QUANTITATIVE METHODS FOR EVALUATING ASSOCIATION BETWEEN MULTIPLE RARE GENETIC VARIANTS AND COMPLEX HUMAN TRAITS

Andrea E. Byrnes

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2013

Approved by:

Ethan M. Lange

Yun Li

Patrick F. Sullivan

Wei Sun

Michael C. Wu

ABSTRACT

Andrea E. Byrnes: Quantitative Methods for Evaluating Association between Multiple Rare Genetic Variants and Complex Human Traits
(Under the direction of Yun Li)

First, we propose two methods for aggregation of rare variants in data from Genome-wide Association Studies (GWAS), a weighted haplotype-based approach and an imputation-based approach, to test for the effect of rare variants with GWAS data. Both methods can incorporate external sequencing data when available. Our methods clearly show enhanced statistical power over existing methods for a wide range of population-attributable risk, percentage of disease-contributing rare variants, and proportion of rare alleles working in different directions. We thus demonstrate that the evaluation of rare variants with GWAS data is possible, particularly when public sequencing data are incorporated.

Second, we present a systematic evaluation of multiple weighting schemes through a series of simulations intended to mimic large sequencing studies of a quantitative trait. We evaluate existing phenotype-independent and phenotype-dependent methods, as well as weights estimated by penalized regression. We find that the difference in power between phenotype-dependent schemes is negligible when high-quality functional annotations are available. When functional annotations are unavailable or incomplete, all methods lose power; however, the variable selection methods outperform the others at a cost of increased computational time. In the absence of highly accurate annotation, we recommend variable selection methods (which can be viewed as "statistical annotation") on top of regions implicated by a phenotype-independent weighting scheme.

Finally, we propose a method to apply the Sequence Kernel Association Test (SKAT), a similarity-based approach for rare variant association, to data from admixed populations by first estimating local ancestry for each variant. In simulations, we find that when the true causal alleles are causal only from only one ancestral population, our proposed approaches show a marked improvement in power over the original SKAT method. In real data, our results support the previously reported European-specific association and illustrate the increased statistical power of the proposed methods to find such associations.

ACKNOWLDGEMENTS

I would like to offer countless thanks to my adviser, Yun Li, for her tireless support and encouragement, and each of my committee members Michael Wu, Wei Sun, Ethan Lange and Patrick Sullivan for their thoughtful contributions to this research. I would like to offer additional thanks to Patrick Sullivan for his mentorship and encouragement over the past seven years. I would also like to thank my collaborators from UNC-Chapel Hill and other organizations for their invaluable comments and suggestions: Mingyao Li from the University of Pennsylvania School of Medicine, Fred Wright, now at the NC State University Center for Bioinformatics; our data collaborators at GlaxoSmithKline, especially Margaret G. Ehm, Matthew R. Nelson, Li Li, and Liling Warren for sharing the targeted sequencing data; our CoLaus collaborators for providing the phenotypic data used in chapters 3 and 4; and Leslie Lange for providing the real data used in the illustrative real data example in chapter 5. Finally, I would like to thank my co-workers, friends and family for their patience and support of this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABCA1	ATP-Binding Cassette Transporter 1
APOA1	High Purity Apolipoprotein A1
APOB	High Purity Apolipoprotein B
ATOM	Association Test by combining Optimally Weighted Markers
BMI	Body Mass Index
CAST	Cohort Allelic Sum Test
CEPH	Centre de'Etude du Polymorphism Humain
CEU	Caucasian European from Utah
CNV	Copy Number Variant
EN	Elastic Net
EREC	Estimated Regression Coefficients
GRR	Genotype Relative Risk
GWAS	Genome-wide Association Study
HDL	High Density Lipoprotein
HG	Haplotype Grouping
IFIH1	Interferon Induced with Helicase 1
Kb	Kilo-base
LCAT	Lecithin—Cholesterol Acyltransferase
LD	Linkage Disequilibrium
LDL	Low Density Lipoprotein
MaCH	Markov Chain Haplotyper

MAF	Minor Allele Frequency
Mb	Mega-base
PAR	Population Attributable Risk
RVC	Rare Variant Collapsing
SKAT	SNP-Set (Sequence) Kernel Association Test
SNP	Single Nucleotide Polymorphism
T1D	Type 1 Diabetes
VT	Variable Threshold
WDS	Weighted Dosage Test
WHG	Weighted Haplotype Test for Genotyped SNPs
WHS	Weighted Haplotype Test
WS	Weighted Sum
WTCCC	Wellcome Trust Case Control Consortium

CHAPTER 1: MOTIVATION AND BIOLOGICAL JUSTIFICATION

In this document, we will discuss statistical methods for assessing association between sets of rare genetic sequence variations and complex human traits. This section provides an overview of the biological problems we are interested in and the some of the statistical strategies employed in an attempt to solve them.

To begin, DNA is a double-stranded molecule consistent of four nucleic acid components: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). DNA is found in the nucleus of the vast majority of plant and animal cells and has been compared to a blueprint for the organism in which it is found. Humans have 22 autosomes, in addition to the sex chromosomes X and Y and mitochondrial DNA, accounting for over 5 billion base pairs in total. We will consider primarily autosomal DNA, for which each individual possesses two copies, one inherited maternally and the other paternally. Over 99% of the DNA sequence is the same across humans (Ohno, 1972), however there are a large number of ways in which human DNA sequence can differ from one another in a single region including microsatellites, copy number variations (CNVs), insertions, deletions, inversions and single nucleotide polymorphisms (SNPs). Any one of these can be called a genetic variant, meaning that it contains a sequence of nucleic acids that is different from the consensus sequence or from what is most common.

A single nucleotide polymorphism is one such genetic variant that occupies only one base pair. As previously stated, much of the genome is shared across humans, however some of these variants, SNPs included, are quite common with variant or minor allele frequency (abbreviated MAF) near 0.5. Because of this, a great deal of genetic variation can be measured by a relatively small subset of these SNPs. In the 1990'smicroarray technologies from companies like Affymetrix and Illumina began to capitalize on these common SNPs in the form of genome-wide SNP platforms. Today these technologies can accurately assess up as many as 1 million pre-selected SNPs [e.g. the Affy Axiom or Illumina 1M], however these technologies are limited in that they cannot discover new variants. Though rare variants outnumber common ones (1000 Genomes Consortium et al., 2010; Mathieson & McVean, 2012), rare variants are seldom included in GWAS panels since they contain little information in relatively small sample sizes. This stands to reason since a marker that is not polymorphic or barely polymorphic in a sample provides little or no statistical power to detect association between the single SNP and the outcome of interest.

GWAS studies yielded promising genetic loci in association with many complex human traits such as coronary artery disease, diabetes, bipolar disorder, rheumatoid arthritis, obesity and height (Frayling et al., 2007; Weedon et al., 2008; WTCCC, 2007), but ultimately could not explain the extent to which these traits appear to be inherited. This phenomenon came to be known as "missing heritability" (Maher, 2008; Manolio et al., 2009). Some investigators proposed that the apparent missing heritability was due to rare or even private mutations and that association testing for these types of variants was not possible without whole sequence data on thousands of individuals, if at all. These authors advocated abandoning large scale GWAS for complex traits (Goldstein, 2011). Others argued that the missing heritability could also be due to several relatively common variants of modest effect size and that current GWAS sample sizes were not large enough to elucidate these associations. (Sullivan, 2012) The GWAS study design remains popular, partially due to the emergence of genome imputation for variants not typed. Though rare variants are not typed by the GWAS chips themselves, imputation methods use information from an outside panel such as HapMap or 1000 Genomes to predict the genotypes of markers not in the GWAS panel, including some rare variants, via Markov Chain Monte Carlo (Y. Li, Willer, Ding, Scheet, & Abecasis, 2010; Y. Li, Willer, Sanna, & Abecasis, 2009; Marchini & Howie, 2010).

From 2009 to the present, "next generation" sequencing technologies have brought the cost of whole exome and whole genome sequencing down to hundreds of dollars per sample. These rapid advances in technology allowed investigators to collect larger samples (hundreds or thousands of individuals) that evaluate every base of the genome (or exome), which was previously unimaginable due to cost. Investigators can now evaluate thousands of known rare variants and discover previously unknown variants with these technologies. However, for rare variants, evaluating each variant independently as in GWAS analysis requires a sample size far greater than even these studies can provide. Further if, a variant is unique to one individual, or "private," no study design will discover this causal relationship if each variant is evaluated one at a time. Because of this, the idea of combining information from multiple variants across a genomic region, gene or pathway became increasingly popular.

Previously, we have discussed the MAF as a fixed quantity; however, data from Hapmap (International Hapmap Consortium, 2005) and 1000 Genomes (1000 Genomes Consortium et al., 2010) demonstrate that MAF can vary greatly across populations. In fact, the existence of a particular variant can be population-specific. When dealing with data from a single ancestral population, this does not change the conclusions greatly; however, when two or more populations are present in the study sample these differences can quickly lead to incorrect results. In particular, populations in which more than one ancestral population's genetic contribution is present in most or all of the individuals, e.g. African American or Hispanic populations, can be especially complicated. Such populations are known as admixed populations and many methods have emerged to adjust for the complexity they bring to genetic studies.

This compilation of projects attempts to survey and compare the methodology of several existing methods for the aggregation of rare variants across a genomic region (e.g. gene, exon, pathway). We also aim to improve upon some of these methods and adapt them for use in admixed populations. The next section deals with the history rare-variant collapsing methods and their evolution from simple counting methods, to more complex systems of weighting that utilize phenotype information, and finally, to similarity-based methods which use more complex statistical techniques to assess genomic similarity between individuals and their trait of interest.

The third chapter proposes two methods, a weighted haplotype-based approach and an imputation-based approach, to test for the effect of rare variants with GWAS data. Both methods can incorporate external sequencing data when available. We evaluated our methods and compared them with methods proposed in the sequencing setting through extensive simulations. Our methods clearly show enhanced statistical power over existing methods for a wide range of population-attributable risk, percentage of disease-contributing rare variants, and proportion of rare alleles working in different directions. We also applied our methods to the *IFIH1* region for the type 1 diabetes GWAS data collected by the Wellcome Trust Case-Control Consortium. Our methods yield p-values on the order of 10^{-3}, whereas the most significant p-value from the existing methods is greater than 0.17. Therefore, we demonstrate that the evaluation of rare variants with GWAS data is possible, particularly when public sequencing data are incorporated. This work was published in the American Journal of Human Genetics in 2010 (Y. Li, Byrnes, & Li, 2010).

The forth chapter presents a systematic evaluation of multiple weighting schemes through a series of simulations intended to mimic large sequencing studies of a quantitative trait. We evaluate existing phenotype- independent and phenotype-dependent methods, as well as weights estimated by penalized regression approaches including Lasso (Tibshirani, 1996), Elastic Net (Zou & Hastie, 2005), and SCAD(Xie & Huang, 2009). We find that the difference in power between phenotype-dependent schemes is negligible when high-quality functional annotations are available. When functional annotations are unavailable or incomplete, all methods suffer from power loss; however, the variable selection methods outperform the others at the cost of increased computational time. Therefore, in the absence of good annotation, we recommend variable selection methods (which can be viewed as "statistical annotation") on top of regions implicated by a phenotype-independent weighting scheme. Further, once a region is implicated, variable selection can help to identify potential causal single nucleotide polymorphisms for biological validation. These findings are supported by an analysis of a high coverage targeted sequencing study of 1,898 individuals.

The final section proposes a combination method to apply the SKAT similarity-based approach (Wu et al., 2011) to data from admixed populations by first estimating local ancestry for each variant using Hapmix (Price et al., 2009) and MaCH-Admix (Liu, Li, Wang, & Li, 2013). We find that when the true causal alleles come only from the less common ancestral population, this approach shows a marked improvement in power over the original SKAT method alone. When the true causal alleles come only from the more prevalent ancestral population, however, the SKAT approach alone seems adequate to capture the association signal. This work is not yet complete, but we outline the proposed next steps in the final part of this section, which include more simulation replicates and the application to a real data set.

CHAPTER 2: LITERATURE REVIEW

This section presents a partial review of many of the papers previously published on the topic of collapsing information across rare variants. It is by no means complete since the number of these papers is quite large; however, it is an attempt to show the development of several methods used to attack this problem. Mathematical notation between the works discussed here is also quite diverse, so n the description of many of these works, some of the mathematical notations (variable names, etc.) has been altered slightly to keep the notation as consistent as possible throughout this document.

2.1 Early Methods

Before the advent of high-throughput genomic sequencing, it had been hypothesized that collections or combinations of rare variants could be responsible for some of the heritability in human complex traits. When GWAS and other studies turned up promising candidate genes, many researchers invested in re-sequencing and other molecular experiments in order to learn more about these candidates. The first methods for association of rare genomic variants arose from the need to analyze these data.

Well before the rise of "next generation" genome sequencing technologies, (Cohen et al., 2004) demonstrated that rare alleles could indeed have a measureable effect on human traits. Cohen et. al. sampled 128 individuals from the top and bottom five percent of the HDL cholesterol distribution in the Dallas Heart Study. All subjects were sequenced for three candidate genes, *ABCA1*, *APOA1* and *LCAT*. Non-synonymous mutations (that is, variation that effects the resulting protein) were considerably more common in the low HDL group as compared to the high HDL group, most notably in *ABCA1*, (p<0.0001). The investigators reported that one in six individuals with low HDL had a rare mutation in *ABCA1* or *APOA1* and went on to replicate these findings in a Canadian sample, thus making a strong case for the involvement of rare sequence variants in complex disease. These investigators did not develop a novel statistical method for assessing the combined effects of rare variants on the genome-wide scale, but they did demonstrate the potential impact of considering associations with combinations of rare variants. Note also that the rare variants were found primarily in the low HDL group, suggesting that these variants exhibit a protective effect. It had been previously suggested that genetic variants can, in most cases, be assumed to have null or deleterious effect. Variants of protective effect were (and are still) considered the exception, rather than the rule. However, the results of this study show the importance of detecting association between genetic variants and human phenotypes in either direction.

As the interest in rare variation grew, so too did the interest in capturing their associations statistically. An approach similar to that described above was formalized by (Morgenthaler & Thilly, 2007). In their manuscript, the authors emphasized the importance of limiting these tests to promising genes and adjusting for many important biological variables such as the number of genes and variants expected to be truly involved in the etiology of the trait. In this work, Morgenthaller and Thilly focused primarily on case-control studies and they suggested using the rare variant count among cases compared that of controls to conduct a T-test for association between the rare variant “burden” and the disease of interest. They named this method cohort allelic sum test (CAST). Note that this test only compares counts of rare alleles, and so information must be pooled across individuals and across markers.

In 2008, Li and Leal proposed the Combined Multivariate and Collapsing (CMC) method to collapse across variants across genomic regions, functional groups, MAF categories or other groupings (B. Li & Leal, 2008). These authors advocated first splitting the *M* markers into *k* groups (for example, by MAF bin) and then using a collapsing method in which, for each individual *i*, in the set of variants under study, Li and Leal suggested computing X_i as follows,

$$X_i = \begin{cases} 1, & \text{rare variants present} \\ 0, & \text{rare variants absent} \end{cases}$$

so that any individual with rare variants present was given weight 1 and all others have weight 0. This reduced the dimension of the multivariate test from *M* to *k*, making a multivariate test feasible where it may not otherwise be. Unlike the approach of (Morgenthaler & Thilly, 2007), Li and Leal collapsed information across markers, within individual, for each group *k*. The authors reported good results for situations in which one or more rare variants in the same group were truly deleterious.

In the same year, (M. Li, Wang, Grant, Hakonarson, & Li, 2009) proposed a method to use the information contained in outside data sets (e.g. HapMap) to better assess association between traits and genotypes called ATOM. Genotype imputation had also been helpful in gaining information about un-typed markers in GWAS studies (Y. Li, Willer, et al., 2010; Y. Li et al., 2009); however, unlike the imputation study design, ATOM did not require that each un-typed variant be explicitly imputed. ATOM aimed to assign weights in markers based on the amount of association they have with the trait locus, which was assumed to be un-typed. It did so by capitalizing on the correlation or linkage disequilibrium (LD) structure of the region. Suppose that the external data set has M_e markers and M_e > M. For each pairwise combination of variants in the original data set, $j \in \{1,2,...,M\}$, and each marker in the external data set $l \in \{1,2,...,M_e\}$, the weight w_j^l is computed,

$$w_j^l = \frac{\Delta_j^l}{q_j(1-q_j)},$$

where q_j is the minor allele frequency at variant *j* from the reference data and Δ_j^l is the linkage disequilibrium (LD) coefficient for markers *l* and *j*. Then, for each of the markers in the external data set, $l \in \{1,2,...,M_e\}$ the authors computed the score,

$$S_{il} = \frac{1}{m}\sum_{j=1}^{M} w_j^l x_{ij}$$

so each individual had M_e scores. Then the authors performed principal component analysis (PCA) on these scores, thus reducing the dimension of the problem significantly. The principal components were then tested for association with the trait by conventional regression methods without permutation. ATOM performed well compared to other previous methods in terms of power and also performed well when compared to simple haplotype approaches not discussed here.

In 2009, Madsen and Browning similarly hypothesized that rarer variants were more likely to have deleterious effects and that a higher burden of rare variants should likewise be more harmful than one lone rare variant. With this motivation, they proposed a simpler method that does not rely on an external dataset called the Weighted Sum (WS) method for case-control data. (Madsen & Browning, 2009) These investigators used the estimated minor allele frequency (MAF) of marker *j* among controls, denoted *q_j*, where, $q_j = \frac{\sum_{i=1}^{N_{control}} x_{ij} + 1}{N_{control} + 2}$ to construct weights, $\hat{w}_j = \sqrt{N_{total} q_j (1 - q_j)}$ so that, the rarer the allele, the smaller the quantity $\hat{w}_j$. Madsen and Browning then computed the genetic scores for each individual, denoted $S_i = \sum_{j=1}^{M} \frac{x_{ij}}{\hat{w}_j}$. Thus, the largest components of this weighted sum came from the variants rarest among controls. This stands to reason since deleterious alleles may confer a selective disadvantage and therefore be less common in the population than neutral or beneficial ones. Madsen and Browning performed a Wilcoxon Rank Sum test on the collection of *S_i*'s to assess significance. The many simulations presented in the Madsen and Browning manuscript show the Weighted Sum (WS) method distantly outperforming CAST, CMC and single variant tests in terms of power to detect rare variant association in a number of situations. They also demonstrated their methods' improved power by applying it to the ENCODE data.

In the following year, Price et. al. proposed a similar method to allow for an unknown threshold on MAF, denoted *T*, below which variants may be substantially more indicative of a functional variant (Price et al., 2010). As motivation for the idea, Price, et. al. first constructed a weighted sum with a fixed threshold, *T*, and constructed genetic scores, $S_i = \sum_{j=1}^{M} \xi_j (2 - x_{ij}) y_i$ where $\xi_j = I(q_j < T)$ and *y_i* is the outcome of interest. The authors evaluated both *T*=0.01 and 0.05. Price et. al. then generalized this approach to their Variable Threshold (VT) method by constructing a Z-scores, *Z(T)*, for a range or *T* values. The Z-score *Z(t)* is constructed, $Z(t) = \sum_{j=1}^{M}\sum_{i=1}^{N}\xi_j^T(2-x_{ij})(y_i - \overline{y})$, where $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the mean of the outcomes y_i. The value of *T* that maximizes *Z(t)* in each case was then chosen as the threshold. Because of this step, the Wilcoxon Rank Sum test was no longer valid and statistical significance needed to be assessed by permutation. The authors found their VT approach had greater power than WS and fixed-threshold methods in simulations for both dichotomous and continuous outcomes. (Price et al., 2010) also applied their VT method to Polyphen-filtered data (Ramensky, Bork, & Sunyaev, 2002) and found a greater improvement in power with the addition of good bioinformatics data.

Also in 2010, a manuscript in Genetic Epidemiology (Morris & Zeggini, 2010) presented a simple experiment demonstrating the importance of the choice of weighting scheme for these types of approaches and the potential for falsely significant results due to non-causal rare variants. The authors tested two scoring methods, which they call RVT1 and RVT2. RVT1 constructed weights for predefined genomic regions according to the *proportion* of rare variant sites at which an individual *i* had 1 or 2 copies of the minor allele, thus fitting the following model.

$y_i = \alpha + \lambda \frac{\sum_{j=1}^{M} I(x_{ij} > 0) I(q_j < T)}{\sum_{j=1}^{M} I(0 < q_j < T)} + \gamma Z_i + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$, Z_i is a vector of covariates and *T* is the threshold for determining the definition of “rare” in this case. Morris and Zeggini’s other model, named RTV2, assigned weights out of an indicator function, which took value 1 when any rare variant sites contained one or more copies of the minor allele in individual *i* and 0 otherwise.

$y_i = \alpha + \lambda I\left(\sum_{j=1}^{M} I(q_j < T)x_{ij} > 0\right) + \gamma Z_i + \varepsilon_i$, where $\varepsilon_i \sim N(0,\sigma^2)$ and Z_i and T are as before.

The authors found that the weighting scheme based on the proportion of rare variant sites (RVT1) with one or more minor alleles had equal or greater power to the method that only considers whether rare variants are present or not. Morris and Zeggini demonstrated that this power difference was most pronounced when the number of non-causal rare mutations increases. Since RVT1 gave larger weight to individuals with a larger burden of rare variants, these same individuals were more likely to carry one or more rare deleterious variants.

2.2 Using the outcome to inform choice of weights

Though diverse, the early methods for rare variant association demonstrated the importance of combining the data in an intelligent way, rather than simply searching for the presence or count of rare variants. Many of these methods were devised for application to GWAS data and GWAS data after imputation using an outside reference panel, as described in the previous section. The advent and refinement of “Next Generation” sequencing technology only made such methods more appealing and, in a relatively short period of time, a plethora of new methods arose, many of them attempted to directly estimate a weight for each marker by explicitly using the outcome measurements. In this section, we will outline several such previously proposed methods for binary and continuous outcomes.

To start, (Han & Pan, 2010) described a method to use the information from individual markers via marginal logistic regression coefficient estimates for case-control data. Han and Pan suggested first fitting a series of univariate regression models of the form,

$$\text{logit Pr}(Y_i = 1) = \beta_0 + x_{ij}\beta_j$$

where Y_i is the case or control status of individual *i*, and x_{ij} is the genotype for individual *i* at locus *j* as before. From each such model, the estimated coefficient, $\tilde{\beta}_j$, and a p-value, p_j, were used to estimate the weight for each marker. First, the genotype data was recoded such that,

$$x_{ij}^* = \begin{cases} x_{ij}, & \text{if } \tilde{\beta}_j \geq 0 \text{ and } p_j < \alpha_0 \\ 2 - x_{ij}, & \text{if } \tilde{\beta}_j < 0 \text{ and } p_j < \alpha_0 \end{cases},$$

where α_0 is a predetermined p-value threshold, so that the coefficient,

$$\hat{\beta}_c = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} x_{ij}^{*2}\tilde{\beta}_j}{\sum_{i=1}^{N}\left(\sum_{j=1}^{k} x_{ij}^*\right)^2}$$

from the model, $\text{logit Pr}(Y_i = 1) = \beta_0 + \sum_{j=1}^{M} x_{ij}\beta_c$, which assumes (probably incorrectly) that all variants with causal effect had the same odds ratio. The authors tested the hypothesis $H_0 : \beta_c = 0$ with the usual score test. Despite the questionable assumption of a constant odds ratio for all causal variants, this method performed well in comparison to previous methods in terms of power and type I error. Also note that, because the data were used to estimate the weight each variant received in the analysis, analytical p-values could not be evaluated here. Instead, the authors advocated a permutation approach in which the Y_i are shuffled randomly to attain a correct p-value.

In 2011, Zhang et. al proposed a similar method that also used the results of single marker tests in to inform the choice of weights for each marker (Zhang, Irvin, Arnett, Province, & Borecki, 2011). Assume single marker tests via linear (for continuous outcomes) or logistic (for binary outcomes) have already been conducted and for each marker, we have an estimate of the coefficient for marker *j*, b_j and the standard error of that estimate, s_{bj}. The investigators proposed fitting the model,

$$Y_i = \alpha + \beta \sum_{j=1}^{M} w_j x_{ij} + \varepsilon_i$$

where x_{ij} in the number of minor alleles at locus *j* for individual *i* and w_j is the weight assigned to marker *j* determined by,

$$w_j = 2\{p(t \le t_j) - 0.5\}, \text{ with } t_j = \frac{b_j}{s_{b_j}}.$$

where the distribution of *t* is determined empirically from all of the single marker tests conducted. The probability $p(t \le t_j)$ is a left tail p-value, and so the test was named the p-value weighted sum test (PWST). In order to assess significance, the authors simply tested the hypothesis, $H_0 : \beta = 0$. PWST was also shown to perform well compared to phenotype-independent approaches.

In 2011, Lin and Tang rigorously showed that the optimal unbiased weight for each variant was proportional to the true coefficient β_j in the limit. (Lin & Tang, 2011) Since the coefficient could be estimated from the data, it seemed sensible to set $\xi_j = \hat{\beta}_j$, where $\hat{\beta}_j$ is the appropriate estimate of the coefficient β_j. However, the authors pointed out two problems with this approach. First, using these weights, their test statistic *T* would not be asymptotically normal and, second, the values of $\hat{\beta}_j$ were relatively unstable considering that the variants in question are rare and the sample size was finite. The authors instead proposed a compromise in which they fit the model, $Y_i = \alpha_{i0} + \alpha_i + \sum_{j=1}^{M} \beta_j x_{ij} + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ is assumed and construct weights $\xi_j = \hat{\beta}_j + \delta$, where delta is a known constant. Then the genetic score was constructed, as in previous methods, $S_i = \sum_{j=1}^{M} \xi_j x_{ij}$, and evaluated with a score statistic. The authors named this method EREC (Estimated Regression Coefficients) and found it performed well is simulations and on real data, though some of the similarity approaches (discussed in the next sub-section) produced smaller p-values when applied to real data. The EREC method cannot be applied in cases where M>N and is not intended to account for variant effects in different directions.

In the same year, a Bayesian method that used the data to directly estimate the weights of the individual markers, in addition to assessing to significance of the association between genotypes and trait (as in EREC), was proposed by (Yi & Zhi, 2011). Since this method was intended for use on case-control data, the investigators ultimately aim to fit the logistic model, $\operatorname{logit} \Pr(Y_i = 1) = \beta_0 + \sum_{j=1}^{M} x_{ij}\beta_j$, which they rewrote as $\operatorname{logit} \Pr(Y_i = 1) = \beta_0 + \beta\sum_{j=1}^{M} x_{ij}\alpha_j$. They then rephrased the problem, first estimating the $\alpha_j, j \in \{1,2,...,M\}$, and then testing, $H_0 : \beta = 0$. Yi and Zhi offer a different solution to the problem of instability for the estimates of the individual weights from (Lin & Tang, 2011) by putting priors on the parameters α_j and β. The authors proposed an informative prior on the α_j,

$$\alpha_j \sim N(\mu_j, \tau_j^2),\ \ \tau_j^2 \sim Inv - \chi^2(1, s_\alpha^2)$$

where s_α^2 was chosen to be a small value such as 0.5. Yi and Zhi point outed that the choice of the Student-t priors on α_j are designed to better deal with disparate effects. The prior parameter μ_j could be manipulated according to the prior knowledge about the variant j, and though the authors did not do this, they suggested using frequency distribution or functional credibility to determine the value of μ_j. Because of the rarity of the alleles in question, the variance of $\sum_{j=1}^{M} x_{ij}\alpha_j$ could be quite low and so, the estimate of β could also become unstable, the authors suggested a weakly informative prior on β

$$\beta \sim N(0, \tau_\beta^2),\ \ \tau_\beta^2 \sim Inv - \chi^2(1, 2.5^2)$$

which was meant to keep the β parameter in a reasonable range. This Bayesian linear model was fitted using Markov Chain Monte Carlo and showed good results for type I error and power far surpassing any of the phenotype-independent methods discussed in the previous sub-section. This gain in power was particularly noticeable when the effects of the causal variants acted in opposite directions. This method was not, however, compared to any other phenotype-dependent methods previously discussed.

2.3 Similarity-based approaches: why weight?

Simultaneously with many of the methods discussed above, many statistical geneticists and biostatisticians began to question if weighting each individual marker was necessary at all. While having weights for individual markers may be helpful in determining the disease etiology in some cases, simply implicating a gene or molecular target can also be helpful to biological and pharmaceutical researchers. The above methods directly estimate a weight for each marker, but the following methods are aimed at implicating a genomic region by collapsing sets of markers that tend to be shared across two individuals when the outcomes are also similar.

One of the first of such methods was proposed well before the popularization of “next generation” sequencing techniques by (Schaid, McDonnell, Hebbring, Cunningham, & Thibodeau, 2005). Schaid and colleagues suggested using a U-statistic of the form,

$$U_{global} = \frac{\sum_{i<i'} K(x_i, x_{i'})}{\binom{N}{2}} = \sum_{j=1}^{M} w_j \frac{\sum_{i<i'} K(x_{ij}, x_{i'j})}{\binom{N}{2}} = \sum_{j=1}^{M} w_j U_j$$

where $K(x_i, x_{i'})$ is a symmetric kernel function that compares the genotype of individual *i* to that of individual *i’*. Specifically, it is the weighted sum of the variant-specific kernels,

$$U_j = \frac{\sum_{i<i'} K(x_{ij}, x_{i'j})}{\binom{N}{2}}.$$

The authors suggested two kernels with which to quantify the similarity of individual *i* and *i’* at locus *j*: first the “allele-match” kernel, which was a simple count of the number of alleles at locus *j* that match between individuals *i* and *i’*. Second, the “linear dosage” kernel added the number of the minor alleles at locus *j* together for individuals *i* and *i’*. A normalized version of these U-statistics were compared to several simpler methods, such as simply taking the maximum signal from the region and the conventional multivariate

T-test with good results for power and type I error under most conditions, particularly when the number of risk loci rose above 5.

(Neale et al., 2011) also suggest a method to test for any association between a set of genotypes and phenotype rather than individually estimating weights for each marker. For motivation, the authors used the example of set of *M* coins that could be either fair or biased. Each coin could land as either a case or a control and, if the coin was fair, it would land case and control with equal probability. If the coin was biased (i.e. if the marker is associated with the trait in either direction), they expected the coin to land preferentially as a case or as a control. Thus, the C_α tests for the presence of biased coins over the *M* markers, rather than a test for single biased coin. The C_α test statistic compares the variance of each observed count with the expected variance and then sums over all variants.

$$T = \sum_{j=1}^{M} [(x_{case,j} - x_{total,j} p_0)^2 - x_{total,j} p_0 (1 - p_0)]$$

where p_0 is the expected number of times the minor allele is expected to turn up in the cases, given the number of total copies of the minor allele in the sample and assuming that the j^{th} marker is like a fair coin, that is $p_0 = \frac{x_{total,j}}{2}$ for $x_{total,j} \geq 2$. The obvious problem with this setup was singleton counts, since they contain no variance information. The authors suggested binning all of the singleton counts into one category and proceeding as if they were all from one marker. The quantity *T* was then normalized and compared to a one-tailed normal distribution. The C_α test performed well in terms of power when compared to simple burden tests like that of (B. Li & Leal, 2008; Madsen & Browning, 2009), however the asymptotic properties of the proposed statistic had heavier tails that expected, particularly when the sample size is small. Further, this test assumed independence between all variants and for all these reasons, the authors suggested using permutations to assess significance particularly in the presence of LD or when sample size is small.

Later in 2011, Wu and Lee et. al. suggested a more general test with an arbitrary weight matrix also using kernel methodology to compare the genomes of all *N* samples called SKAT (Wu et al., 2011). As with many of the methods considered above, SKAT aimed to fit a model of the form, $Y_i = \mu + \beta X$ that may or may not also have covariates included in the μ component. They proposed a statistic *Q* of the form,

$$Q = (Y - \hat{\mu})' K (Y - \hat{\mu}), \text{ where } K = XWX'$$

in which *X* is the $N \times M$ matrix of minor allele counts, as before and W is a matrix of weights that can be specified by the users. Wu, Lee and colleagues explored several kernels to incorporate information from various data types to provide a more powerful test. The matrix *K* was constructed to measure the pairwise genotypic similarity between every two individuals in the sample. The matrix *W* quantifies the degree of importance each variant. In the absence of a user derived weight matrix, the SKAT authors recommend using a matrix *K* such that $K(x_i, x_{i'}) = \sum_{j=1}^{M} w_j x_{ij} x_{i'j}$ when no interactions between variants are present (linear kernel) and $K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{M} w_j x_{ij} x_{i'j}\right)^2$ for when there are interactions between variants (quadratic kernel). They suggested choosing the weights according to $w_j = Beta(q_j, 1, 25)$ where q_j is the MAF of variant *j*, as before. The SKAT authors saw power that was much improved compared to the burden tests and advocate this method over the C_α test because it can easily account for covariates and interactions between variants.

Interestingly, the SKAT authors noticed a drop in power when the true situation (unknown in the case of a real study) was similar to that in which the Madsen & Browning method (Madsen & Browning, 2009) is ideal. That is, situations where the trait was influenced, not by a particular subset of rare genetic variants, but by the number of deleterious, rare “hits” observed in the region. In 2011, Lee and colleagues proposed SKAT-O to optimize the test, even if this was the case (Lee et al., 2012). SKAT-O performed both the SKAT test and the burden test and produces a weighted sum of the two. In situations where the SKAT statistic would be most powerful, SKAT-O lost very little power in comparison to SKAT; however, SKAT-O demonstrated a marked improvement over the original SKAT approach when the true association was a series of very low frequency hits.

CHAPTER 3: WHAIT

3.1 Introduction

In this chapter, we propose two methods to search for the aggregated effect of rare variants with GWAS data. Our approaches do not rely on the availability of external sequencing data, but they can incorporate such information when available. Moreover, our methods make no assumption on the direction of association of rare alleles with disease risk. We applied our methods, along with existing methods proposed in the sequencing context, to simulated data sets. Our methods demonstrated better performance across a wide range of scenarios with an average power improvement of 8.6% (31.6%) in the absence (presence) of external sequencing data. We also applied our methods to the Wellcome Trust Case-Control Consortium (WTCCC) type 1 diabetes (T1D) GWAS data set in the *IFIH1* gene region, where both common and multiple rare variants have been found to influence the risk of T1D (Barrett et al., 2009; Nejentsev, Walker, Riches, Egholm, & Todd, 2009; Smyth et al., 2006).

3.2 Methods

3.2.1 Weighted haplotype score test

Our first test is a weighted haplotype test. Assume a sample of N diploid individuals is collected, among which N_{cs} are affected cases and N_{ct} are unaffected controls. Let m denote the number of genotyped markers in a region of interest. Further denote haplotypes of the N individuals by $H = (H_1, H_2, ..., H_i, ..., H_N)^t$, where

H_i = $\{H_{i,1}, H_{i,2}\}$ are the two haplotypes carried by the i^{th} individual, consisting of the m markers in the region. For each individual *i*, we define a weighted haplotype score as follows:

$$WHS_i = \sum_{j=1}^{2} W_{H_{ij}} ,$$

in which the sum is taken over the two haplotypes of individual *i*. W_h stands for the weight of haplotype *h* and is defined as

$$W_h = I(h \in \mathbf{C}) \cdot (-1)^{I(h \in \mathbf{P})} \cdot S_h ,$$

in which **C** is the set of disease-contributing haplotypes including both risk and protective haplotypes, **P** is the set of disease-protective haplotypes (note that **P** is a subset of **C**), and S_h is a score assigned to haplotype *h*. Following the weighting scheme proposed by Madsen and Browning (Madsen & Browning, 2009) for SNPs, we define S_h as

$$S_h = \sqrt{N_{ct} \cdot f_{ct,h} \cdot (1 - f_{ct,h})} ,$$

in which $f_{ct,h}$ denotes the adjusted frequency of haplotype *h* among controls and is defined as

$$f_{ct,h} = \frac{C_{ct,h} + 1}{2(N_{ct} + 1)} ,$$

in which $C_{ct,h}$ is the number of haplotype *h* among controls. The rationale of using such a score is that a rare variant (most likely untyped in GWAS) is more likely to be tagged by a rare haplotype than by a common haplotype, and thus rare haplotypes should receive more weight in the analysis. To define the sets of the disease-contributing and disease-protective haplotypes, we first split the data into a testing set and a training set and then compared the haplotype frequencies between cases and controls in the training set according to the formula below:

$$\begin{cases} h \in C \quad \mathit{if} & \left| f_{cs,h}^{tr} - f_{ct,h}^{tr} \right| > \mu \sqrt{\dfrac{f_{ct,h}^{tr}(1 - f_{ct,h}^{tr})}{2N_{ct}^{tr}}}, \\ h \in P \quad \mathit{if} & f_{cs,h}^{tr} - f_{ct,h}^{tr} < -\mu \sqrt{\dfrac{f_{ct}^{tr}(1 - f_{ct}^{tr})}{2N_{ct}^{tr}}}, \end{cases} \qquad \text{(Equation 1)}$$

with *tr* standing for "training set." Here, μ is a constant that is determined by a pre-specified type I error rate. For example, $\mu = 1.28$ (1.64) corresponds to a type I error of 0.2 (0.1). Following (Zhu, Feng, Li, Lu, & Elston, 2010) we set $\mu = 1.28$ and randomly selected 30% of the samples for training in the analysis.

We note that by explicitly modeling the two sets of haplotypes as described above, we do not need to make assumptions about the direction of association between rare alleles and disease risk. Weighted haplotype scores are calculated in the testing set after identifying the two sets of haplotypes with the training set. To assess whether the rare variants are significantly associated with the disease, we can perform a standard Wilcoxon (Wilcoxon, 1945) test on the weighted haplotype scores and assess the significance of the test by permutations. For each permuted data set, the training set and the testing set will be obtained in a similar fashion as the original data set. Because typical GWAS data consist of genotypes rather than haplotypes, we need to infer haplotypes from unphased genotypes. This step can be done via standard phasing methods, including PHASE, fastPHASE, MaCH, and Beagle (Browning, 2006; Y. Li et al., 2009; Scheet & Stephens, 2006; Stephens & Scheet, 2005). We used MaCH, which allows the incorporation of external genotyping, haplotyping, or sequencing data. Our weighted haplotype approach can be applied to haplotypes consisting of GWAS markers alone or to haplotypes including additional markers via incorporation of external reference data.

3.2.2 Weighted dosage score test

Our second test is a weighted imputation dosage test. Following the notations defined above, we assume that there are a total of M markers genotyped or sequenced after the incorporation of one or more external data sets, e.g. the International HapMap Project (Frazer et al., 2007; International Hapmap Consortium, 2005) or the 1000 Genomes Project (Kaiser, 2008). We have previously described a hidden Markov model-based method that imputes untyped markers in study samples by exploiting external data as reference, which was implemented in software MaCH and has become standard in GWAS analysis (de Bakker et al., 2008). Let $D = (D_1, D_2, ..., D_i, ..., D_N)^t$ denote the dosage matrices across M markers for the N study subjects, in which $D_i = (D_{i,1}, D_{i,2}, ..., D_{i,j}, ..., D_{i,M})$ denotes the dosages of the i^{th} individual. Here D_{ij} is the dosage for the i^{th} individual at marker j, which is defined as the expected number of the rare allele at marker j. Now we define the weighted dosage score for each individual i as

$$WDS_i = \sum_{j=1}^{M} I(j \in \mathbf{M_C}) \cdot (-1)^{I(j \in \mathbf{M_P})} \cdot D_{i,j} ,$$

in which the summation is taken over all M markers with genotype dosage scores. Here $\mathbf{M_C}$ is the set of markers with the rare allele that contributes to disease risk, and $\mathbf{M_P}$ is the set of markers with the rare allele that decreases disease risk. We define these two sets by examining frequency difference between cases and controls, similar to Equation 1, for the weighted haplotype test. After obtaining the scores, the standard Wilcoxon test is applied to test for association with the disease, and its significance is assessed via permutation. We compared our proposed methods with the following three methods proposed in the sequencing context. (1) Weighted SNP Test (denoted by WS) (Madsen & Browning, 2009) is a weighted- sum method in which rare alleles are aggregated and weighted according to a function of minor allele frequency among controls. Despite the fact that the method was proposed as a test for ‘‘rare mutations,’’ it indeed sums over all markers by giving smaller weight to alleles with higher frequency. Although an omnibus regional-based test that evaluates both common and rare variants is some- times desired, here we are interested in a regional-based test for rare variants only, assuming that common variants have been thoroughly evaluated by large-scale GWAS. Because of this, we compared our methods with both the originally proposed test (denoted by WS_{all}) and a modified version of it (denoted by WS_{rare}), in which only markers with minor allele frequency (MAF) < 5% are included. (2) (Zhu et al., 2010) proposed a haplotype grouping method (denoted by HG) that counts the number of rare risky haplotypes for each individual and uses a Fisher's exact test for testing. (3) We also applied the rare variant collapsing method (denoted by RVC) proposed by (B. Li & Leal, 2008) which groups each individual into one of two groups: carrying any rare allele or not. Together with case-control status, a 2×2 table is generated, and a standard test for contingency table (e.g., chi-square test for independence) is applied. Table 3.1 lists the above-described tests and their abbreviations.

Table 3.1. Abbreviation and description of tests applied

Test Abbreviation	Description
WDS	Weighed dosage test on genotyped plus imputed SNPs with external sequencing data
WHS	Weighted haplotype test on genotyped plus imputed SNPs with external sequencing data
WHG	Weighted haplotype test on genotyped SNPs only
HG	Haplotype grouping test orioised by Zhu et. al.
WS_{all}	Original weighted SNP test aggregating evidence over all (regardless of MAF) SNPs proposed by Madsen and Browning
WS_{rare}	Modified weighted SNP test aggregating evidence over rare (MAF<5%) SNPs only
RVC	Rare variant collapsing method proposed by Li and Leal

3.2.3 Simulation Setup

We simulated 10,000 chromosomes for a series of 100 1 Mb regions with a coalescent model that mimics linkage disequilibrium (LD) in real data, accounts for variations in local recombination rates, and models population history, consistent with the HapMap CEU (CEPH people from Utah, USA) samples (Schaffner, Foo, & Gabriel, 2005). We then took a random subset of 1000 simulated chromosomes (i.e., 500 individuals) to serve as the external reference, mimicking the targeting sample size for the 1000 Genomes Project. To generate a set of GWAS markers in each region, we first randomly picked 120 chromosomes, mimicking Phase II HapMap CEU data. We then ascertained and thinned polymorphic sites to match marker density and allele frequency spectrum of their real-data counterparts. Based on LD measures calculated with the 120 chromosomes, we selected a set of 100 SNPs for each region that included 90 tagSNPs tagging the largest number of SNPs and 10 additional SNPs picked at random among the remaining SNPs. The final set of retained SNPs (GWAS markers in the region) captured ~78% of the common variants (MAF > 5%) at a conventional r^2 cutoff of 0.8, similar to the real-data performance of the Illumina HumanHap300 BeadChip SNP genotyping platform.

Within each simulated 1Mbregion, we picked an ~50 kb region as the causal region in which we assume only rare variants (variants with population MAF between 0.1% and 5%) contribute to the disease risk. We randomly selected *d*% of the rare variants in the causal region to be causal, i.e., to influence disease risk. Among these rare variants, we further assume that *r*% of them increase disease risk, whereas the remaining (100 – *r*)% decrease disease risk. To ensure that each variant only has a small contribution to the overall disease risk, we followed a model similar to that proposed by (Madsen & Browning, 2009). Specifically, the contribution of each causal variant j to the overall genotype relative risk (GRR) is defined as:

$$GRR_j = \left(\frac{PAR}{(1-PAR)\cdot MAF_j} + 1 \right)^{(-1)^{I(\xi_j=1)}} ,$$

in which *PAR* is the population attributable risk and $\xi_j = 1$ indicates that the rare allele of marker *j* decreases disease risk. Following (Madsen & Browning, 2009), we used the same marginal PAR for each causal variant, which intrinsically assumes that alleles with lower frequency have higher GRR than alleles with higher frequency. In our 50 kb core region, there are ~500 SNPs with MAF < 5%. To generate the chromosomes for an individual, we randomly selected two chromosomes $\{H_1, H_2\}$ from the remaining 9000 chromosomes that were not selected as external reference. The disease status of the individual was assigned according to

$$P(affected \mid \{H_1, H_2\}) = f_0 \times \prod_{k=1}^{2} \prod_{j=l}^{m_c} GRR_j^{I(H_{k,j}=a_j)} ,$$

in which f_0 is the baseline penetrance and was fixed at 10% in our simulations (1% and

5% were also evaluated and resulted in similar patterns but with slight power loss), m_c is the number of causal SNPs, and a_j is the rare allele of SNP j. Sampling was repeated until the desired number of cases and controls was reached. In our simulations, d took values from 10% to 50% by an increment of 10%. Among the disease risk influencing loci, we set the value of r, the percentage of rare alleles increasing disease risk, at 5%, 20%, 50%, 80%, and 100%, respectively.

For each of the 100 regions, two independent data sets with 1000 cases and 1000 controls were simulated with the model described above. In addition, five independent null data sets of the same sample size were simulated, assuming no genetic effect by randomly sampling 4000 chromosomes (i.e., 2000 individuals) from the pool of 9000 chromosomes. Average power was estimated based on the 100 regions, which represent a wide range of LD patterns. To account for local LD differences, we permuted each of the null sets 200 times to obtain region-specific empirical significant threshold. For the weighted haplotype analysis, we considered two versions: WHG, which uses haplotypes consisting of GWAS SNPs only, and WHS, which uses haplotypes encompassing both genotyped and imputed SNPs. For both the weighted haplotype tests and the weighted dosage test, untyped SNPs with Rsq (estimated imputation quality) < 0.3 were discarded from subsequent analysis (Y. Li et al., 2009). In all analyses, we used haplotypes reconstructed from the unphased genotypes and imputed genotypes for markers that are not included on the GWAS chip. Our methods (WHG, WHS, and WDS), together with $\mathrm{WS}_{\mathrm{all}}$, $\mathrm{WS}_{\mathrm{rare}}$, HG, and RVC, were applied to the 1000 null data sets within each region to determine the region-specific empirical significance threshold, ensuring the correct type I error rate of 0.05 for all tests.

3.3 Results

Figure 3.1 shows the empirical power of our methods relative to the other four methods proposed in the sequencing context as a function of *r*, the proportion of rare alleles increasing disease risk, which ranges from 5% to 100%. We fixed PAR at 0.5% and *d* (percent of disease-influencing rare variants) at 50%. Although the synergy assumption is more reasonable for rarer alleles than for common alleles because rarer alleles tend to disrupt gene function, our knowledge regarding the direction of rarer alleles is still limited. Therefore, methods robust to such an assumption are desirable. Although all methods have decreased power when rare alleles work in different directions, our methods performed better by explicitly modeling the direction of association. For example, compared with the haplotype grouping (HG) method, the advantage of our weighted haplotype method (WHG, on GWAS SNPs only without the aid of external sequencing data) manifests more when a larger proportion of the rare alleles is protective: power gain is 9.1% when all of the rare alleles at disease-contributing loci increase disease risk, and the power gain increases to 20.7% when only 5% of the rare alleles increase disease risk.

Our proposed tests increase power through two different mechanisms: by using haplotypes to better capture information for rare variants (mostly untyped in GWAS) and by using external sequencing data to impute rare variants. Let us consider the first mechanism by examining tests on GWAS data alone, namely WHG, HG, WS_{all}, WS_{rare}, and RVC. At GWAS level, haplotype-based methods clearly manifest their advantages. Among the five methods, the two haplotype-based methods (WHG and HG) rank as the best two across the five scenarios presented in Figure 1. Note that WS_{all} and WS_{rare} can be viewed as special cases of WDS, where the dosages only take values 0, 1, or 2 at directly genotyped markers. Therefore, at the GWAS level, haplotype-based methods are preferred over single-marker dosage-based tests. This is because causal rare variants are better captured by haplotypes constructed from GWAS SNPs than by those SNPs themselves. Between the two haplotype-based methods, our weighted haplotype method (WHG) increases power by an average of 13.2% over HG by weighting individual haplotypes (instead of lumping them together into groups) and by explicitly modeling the direction of association.

Next we consider the second mechanism by looking at tests that incorporate external sequencing data, namely WHS and WDS. Both are more powerful than WHG, the best test based on GWAS data alone. The average power gain of WHS and WDS over WHG is 3.8% and 22.0%, respectively. At this pseudo-sequencing level (i.e., study subjects imputed with SNPs of sequencing density), a single-marker dosage-based test is more powerful than haplotype-based methods. This is not surprising because, at the pseudo-sequencing level, causal rare variants are better captured by their imputed counterpart than by haplotypes. The same applies to data at the sequencing level (i.e., when study subjects are directly sequenced). Of course, if there are genuine haplotype effects, we anticipate that WHS will perform better. To quantify the extent of better performance, we need more empirical data on the distribution of genuine haplotype effects, which is beyond the scope of this paper. Currently, we have little evidence even to convincingly conclude the presence of genuine haplotype effects. Therefore, with the presence of external sequencing data and under the assumption that single variants, cumulatively, contribute to disease risk, we recommend WDS over WHS.

Figure 3.1. Comparison of Power by r: Percent of Rare Alleles in the Causal Region that Increase Disease Risk Power of all tests was assessed at the 5% level by using empirical significance threshold determined by 1000 null data sets per region. 50% of the rare alleles in the causal region were assumed to contribute to disease risk (i.e., d fixed at 50%), and the PAR of each contributing SNP was fixed at 0.5%.

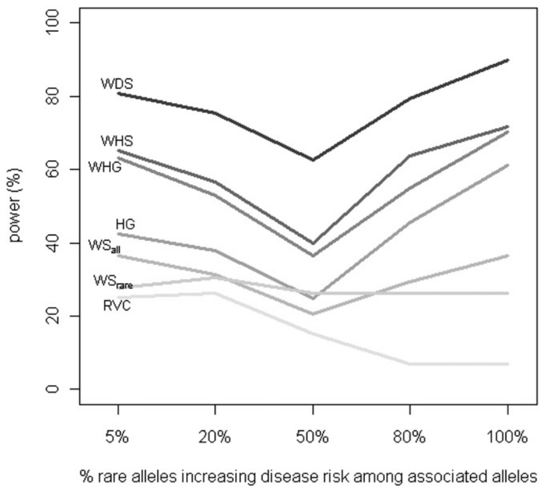
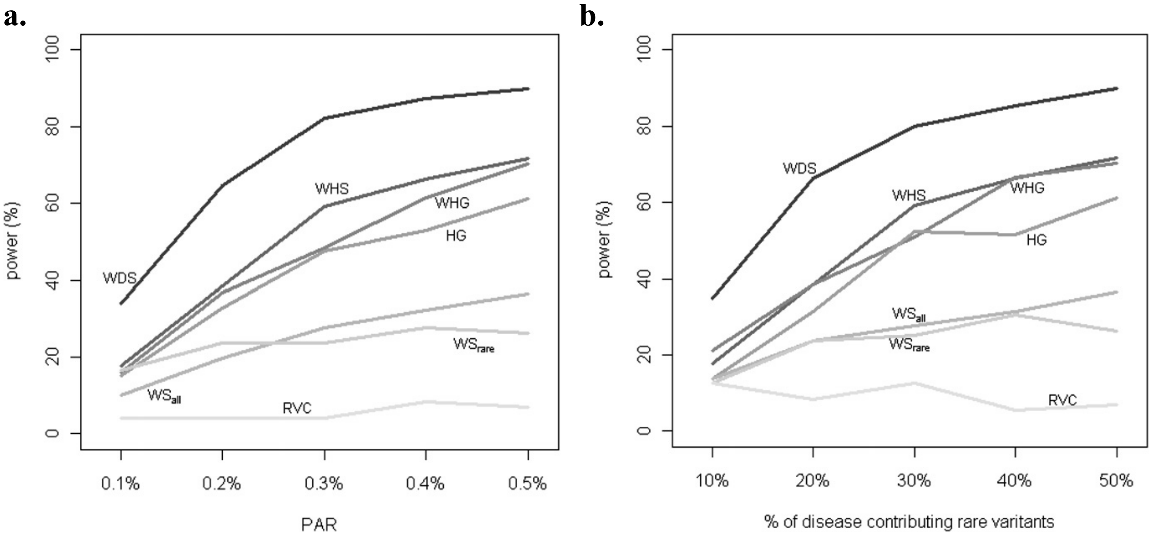


Figure 3.2 shows the power of different tests under situations with varying PAR and varying percentage of disease-contributing rare variants. We fixed the value of *d* (percentage of rare alleles influencing disease risk) at 100%. The value of *r* (percent of causal alleles increasing disease risk) was fixed at 50% for Figure 3.2a, and the per SNP PAR was fixed at 0.5% for Figure 3.2b. Although the power decreases with decreasing PAR or decreasing percentage of disease-contributing variants for all methods, our WHG and WHS are comparable, if not slightly better, than other alternatives, and our WDS is more powerful than the other methods by utilizing sequencing information from external data and explicitly modeling the SNP-level dosages**.**

We note that tests on rare GWAS SNPs only (WS$_{rare}$ and RVC) are less powerful in general, because at GWAS marker density, a typical gene region may contain few, if any, directly genotyped rare variants. In our simulations, 64 out of the 100 regions have no rare variants within the ~50 kb core causal regions. These tests, proposed in the sequencing context, are thus not suitable for analyzing GWAS data.

Figure 3.2. Comparison of Power by PAR and *d*: In figure 3.2a, power of all tests was assessed at the 5% level by using empirical significance threshold determined by 1000 null data sets per region. 50% of the rare alleles in the causal region were assumed to contribute to disease risk (i.e., d fixed at 50%), and all contributing rare alleles were assumed to increase disease risk (i.e., r fixed at 100%). In figure 3.2b, power of all tests was assessed at the 5% level by using empirical significance threshold determined by 1000 null data sets per region. All rare alleles in the causal region were assumed to increase disease risk, and the PAR of each contributing SNP was fixed at 0.5%.



Encouraged by results from simulations, we applied our methods to real data. Multiple common and rare variants in *IFIH1*, a cytoplasmic helicase that mediates induction of interferon response to viral RNA, have been established to influence risk of T1D. In particular, variants disrupting *IFIH1* function have been suggested to confer protection from T1D (Nejentsev et al., 2009). We took the WTCCC T1D data to search for rare variants associated with T1D susceptibility. In the WTCCC GWAS data set, 10 SNPs were found in the *IFIH1* region, with four being monomorphic in both the T1D set and the two control sets (NBS and 58C), leaving six SNPs for analysis. These six SNPs and their allele frequencies among cases and controls are tabulated in Table 3.2. We applied our methods, along with the others, to this data set. Because the common SNP rs1990760 (MAF > 30%) in *IFIH1* has been found to influence T1D risk (Barrett et al., 2009; Smyth et al., 2006), we restricted our analysis to SNPs or haplotypes with frequency < 5% to rule out signals due to LD with rs1990760. Our goal is to assess whether there is any residual association with T1D because of rare variants, which have been ignored in the previous GWAS analysis. We used the March 2010 release of 60 CEU individuals from the 1000 Genomes Project as reference for imputation. We used SNPs in the ~50 kb *IFIH1* gene region plus 2 Mb flanking on each side for phasing and imputation. Again, we discarded imputed SNPs with Rsq < 0.3. For the haplotype grouping method, the original test failed in this data set because rare alleles in *IFIH1* are associated with decreased risk of T1D. P-values based on 100,000 permutations are shown in Table 3.3. The p values from our methods are in the order of 10^{-3}, whereas the most significant p value from existing methods is >0.17. This example clearly demonstrates the importance of using appropriate methods when searching for the effect of rare variants with GWAS data.

Table 3.2. Allele frequencies of six polymorphic SNPs in *IFIH1*

SNP	**58C**	**NBS**	**T1D**
rs3747517	27.66%	26.31%	24.16%
rs41463049	1.12%	1.06%	1.02%
rs6432714	1.18%	1.06%	1.02%
rs13023380	48.88%	47.46%	45.24%
rs7559193	0.17%	0.10%	0.00%
rs12479125	1.18%	1.06%	1.02%

Table 3.3. Permutation p-values based on 10,000 permutations, for the association of rare variants in *IFIH1* with T1D risk in WTCCC data set

Test	**p-value**
WDS	0.00431
WHS	0.00738
WHG	0.00746
HG	1.000
WS_{rare}	0.329
RVC	0.179

3.4 Discussion

In summary, we have proposed two tests to assess the impact of multiple rare variants on disease risk. We show through simulations and a real-data example that by maximally extracting information from GWAS data, as well as the incorporation of publicly available sequencing data, our methods provide an intermediate solution for the analysis of rare variants before study-specific sequencing data become available. Our results suggest that at the GWAS level, haplotype-based methods are more powerful, but at the pseudo-sequencing level (i.e., GWAS data imputed with publicly available sequencing data), a test based on weighted sum of single-marker dosages is more powerful.

By assuming that we know the 50 kb causal region a priori, we may have overestimated the power in the simulations. We thus repeated the experiment by extending the test region to 100 kb (25 kb flanking region on either side of the core region) and to 200 kb (75 kb flanking on either side) to mimic the lack of knowledge on the lengths of regulatory regions flanking a gene or an exon. We found that the power difference is within 2%. In most situations, power was slightly lower, but in a few situations, power was slightly higher, because some variants in the non-causal flanking region happen to tag the causal variants better because of LD. These results are not surprising, because our methods can eliminate irrelevant SNPs or haplotypes by comparing frequency differences between cases and controls in the training data set. The analysis of rare variants with GWAS data is challenging because of several reasons. First, SNPs picked by the commonly used GWAS genotyping platforms have poor coverage for rare variants in general. Second, we have no catalog of rare variants in our genome, and our knowledge regarding their impact on phenotypic variations is still limited. Third, traditional association tests are suitable for the analysis of common variants but are generally underpowered for the analysis of rare variants. By utilizing LD information and incorporating publicly available sequencing data, we show that hunting for rare variants with GWAS data is possible.

Our methods are proposed for GWAS data, which are still the most commonly available type of data for gene mapping studies. In both our simulations and the real data analysis of T1D with gene *IFIH1*, we only have GWAS data on the study subjects. We compared our methods with alternatives proposed for sequencing data and demonstrated that methods that are specifically targeted for the analysis of rare variants in GWAS settings such as ours perform much better than methods that are developed for sequencing data. We note that our targeted ‘‘rare’’ variants (MAF 0.1%–5%) differ from those in methods developed in the sequencing context (including extremely rare variants with MAF < 0.5% or 0.1%). For extremely rare variants (MAF < 0.5%), our methods are expected to have low power because of low phasing and imputation quality with GWAS data. Although our methods are proposed for GWAS data, they can be applied directly to sequence data or to partially sequenced data in which selected individuals under study are sequenced. Therefore, our methods provide a useful alternative but are not meant to replace existing methods, given fundamental differences in their targeted data type (GWAS versus sequencing) and targeted MAF range. Because the performance of our weighted imputation dosage test depends critically on the imputation quality of rare variants (MAF < 5%), we decided to evaluate the quality in real data from the FUSION project (Scott et al., 2013) by masking and imputing all rare variants in a subset of individuals with constructed haplotypes encompassing both common and rare variants from an independent set of FUSION individuals (of varying sizes) as reference. We found that imputation quality for rare variants improves when the sample size in the reference panel increases. For example, the accuracy among the heterozygotes (r^2) increases from 83.4% (74.3%) to 97.0% (92.9%) when the number of reference haplotypes increases from 60 to 1000.

Our methods and others evaluated in this study were developed for the analysis of rare variants, but we have found that inclusion of common variants can increase the power (data not shown). This is demonstrated by the superior performance of WS_{all} (test that includes all variants) over WS_{rare} (test that only includes rare variants), even though only rare variants that contribute to disease risk were included in our simulations. This is not entirely surprising, because common variants or haplotypes can carry some information of untyped rare variants. One major issue of including common variants in testing is misclassification, that is, inclusion of variants that do not contribute to disease risk. However, by searching for frequency difference in a training set, our methods can alleviate this misclassification issue. In general, we recommend testing common variants first, for instance, via standard single-marker test. If there is no evidence of association with common variants, we then search the entire MAF space for the effect of rare variants. When common variants are found to be associated (such as in the *IFIH1* example), we should restrict our attention to rare variants or haplotypes only to alleviate the residual effects of common variants.

Both of our tests assess the effect of multiple variants in aggregate in a predefined genomic region, typically a known gene annotated by RefSeq or other gene annotations.

For real-life GWAS data, we recommend performing the tests for all known genes if no prior knowledge exists or for a list of one or more candidate genes in the presence of such knowledge. We note that the weighted dosage-based test is more flexible than the haplotype-based test in that it can be used to test for an arbitrary set of SNPs (for example, non-synonymous rare SNPs in a pathway), which may involve SNPs on different chromosomes.

One issue with the haplotype-based test is that the haplotypes are not known but instead are inferred with uncertainty. Fortunately, most phasing methods, including PHASE and MaCH, can estimate the probabilities of possible haplotype configurations for each individual in addition to providing the best-guess haplotypes. With these estimates, we can easily model the phasing uncertainty into our weighted haplotype test by allowing possible haplotype configurations of each individual to contribute to the haplotype frequency estimates, as well as to the weighted haplotype score, according to their estimated probabilities. An alternative approach is to perform multiple imputation on 5–10 imputed data sets (Little & Rubin, 2010). Note that each imputed data set has to be drawn from a different posterior distribution to ensure proper multiple imputation. This can be achieved either by imputing from different reference sets (for example, from bootstrap samples of the HapMap or 1000 Genomes reference set) or by drawing from different iteration in a full Bayesian framework in which the model parameters are also up- dated in each iteration. Neither approach had noticeable impact on the *IFIH1* real data set, but further work is warranted.

Both of our proposed tests can be extended to analyze quantitative traits and to accommodate covariates. Both of our tests, in a nutshell, derive one ‘‘genetic score’’ for each individual and assess the association between the genetic score and phenotype of interest. The genetic score is a weighted sum of contributing SNP dosages or haplotypes. Although the weights are defined for dichotomous trait in this work, we can easily extend the work to quantitative traits by first estimating the weights, for the very simple example, via regression, then deriving the genetic score accordingly, and finally performing the association testing. In the above general setting, covariates can be conveniently incorporated.

CHAPTER 4: FUNCTIONAL AND STATISTICAL ANNOTATION

4.1 Introduction

In this chapter, we present an evaluation of multiple weighting schemes through a series of simulations. We evaluate several existing phenotype-independent (Cohen et al., 2004; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007) and -dependent weighting schemes (Wu et al., 2011; Xu et al., 2012), as well as weighting schemes determined by linear regression, penalized regression and variable selection methods, including Lasso (Tibshirani, 1996), Elastic Net (Zou and Hastie, 2005) and SCAD (Xie and Huang, 2009). We conduct simulations under a variety of scenarios with different numbers of true causal variants, mixtures of direction of effect and availability of functional information, mimicking sequencing studies of a quantitative trait. We then apply each of these methods to a set of high coverage targeted sequencing data (Nelson et al., 2012) of 1898 individuals from the CoLaus population-based cohort (Firmann et al., 2008).

4.2 Methods

4.2.1 Statistical Methods

Over the last few years, numerous sensible weighting schemes have been proposed. In most of these methods a genomic region or variant set is assigned a weighted sum over the variants meant to describe the burden of potentially influential variants carried by each individual. We call this weighted sum S_i. Further, we assume there are *N* individuals under study, indexed by *i*, and for each individual we have *M* variants in the region or variant set, indexed by *j*.

First, we examine three approaches that are independent of the observed phenotype. The first of these is a simple indicator of whether or not rare variants (minor allele frequency, MAF < 0.01) are present in the region (Cohen et al., 2004). That is,

$$S_i = I\left(\sum_{j=1}^{M} I(\hat{q}_i < Q)x_{ij} > 0\right)$$

where x_{ij} is the number of minor alleles observed for individual *i* at variant *j*.

$\hat{q}_i = \frac{\sum_{i=1}^{N} x_{ij} + 1}{2N+2}$ is the estimated MAF of variant *j* in the data with pseudo counts and Q is the MAF threshold. In this work, we consider *Q=0.05*.

Second, we examine a count approach which assigns a higher score to individuals carrying a larger number of rare alleles (Morgenthaler and Thilly, 2007);

$$S_i = \sum_{j=1}^{M} I(\hat{q}_j < Q)x_{ij} .$$

with x_{ij} being the count of rare alleles for individual *i* at variant *j* and $\hat{q}_j$ being the estimated MAF, as defined above.

We also consider the approach proposed by Madsen and Browning (Madsen and Browning, 2009) where the weight for variant *j* is a function of the minor allele frequency (MAF):

$$S_i = \sum_{j=1}^{M} \xi_j x_{ij} \text{ , where } \xi_j = \frac{1}{\sqrt{N \times \hat{q} \times (1-\hat{q})}}$$

with x_{ij} and $\hat{q}_j$ as above. In the original Madsen and Browning framework for case-control studies, MAFs are estimated using controls only. However, in this paper, the outcome of interest is quantitative and we estimate MAF using the entire sample, which makes the method phenotype-independent in this context.

We also consider phenotype-dependent regression-based methods. First, we examine the performance of marginal regression coefficients. That is, we fit the simple linear regression model $Y = x_j\beta + \varepsilon$ for each variant j separately and independently and then take the fitted values $\tilde{\beta}_j$ to be our weights.

$$S_i = \sum_{j=1}^{M} \xi_j x_{ij} \text{ , where } \xi_j = \tilde{\beta} \text{, the MLE of } \beta \text{ for the model above.}$$

Though imperfect, this weighting scheme allows investigators to test for associations with multiple rare variants in cases where $N < M$ and begin to follow up on individual variants that may potentially be of interest.

Second, we consider weights from ordinary multiple regression, modeling all of the M variants simultaneously. That is, we fit the model $Y = X\beta + \varepsilon$, where the $(i, j)^{\text{th}}$ element of the matrix $X = x_{ij}$, the minor allele count for individual i at variant j. We then take S_i to be as above, with the fitted values from this multiple regression, $\hat{\beta}_j = \xi_j$ (Lin and Tang, 2011; Xu et al., 2012).

We also consider weights from several variable selection methods. Such methods are appealing since we expect the majority of rare variants not to influence the quantitative trait of interest. Use of penalized regression is therefore expected to reduce the number of non-zero weights. Similar strategies were recently proposed in the context of rare variant association testing (Turkmen & Lin 2012; Zhou, 2010). In penalized regression, we solve for the $\hat{\beta}'s$ which best fit the data, subject to some constraint(s) or penalty. That is, instead of minimizing the sum of squared error, $(Y-\beta X)'(Y-\beta X)$, we aim to minimize the sum of squared errors and an additional penalty term, $(Y-\beta X)'(Y-\beta X)+P(\lambda,\beta)$. In general, the greater the number of parameters included in the model, the greater the penalty. A number of penalty functions have been proposed and extensively studied in the recent statistical literature (Heckman & Ramsay 2000; Hesterberg et al., 2008; Kyung et al., 2010; Wu & Lange 2008). Of these, we chose three: the Lasso which imposes a linear penalty (Tibshirani 1996), Elastic Net (EN) which imposes a quadratic penalty (Zou and Hastie 2005) and SCAD which is designed to penalize smaller coefficients more heavily than larger coefficients (Xie & Huang 2009).

For Lasso and SCAD, only one tuning parameter, λ, is required. We used the R packages *lars* (Efron, Hastie, Johnstone, & Tibshirani, 2004) and *ncvreg* (Breheny Huang 2011) with default parameter values, which is to choose the optimal λ among a grid of 100 possible values equally spaced on the log-scale. For Elastic Net, there are two tuning parameters, one for the linear component and one for the quadratic component. The linear term, λ_1, is chosen in the same way as the λ parameter for the Lasso and SCAD methods, discussed above. The quadratic parameter, λ_2, was set to 1 in all simulations and for the real data. We used the R package *elasticnet* to fit the EN models (Zou & Hastie, 2005). After model fitting, we then use estimated coefficients from each of these variable selection methods as weights. The number of non-zero coefficients included is upper-bounded by 100 for each of these schemes throughout this work.

Under each weighting scheme examined, we determine the significance of a genomic region using a score test of the following form: $U = \sum_{i=1}^{N} (Y_i - \bar{Y}) S_i$ where

$S_i = \sum_{j=1}^{M} \xi_j x_{ij}$ in which N is the number of individuals under study, and Y_i is the quantitative trait value for the i^{th} individual. S_i is the genetic score for the i^{th} individual, a weighted sum across multiple variants. Specifically, x_{ij} is the number of minor alleles observed for individual i at variant j where x_{ij} are not normalized. M is the number of variants in the region under study (discovered through sequencing in our context) and ξ_j is the weight of variant j under one of the above weighting schemes. The analytical distribution for this statistic is not generally known in this context, so significance must be assessed empirically by permutation.

Additionally, we apply the similarity-based method SKAT (Wu et al., 2011) to each of our simulated data sets and the real data set for comparison. We use weights based on the default Beta distribution implemented in the SKAT package, version 0.79. We will comment in the Discussion section on the conceptual differences between the weighting schemes we consider in this work and the SKAT methodology.

4.2.2 Simulation Setup

We simulate 45,000 chromosomes for a series of 100 50Kb regions with a coalescent model [Schaffner et al. 2005] that mimics linkage disequilibrium (LD) in real data, accounts for variations in local recombination rates and models population history consistent with the CEU samples. We then randomly select 2,000 simulated chromosomes (forming 1,000 diploid individuals) to mimic a large sequencing study. For each region, we simulate one single pool of 45,000 chromosomes instead of multiple pools of 2,000 chromosomes so that the causal variants in each region can be determined by population MAFs (MAFs calculated using the entire population of 45,000 chromosomes) and thus retained across replicates from the same region. We assume only rare variants (0.001< population MAF <0.05) influence the value of the quantitative trait and we randomly select *m* variants that truly influence the quantitative trait value. For each variant, we independently assign the direction of influence according to *r*, the probability that a causal variant will increase the trait value. Following (Wu et al., 2011), we then simulate quantitative traits under the null model:

$$y_i = 0.5E_{1i} + 0.5E_{2i} + \varepsilon_i \qquad \text{(Null model)}$$

where E_{1i}, E_{2i} and ε_i are independent with $E_{1i} \sim Bernoulli(0.5)$ to mimic a binary covariate, $E_{2i} \sim Normal(0,1)$ to mimic a continuous covariate, and $\varepsilon_i \sim Normal(0,1)$. We also simulate quantitative traits under an alternative model:

$$y_i = 0.5E_{1i} + 0.5E_{2i} + \sum_{j=1}^{m} \beta x_{ij}^{C} + \varepsilon_i \qquad \text{(Alternative Model)}$$

where $\beta_j = r_j\ |k \times F(MAF_j)|$ and $r_j = 1$ with probability *r* and $r_j = -1$ with probability (1-*r*). E_{1i}, E_{2i} and ε_i are as before, *j* indexes the truly causal variants and x_{ij}^{C} is the number of minor alleles individual *i* has at causal variant *j*. The link function *F* takes one of the following forms:

$$F_{\log}(q) = k \times \log(q),\ F_{\log}(q) = k \times \log\left(\frac{q}{1-q}\right),\ F_{MB}(q) = k \times \frac{1}{\sqrt{q(1-q)}},$$

where *N* is the number of individuals sequenced. We call the first link function log, the second logit, and the third Madsen-Browning (MB). In addition, we also consider $F_{random}(q)$, a random value chosen from the *exponential*(1) distribution, independent of *q* and multiplied by *k*. The constant *k* is a scaling factor to control the magnitude of the change in quantitative trait due to truly causal genetic variants. In our simulations *k* is set to 0.2, which keeps the heritability h^2, between 0.1% and 2.5%. Complex human quantitative traits are thought to have heritability estimates in this range (Manolio et al., 2009). In the Results section, we report the results for the logit link function; results for all four link functions are given in the Appendix A.

To assess significance in each simulated setting, score test statistic from each weighting scheme is compared to the empirical distribution of the test statistic obtained under the null simulations. We assess the significance of each test at the α=0.01 level using the empirical null distribution, which we approximate using 100,000 data sets simulated under the null hypothesis of no variant contributing to the quantitative trait.

For each of the 100 regions we simulate, we randomly select 100 samples of 2,000 chromosomes (forming 1,000 diploid individuals). We then assign quantitative trait values under the null model specified above. Using these 100 × 100=10,000 data sets simulated under the null hypothesis, we obtain the empirical null distribution of the test statistics for each method.

We also simulate data under several null hypotheses. For each choice of *r*, *m* and *F*(.), we select 2,000 chromosomes from the population of 45,000 chromosomes again via simple random sampling. Again, we randomly pair these chromosomes to form diploid individuals and replicate 100 times for each region. For each replicate, we randomly select *m* rare variants to be causal. Each causal variant is assigned a direction in which to exert its effect (positive with probability *r* and negative with probability 1 – *r*).

In order to evaluate the effects functional annotation, we must also simulate these annotations. In each simulated data set, we annotate variants as "functional" or "non-functional". We assume that we have a reasonably good bioinformatics tool such that a true causal variant has 90% probability to be annotated as "functional". Even a perfect bioinformatics tool can only predict functionality, not causality or association with a particular trait of interest. Because of this, we annotate an additional random number of *W* non-causal variants as "functional". Kryukov and colleagues (Kryukov et al., 2009) have estimated that approximately one third of *de novo* missense mutations (that would be predicted as functional by a sensible bioinformatics tool) have no effect on phenotypic traits. We therefore used 1/3 as the lower bound for the fraction of non-causal variants annotated and simulated $W \sim N(25,5)$, rounded to the nearest integer. We evaluate the performance of each of these weighting schemes both using all variants without the help of the bioinformatics tool, and using only the "functional" variants annotated. Under the null distribution, *W* variants are selected at random.

In order to simulate GWAS data sets, we use the same choice of causal variants in each region as in the simulated sequencing data. Consequently, the direction of association and true effect size of each of these are unchanged. In order to simulate GWAS SNPs, we select 1000 chromosomes from the total 45,000 to mimic the 1000 Genomes (Abecasis et al., 2012) sample. The simulated 1000 Genomes sample is used to define LD, based on which GWAS SNPs are selected. For each region, we choose 75 GWAS SNPs consisting of the first 70 tagSNPs (SNPs with the highest number of LD buddies where an LD buddy is a SNP with which the $r^2 > 0.8$) and 5 SNPs at random from the remaining set of SNPs, mimicking the Illumina Omni5 or Affymetrix Axiom high-density SNP genotyping platforms.

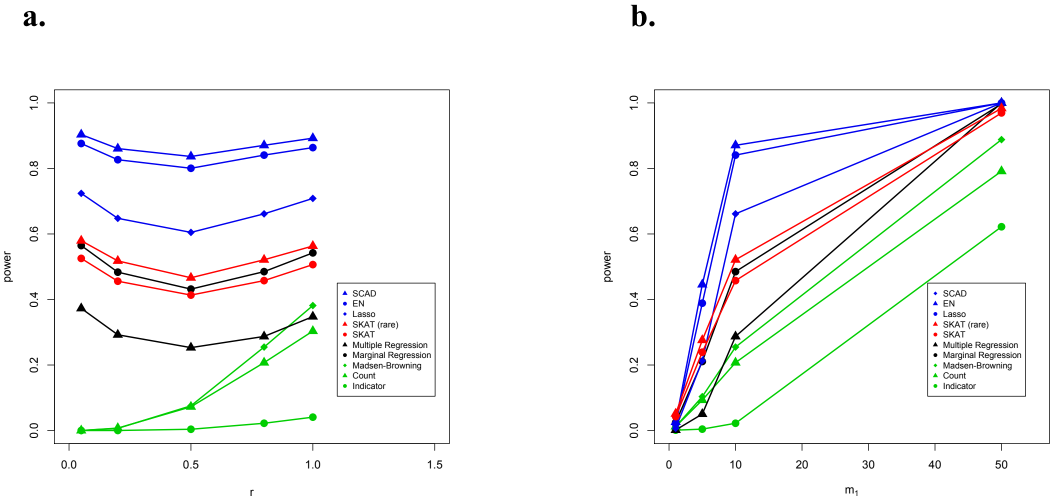
4.3 Results

4.3.1 Results with Simulated Sequencing Data Sets

First, we compare these methods discussed previously in the absence of a bioinformatics tool. Throughout our simulations, we observe several consistent patterns. First, when we apply these methods in the absence of a Bioinformatics tool (thus, all variants are included in analysis), variable selection schemes (most noticeably Lasso and EN) outperform other methods, including SKAT, in nearly all situations (notable exceptions are discussed below). For example, under the simulated setting of 10 causal variants, among which we expect to five increase quantitative trait value, the power is 80.0% and 83.7% for Lasso and EN, and is 0.4%, 7.3%, 7.6%, 43.2%, 25.3%, 60.5%, 41.3%, and 46.6% for Indicator, Count, Madsen-Browning, Marginal Regression, Multiple Regression, SCAD, SKAT (all variants), and SKAT (rare variants only) respectively (Figure 4.1a). Under the simulated setting of 50 causal variants among which 40 are expected to increase quantitative trait value, power is 100% for both Lasso and EN, and is 0.03%, 0.19%, 0.07%, 99.63%, 100%, 100%, 96.9%, and 98.5% for Indicator, Count, Madsen-Browning, Marginal Regression, Multiple Regression, SCAD, SKAT (all variants), and SKAT (rare variants only) respectively (Figure 4.1b).

In the presence of a good bioinformatics tool (as introduced in the Methods section) the power increases for each of the methods previously discussed. Most notably, the phenotype-independent methods show a substantial gain in power once the bioinformatics tool is applied. For example, under the simulated setting of 10 causal variants, among which five are expected to increase the quantitative trait value, the power is 99.83% and 99.80% for Lasso and EN, and is 23.91%, 17.15%, %, 18.85%, 97.87%, 99.73%, 99.76%, 98.49%, and 98.34% for Indicator, Count, Madsen-Browning, Marginal Regression, Multiple Regression, SCAD, SKAT (all variants), and SKAT (rare variants only) respectively (Figure 4.2a). Under the simulated setting of 50 causal variants, among which 40 increase quantitative trait value, power is 100% for both Lasso and EN, and is 99.38%, 98.89%, 96.51%, 100%, 100%, 100%, 100%, and 100% for Indicator, Count, Madsen-Browning, Marginal Regression, Multiple Regression, SCAD, SKAT (all variants), and SKAT (rare variants only) respectively (Figure 4.2b). Although power increases for all methods, the relative performance of the methods changes little from that under the absence of a bioinformatics tool.

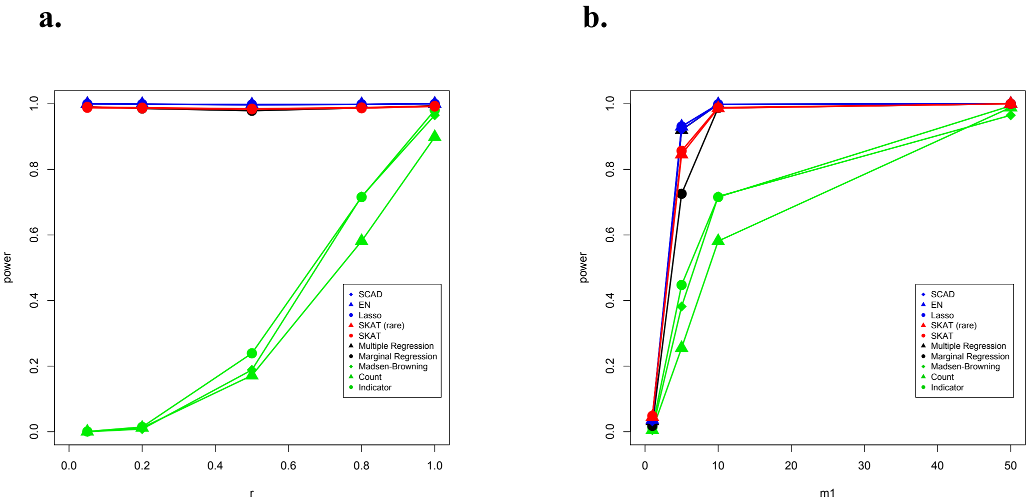
Figure 4.1: Power Comparison in the Absence of a Bioinformatics Tool. Figure 4.1 shows the power (Y-axis) of the different methods across a wide spectrum of *m* (the number of true causal variants) and *r* (the proportion of variants that contribute to our quantitative trait in a positive direction) in the absence of a bioinformatics tool. In Figure 4.1a, we fix *m* at 10 and show power comparisons across the entire spectrum of *r* (X-axis). Figure 4.1b shows how power changes as a function of *m* (X-axis) with *r* fixed at 0.8. Here we use the logit link function.



a.

b.

Figure 4.2: Power Comparison in the Presence of the Good Bioinformatics Tool. Figure 4.2 shows the power (Y-axis) of the different methods across a wide spectrum of *m* (the number of true causal variants) and *r* (the proportion of variants that contribute to our quantitative trait in a positive direction) in the presence of the good bioinformatics tool described in the Method section. Like in Figure 4.1a, we fix *m* at 10 and show power comparisons across the entire spectrum of *r* (X-axis) in Figure 4.2a. Similarly, Figure 4.2b how power of the methods changes as a function of *m* (X-axis) with *r* fixed at 0.8. Again the logit link function is used.



As the number of true causal variants (*m*) increases, so does power for all methods. This is to be expected since adding more causal variants increases the signal-to-noise ratio. When the number of true causal variants is very small, none of the methods have adequate power. Interestingly, it is in these situations where *m* is very small that SKAT manifests its advantage over other methods examined. As *r* gets smaller (that is, the probability that a causal variant will contribute positively to the quantitative trait values gets smaller), the power of the phenotype-independent methods decreases. For example, the phenotype-independent methods have close to 0 power when *r*=0.05; while the phenotype-dependent methods are relatively unaffected by changing values of *r* (Figure 4.1a and Figure 4.2a). We also observe a slight dip in power in all of the phenotype-dependent schemes when *r*=0.5 and no bioinformatics information is used (Figure 4.1a), which is to be expected since the signals from different directions are canceling one another. Similar trends are seen in all simulations with all four link functions (shown supplementary figures 1 and 2).

4.3.2 Weight Estimation & Identification for Individual Variants

Table 4.1 shows the correlation between the true and estimated values of the weights for each method under the simulation settings in which the number of truly causal variants, *m*, is 10 and the proportion of variants contributing in the positive direction, *r*, is 80%. Of note, the correlation between true and estimated weights increases for all methods with the addition of bioinformatics filtering. The Elastic Net and Lasso yield the highest correlations between estimated and true weights, both in situations where we restrict to variants that are likely to be functional (Pearson correlations of 0.285 and 0.355), and when we do not (Pearson correlations of 0.744 and 778).

When using variable selection schemes, we have the opportunity to identify individual causal variants within the region or variant set under study. Figure 4.3 illustrates the accuracy with which the causal variant(s) can be identified by each weighting scheme. Note that the causal variant(s) are not always 100% identified, but in many cases, the causal variant, or a variant in high LD ($r^2 > 0.8$), have estimated non-zero weights. For example, if we fix *m=10*, *r=0.8* and the logit link function, without considering LD buddies, we need to consider the top 696 (109 and 12) variants in order to detect 90% (60%, 30%) of the causal variants using EN (Figure 4.3a); taking LD buddies into consideration, the numbers decrease to 378 (14 and 4) (Figure 4.3b). When we also consider functional information we consider fewer variants and narrow the field to include a higher proportion of truly causal variants. In this case, we need to consider the top 408 (16 and 4) variants in order to detect 90% (60%, 30%) of the causal variants

(Figure 4.3c) without considering LD buddies; with LD buddies taken into consideration, the numbers decrease to 374 (13 and 3) (Figure 4.3d).

Figure 4.3: How Far Down the Ranked List are the Truly Causal Variants when All Variants are Included? Figure 4.3a shows the number of variants that must be considered (Y-axis) in order to catch the top 10%, 20% … 100% of truly causal variants (X-axis) in simulation when all variants are considered. We assume that the variants are ranked in order of significance. These plots aggregate true and estimated weights from all 10,000 replicates of the experiment and once again, we fix *r* at 0.8, *m* at 10 and use the logit link function. Figure 4.3b. takes LD buddies (variants with $r^2 > 0.8$ with causal variant) into consideration. Figure 4.3c. restricts the results from 4.3a. to functional variants only using a good bioinformatics tool. Figure 4.3d. is restricted to functional variants only and takes LD buddies into account.

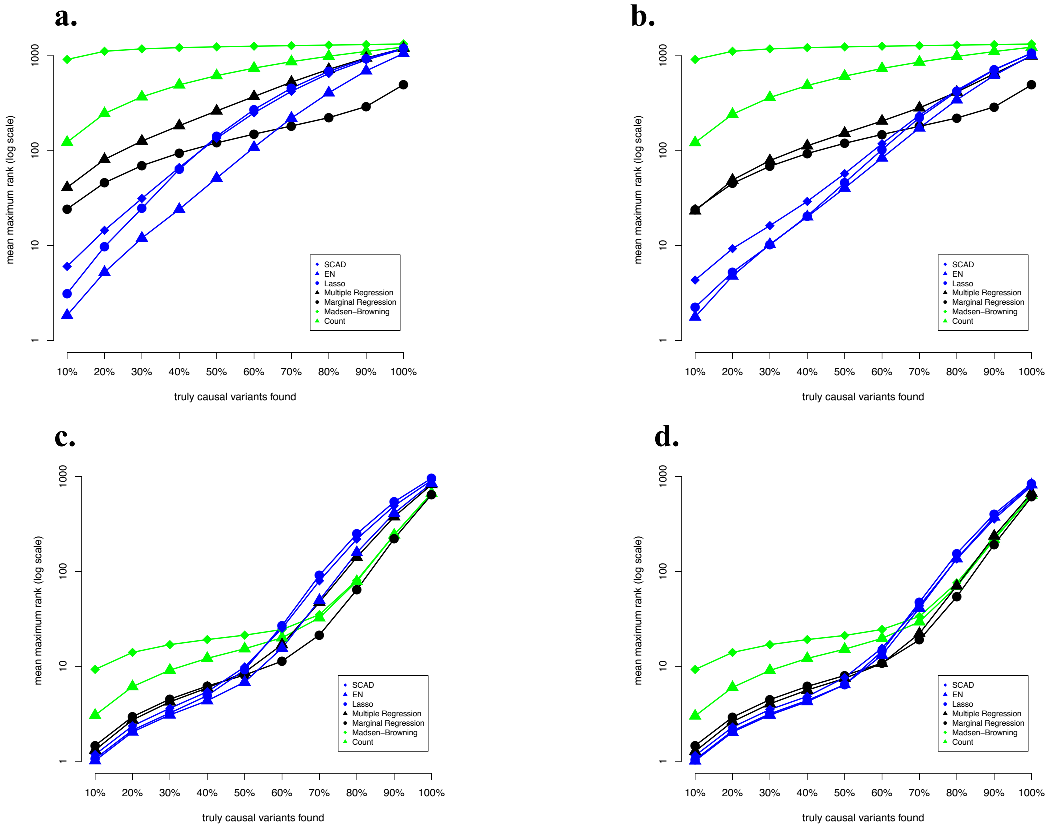


Table 4.1: Average Pearson Correlation of True and Estimated Weights (*m*=10 and *r*=0.8)

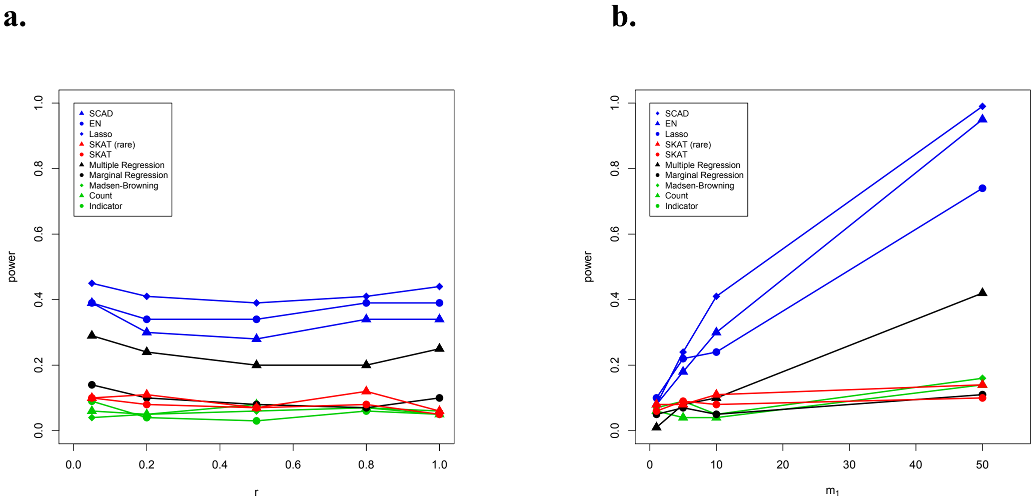
Method	All markers	Limited to functional markers
Indicator	-	-
Count	0.0126	0.2386
Madsen-Browning	0.0591	0.1225
Marginal Regression	0.1588	0.6490
Multiple Regression	0.0883	0.6537
Lasso	0.2852	0.7436
EN	0.3555	0.7787
SCAD	0.2301	0.7344
SKAT (all)	-	-
SKAT (rare only)	-	-

4.3.3 Results with Simulated GWAS Data Sets

Studies that sequence a portion or the entirety of the genome are becoming increasingly common, but still much more GWAS data exist than sequencing data. Imputation has been shown to accurately predict genotypes at untyped variants from GWAS data in a variety of circumstances [Auer et al., 2012; de Bakker et al., 2008; Li et al., 2010a; Li et al., 2009b; Liu et al., 2012; Marchini and Howie, 2010]. Using our simulated GWAS data and simulated reference, we observe that variable selection can improve power for GWAS data as well. However, the power is consistently lower than that under the sequencing setting due to the imperfect rescue of information through imputation (comparing Figure 4.1 with Figure 4.4). In our simulations, the imputation accuracy is 99.66% for all variants and 99.98% for rare variants, but most of the inaccuracies are due to missed rare variants. In fact, among variants with MAF < 0.001 nearly all inaccuracies are due failure to identify the minor allele. Specifically, the squared Pearson correlation between the imputed genotypes (continuous, ranging from 0 to 2) and the true underlying genotypes (coded as 0, 1 and 2) is only 0.2397 for variants with MAF < 0.001.

Supplementary Figure 3 shows the relative power of these weighting schemes over a range of *r* (Figure 4.4a) and *m* (Figure 4.4b).

Figure 4.4: Power Comparison for simulated GWAS data under imputation. Figure 4.4 shows the power achieved in simulated GWAS data (Y-axis) of the different methods across a wide spectrum of *m* (the number of true causal variants) and *r* (the proportion of variants that contribute to our quantitative trait in a positive direction) in the absence of a bioinformatics tool. In Figure 4.4a, we fix *m* at 10 and show power comparisons across the entire spectrum of *r* (X-axis). Supplementary Figure 4.4b shows how power changes as a function of *m* (X-axis) with *r* fixed at 0.8. Here we use the logit link function.



4.3.4 Results with Real Data Set

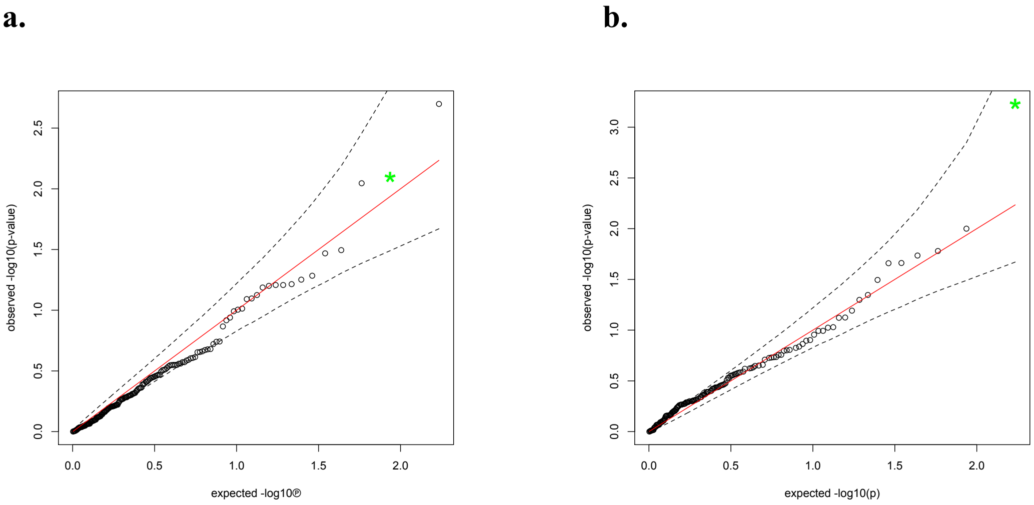
Of the over 6,000 individuals in the CoLaus cohort [Firmann et al., 2008], 1,898 had recorded total cholesterol and targeted sequence data in 202 drug target genes [Nelson et al., 2012]. Sequencing was done at moderately high coverage (with median coverage 27X) and genotype calls were obtained using *SOAP-SNP* [Li et al., 2009a]. Sporadic missing genotypes were imputed with MaCH [Li et al., 2010b]. One gene previously known to be associated with total cholesterol in these data is used as a positive control. We test each of the 172 autosomal genes with and without removing non-functional variants using ANNOVAR [Wang et al., 2010]. For each method, we estimate weights in association with total cholesterol and, for the methods that accommodate covariates, we adjust for age, age^2, sex and the first five principal components. For the phenotype-independent methods, no covariate adjustment is performed and significance is assessed by permutation of the Y_i's. For methods allowing covariates (marginal and multiple regression, Lasso, EN and SCAD), permutation of outcomes alone is not appropriate. For these methods, we fit a regression model, $Y_i \sim Z_i$, where Z is the matrix of covariates and then obtain residuals, ε_i. The ε_i's are then randomly permuted to obtain a set of ε^*_i's, the permuted residuals. For each permutation, we fit the model $\varepsilon^*_i \sim X_i$ in order to re-estimate the weights ξ_j and scores S_i as in [Davidson and Hinkley, 1997]. We do 10,000 such permutations and, from these, obtain a null distribution of statistics with which to assess significance. Since SKAT produces analytical p-values shown to preserve type I error [Wu et al., 2011], we use the SKAT analytical p-values without permutation.

When all variants regardless of bioinformatics prediction are included, the variable selection methods Lasso and EN yield the smallest p-values compared to other methods for the previously implicated gene. However, the previously implicated gene is not the most significant among the 172 genes tested. Using *ANNOVAR* annotations [Wang et al., 2010], we restrict to non-synonymous variants in coding regions of the genome only. When considering only these functional variants, most weighting schemes identify the correct gene with highly significant p-values (Table 4.2 and Figure 4.5).

Table 4.2: Permuted p-values on positive control gene in the real data set

Method	All variants (491)	Limited to functional variants (13)
Indicator	0.208	0.00057
Count	0.068	***0.00017***
Madsen-Browning	0.090	0.00041
Marginal Regression	0.166	0.00420
Multiple Regression	0.136	0.00395
Lasso	0.017	0.00053
EN	0.008	0.00059
SCAD	0.111	0.00078
SKAT (all)	0.329	0.00142
SKAT (rare only)	0.348	0.00142

Figure 4.5: QQ-Plots for p-values in real data. Figure 4.5 shows observed (Y-axis) vs. expected (X-axis) $-\log_{10}(p)$ values for 172 genes in real data. These p-values are computed from the Elastic Net weighting scheme and 10,000 permutations. The gene previously implicated in total cholesterol is shown with a green star, all other genes are represented by black circles.



4.4 Discussion

In summary, through extensive simulation studies with varying number, model, and direction of causal variant(s) contributing to a quantitative trait, we find that functional annotations derived from good set of bioinformatics tools can substantially boost power for rare variant association testing. In the absence of good bioinformatics tools, "statistical" annotation based on phenotype-dependent weighting of the variants, particularly through variable selection based methods to both select potentially causal/associated variants and estimate their effect sizes, manifests advantages. This observation holds for both sequencing-based studies or studies based on a combination of genotyping, sequencing, and imputation. We find additional supporting evidence from application to a real sequencing-based data set.

The price one has to pay for adopting phenotype-dependent methods is the necessity of permutation, which can be easily performed through permuting of residuals for the analysis of quantitative traits (Davidson and Hinkley 1997; Lin 2005) or using the BiasedUrn method (Epstein et al., 2012) recently proposed for binary traits. This, in turn, increases computational costs. Therefore, we recommend primarily using phenotype-dependent weighting for refining the level of significance. That is, we recommend applying phenotype-dependent weighting only to genomic regions or variant sets that have strong evidence of association (but not necessarily reaching genome-wide significance) from methods that do not require permutation (for example, SKAT (Wu et al., 2011)).

We note that testing over a region by aggregating information across variants is a different task from estimating effect sizes of individual variant (as measured by the variant weights in our work). Perfection in the latter (that is, being able to estimate weights for each individual variants accurately) leads to perfection in the former (that is, maximal testing power over the region harboring those variants), but not vice versa. Based on our simulations where we know the true contribution (effect size) of each individual variant, we find that individual effect sizes cannot be well estimated (Pearson correlation between true and estimated effect sizes < 0.5 even for the best variable selection based methods). However, these methods can still increase power of region or variant set association analysis without accurate estimation of individual variant effect sizes. In addition, these methods are able to identify the vast majority of the causal variants, particularly when LD buddies are considered.

In this chapter, we mainly consider aggregation of information at the genotype level (where we first obtain a regional genotype score via a weighted sum of genotype scores for individual variants and then assess the association between the regional genotype score and the phenotype of interest), which underlies the largest number of rare variant association methods published. In contrast, there are methods that aggregate information at the effect size level (for example, SKAT (Wu et al., 2011) where the final regional score test statistic is a weighted sum of the test statistics for individual variants) or at the p-value level, for example in (Cheung et al., 2012). Our comparisons with SKAT suggest that the same conclusions apply to aggregation methods at levels other than genotype.

Lastly, although one could potentially argue that the phenotype-dependent methods require an undesirable computing-power trade-off in the presence of good bioinformatics tools, in practice, we rarely (if ever) get perfect bioinformatics tools. In addition, even perfect bioinformatics tools can only predict functionality but NOT causality or association with particular phenotypic trait(s) of interest. Therefore, we view that the application of "statistical annotation" through phenotype-dependent weighting, particularly using variable selection based methods, to top regions or variant sets implicated by computationally efficient phenotype-independent methods, is valuable.

CHAPTER 5: SKAT-ADMIX

5.1 Introduction

Allele frequencies can differ greatly between populations. This phenomenon is called population stratification and, if not properly controlled for, population stratification can lead to either false positive or false negative findings. (Choudhry et al., 2006; Freedman et al., 2004) A plethora of methods for adjusting for population stratification among common variants have been proposed (Epstein, Allen, & Satten, 2007; M. Li, Reilly, Rader, & Wang, 2010; Montana & Pritchard, 2004; Price et al., 2006); however, (Mathieson & McVean, 2012) show significant evidence that rare variants show stronger population stratification than common variants. Further, (Mathieson & McVean, 2012) also demonstrate that current methods do not adequately account for population stratification among rare variants.

When dealing with genetic data from an admixed population (that is, a population composed of two or more distinct ancestral populations, e.g. African American or Hispanic populations), adjusting for population stratification is is especially crucial. In admixed populations, information from more than one ancestral population in contained within a single individual. Earlier this year, Mao et. al. developed an approach based on the WHaIT method discussed in chapter 3 (Y. Li, Byrnes, et al., 2010) to account take admixture into account when assessing association between a phenotype and a genomic region (Mao, Li, Liu, Lange, & Li, 2013). Mao et. al. combine WHaIT with the Hapmix approach for estimating “local ancestry,” that is, the ancestry for each individual at each variant locus. This process results in a SKAT statistic and p-value for each ancestral population. If desired, the SKAT statistics can be combined using a process similar to that used in MetaSKAT (Ionita-Laza, Lee, Makarov, Buxbaum, & Lin, 2013) and SKAT-O (Lee et al., 2012)

5.2 Methods

5.2.1 Splitting SKAT

Consider a genomic region with M variants typed on N individuals. The $N \times M$ genotype matrix can be denoted X with *(i,j)*th element x_{ij}, the number of minor alleles at locus *j* in individual *i*. However, in an admixed population of P ancestral populations, $X = \sum_{l=1}^{P} X_l$, where X_l has elements x_{lij}, the number of minor alleles from ancestral population *l*. Throughout this chapter, we will consider only two-way admixture in African Americans, so we will consider $X = X_A + X_E$, where X_A is the minor allele count matrix for alleles of African origin and X_E for the alleles of European origin.

Using a variety of previously proposed ancestry estimation methods, we estimate the minor allele counts from each parent population at each locus for each individual. Where, in truth, the elements of X_A and X_E are 0, 1 or 2, Hapmix outputs a probability for each possible combination of ancestry. In order to estimate the minor allele counts from each parent population *l*, we consider,

$\hat{x}_{lij} = P[\text{exactly one minor allele from population l}] + 2P[\text{two minor alleles from population l}]$

which is continuous in [0,2]. Note that for these preliminary findings, no threshold on the probability has been imposed. We then use the estimated matrices $\hat{X}_{Aij}$ and $\hat{X}_{Eij}$ as input for SKAT (Wu et al., 2011), rather than the genotype matrix.

5.2.2 Ancestry Estimation

The first method examined to estimate local ancestry, “diploid Hapmix,” directly estimates the probability of each possible genotype and local ancestry from the genotype data using a panel of reference haplotypes from each ancestral population by Hapmix. (Price, et. al. 2009) Second, we restrict the diploid Hapmix probabilities to include only those that are compatible with the input genotypes. This strategy is motivated by the tendency of Hapmix to occasionally miss rare variants because they are indistinguishable from sequencing errors. Next, we add a phasing step prior to local ancestry estimation. We use MaCH (Y. Li, Willer, et al., 2010) to infer phase information, using an external reference haplotype panel. After phasing is complete, we use Hapmix (this time in Haploid mode) to estimate the local ancestry of each locus. For comparison, we also use MaCH-Admix (Liu et al., 2013) to estimate local ancestry of each locus for the phased data only.

5.2.3 Simulation Setup

In order to completely evaluate our simulation, we must simulate African American chromosomes for which the ancestral population is known for each locus. To accomplish this, we first simulate 3000 European and 3000 African chromosomes using COSI (Schaffner et al., 2005). Of these, 1000 from each group are set aside to serve as a reference panel for the ancestral populations. The remaining chromosomes are used to construct simulated 2000 African American chromosomes in concordance with the population history of African Americans. First, we construct 425 African American chromosomes containing “switchpoints,” that is cross overs between African and European ancestry chromosomes. 425 chromosomes from each ancestry group are combined at a randomly determined locus in the region. The crossover map, provided by COSI, provides weights for these random assignments and the ancestry of the “first” chromosome segment is also determined at random. The remaining 1575 African American population chromosomes are pulled from the pool of unused chromosomes so far, 1408 African and 167 European. These rates of “switchpoint” occurrence and proportion European Ancestry are consistent with the findings presented in (Wegmann et al., 2011) and (Parra et al., 1998). These 2000 African American chromosomes are randomly paired to form 1000 diploid individuals. 100 replicates of this process were preformed.

Causal variants are chosen from rare variants (MAF<0.05) within one ancestral population, either African or European. Though some variants are rare in one ancestral population, and not in the other, we consider a variant to be rare, and thus eligible for being causal, if it is rare in the population in question. We choose *m* such variants from each ancestral population; values of *m* considered are 1, 2, 3, 4, 5, 10, 20, 30, 40 and 50. Each variant can also contribute to the quantitative trait of interest in the positive or negative direction. The probability that a given causal variant contributes in the positive direction is denoted by *r*; values of *r* considered are 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0.

Let the true weight, $w_j = k \times \log\left(\frac{q_j}{1-q_j}\right)$ where q_j is the MAF of causal marker j. Let the set of causal markers be denoted M^C and the population from which the causal variants come be denoted *k*. So, we simulate quantitative trait under the alternative for individual *i* to be, $QT_i = 0.5E_1 + 0.5E_2 + \sum_{j \in M^C} w_j x_{lik}$, where E_1 ~ Bernoulli(0.5) and E_2 ~ Normal(0,1). Where E_1 and E_2 are independent of one another and across individuals. Similarly, for the null simulations used to assess type I error, we simulate $QT_i = 0.5E_1 + 0.5E_2$ as in (Wu et al., 2011).

5.2.4 Real Data

We also apply these methods, along with the original SKAT, to the African American samples from the HeartGO data, part of the Exome Sequencing Project, in the *APOB* gene and 1kb flanking region, which has been previously shown to have a population-specific association to LDL cholesterol levels. (Mao et al., 2013) We use 1000 Genomes reference panels (1000 Genomes Consortium et al., 2010) for European (758 samples) and African (492 samples) populations for ancestry estimation. We adjust for the covariates, age, sex, BMI, smoking status, and the first 10 principal components. There are 7075 samples with valid LDL cholesterol and covariate information and 895 polymorphic SNPs.

5.3 Results

5.3.1 Type I Error

First, we examine type I error for this method in simulation using the 100 simulated replicates, with 100 null simulations per replicate, resulting in 10,000 data sets. We find that type I error is conserved for both European and African causal alleles as illustrated in Table 5.1.

Table 5.1: Type I Error rates of separate SKAT tests for all methods.

Test For	SKAT	Diploid Hapmix + SKAT	Diploid Hapmix + Genotype Adjustment + SKAT	MaCH Phasing + Hapmix + SKAT	MaCH Phasing + MaCH-Admix + SKAT
European Causal Alleles	0.0491	0.0469	0.0384	0.0466	0.0562
African Causal Alleles		0.0495	0.0500	0.0447	0.0538

5.3.2 Power

We compare power of this ancestry-specific analysis compared to the original SKAT approach. We find that the power to detect causal alleles of European origin is greatly improved compared to the original SKAT method (Figure 5.1), however, when the causal alleles are of African origin, the methods are more comparable. It appears that the causal African alleles produce signal enough that they can be picked up by the SKAT approach without adjusting the data for ancestry (Figure 5.2). The complete results for these experiments are shown in Appendix B, figures 1 through 4.

Figure 5.1: Power for European causal variants. Figure 5.1a shows power (Y-axis) for SKAT Admix and the proposed methods over the number of causal variants m (X-axis) when r is fixed at 0.8 and all of the causal variants are of European origin. Figure 5.1b similarly shows power (Y-axis) over the proportion of variants contributing in the positive direction r (X-axis) for 10 causal variants. For the proposed methods, α=0.025 to adjust for multiple comparisons and for SKAT α=0.05.

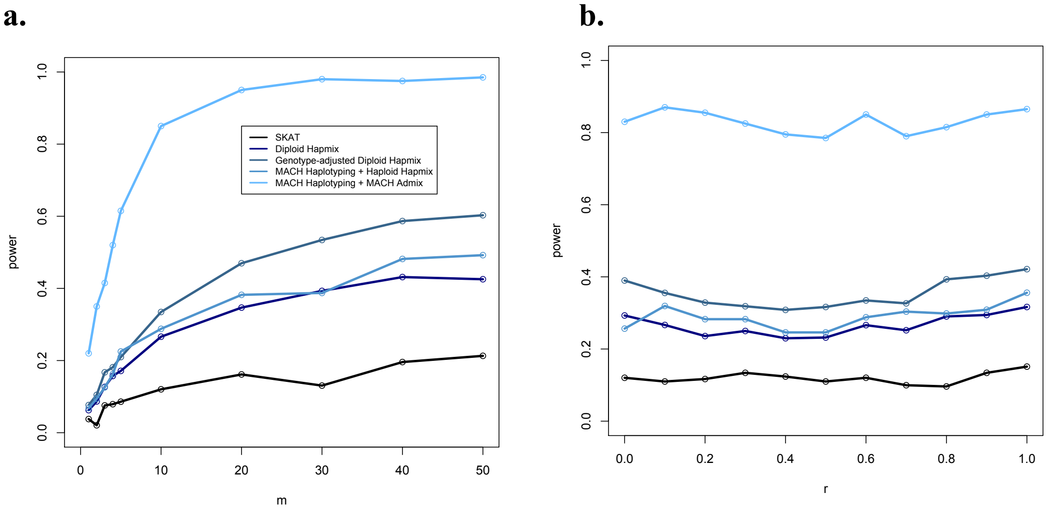
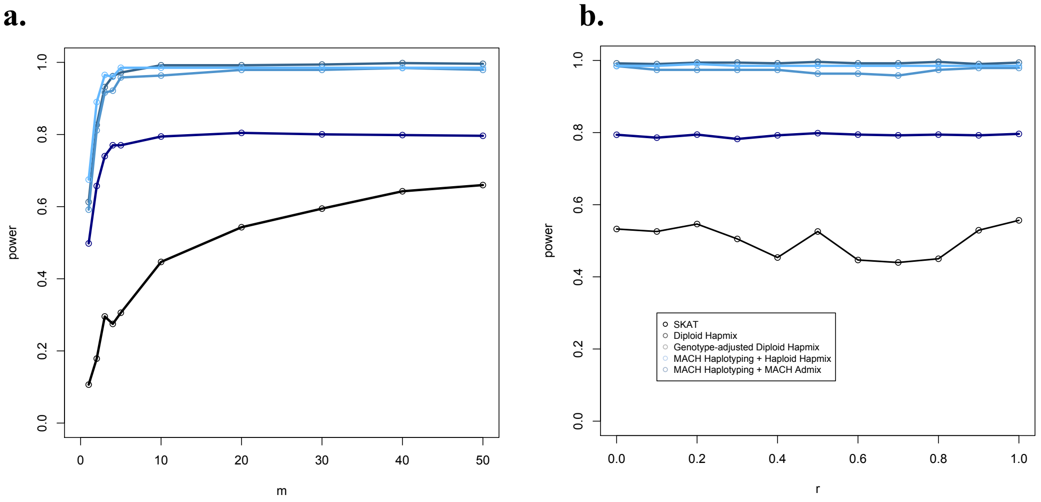


Figure 5.2: Power for African causal variants. Figure 5.2a shows power (Y-axis) for SKAT and the proposed methods over the number of causal variants m (X-axis) when the proportion of causal variants that contribute in the positive direction, r, is fixed at 0.8 and all of the causal variants are of African origin. Figure 5.2b similarly shows power (Y-axis) over the proportion of variants contributing in the positive direction r (X-axis) for 10 causal variants. For the proposed methods, α=0.025 to adjust for multiple comparisons and for SKAT α=0.05.



5.3.3 Real Data

The *APOB* gene has been previously implicated in LDL-cholesterol levels in African Americans (Mao et al., 2013) and the effect is suspected of being European-specific. (Mao et al., 2013) Here, we implement each of the proposed strategies in the *APOB* gene and 10kb flanking region to 7,075 of the African American samples in the HeartGO data for association with LDL cholesterol. The HeartGO data are whole exome data with >10x coverage exome-wide. We adjust for sex, age, BMI, smoking status and the first 10 principal components. P-values for each of the proposed methods are shown in Table 5.2.

Table 5.2: P-values for individual SKAT tests for *APOB* in African Americans in association with LDL-cholesterol.

Test For	SKAT	Diploid Hapmix + SKAT	Diploid Hapmix + Genotype Adjustment + SKAT	MaCH Phasing + Hapmix + SKAT	MaCH Phasing + MaCH-Admix + SKAT
European Causal Alleles	0.0818	2.61×10^{-8}	1.40×10^{-8}	2.31×10^{-8}	8.23×10^{-9}
African Causal Alleles		0.311	0.300	0.327	0.331

5.4 Discussion

The results presented above demonstrate that the power of SKAT to detect ancestral population-specific associations is greatly improved by first directly estimating local ancestry and running SKAT, either separately or in combination. Type I error is well preserved in the individual tests and in the proposed combined test. The real data example of LDL-cholesterol supports these findings and was able to significantly replicate the findings of (Mao et al., 2013) for a European ancestry-specific effect of rare variants in the gene *APOB*. As we might expect, association with rare variants from the European population is more difficult than association with rare variants from the African population since the majority of the samples have primarily African ancestral alleles. Further, when the true causal variants are from the European population, haplotyping and using MaCH-Admix to estimate ancestry also improves power a great deal by providing more accurate ancestry estimation. The observed gain in power comes at a computational cost, since we must use MaCH to haplotype the genotype data and then run MaCH-Admix to estimate ancestry, which takes approximately two times the computational time as running Hapmix. Looking forward, combining the processes of phasing and ancestry estimation may improve computational time and allow for the implementation of MaCH-Admix in this context on a larger scale.

CHAPTER 6: CONCLUDING REMARKS

This document presents several novel methods for aggregation of rare variants within a genomic region and makes reference to many more. While each method is intended for a specific study design and involves a variety of statistical and computational tools, the central goal remains the same: to maximize and aggregate signal from truly associated rare variants and to minimize noise from sequence variation that does not contribute to the trait of interest. We have demonstrated that various statistical methodology can be used to aggregate across rare variants to improve power to detect associations while maintaining acceptable type I error rates. In GWAS data, we found that adding information from external sequencing data via imputation can improve power to detect associations between human traits and rare variants not typed by GWAS, particularly when we restrict the variants considered to those which are likely to have an effect and when we adjust the sign. In sequence data, we found that directly estimating the contribution of each variant greatly improves power over existing methods, however the estimation of this contribution is not generally accurate at the marker level and is comparatively quite computationally intensive. Similarly, we adapted the SKAT approach for admixed populations and found that the most computationally intensive methods of estimating ancestry perform best and consequently lead to the highest power, however these methods are also the most computationally intensive.

In general, the proposed methods which control for more confounders and estimate genetic effects most directly have the most statistical power to detect associations. However, while more computationally intensive methods tend to perform better, they generally come at a computational cost. As molecular technologies become more accurate and less cost-prohibitive, many investigators are aiming to run tests like these on many regions on the genome. For this reason, the balance between computational price and method performance must be considered for these methods in order to ensure that they get used and serve their intended purpose.

APPENDIX A: SUPPLEMENTARY FIGURES FROM CHAPTER 4

Figure A.1: Complete power results for all link functions and all values of *m*

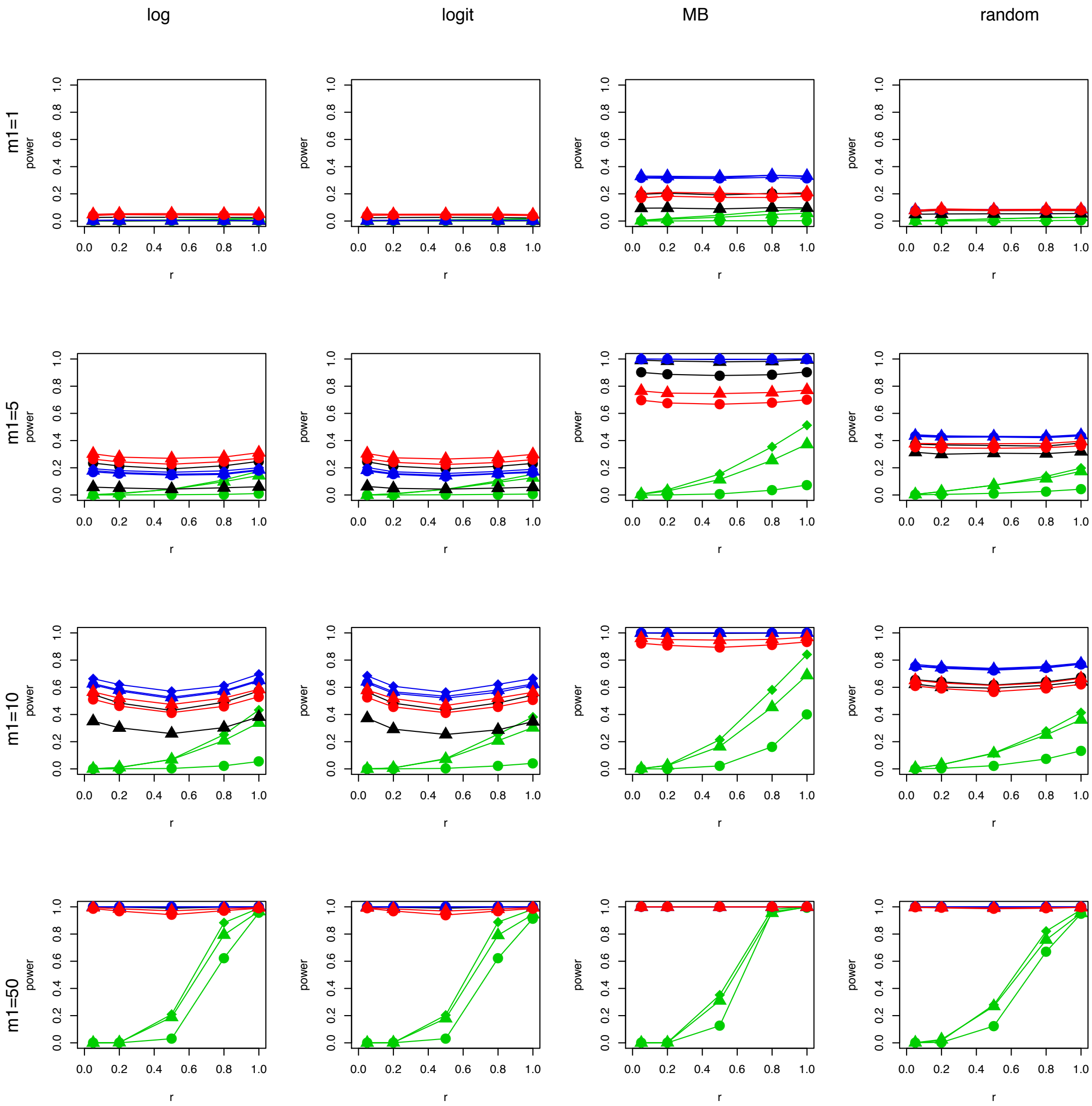
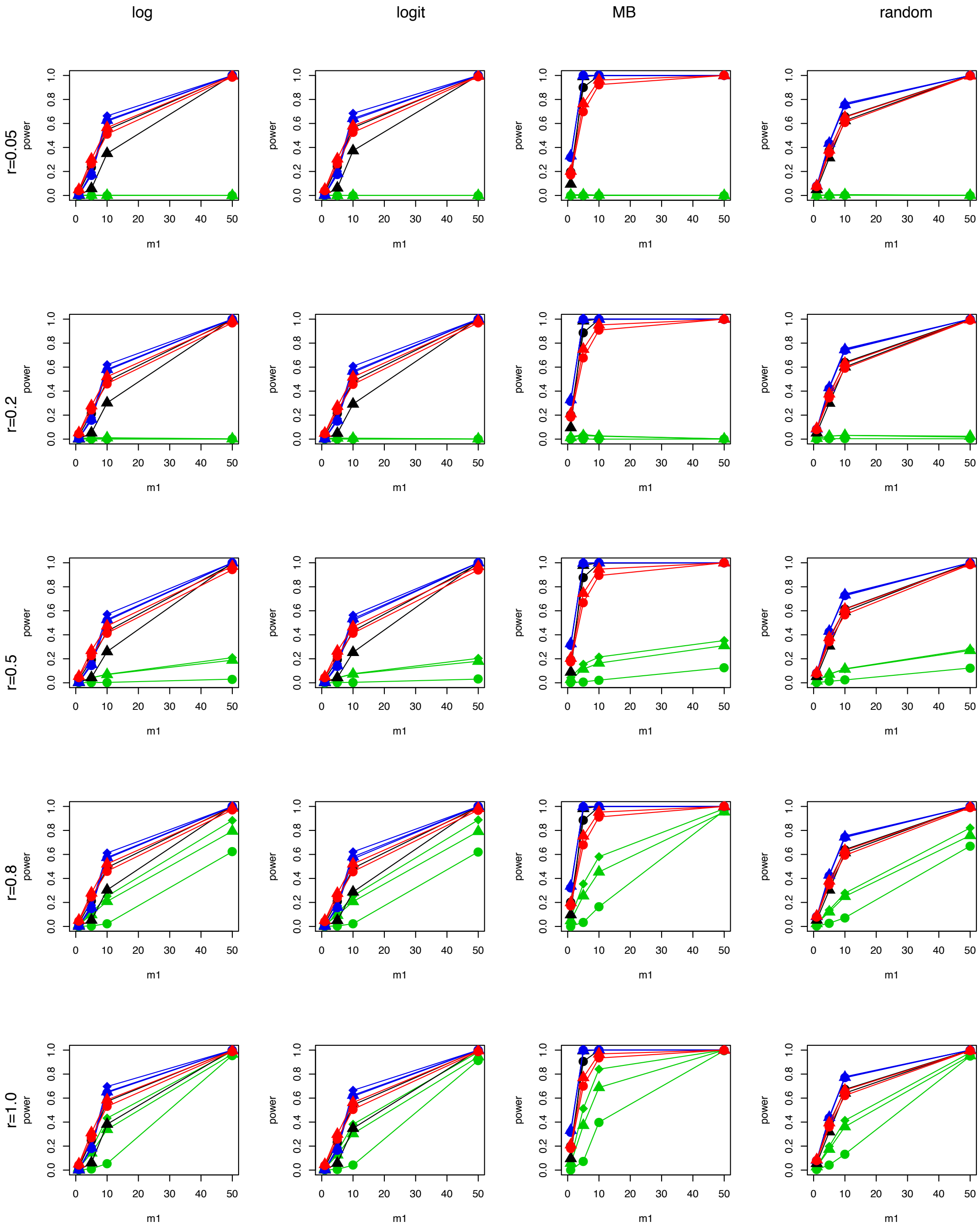


Figure A.2: Complete power results for all linking functions and all values of *r*.



APPENDIX B: SUPPLEMENTARY FIGURES FROM CHAPTER 5

Figure B.1: Complete power results for African causal alleles over all values of *m*.

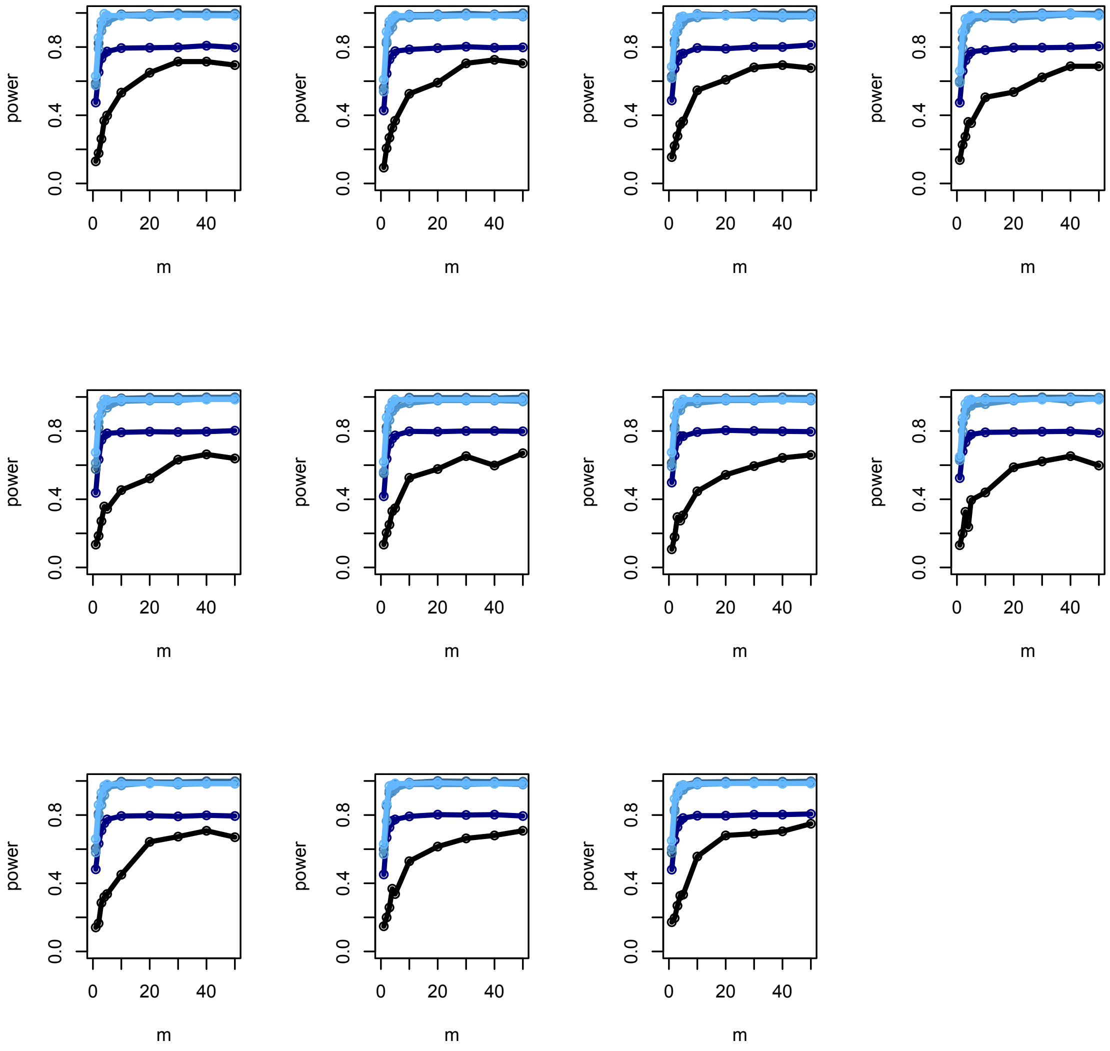


Figure B.2: Complete power results for African causal alleles over all values of *r*.

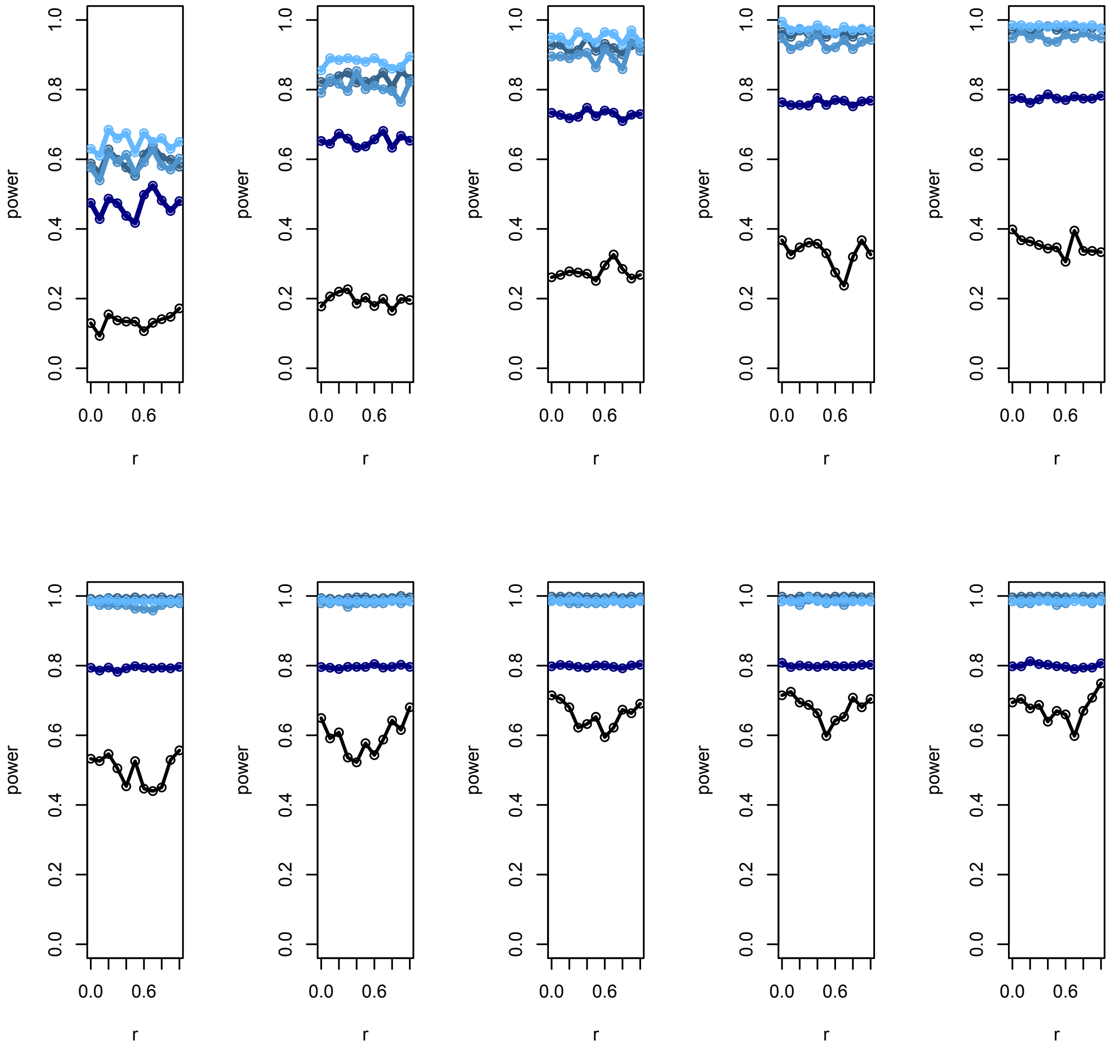


Figure B.3: Complete power results for European causal alleles over all values of *m*.

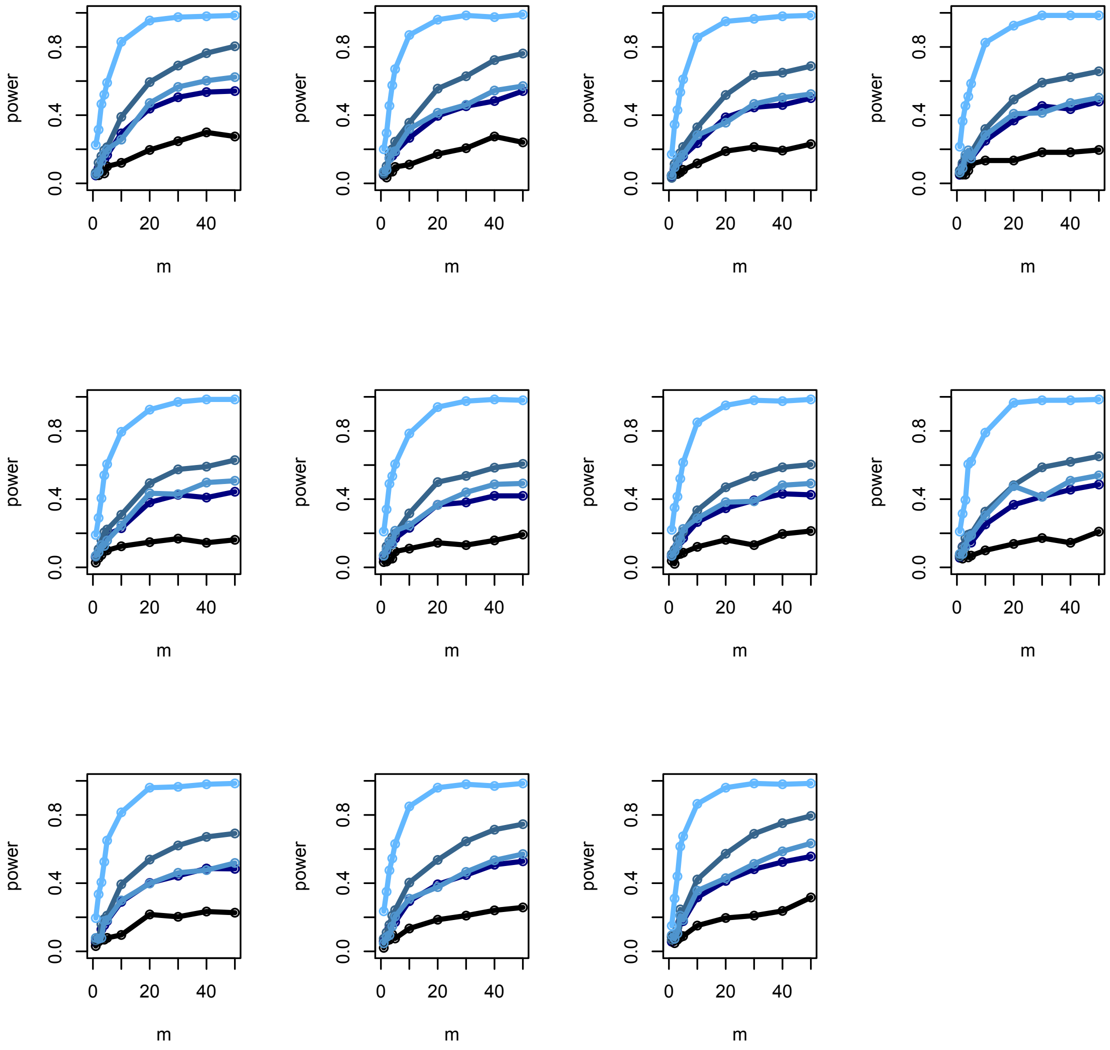
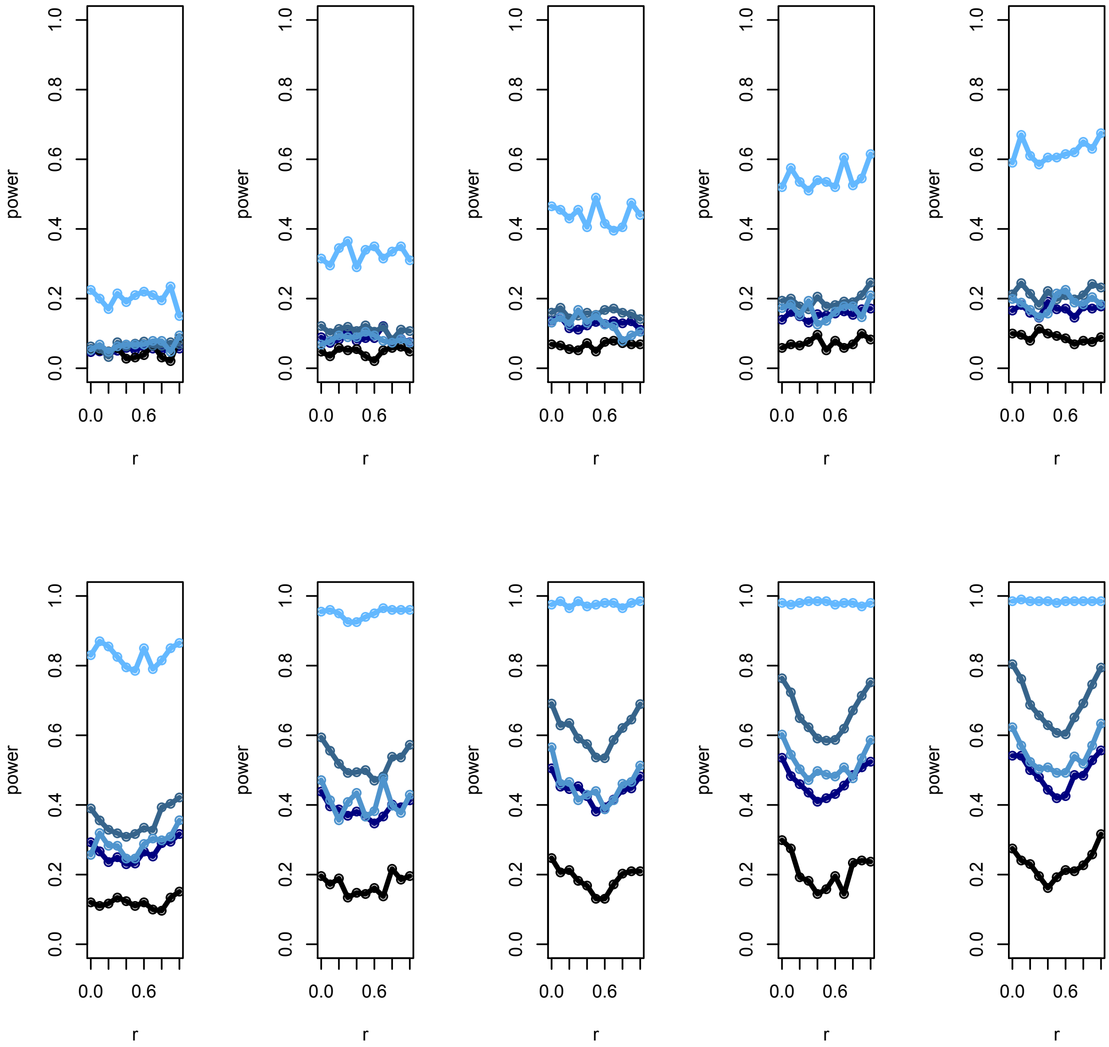


Figure B.4: Complete power results for European causal alleles over all values of *r*.



REFERENCES

1000 Genomes Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., … McVean, G. a. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–73. doi:10.1038/nature09534

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. a, … Rich, S. S. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics*, *41*(6), 703–7. doi:10.1038/ng.381

Breheny, P., & Huang, J. (2011). Coordinate Descent Algorithms for Nonconvex Penalized Regression, With Applications To Biological Feature Selection. *The annals of applied statistics*, *5*(1), 232–253. doi:10.1214/10-AOAS388

Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *American journal of human genetics*, *78*(6), 903–13. doi:10.1086/503876

Choudhry, S., Coyle, N. E., Tang, H., Salari, K., Lind, D., Clark, S. L., … Burchard, E. G. (2006). Population stratification confounds genetic association studies among Latinos. *Human genetics*, *118*(5), 652–64. doi:10.1007/s00439-005-0071-3

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., & Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science (New York, N.Y.)*, *305*(5685), 869–72. doi:10.1126/science.1099870

Davidson, A. C., & Hinkley, D. V. (1997). *Bootstrap*. doi:10.1007/SpringerReference_9179

De Bakker, P. I. W., Ferreira, M. a R., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human molecular genetics*, *17*(R2), R122–8. doi:10.1093/hmg/ddn288

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). LEAST ANGLE REGRESSION. *The Annals of Statistics*, *32*(2), 407–499.

Epstein, M. P., Allen, A. S., & Satten, G. a. (2007). A simple and improved correction for population stratification in case-control studies. *American journal of human genetics*, *80*(5), 921–30. doi:10.1086/516842

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., … McCarthy, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (New York, N.Y.)*, *316*(5826), 889–94. doi:10.1126/science.1141634

Frazer, K. a, Ballinger, D. G., Cox, D. R., Hinds, D. a, Stuve, L. L., Gibbs, R. a, … Leal, S. M. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–61. doi:10.1038/nature06258

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. a, Patterson, N., … Altshuler, D. (2004). Assessing the impact of population stratification on genetic association studies. *Nature genetics*, *36*(4), 388–93. doi:10.1038/ng1333

Goldstein, D. B. (2011). The importance of synthetic associations will only be resolved empirically. *PLoS biology*, *9*(1), e1001008. doi:10.1371/journal.pbio.1001008

Han, F., & Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human heredity*, *70*(1), 42–54. doi:10.1159/000288704

International, T., & Consortium, H. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–320. doi:10.1038/nature04226

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., & Lin, X. (2013). Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. *American journal of human genetics*, *92*(6), 841–853. doi:10.1016/j.ajhg.2013.04.015

Kaiser, J. (2008). A Plan to Capture Human Diversity in 1000 Genomes. *Science (New York, N.Y.)*, *319*(January), 395.

Kryukov, G. V, Shpunt, A., Stamatoyannopoulos, J. a, & Sunyaev, S. R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(10), 3871–6. doi:10.1073/pnas.0812824106

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. a, … Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American journal of human genetics*, *91*(2), 224–37. doi:10.1016/j.ajhg.2012.06.007

Li, B., & Leal, S. M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases : Application to Analysis of Sequence Data. *American Journal of Human Genetics*, *83*, 311–321. doi:10.1016/j.ajhg.2008.06.024.

Li, M., Reilly, M. P., Rader, D. J., & Wang, L.-S. (2010). Correcting population stratification in genetic association studies using a phylogenetic approach. *Bioinformatics (Oxford, England)*, *26*(6), 798–806. doi:10.1093/bioinformatics/btq025

Li, M., Wang, K., Grant, S. F. a, Hakonarson, H., & Li, C. (2009, February 15). ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics (Oxford, England)*. doi:10.1093/bioinformatics/btn641

Li, Y., Byrnes, A. E., & Li, M. (2010). To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. *American journal of human genetics*, *87*(5), 728–35. doi:10.1016/j.ajhg.2010.10.014

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, *34*(8), 816–34. doi:10.1002/gepi.20533

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual review of genomics and human genetics*, *10*, 387–406. doi:10.1146/annurev.genom.9.081307.164242

Lin, D.-Y., & Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *American journal of human genetics*, *89*(3), 354–67. doi:10.1016/j.ajhg.2011.07.015

Liu, E. Y., Li, M., Wang, W., & Li, Y. (2013). MaCH-admix: genotype imputation for admixed populations. *Genetic epidemiology*, *37*(1), 25–37. doi:10.1002/gepi.21690

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, *5*(2), e1000384. doi:10.1371/journal.pgen.1000384

Maher, B. (2008). The case of the missing heritability. *Nature*, *456*(November). Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=20806391

Manolio, T., Collins, F., Cox, N., Goldstein, D., Hindorff, L., Hunter, D., … Visscher, P. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. doi:10.1038/nature08494.Finding

Mao, X., Li, Y., Liu, Y., Lange, L., & Li, M. (2013). Testing genetic association with rare variants in admixed populations. *Genetic epidemiology*, *37*(1), 38–47. doi:10.1002/gepi.21687

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature reviews. Genetics*, *11*(7), 499–511. doi:10.1038/nrg2796

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, *44*(3), 243–6. doi:10.1038/ng.1074

Montana, G., & Pritchard, J. K. (2004). Statistical tests for admixture mapping with case-control and cases-only data. *American journal of human genetics*, *75*(5), 771–89. doi:10.1086/425281

Morgenthaler, S., & Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation research*, *615*(1-2), 28–56. doi:10.1016/j.mrfmmm.2006.09.003

Morris, A. P., & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, *34*(2), 188–93. doi:10.1002/gepi.20450

Neale, B. M., Rivas, M. a, Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., … Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, *7*(3), e1001322. doi:10.1371/journal.pgen.1001322

Nejentsev, S., Walker, N., Riches, D., Egholm, M., & Todd, J. A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, Protect Against Type 1 Diabetes, *333*(April), 387–389.

Ohno, S. (1972). An argument for the genetic simplicity of man and other mammals. *Journal of Human Evolution*, *1*(6), 651–662. doi:10.1016/0047-2484(72)90011-5

Parra, E. J., Marcini, a, Akey, J., Martinson, J., Batzer, M. a, Cooper, R., … Shriver, M. D. (1998). Estimating African American admixture proportions by use of population-specific alleles. *American journal of human genetics*, *63*(6), 1839–51. doi:10.1086/302148

Price, A. L., Kryukov, G. V, de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J., & Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *American journal of human genetics*, *86*(6), 832–8. doi:10.1016/j.ajhg.2010.04.005

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. a, & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904–9. doi:10.1038/ng1847

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., … Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, *5*(6), e1000519. doi:10.1371/journal.pgen.1000519

Ramensky, V., Bork, P., & Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. *Nucleic acids research*, *30*(17), 3894–900. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=137415&tool=pmcentrez&rendertype=abstract

Schaffner, S., Foo, C., & Gabriel, S. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome …*, *15*, 1576–1583. doi:10.1101/gr.3709305.

Schaid, D. J., Mcdonnell, S. K., Hebbring, S. J., Cunningham, J. M., & Thibodeau, S. N. (2005). Nonparametric Tests of Association of Multiple Genes with Human Disease. *American journal of human genetics*, *76*, 780–793.

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics*, *78*(4), 629–44. doi:10.1086/502802

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., … Jackson, A. U. (2013). A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science (New York, N.Y.)*, *316*, 1341–1345.

Smyth, D. J., Cooper, J. D., Bailey, R., Field, S., Burren, O., Smink, L. J., … Todd, J. a. (2006). A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nature genetics*, *38*(6), 617–9. doi:10.1038/ng1800

Stephens, M., & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American journal of human genetics*, *76*(3), 449–62. doi:10.1086/428594

Sullivan, P. (2012). Don't give up on GWAS. *Molecular psychiatry*, *17*(1), 2–3. doi:10.1038/mp.2011.94

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *58*(1), 267–288. Retrieved from http://www.jstor.org/stable/10.2307/2346178

Turkmen, A., & Lin, S. (2012). An optimum projection and noise reduction approach for detecting rare and common variants associated with complex diseases. *Human heredity*, *74*(1), 51–60. doi:10.1159/000343797

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, *38*(16), e164. doi:10.1093/nar/gkq603

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., … Frayling, T. M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics*, *40*(5), 575–83. doi:10.1038/ng.121

Wegmann, D., Kessner, D. E., Veeramah, K. R., Mathias, R. a, Nicolae, D. L., Yanek, L. R., … Novembre, J. (2011). Recombination rates in admixed individuals identified by ancestry-based inference. *Nature genetics*, *43*(9), 847–53. doi:10.1038/ng.894

WTCCC. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–78. doi:10.1038/nature05911

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of human genetics*, *89*(1), 82–93. doi:10.1016/j.ajhg.2011.05.029

Xie, H., & Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, *37*(2), 673–696. doi:10.1214/07-AOS580

Xu, C., Ladouceur, M., Dastani, Z., Richards, J. B., Ciampi, A., & Greenwood, C. M. T. (2012). Multiple regression methods show great potential for rare variant association tests. *PloS one*, *7*(8), e41694. doi:10.1371/journal.pone.0041694

Yi, N., & Zhi, D. (2011). Bayesian analysis of rare variants in genetic association studies. *Genetic epidemiology*, *35*(1), 57–69. doi:10.1002/gepi.20554

Zhang, Q., Irvin, M. R., Arnett, D. K., Province, M. a, & Borecki, I. (2011). A data-driven method for identifying rare variants with heterogeneous trait effects. *Genetic epidemiology*, *35*(7), 679–85. doi:10.1002/gepi.20618

Zhou, H., Sehl, M. E., Sinsheimer, J. S., & Lange, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Bioinformatics (Oxford, England)*, *26*(19), 2375–82. doi:10.1093/bioinformatics/btq448

Zhu, X., Feng, T., Li, Y., Lu, Q., & Elston, R. C. (2010). Detecting rare variants for complex traits using family and unrelated data. *Genetic epidemiology*, *34*(2), 171–87. doi:10.1002/gepi.20449

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series …*, *67*(2), 301–320. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x/full