval. The results can be integrated using a model averaging procedure. It has been shown that it is an effective approach to further reduce the impact of mis-specified prevalence [85].

There can be heterogeneities of sequence coverage depth within and between different studies. Coverage depth differences within a single study may cause inflated type I errors. Possible strategies to reduce the bias include incorporating the mean coverage depth of each individual in the analysis as a covariate[86]. The method can be used with the MULTI-TRAIT-MAP model. In order to be robust against between-study heterogeneities,

meta-analyses procedure should be implemented for the joint analysis, instead of performing mega-analysis that combines individual participant data[29,30].

When multiple phenotypes are analyzed, to avoid inflated type I error due to testing multiple hypotheses, a stringent significance level must be specified. Due to phenotypic correlations, Bonferroni corrections for testing multiple genes and phenotypes can be overly conservative. Instead, the spectral decomposition based method in Nyholt et al[87] can be used. In addition to correctly controlling for family-wise error rates, it is important that the findings can be replicated using independent samples[23].

With large scale implementation of sequence based genetic association studies, the capability for mapping complex traits will be further elevated. Detecting associations with rare variants and jointly investigating multiple phenotypes together can be an ambitious and difficult task given the moderate sample sizes of existing studies. Taking advantage of multiple studies and mapping commonly measured phenotypes using MULTI-TRAIT-MAP is therefore highly beneficial and will greatly accelerate the process of dissecting complex trait genetic etiologies.

Table 8: Definitions of Selection Mechanisms

Study designs	Definition
Case-control	Cases and controls are sampled based upon the binary primary phenotype A. The trait status is determined by $A = \delta(A^* \geq a^C)$, where a^C is the 90th percentile of the liability trait A^*. A total of 500 cases and 500 controls are sequenced.
Extreme Trait	One thousand individuals with quantitative trait B values in the upper and lower 10% were sequenced from a cohort of 5,000 individuals.
Multiple-trait	The affection status is defined by $C = \delta(C^* \geq c^C)$, where c^C is the 90th percentile for the liability trait C^*. Five hundred affected individuals with trait T values >65th percentile are sequenced and 500 unaffected individuals are also sequenced regardless of their T values.
Extreme-trait study where T is sampled and analyzed as primary trait	In an extreme-trait study, individuals with extreme T values in the upper and lower 2%, 6% and 10% are sampled and sequenced from a cohort of 5,000 individuals.
Population Based Study Design	One thousand, 2,000 and 3,000 individuals are randomly sampled from the general population regardless of their phenotypes.

Table 9 Power to detect secondary trait *T* associations using case-control, extreme-trait, and multiple-trait study design.

Genetic Parameters			Power[d]
$\tilde{\beta}_{A^*}(\tilde{\beta}_B, \tilde{\beta}_{C^*})$	$\tilde{\beta}_T$ [b]	$\rho_{A^*,T}(\rho_{B,T}, \rho_{C^*,T})$	CC[e]/ET[f]/MT[g]
0.5	-0.5	-0.3	0.536/0.562/0.316
0.5	-0.5	0.3	0.548/0.605/0.418
0.5	0.5	-0.3	0.557/0.582/0.448
0.5	0.5	0.3	0.535/0.557/0.506
0.5	-0.5	-0.6	0.533/0.582/0.292
0.5	-0.5	0.6	0.556/0.654/0.471
0.5	0.5	-0.6	0.565/0.667/0.391
0.5	0.5	0.6	0.545/0.589/0.562
0	-0.5	-0.3	0.510/0.555/0.325
0	-0.5	0.3	0.499/0.557/0.412
0	0.5	-0.3	0.508/0.544/0.414
0	0.5	0.3	0.517/0.555/0.497
0	-0.5	-0.6	0.527/0.598/0.315
0	-0.5	0.6	0.513/0.609/0.447
0	0.5	-0.6	0.521/0.606/0.373
0	0.5	0.6	0.531/0.602/0.549

[a]Causal variant phenotypic effect for liability trait A^*, trait B and liability trait C^*.

[b]Causal variant effect for secondary trait T

[c]Residual correlation between the primary (liability) trait and secondary trait T

[d]Power was empirically estimated using 5,000 replicates under a significance level $\alpha = 0.05$

[e]Power for case-control study. A case-control study sample consists of 500 cases and 500 controls.

[f]Power for extreme-trait study. An extreme-trait study sample consists of 1,000 individuals with extreme trait B values selected from a cohort of 5,000.

[g]Power for multiple-trait study. A multiple-trait study sample is obtained based upon both trait C and trait T. The affection status is determined by C. Five hundred affected individuals with T values > 65th percentile as well as 500 unaffected individuals are sequenced.

Table 10: Power to detect secondary trait T associations for individual studies (case-control and multiple-trait) and the combined analysis.

Parameters						Power[g]		
	$\widetilde{\beta}_T^{CC}$ [b]	$\rho_{A^*,T}$ [c]	$\widetilde{\beta}_{C^*}$ [d]	$\widetilde{\beta}_T^{MT}$ [e]	$\rho_{C^*,T}$ [f]	Case-control[h]	Multiple –Trait[i]	Combined-Analysis
0	-0.5	0.3	0.5	-0.5	0.3	0.510	0.418	0.690
0	-0.5	03	0.5	-0.5	0.3	0.499	0.418	0.680
0	0.5	-0.3	0.5	0.5	0.3	0.508	0.526	0.726
0	0.5	0.3	0.5	0.5	0.3	0.517	0.526	0.732
0	-0.5	-0.6	0.5	-0.5	0.3	0.527	0.418	0.703
0	-0.5	0.6	0.5	-0.5	0.3	0.513	0.418	0.685
0	0.5	-0.6	0.5	0.5	0.3	0.521	0.526	0.731
0	0.5	0.6	0.5	0.5	0.3	0.531	0.526	0.741

[a]Causal variant phenotypic effect for liability trait A^* in the case-control study sample

[b]Causal variant phenotypic effect for trait T in the case-control study sample

[c] Residual correlations between liability trait A^* and trait T in the case-control study sample

[d]Causal variant phenotypic effect for liability trait C^* in the multiple-trait study sample

[e] Causal variant phenotypic effect for trait T in the multiple-trait study sample

[f] Residual correlations between liability trait C^* and trait T in the multiple-trait study sample

[g] Power was empirically estimated using 5,000 replicates under a significance level $\alpha = 0.05$

[h] Case-control sample consists of 500 cases and 500 controls.

[i]A multiple-trait dataset is obtained based upon both trait C and trait T .The affection status is determined by C . Five hundred affected individuals with trait T values >65th percentile are sequenced, as well as 500 unaffected individuals.

Table 11: Results for the secondary phenotype analyses using sequence data from the *ANGPTL3, ANGPTL4, ANGPTL5* and *ANGPTL6* genes.

Primary Phenotype	P-values for Analyzing Secondary Phenotypes[a]							
	BMI	DiasBP	SysBP	TCL	LDL	HDL	TG	Gluc
	ANGPTL 3							
BMI	-	0.649	0.766	0.429	0.681	0.717	0.121	0.114
DiasBP	0.941	-	0.889	0.580	0.745	0.309	0.441	0.398
SysBP	0.550	0.509	-	0.371	0.223	0.689	0.073	0.222
TCL	0.988	0.955	0.327	-	0.971	0.289	0.163	0.151
LDL	0.871	0.372	0.349	0.114	-	0.116	0.183	0.024*
HDL	0.945	0.616	0.312	0.825	0.668	-	0.561	0.639
TG	0.910	0.883	0.437	0.945	0.418	0.863	-	0.148
Gluc	0.652	0.208	0.351	0.982	0.475	0.692	0.335	-
	ANGPTL 4							
BMI	-	0.292	0.268	0.733	0.440	0.497	0.025*	0.972
DiasBP	0.965	-	0.380	0.361	0.363	0.121	0.137	0.389
SysBP	0.993	0.551	-	0.728	0.754	0.099	0.012*	0.405
TCL	0.861	0.532	0.571	-	0.052	0.759	0.065	0.933
LDL	0.281	0.894	0.269	0.135	-	0.053	0.010*	0.999
HDL	0.708	0.904	0.286	0.318	0.262	-	0.107	0.874
TG	0.310	0.364	0.584	0.629	0.326	0.784	-	0.845
Gluc	0.824	0.524	0.084	0.848	0.561	0.479	0.118	-
	ANGPTL 5							
BMI	-	0.920	0.114	0.521	0.233	0.056	0.377	0.797
DiasBP	0.118	-	0.096	0.451	0.803	0.092	0.616	0.367
SysBP	0.203	0.887	-	0.117	0.160	0.304	0.791	0.294
TCL	0.107	0.536	0.923	-	0.399	0.014*	0.221	0.488
LDL	0.084	0.735	0.587	0.202	-	0.002*	0.147	0.458
HDL	0.387	0.866	0.917	0.463	0.991	-	0.569	0.900
TG	0.044*	0.871	0.074	0.296	0.597	0.185	-	0.448
Gluc	0.030*	0.779	0.957	0.546	0.717	0.002*	0.451	-
	ANGPTL 6							
BMI	-	0.300	1.000	0.606	0.457	0.324	0.401	0.419
DiasBP	0.008*	-	0.385	0.459	0.690	0.478	0.721	0.197
SysBP	0.773	0.816	-	0.622	0.853	0.668	0.338	0.490
TCL	0.024*	0.530	0.992	-	0.823	0.324	0.702	0.940
LDL	0.089	0.383	0.850	0.485	-	0.429	0.801	0.314
HDL	0.034*	0.101	0.873	0.800	0.870	-	0.393	0.215
TG	0.210	0.735	0.974	0.357	0.695	0.561	-	0.811
Gluc	0.153	0.402	0.897	0.340	0.531	0.267	0.905	-

[a] For each phenotype, individuals were selected with trait values in the upper and lower quartiles and the remaining seven phenotypes were analyzed as secondary traits using MULTI-TRAIT-MAP.

References

1 Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP: Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 2008;40:592-599.
2 Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hebert S, Doelle H, Ersoy B, Kryukov G, Schmidt S, Yosef N, Ruppin E, Sharan R, Vaisse C, Sunyaev S, Dent R, Cohen J, McPherson R, Pennacchio LA: Medical sequencing at the extremes of human body mass. Am J Hum Genet 2007;80:779-791.
3 Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: Multiple rare alleles contribute to low plasma levels of hdl cholesterol. Science 2004;305:869-872.
4 Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, Hobbs HH: Multiple rare variants in npc1l1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A 2006;103:1810-1815.
5 Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC: Population-based resequencing of angptl4 uncovers variations that reduce triglycerides and increase hdl. Nat Genet 2007;39:513-516.
6 Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC: Rare loss-of-function mutations in angptl family members contribute to plasma triglyceride levels in humans. J Clin Invest 2009;119:70-79.
7 Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI: Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. Am J Hum Genet 2008;82:100-112.
8 Bodmer W, Bonilla C: Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 2008;40:695-701.
9 Ng PC, Henikoff S: Sift: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812-3814.
10 Ramensky V, Bork P, Sunyaev S: Human non-synonymous snps: Server and survey. Nucleic Acids Res 2002;30:3894-3900.
11 Karchin R: Next generation tools for the annotation of human snps. Brief Bioinform 2009;10:35-52.
12 Moore JH, Williams SM: New strategies for identifying gene-gene interactions in hypertension. Ann Med 2002;34:88-95.
13 Zhang Y, Liu JS: Bayesian inference of epistatic interactions in case-control studies. Nat Genet 2007;39:1167-1173.
14 Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001;69:138-147.
15 Culverhouse R, Klein T, Shannon W: Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 2004;27:141-152.

16 Nelson MR, Kardia SL, Ferrell RE, Sing CF: A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res 2001;11:458-470.
17 Stratton MR, Rahman N: The emerging landscape of breast cancer susceptibility. Nat Genet 2008;40:17-22.
18 Fitze G, Appelt H, Konig IR, Gorgens H, Stein U, Walther W, Gossen M, Schreiber M, Ziegler A, Roesner D, Schackert HK: Functional haplotypes of the ret proto-oncogene promoter are associated with hirschsprung disease (hscr). Hum Mol Genet 2003;12:3207-3214.
19 Fitze G, Schierz M, Kuhlisch E, Schreiber M, Ziegler A, Roesner D, Schackert HK: Novel intronic polymorphisms in the ret proto-oncogene and their association with hirschsprung disease. Hum Mutat 2003;22:177.
20 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. Am J Hum Genet 2008;83:311-321.
21 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 2009;5:e1000384.
22 Liu DJ, Leal SM: A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet 2010;6:e1001156.
23 Liu DJ, Leal SM: Replication strategies for rare variant complex trait association studies via next-generation sequencing. Am J Hum Genet 2010;87:790-801.
24 Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR: Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci U S A 2009;106:3871-3876.
25 Cauchi S, Nead KT, Choquet H, Horber F, Potoczna N, Balkau B, Marre M, Charpentier G, Froguel P, Meyre D: The genetic susceptibility to type 2 diabetes may be modulated by obesity status: Implications for association studies. BMC Med Genet 2008;9:45.
26 Cauchi S, Meyre D, Dina C, Choquet H, Samson C, Gallina S, Balkau B, Charpentier G, Pattou F, Stetsyuk V, Scharfmann R, Staels B, Fruhbeck G, Froguel P: Transcription factor tcf7l2 genetic study in the french population: Expression in human beta-cells and adipose tissue and strong association with type 2 diabetes. Diabetes 2006;55:2903-2908.
27 Lin DY, Zeng D: Proper analysis of secondary phenotype data in case-control association studies. Genet Epidemiol 2009;33:256-265.
28 Richardson DB, Rzehak P, Klenk J, Weiland SK: Analyses of case-control data for additional outcomes. Epidemiology 2007;18:441-445.
29 Ioannidis JP, Thomas G, Daly MJ: Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 2009;10:318-329.
30 McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356-369.

31 Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010;11:415-425.
32 Hernandez RD: A flexible forward simulator for populations subject to selection and demography. Bioinformatics 2008;24:2786-2787.
33 Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD: Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 2008;4:e1000083.
34 Gudmundsdottir K, Ashworth A: The roles of brca1 and brca2 and associated proteins in the maintenance of genomic stability. Oncogene 2006;25:5864-5874.
35 Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton D, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles R, Evans DG, Houlston R, Murday V, Narod S, Peretz T, Peto J, Phelan C, Zhang HX, Szabo C, Devilee P, Goldgar D, Futreal PA, Nathanson KL, Weber B, Rahman N, Stratton MR: Low-penetrance susceptibility to breast cancer due to chek2(*)1100delc in noncarriers of brca1 or brca2 mutations. Nat Genet 2002;31:55-59.
36 Victor RG, Haley RW, Willett DL, Peshock RM, Vaeth PC, Leonard D, Basit M, Cooper RS, Iannacchione VG, Visscher WA, Staab JM, Hobbs HH: The dallas heart study: A population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. Am J Cardiol 2004;93:1473-1480.
37 Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG: Recent and ongoing selection in the human genome. Nat Rev Genet 2007;8:857-868.
38 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904-909.
39 Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD: Genes mirror geography within europe. Nature 2008;456:98-101.
40 Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann HE, Ruther A, Schreiber S, Becker C, Nurnberg P, Nelson MR, Krawczak M, Kayser M: Correlation between genetic and geographic structure in europe. Curr Biol 2008;18:1241-1248.
41 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. Genetics 2000;155:945-959.
42 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. Am J Hum Genet 2000;67:170-181.
43 Webster RJ, Warrington NM, Weedon MN, Hattersley AT, McCaskie PA, Beilby JP, Palmer LJ, Frayling TM: The association of common genetic variants in the apoa5, lpl and gck genes with longitudinal changes in metabolic and cardiovascular traits. Diabetologia 2009;52:106-114.

44 Koster A, Chao YB, Mosior M, Ford A, Gonzalez-DeWhitt PA, Hale JE, Li D, Qiu Y, Fraser CC, Yang DD, Heuer JG, Jaskunas SR, Eacho P: Transgenic angiopoietin-like (angptl)4 overexpression and targeted disruption of angptl4 and angptl3: Regulation of triglyceride metabolism. Endocrinology 2005;146:4943-4950.
45 Yagyu H, Lutz EP, Kako Y, Marks S, Hu Y, Choi SY, Bensadoun A, Goldberg IJ: Very low density lipoprotein (vldl) receptor-deficient mice have reduced lipoprotein lipase activity. Possible causes of hypertriglyceridemia and reduced body mass with vldl receptor deficiency. J Biol Chem 2002;277:10037-10043.
46 Nevin DN, Zambon A, Furlong CE, Richter RJ, Humbert R, Hokanson JE, Brunzell JD: Paraoxonase genotypes, lipoprotein lipase activity, and hdl. Arterioscler Thromb Vasc Biol 1996;16:1243-1249.
47 Li B, Ge D, Wang Y, Zhao W, Zhou X, Gu D, Chen R: Lipoprotein lipase gene polymorphisms and blood pressure levels in the northern chinese han population. Hypertens Res 2004;27:373-378.
48 Dudbridge F, Gusnanto A: Estimation of significance thresholds for genomewide association scans. Genet Epidemiol 2008;32:227-234.
49 Chapman J, Whittaker J: Analysis of multiple snps in a candidate gene or region. Genet Epidemiol 2008;32:560-566.
50 Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 2001;69:124-137.
51 Eyre-Walker A, Keightley PD: High genomic deleterious mutation rates in hominids. Nature 1999;397:344-347.
52 Browning JD, Szczepaniak LS, Dobbins R, Nuremberg P, Horton JD, Cohen JC, Grundy SM, Hobbs HH: Prevalence of hepatic steatosis in an urban population in the united states: Impact of ethnicity. Hepatology 2004;40:1387-1395.
53 Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK: High-resolution mapping of expression-qtls yields insight into human gene regulation. PLoS Genet 2008;4:e1000214.
54 Bouatia-Naji N, Rocheleau G, Van Lommel L, Lemaire K, Schuit F, Cavalcanti-Proenca C, Marchand M, Hartikainen AL, Sovio U, De Graeve F, Rung J, Vaxillaire M, Tichet J, Marre M, Balkau B, Weill J, Elliott P, Jarvelin MR, Meyre D, Polychronakos C, Dina C, Sladek R, Froguel P: A polymorphism within the g6pc2 gene is associated with fasting plasma glucose levels. Science 2008;320:1085-1088.
55 Elliott P, Chambers JC, Zhang W, Clarke R, Hopewell JC, Peden JF, Erdmann J, Braund P, Engert JC, Bennett D, Coin L, Ashby D, Tzoulaki I, Brown IJ, Mt-Isa S, McCarthy MI, Peltonen L, Freimer NB, Farrall M, Ruokonen A, Hamsten A, Lim N, Froguel P, Waterworth DM, Vollenweider P, Waeber G, Jarvelin MR, Mooser V, Scott J, Hall AS, Schunkert H, Anand SS, Collins R, Samani NJ, Watkins H, Kooner JS: Genetic loci associated with c-reactive protein levels and risk of coronary heart disease. JAMA 2009;302:37-48.
56 Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ: Target-enrichment strategies for next-generation sequencing. Nat Methods;7:111-118.

57 Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA: Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 2009;10:R32.
58 Ionita-Laza I, Lange C, N ML: Estimating the number of unseen variants in the human genome. Proc Natl Acad Sci U S A 2009;106:5008-5013.
59 Li B, Leal SM: Discovery of rare variants via sequencing: Implications for the design of complex trait association studies. PLoS Genet 2009;5:e1000481.
60 Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ, Frazer KA: Accurate detection and genotyping of snps utilizing population sequencing data. Genome Res
61 Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, Yu F: A snp discovery method to assess variant allele probability from next-generation resequencing data. Genome Res 2009
62 Scheet P, Stephens M: Linkage disequilibrium-based quality control for large-scale genetic studies. PLoS Genet 2008;4:e1000147.
63 Leal SM: Detection of genotyping errors and pseudo-snps via deviations from hardy-weinberg equilibrium. Genet Epidemiol 2005;29:204-214.
64 Douglas JA, Boehnke M, Lange K: A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. Am J Hum Genet 2000;66:1287-1297.
65 Kerem B, Chiba-Falek O, Kerem E: Cystic fibrosis in jews: Frequency and mutation distribution. Genet Test 1997;1:35-39.
66 King MC, Rowell S, Love SM: Inherited breast and ovarian cancer. What are the risks? What are the choices? JAMA 1993;269:1975-1980.
67 Ewing B, Green P: Base-calling of automated sequencer traces using phred. Ii. Error probabilities. Genome Res 1998;8:186-194.
68 Ewing B, Hillier L, Wendl MC, Green P: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8:175-185.
69 Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, Daly MJ, Goldstein DB, John S, Nelson MR, Graham J, Park BK, Dillon JF, Bernal W, Cordell HJ, Pirmohamed M, Aithal GP, Day CP: Hla-b*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nat Genet 2009;41:816-819.
70 Mitsui J, Mizuta I, Toyoda A, Ashida R, Takahashi Y, Goto J, Fukuda Y, Date H, Iwata A, Yamamoto M, Hattori N, Murata M, Toda T, Tsuji S: Mutations for gaucher disease confer high susceptibility to parkinson disease. Arch Neurol 2009;66:571-576.
71 Plomin R, Haworth CM, Davis OS: Common disorders are quantitative traits. Nat Rev Genet 2009;10:872-878.
72 Lange C, Silverman EK, Xu X, Weiss ST, Laird NM: A multivariate family-based association test using generalized estimating equations: Fbat-gee. Biostatistics 2003;4:195-206.

73 Liu J, Pei Y, Papasian CJ, Deng HW: Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet Epidemiol 2009;33:217-227.
74 Allison DB, Thiel B, St Jean P, Elston RC, Infante MC, Schork NJ: Multiple phenotype modeling in gene-mapping studies of quantitative traits: Power advantages. Am J Hum Genet 1998;63:1190-1201.
75 Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR: Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 2010;86:832-838.
76 Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 2009
77 Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered;70:42-54.
78 Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ: Testing for an unusual distribution of rare variants. PLoS Genet 2010;7:e1001322.
79 Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, Bafna V: A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput Biol 2010;6:e1000954.
80 Aitken AC: Notes on selection from a multivariate normal population. Proceedings of the Edinburgh Mathematical Society 1934;4:106-110.
81 Munafo MR, Flint J: Meta-analysis of genetic association studies. Trends Genet 2004;20:439-444.
82 Skol AD, Scott LJ, Abecasis GR, Boehnke M: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 2006;38:209-213.
83 Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M, Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST: The ncbi dbgap database of genotypes and phenotypes. Nat Genet 2007;39:1181-1186.
84 Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P: A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007;445:881-885.
85 Li M, Li C: Assessing departure from hardy-weinberg equilibrium in the presence of disease association. Genet Epidemiol 2008;32:589-599.
86 Garner C: Confounded by sequencing depth in association studies of rare alleles. Genet Epidemiol 2011
87 Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 2004;74:765-769.