May 2016

Detecting Association of Gene-Environment Interactions in Common and Rare Variants for Hypertension

Miguelangel Diaz Medina
University of Wisconsin-Milwaukee

DETECTING ASSOCIATION OF GENE-ENVIRONMENT INTERACTIONS IN COMMON AND RARE VARIANTS FOR HYPERTENSION

by

Miguelangel Diaz Medina

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Mathematics

at

The University of Wisconsin-Milwaukee

May 2016

ABSTRACT

DETECTING ASSOCIATION OF GENE-ENVIRONMENT INTERACTIONS IN COMMON AND RARE VARIANTS FOR HYPERTENSION

by

Miguelangel Diaz Medina

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professor Xuexia Wang and Daniel Gervini

Subsequent malignant neoplasms (SMNs) or secondary cancers are one of the most negative effects resulting from cancer treatment such as chemotherapy or radiation. Given the severity and high incidence of mortality faced by cancer survivors, it is critical that we understand the cause of SMNs so that preventive measures or intervention can be done for individuals facing a higher risk of SMN incidence. The purpose of this thesis is to test the efficacy of newly developed statistical methods used to identify gene-environment interactions that are associated with a specific disease, in this case, SMNs, considering both common and rare variants, using optimally weighted combinations and generalized linear models.

The models proposed are a variation of the model to Test the effect of an Optimally Weighted combination of variants (TOW) and the Variable Weight TOW (VW-TOW). Two newly proposed weighting schemes, Inverse Standard Deviation (ISD) and the Correlation Coefficient Method (CCM) are tested. In order to test the models, real life data from previous studies is analyzed to target and identify genetic variants that have been shown to have an association with a disease, in this case, hypertension, comparing the analyses and results to a study done in testing rare variants for hypertension using family-based tests with different weighting schemes. The study focuses on data from Chromosome 3 genotyped during the Genetic Analysis Workshop 18 (GAW18), obtaining similar

results to those in the hypertension study and the GAW18 study. Partial results from simulated studies are shown to support the methods' development and preliminary analyses. Comparisons are then done with existing methods to show when they exceed current standards.

TABLE OF CONTENTS

Contents

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to begin by thanking my main advisor Dr. Xuexia Wang for allowing me to work with her for my thesis and support our ongoing research through the work done in the thesis. I would like to thank my co-advisor Dr. Daniel Gervini for allowing me to work with him along with a professor from a different department and for giving me useful feedback. Thanks to both Dr. Wang and Dr. Gervini I was introduced into a field I had never thought of studying and discovered how interesting it could be. I would like to thank Dr. Jugal Ghorai for being part of my committee and also providing useful feedback.

1 Introduction

In the study of statistical genetics, statistical models are used to analyze, in a broad sense, inherited traits and genetic data. Genetic data refers mainly to biological material that is inherited during reproduction through sperm and egg cells (Laird et al, 2011). In the past, it was difficult to perform such analyses, having mostly statistical experimental studies in plants and animals. However, as technological power increases, our ability to gather and manipulate data has become more efficient, hence allowing us to reach milestones that in the past may have seen impossible, such as mapping up to 90% of the humane genome (National Human Genome Research Institute, 2015), comparing the genetic sequences of individuals in order to identify chromosomal regions where genetic variants are shared (International Hapmap Project, 2016), identifying and categorizing the functions of specific genes. (National Center for Biotechnology Information, 2016), and performing Genome Wide Association Studies(GWAS); studies which aim to determine genetic variation associated with disease traits. Due to the increase of technological power, we are in an era where statistical genetic studies will provide significant insight into disease etiology, aiding us in developing effective preventive measures as well as more successful disease treatments.

In this thesis, the focus is to test the efficacy of newly developed statistical methods to identify gene-environment interactions that are associated with a specific disease, considering both common and rare variants, using optimally weighted combinations and generalized linear models proposed by Dr. Xuexia Wang.

1.1 Basics of Biology and Statistical Genetics

In order to yield a proper understanding of the material being discussed in this thesis, it is important to clarify some terminology that will be used through the text; terminology that might be foreign to the reader. The thesis is focused on the analysis of the association between *gene-environment* interactions and a disease trait, in both common and rare *variants*. For the context of this thesis, *Gene* refers to a single-nucleotide polymorphism (SNP); however, a gene can be any segment of DNA within a chromosome possessing a specific genetic function. *Environment* refers to any variable, continuous or discrete, that is applied to the individual, such as smoking status, dosage of medicine applied for treatment, age at disease diagnose, etc.

As it is known, the genetic information of a human individual is contained in 23 pairs of chromosomes, 22 autosomal (homologous) pairs, and one sex chromosome pair. Furthermore, there are two DNA chains or sequences in each chromosome, they have a direction; one end is called 5' and the other end is called

Figure 1: DNA Chains and Base Pairs

3'; they are defined according to the asymmetrical bonding of sugar and phosphate, and they are read, by convention, from left to right, beginning at the 5' strand. DNA itself has four bases: Adenine(A), Cytosine(C), Guanine(G), and Thymine(T). Adenine pairs with Thymine and Guanine pairs with Cytosine. We define a base pair (bp) as the two bases from the two DNA chains in a chromosome, hence a base pair can be (AT) or (CG) (See Figure 1). Base pairs are used as the unit of length of chromosomes or DNA sequences. A *marker or locus* is a specific position in a chromosome. It could range from 1 bp to hundreds of bp. An *allele* is a DNA sequence within a marker; however, the terms gene, allele, and sometimes base, are usually interchangeable.

There are several types of markers, such as SNPs, Indels, Variable Numbers of Tandem Repeats (VNTRs), and Structural Variants, but SNPs are one of the main markers studied nowadays, since they explain a large portion of genetic variation in the human genome. A SNP is a single base pair marker that has two bases for the whole human population. The two bases can be from any of the four, not taking into consideration the usual AT or CG pairing (See Table 1).

Since SNPs have two alleles, we can determine their occurrence within a population. Consider SNP_A in a

Table 1: Example of SNPs

Individual	SNP 1	SNP 2	SNP 3	SNP 4
1	A	T	T	A
2	A	G	T	C
3	G	G	T	C
4	A	T	A	C

population of n individuals. Let A_1 and A_2 be the alleles corresponding to SNP_A. Let a_1 be the number of A_1 alleles and a_2 be the number of A_2 alleles. Then A_1 is the minor allele if $a_1 < a_2$ and it's frequency, denoted as the minor allele frequency MAF, is $\frac{a_1}{n}$. The concept of Minor Allele Frequency (MAF) is used to determine whether a SNP is a common or rare variant. Usually, SNPs with a $MAF > .05$ are considered common variants, and SNPs with $MAF < .05$ are considered rare variants. However, this is not a set rule,

and it can vary depending on the researcher and methods being used to analyze the data.

1.2 Data Management

Since there are around 10million SNPs in the human genome, data management can be a difficult task. For example, considers a study on 200 individuals at $300,000$ variants. Depending on the kind of analysis desired, the processing time for this amount of data can be extensive. Fortunately, there are tools that allow us to perform analyses in an efficient way. Notwithstanding, some knowledge in statistical software and basic Unix scripting is fundamental when deciding the right approach in a study.

Genomic data can be inconvenient to manipulate if it is analyzed as the letters corresponding to the alleles in the SNPs, which is why it is ideal to recode the data, and this can be done in several manners, such as additive recoding, recessive recoding, or dominant recoding. Consider again SNP_A with alleles A_1 and A_2. Now assume that allele A_1 is suspected to be the disease allele. If data is recoded additively, then it takes the values $i \in \{0, 1, 2\}$ where i denotes the number of disease alleles. If data is recoded recessively, then a recessive disease model is considered; it is assumed that only individuals with two disease alleles in the marker will have the disease, and the data is recoded to take the values $i \in \{0, 1\}$ where i denotes whether there are two disease alleles in the marker. If data is recoded dominantly, then a dominant disease model is considered; it is assumed that individuals with either 1 or 2 disease alleles will have the disease, and the data is recoded to take the values $i \in \{0, 1\}$ where i denotes whether there is at least 1 disease allele in the marker (See Table 2). The data used in this thesis is coded additively, as it provides the most information out of the three methods.

Table 2: Data Coding: A_1 Considered to be the Disease Allele

Alleles	Additive Coding	Recessive Coding	Dominant Coding
A_1A_1	2	1	1
A_1A_2	1	0	1
A_2A_2	0	0	0

2 The Methods: TOW-SE & VW-TOW-SE

The Methods TOW-SE and VW-TOW-SE are based on previously developed methods TOW and VW-TOW by Sha et al, 2012. The hypothesis stated is that the risk of treatment-related SMNs is associated with joint effects of therapeutic exposures and susceptible genes such as drug-metabolizing genes, drug transport genes, and DNA repair genes. (Wang, 2015).

Consider a sample of n individuals that have been genotyped at M variants (SNPs). Let y_i denote the disease trait for the i^{th} individual (discrete or continuous), E_i as the environmental variable (discrete or continuous), Z_{ic} as the C potential covariates, and G_{im} as the genotypic scores, coded additively, at the M variants.

2.1 Testing the effect of an Optimally Weighted combination of variants (TOW).

The Test for testing the effect of an Optimally Weighted combination of variants (TOW) derives a combination of optimal weights to test the effect of $\sum_m^M w_m^0 g_{im}$. Consider the following generalized linear model:

$$h(E(y_i \mid G_i)) = \beta_0 + \beta_1 g_{i1} + ... + \beta_M g_{iM} \tag{1}$$

We use the generalized linear model (GLM) to model the relationship between disease traits y_i and genotypes G_i, where $h(\cdot)$ is a monotone link function, β_j are the regression parameters, $j \in \{0, 1, ..., M\}$. Depending on whether the disease trait is discrete or continuous, two models, the logistic regression model with the Logit link for a binary trait or the linear model with the identity link for continuous or quantitative traits, can be used. We consider the following score test statistic for the null hypothesis $H_0 : \boldsymbol{\beta} = 0$ [Sha et al, 2012]:

$$S = U^T V^{-1} U \tag{2}$$

where $U = \sum_{i=1}^n (y_i - \bar{y})(g_i - \bar{g})$ and $V = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (g_i - \bar{g})(g_i - \bar{g})^T$. The score test statistic S follows asymptotically a chi-square distribution with $rank(V)$ degrees of freedom (Sha et al, 2012). Although powerful when testing common variants, this test loses power when rare variants are introduced into the model, which is why the weighted combination of variants is introduced. In order to test the effect of $g_i^w = \sum_m^M w_m^0 g_{im}$ the test score statistic becomes:

$$S(w_1, ..., w_M) = n \frac{(\sum_{i=1}^n (y_i - \bar{y})(g_i - \bar{g}))^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (g_i - \bar{g})^2} = n \frac{(\sum_m^M w_m \sum_{i=1}^n (y_i - \bar{y})(g_{im} - \bar{g}_m))^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (g_i - \bar{g})^2} \tag{3}$$

Since rare variants can be assumed to be independent, we have that:

$$\sum_{i=1}^{n}(g_i-\bar{g})^2=\sum_{m=i}^{M}\sum_{l=1}^{M}w_m w_l\sum_{i=1}^{n}(g_{im}-\bar{g}_m)(g_{il}-\bar{g}_l)\approx\sum_{m=1}^{M}w_m^2\sum_{i=1}^{n}(g_{im}-\bar{g})^2 \tag{4}$$

If we let:

$$a_m=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(g_{im}-\bar{g}_m)}{\sqrt{\sum_{i=1}^{n}(g_{im}-\bar{g}_m)^2}} \tag{5}$$

and

$$u_m=w_m\sqrt{\sum_{i=1}^{n}(g_{im}-\bar{g}_m)^2} \tag{6}$$

this yields a score test statistic

$$S_0(w_1,...,w_M)=n\frac{(\sum_{m=1}^{M}a_m u_m)^2}{\sum(y_i-\bar{y})^2\sum_{m=1}^{M}u_m^2}. \tag{7}$$

Since the goal is to obtain the optimal weight, we consider the maximum of $S_0(w_1,...,w_M)$, considering it as a function of $(u_1,...u_M)$ it would reach its maximum at $u_M=a_M$ or

$$w_m\sqrt{\sum_{i=1}^{n}(g_{im}-\bar{g}_m)^2}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(g_{im}-\bar{g}_m)}{\sqrt{\sum_{i=1}^{n}(g_{im}-\bar{g}_m)^2}}$$

$$\implies w_m=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(g_{im}-\bar{g}_m)}{\sum_{i=1}^{n}(g_{im}-\bar{g}_m)^2},\text{ for } m\in\{1,...,M\} \tag{8}$$

Let w_m^0 denote the optimal weights given by (8) and let $g_i^0=\sum_{m=1}^{M}w_m^0 g_{im}$. Then

$$S_0(w_1^0,...,w_M^0)=n\frac{\sum_{i=1}^{n}(y_i-\bar{y})(g_i^0-\bar{g}^0)}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \tag{9}$$

Then the statistic to Test the effect of the Optimally Weighted combination (TOW) of variants $\sum_{m=1}^{M}w_m^0 g_{im}$ is defined as

$$T_T=\sum_{i=1}^{n}(y_i-\bar{y})(g_i^0-\bar{x}^0). \tag{10}$$

By using a permutation method to evaluate the *P-values*, the term $\sum_{i=1}^{n}(y_i-\bar{y})^2$ can be considered as a constant (Sha et al, 2012) and hence the statistic T_T is equivalent to $S_0(w_1^0,...,w_M^0)$.

Notice that the optimal weight w_m^0 is essentially $\frac{\rho(y,g_m)}{\sum_{i=1}^{n}(g_{im}-\bar{g}_m)^2}=w_m^{0*}$ where $\rho(y,g_m)$ is the correlation coefficient between $y=(y_1,...,y_n)$ and $g_m=(g_{1m},...,g_{nm})$. It is clear then that since w_m^{0*} is proportional to $\rho(y,g_m)$, w_m^0 will assign heavy weights to the variants that have strong association with the disease trait of interest and will also adjust the direction of the association, allowing us to consider both causal and

protective variants. Also, since w_m^{0*} is proportional to $(\sum_{i=1}^{n}(g_{im} - \bar{g}_m)^2)^{-1}$, w_m^0 will assign heavy weights to rare variants. As Sha et al mention, similarly to most methods that target rare variants, TOW will lose power when testing the effects of common and rare variants together, which is why the method VW-TOW was proposed.

2.2 VW-TOW

In order to preserve the power of the analysis when dealing with both common and rare variants at the same time, the Variable Weight to Test the effects of the Optimally Weighted combination (VW-TOW) of variants is i proposed. We begin by dividing the variants into common and rare, using a rare variant threshold (RVT), usually considered to be 0.05. Variants with a $MAF < RVT$ are considered rare variants, and those with $MAF > RVT$ are considered common variants. After separating the variants into common and rare, we apply the method TOW to each group and obtain the two test statistics T_r and T_c, representing the TOW test statistic for the rare and common variant groups, respectively (Sha et al, 2012). Then consider $T_\lambda = \lambda\frac{T_r}{\sqrt{var(T_r)}} + (1-\lambda)\frac{T_c}{\sqrt{var(T_c)}}$ and let p_λ denote the *P-value* of T_λ. Then the VW-TOW test statistic is defined as

$$T_{VW-T} = \min_{0\leq\lambda\leq 1} p_\lambda. \tag{11}$$

In order to evaluate the minimization, a simple method was used. The interval $[0,1]$ is divided into K equivalent non-overlapping intervals and we let $\lambda = k/K$ for $k \in \{0,1,...,K\}$. Then we get that $\min_{0\leq\lambda\leq 1} p_\lambda = \min_{0\leq k\leq K} p_{\lambda_k}$. The standard permutation test is used to evaluate the *P-value* of the TOW test statistic T_T, but a variation is used to evaluate the *P-value* of the $VW-TOW$ test statistic T_{VW-T}. Consider a number of Q permutations, and let $T_r^{(q)}$ and $T_c^{(q)}$ be the values of T_r and T_c for the q^{th} permutation, for $q \in \{0,1,...,Q\}$, where $q = 0$ denotes the values from the original data. Then we proceed to calculate the value of $T_{\lambda_k}^{(q)}$ for all values of q and k, estimating $var(T_r)$ and $var(T_c)$ using $T_r^{(q)}$ and $T_c^{(q)}$ (Sha et al, 2012). Lastly, we obtain $p_{\lambda_k}^{(b)}$ using

$$p_{\lambda_k}^{(q)} = \frac{\#\{T_{\lambda_k}^{(d)} > T_{\lambda_k}^{(q)} \mid d \in \{0,1,...,Q\}\}}{Q}. \tag{12}$$

Then considering $p^{(q)}$ as $\min_{0\leq k\leq K} p_{\lambda_k}^{(q)}$ we determine the *p-value* of T_{VW-T} by

$$\frac{\#\{p_{\lambda_k}^{(q)} > p_{\lambda_k}^{(0)} \mid q \in \{1,...,Q\}\}}{Q}. \tag{13}$$

2.3 Adjusting for Confounder Covariates

In order to consider the model with the C potential covariates Z_c we need to take into account certain aspects before applying the methods. We begin by adjusting both the disease traits y_i and the genotypic scores g_{im} by applying a simple linear regression and obtaining the residuals (Sha et al, 2012). We get

$$y_i = \alpha_0 + \alpha_1 z_{i1} + ... + \alpha_c z_{ic} + \epsilon_i \tag{14}$$

and

$$g_{im} = \alpha_0 + \alpha_1 z_{i1} + ... + \alpha_c z_{ic} + \tau_i \tag{15}$$

Obtaining the residuals $\tilde{y}_i$ and $\tilde{g}_{im}$ for the disease trait and the genotypic scores, respectively, we proceed to apply the TOW and VW-TOW methods, defining their test score statistics as

$$T_{TOW} = T_T \mid_{y_i=\tilde{y}_i, g_{im}=\tilde{g}_{im}} \tag{16}$$

and

$$T_{VW-TOW} = T_{WV-T} \mid_{y_i=\tilde{y}_i, g_{im}=\tilde{g}_{im}} \tag{17}$$

respectively. Using this approach is equivalent to applying the linear model directly to the disease trait including the confounder covariates

$$y_i = \alpha_0 + \alpha_1 z_{i1} + ... + \alpha_c z_{ic} + \beta_1 g_{i1} + ... + \beta_m g_{im} + \epsilon_i = \boldsymbol{\alpha}^T Z_i + \boldsymbol{\beta}^T G_i + \epsilon_i \tag{18}$$

where $\boldsymbol{\alpha} = (\alpha_0, ..., \alpha_c)^T$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)^T$, $G_i = (g_{i1}, ..., g_{iM})^T$, and $Z_i = (z_{i1}, ..., z_{ic})^T$. Then the score test statistic for the null hypothesis $H_0 : \boldsymbol{\beta} = 0$ becomes

$$SC = \tilde{U}^T \tilde{V}^{-1} \tilde{U} \tag{19}$$

where $\tilde{U} = \sum_{i=1}^n \tilde{y} \tilde{G}_i$, $\tilde{V} = \frac{1}{n} \sum_{i=1}^n \tilde{y}_i{}^2 \sum_{i=1}^n \tilde{G}_i \tilde{G}_i{}^T$.

The proof of this statement is the following.

Let $Y = (y_1, ..., y_n)^T$ $\epsilon = (\epsilon_1, ..., \epsilon_n) \sim^{iid} N(0, \sigma^2)$, then the log-likelihood of (18) is given by

$$\log l = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (Y - Z\alpha - G\beta)^T (Y - Z\boldsymbol{\alpha} - G\boldsymbol{\beta}). \tag{20}$$

Then

$$\frac{\delta \log l}{\delta \boldsymbol{\beta}} = \frac{1}{\sigma^2}(Y - Z\boldsymbol{\alpha} - G\boldsymbol{\beta})^T G, \tag{21}$$

$$\frac{\delta \log l}{\delta \boldsymbol{\alpha}} = \frac{1}{\sigma^2}(Y - Z\boldsymbol{\alpha} - G\boldsymbol{\beta})^T Z, \tag{22}$$

$$\frac{\delta^2 \log l}{\delta \boldsymbol{\beta}\boldsymbol{\beta}^T} = -\frac{1}{\sigma^2}G^T G,\ \frac{\delta^2 \log l}{\delta \boldsymbol{\alpha}\boldsymbol{\alpha}^T} = -\frac{1}{\sigma^2}Z^T Z, \text{ and } \frac{\delta^2 \log l}{\delta \boldsymbol{\alpha}\boldsymbol{\beta}^T} = -\frac{1}{\sigma^2}Z^T G. \tag{23}$$

Now let $\hat{\alpha}$ and $\hat{\sigma}^2$ denote the maximum likelihood estimators of α and σ^2 under $H_0 : \boldsymbol{\beta} = 0$. Then

$$\hat{\alpha} = (Z^T Z)^{-1} Z^T Y \text{ and } \hat{\sigma}^2 = \frac{1}{n} Y^T (\boldsymbol{I} - P) Y = \frac{1}{n} \tilde{Y}^T \tilde{Y} \tag{24}$$

where $P = Z(Z^T Z)^{-1} Z^T$ and $\tilde{Y} = (\tilde{y}_1, ..., \tilde{y}_n)$ is the vector of the residuals obtained from (14). Let $\theta = (\alpha^T, \beta^T)^T$, then we obtain the following score and information matrix

$$S = \frac{\delta \log l}{\delta \theta}_{|\alpha=\hat{\alpha},\beta=0} = \frac{1}{\sigma^2 (0, U^T)^T} \tag{25}$$

and

$$I = -E\frac{\delta^2 \log l}{\delta \theta \theta^T}_{|\alpha=\hat{\alpha},\beta=0} = \frac{1}{\sigma^2} \begin{pmatrix} Z^T Z & Z^T G \\ G^T Z & G^T G \end{pmatrix}, \tag{26}$$

where $U = \tilde{Y}^T G$. Note that $(\boldsymbol{I} - P)$ is idempotent. Hence $U = \tilde{Y}^T G = \tilde{Y}^T (\boldsymbol{I} - P) G = \tilde{Y}^T \tilde{G}$ and $G^T (\boldsymbol{I} - P) G = \tilde{G}^T \tilde{G}$ where $\tilde{G} = (\tilde{G}_1, ..., \tilde{G}_n)$ is the $(n \text{x} M)$ matrix of the residuals obtained from (15). Then the test score statistic is

$$T_{linear} = \frac{1}{\sigma^2} U^T V^{-1} U \tag{27}$$

where $U = \tilde{Y}^T \tilde{G} = \sum_{i=1}^n \tilde{y}_i \tilde{G}_i$ and $V = \tilde{G}^T \tilde{G} = \sum_{i=1}^n \tilde{G}_i \tilde{G}_i^T$

Hence, T_{linear} and SC from (19) are proportional, which completes the proof

Now, similarly to the main model, the score test statistic to test the effect of the weighted combination of variants $g_i = \sum_{m=1}^M w_m g_{im}$ is given by $SC(w_1, ..., w_m) = n\frac{(\sum_{i=1}^n \tilde{y}_i \tilde{g}_i)^2}{\sum_{i=1}^n \tilde{y}_i^2 \sum_{i=1}^n \tilde{g}_i^2}$, and following the same procedure used in the non-covariate method we have that $SC(w_1, ..., w_M)$ reaches its maximum when $w_m = \frac{\sum_{i=1}^n \tilde{y}_i \tilde{g}_{im}}{\sum_{i=1}^n \tilde{g}_{im}^2}$ and hence the maximum of $SC(w_1, ..., w_M)$ is equivalent to T_{TOW}.

2.4 Testing the effect of an Optimally Weighted combination of variants considering SNP-Environment interaction (TOW-SE).

Considering the same set up form the beginning of the section, now we will use the following generalized linear model, which include the interaction term between the SNPs and the Environmental trait. Consider

$$f(E(y_i \mid G_i, E_i, Z_i)) = \alpha_0 + Z_i\boldsymbol{\alpha} + E_iG_i\boldsymbol{\beta} + G_i\boldsymbol{\xi} + \boldsymbol{\eta}E_i \tag{28}$$

and just as in the TOW method, for continuous or quantitative disease traits, $f(\cdot)$ will be the monotone link function, while for binary traits, the logit link function will be used. The parameters α_0, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\eta}$ are the respective regression coefficients of each term, and the corresponding null hypothesis becomes $H_0 : \boldsymbol{\beta} = 0$. However, since we are testing the SNP-Environment interaction, when we adjust for covariates we are interested in obtaining $\tilde{y}_i$ as the residual of y_i and $\tilde{X}_i = (\tilde{x}_{i1}, ..., \tilde{x}_{iM})$ as the residuals of $E_iG_i = (E_ig_{i1}, ..., E_ig_{iM})$ (Sha et al, 2015). Then the relationship between $\tilde{y}_i$ and $\tilde{X}_i = (\tilde{x}_{i1}, ..., \tilde{x}_{iM})$ is modeled by the general linear model

$$f(E(\tilde{y}_i \mid \tilde{X}_i)) = \beta_0^* + \tilde{X}_i\boldsymbol{\beta}^*. \tag{29}$$

Then the initial null hypothesis is equivalent to $H_0 : \boldsymbol{\beta}^* = 0$. Since there is some particular interested in accurately identifying interactions between rare variants and the environmental trait, it is desired to efficiently deal with the data, to avoid losing power due to large degrees of freedom or due to sparse data. Hence, three different weighting schemes of the form $\sum_{m=1}^{M} w_i\tilde{x}_{im}$ are introduced as a solution to the problem (Wang et al,2015).

Optimal Weight and TOW-SE: The first weighting scheme uses the same score test as the initial TOW Method:

$$S(w_1, ..., w_M) = n\frac{(\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})(\tilde{x}_i - \bar{\tilde{x}}))^2}{\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})^2\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})^2} = n\frac{(\sum_{m}^{M} w_m\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})(\tilde{x}_{im} - \bar{\tilde{x}}_m))^2}{\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})^2\sum_{i=1}^{n}(\tilde{x}_i - \bar{\tilde{x}})^2} \tag{30}$$

which reaches its maximum at $S_0(w_1^0, ..., w_M^0) = n\frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})(\tilde{x}_i^0 - \bar{\tilde{x}}^0)}{\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})^2}$ when $w_m = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})(\tilde{x}_{im} - \bar{\tilde{x}}_m)}{\sum_{i=1}^{n}(\tilde{x}_{im} - \bar{\tilde{x}}_m)^2}$ and $\tilde{x}_i^0 = \sum_{m=1}^{M} w_m^0\tilde{x}_{im}$. Then the test statistic is defined as in TOW as $T_{T-SE} = \sum_{i=1}^{n}(\tilde{y}_i - \bar{\tilde{y}})(\tilde{x}_i^0 - \bar{\tilde{x}}^0)$. In order to make T_{T-SE} equivalent to $S_0(w_1^0, ..., w_M^0)$ we use, once again, a permutation test to evaluate the *P-values*. Just as it TOW, the optimal weight w_m^0 assigns heavy weights to gene-environment interactions that have strong association with the studied disease trait, as well as adjusting for the direction of the interaction.

In order to maintain the power when testing for both common and rare variants at the same time, the method VW-TOW-SE is proposed.

VW-TOW-SE This method applies the procedures from VW-TOW in the exact same manner. We divide the variants into common and rare, using the RVT of 0.05. After separating the variants into common and rare, we apply the method TOW-SE to each group and obtain the two test statistics T_r and T_c, representing the TOW-SE test statistic for the rare and common variant groups, respectively. Then we consider $T_\lambda = \lambda \frac{T_r}{\sqrt{var(T_r)}} + (1-\lambda)\frac{T_c}{\sqrt{var(T_c)}}$ and let ρ_λ denote the *P-value* of T_λ. Then the VW-TOW-SE test statistic is defined as $T_{VW-TOW-SE} = \min_{0 \leq \lambda \leq 1} p_\lambda$. Then the standard, and the modification of the permutation methods mentioned above are used to obtain the *P-values* for both T_{T-SE} and $T_{VW-T-SE}$.

Inverse Standard Deviation (ISD) Method The second weighting scheme proposes a weight w_m based on the inverse standard deviation of $p_m = MAF_m$, where MAF_m is the Minor Allele Frequency of the m^{th} variant. Then the weight assigned to each variant is $w_m = \frac{1}{\sqrt{np_m(1-p_m)}}$. The focus of this weighting scheme is to put heavier weights to gene-environment interactions of rare variants.

Correlation Coefficient Method (CCM) Given the evidence that shows that there exists a positive correlation between environmental exposures and genetic factors (Wang, 2015), a weighting scheme using the correlation coefficient ρ_m between the genotypic score at the m^{th} variant and the environmental variable in individuals that have been diagnosed (i.e. cases). Then we define the weight w_m as $w_m = \rho_m$. Using this weighting scheme, we get that whenever ρ_m is positive and close to 1, it puts a heavy weight to the gene-environment interactions that have a strong and positive association with the disease trait; and if ρ_m is negative and close to -1, it puts a heavier weight to the gene-environment interactions that have strong and negative association with the trait of interest. This is also a good weighting scheme for adjusting for the direction of the gene-environment interaction.

2.5 Comparison Methods: iSKAT & MinP

In order to determine the usefulness of the newly proposed methods, a comparison with existing methods was necessary. The two methods chosen for comparison were the Test for Rare Variants by Environment Interactions using interaction Sequence Kernel Association Test (iSKAT) (Lin et al, 2015) and the Minimum P-value method.

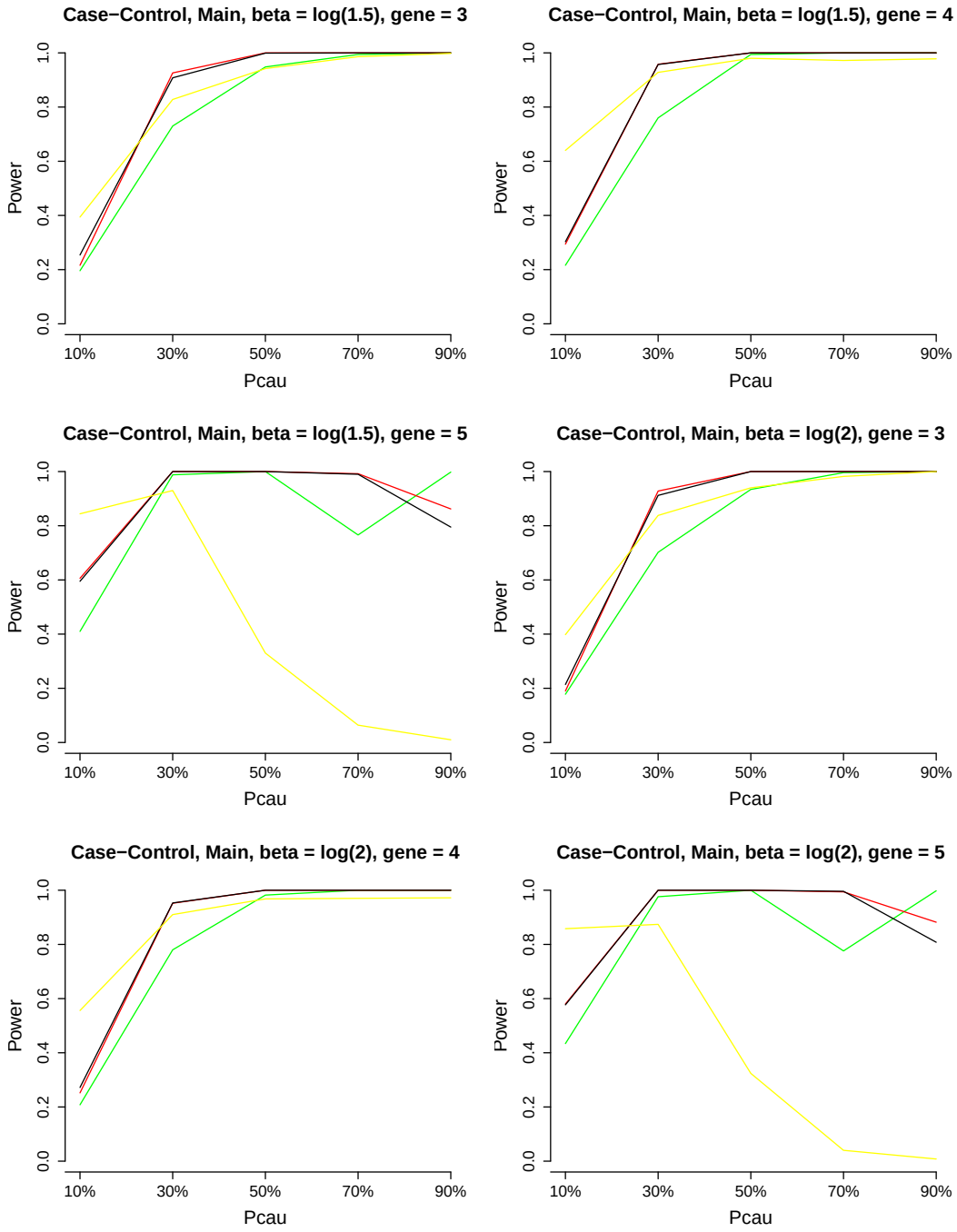
3 Simulations and Partial Preliminary Results

The empirical Mini-Exome genotype data provided by the GAW17 was used for the performed simulation. The GAW17 dataset contains the haplotypes of 697 unrelated individuals on 3,205 genes. The four genes: ELAVL4 (gene1), MSH4 (gene2), PDE4B (gene3), and ADAMTS4 (gene4) with 10, 20, 30, and 40 variants were used, respectively, to simulate the data for the study. The four genes were merged to form a super gene (Sgene) with 100 variants. The distributions of MAFs in the 100 variants in the Sgene and in the 24,487 variants in all the 3,205 genes are given in Figure 2 (Sha et al, 2012). During the simulation studies, we generate genotypes based on the haplotypes of the 697 individuals in the Sgene. The haplotypes were provided from the initial study during the development of the TOW and VW-TOW methods (Sha et al, 2012). and all of the data simulated was done in the same manner as the data simulated for the TOW and VW-TOW simulations analyses (See Shat et al, 2012 for more information).

In order to determine the efficiency of the models proposed, a set of different combinations of important parameters was used. These initial parameter combinations were: the location of the gene ($gene \in \{3, 4, 5\}$), the proportion of causal variants ($pcau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$), the proportion of protective variants ($nprot \in \{0, 0.2, 0.4\}$), whether there was a main effect from the SNP ($maineff \in \{0, 1\}$), the mean of the environmental trait, simulated to be $Normal(envmean, 1)$ ($envmean \in \{0, 50, 100, 150, 200, 250\}$), the coefficient of the gene-environment interaction ($beta = \{\log(1.5), \log(2)\}$), and whether the disease trait was binary or continuous ($quan \in \{0, 1\}$, 0 for binary, 1 for continuous). In total, we considered 2160 combinations of variables to simulate potential situations where the method might be applied. For simplicity, we reduced the number of combinations by half, to 1080, by considering only a binary disease trait model.

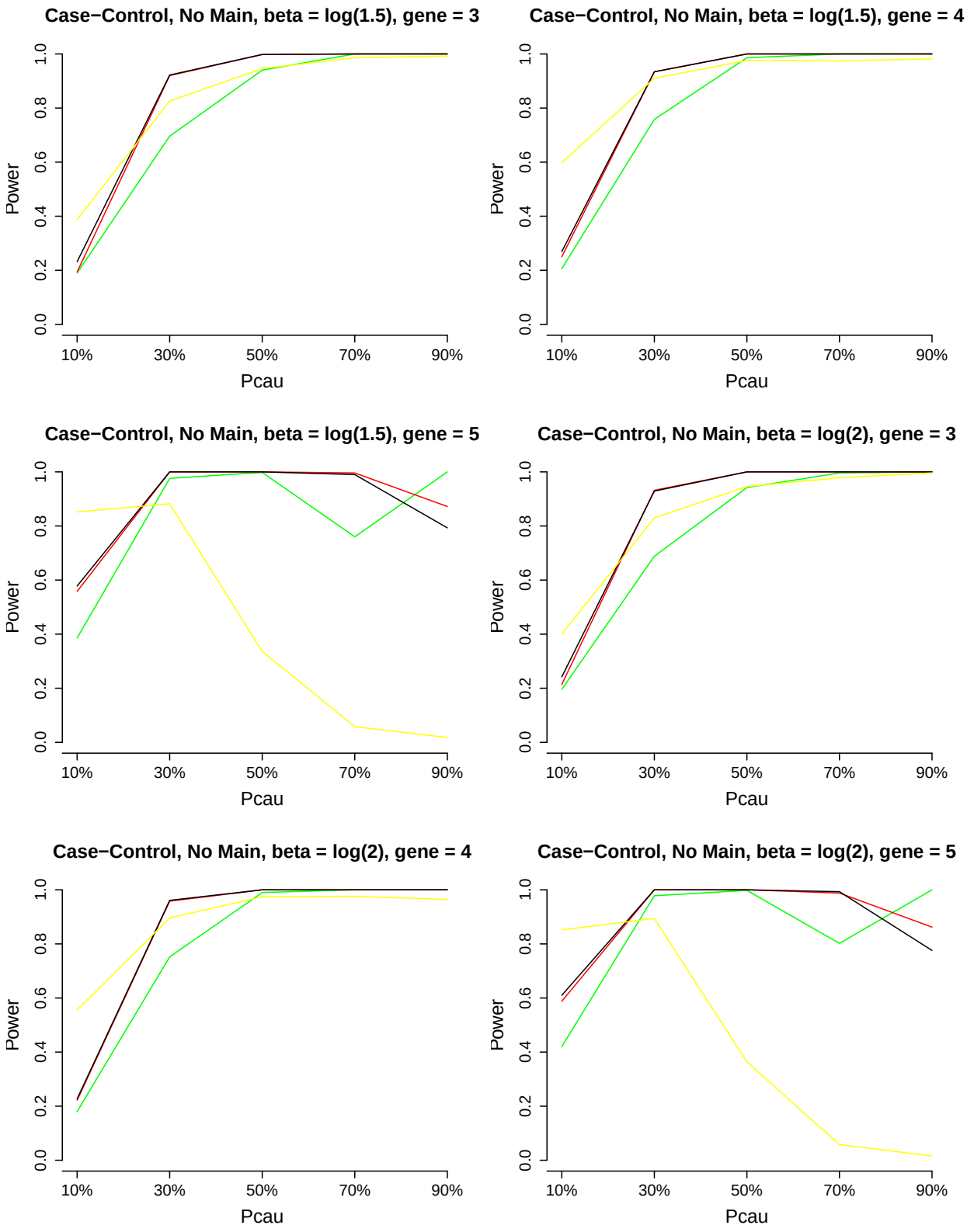
For each combination, 500 replications of the model were done, in order to use the permutation tests described above. Below are some preliminary results from some of the most significant combinations. Consider Figure 2. We can see that as the proportion of causal variants increases, so does the power of each method. However, in most settings, the methods TOW-SE and VW-TOW-SE are consistently more powerful than the other methods, and although the Minimum P-value method is also higher, this method fails to keep the type I error under control, unless a permutation method is applied to it, which makes the method significantly more computationally intensive than the other methods, hence inconvenient to use in most settings. Notice that when the methods were applied in gene 5, the results were not as straight forward as in genes 3 and 4. Regarding this disparity in comparison with the other genes, min depth analysis needs to be done in gene 5 to discover the cause of the drastic results. Regarding Figure 3, where no main effect from the SNP was considered we can see that the results are almost identical to the ones where main effect was considered, and once again, the results from gene 5 seem to need further investigation.

Figure 2: Power Comparison as the Proportion of Causal Variants Increases, when There is a Main Effect from the SNP



Green represents iSKAT, Red represents TOW, Black Represents VW-TOW, and Yellow represents MinP

Figure 3: Power Comparison as the Proportion of Causal Variants Increases, when There is No Main Effect from the SNP



Green represents iSKAT, Red represents TOW, Black Represents VW-TOW, and Yellow represents MinP

From a different perspective, we present the power of the different methods from the simulated data as the proportion of protective variants increased. The results agree with the ones from above (See Figures 4 and 5). Once again, we considered the two settings, main effect from the SNP and no main effect, and we show the results below. Since the results from gene 5 seem to need more information, only results from genes 3 and 4 are shown.

Figure 4: Power Comparison as the Proportion of Protective Variants Increases, when There is a Main Effect from the SNP, on gene 3

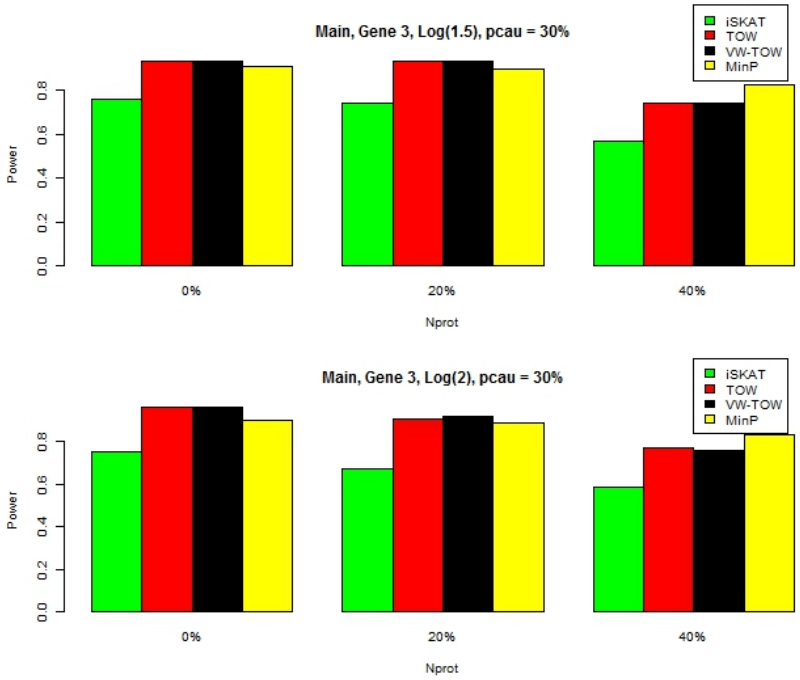


Figure 5: Power Comparison as the Proportion of Protective Variants Increases, when There is a Main Effect from the SNP, on gene 4

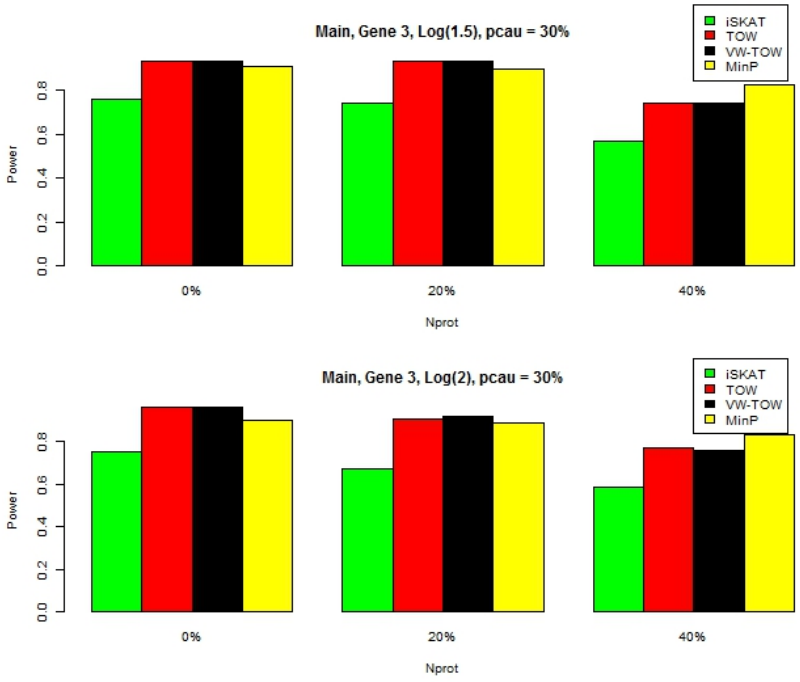


Figure 6: Power Comparison as the Proportion of Protective Variants Increases, when There is No Main Effect from the SNP, on gene 3

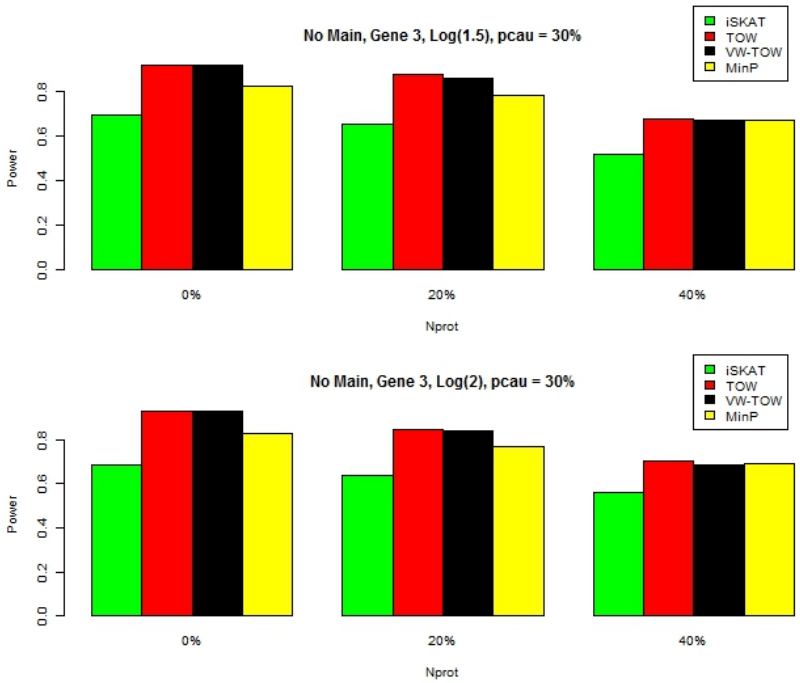
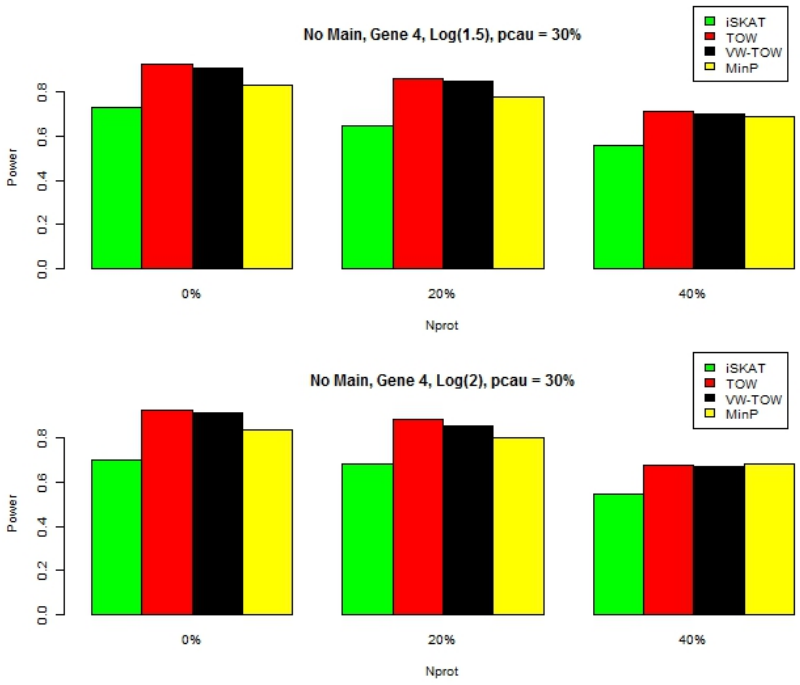


Figure 7: Power Comparison as the Proportion of Protective Variants Increases, when There is No Main Effect from the SNP, on gene 4



4 Real Data Analysis: GAW18

The methods TOW-SE and VW-TOW-SE seem to be useful in some settings, but in order to test accurate identification of SNPs that have an association with a disease trait, it is desired to test the methods in real data. Part of the data from the Genetic Analysis Workshop 18 (GAW18) was provided to test the methods and compare the results with variants identified by Wang et al in a previously released paper. The data provided was a set of 142 unrelated individuals genotyped at $1,215,399$ variants. PLINK was used to further clean the data in order to remove noise and erroneous data. After cleaning the data for discarding genotypes that had less than 5% genotyping rate and non SNP data, only 585397 SNPs remained. The remaining SNPs were then divided into non-overlapping windows of 100 SNPs each in order to apply the methods. The goal was to identify regions that contained SNPs that have a strong association with hypertension. Two environmental traits were considered, a quantitative trait, age at disease diagnose, and a binary trait, smoker status, with 0 denoting non-smoker status. Since the Bonferroni correction would have been considerably conservative for this particular dataset ($\frac{0.05}{585397} = (0.085)10^{-6}$), a different method was used to identify significant windows. Using PLINK, the number of independent windows was identified $(72,217)$, and that number was used to consider as the threshold for significance ($\frac{0.05}{72217} = (0.0692)10^{-5}$). Below we can see the top ten windows according to the p-values obtained from each of the compared methods.

4.1 Results for Quantitative Environmental Variable: Age at Diagnose

Table 3: P-values from the Considered Methods for Age at Diagnose

"Window"	"P-value ISD"	"Window"	"P-value CCM"	Window	P-value TOW
4353	0.000127796260687041	2670	0.00332227402518026	1900	0.00010651
4458	0.000142648480121133	4927	0.00936317695426392	2635	0.000107272
3552	0.000161628226907351	5452	0.00954981106807906	1912	0.000107821
448	0.000197890706833959	2623	0.012099117037367	435	0.000108369
146	0.000258317337306635	105	0.0126957173768499	358	0.000108533
5452	0.000265360363298828	3711	0.013338649924398	491	0.000108851
2671	0.000276883896236702	4796	0.0139377574657057	4452	0.000112642
2669	0.00028800011006147	2762	0.0141387599509206	5721	0.000114368
2762	0.000303103474680544	4576	0.0147110228354675	146	0.000115835
4927	0.000307281789389546	3421	0.0153946404175944	5536	0.000126842

"Window"	"P-value iSKAT"	"Window"	"P-value MinP"
3103	NA	1890	0.000100122073315671
2640	0.000138192609795085	4004	0.000100856751922798
3150	0.000312628808616267	2918	0.000100947759807247
2613	0.000329452154594012	2789	0.000100980750696368
1899	0.000374110321790022	2691	0.000101078209077982
2612	0.000494980527388722	1136	0.000101209016525324
5257	0.000665823643444807	283	0.000101263399734529
3491	0.00138761693100742	941	0.000101347904547539
2607	0.00139720380658059	2886	0.000101352426655999
2603	0.00141196166729574	4049	0.000101393166343222

We can see that none of the obtained *P-values* imply significance in any window according to the current threshold $(0.0692)10^{-5}$. However, we should still consider the windows with the lowest *P-values*. We have at most 50 different windows, for each phenotype considered, that could potentially hold significance.
Tables 5 and 6 show the windows that are significant when compared to the GAW18 and family based hypertension studies results. From these results we can see that indeed the newly developed methods succeed at identifying $gene - environment$ associations with the disease. Furthermore, out of all the methods, the Correlation Coefficient Method signaled the window that have the variants that explain the most variation in the genome, when considering the discrete environmental variable (Table 6, highlighted in bold).

Further Research

After showing the methods' efficacy and superiority to current standards in certain cases, the next step is to analyze real data on Secondary Malignant Neoplasms (SMNs) in order to identify variants that have either a positive or negative association with the disease. The implication of this discovery could help effectively treat cancer patients and decrease the rate of SMN incidence.

4.2 Results for Binary Environmental Variable: Smoker Status

Table 4: P-values from the Considered Methods for Smoker Status

"Window"	"P-value ISD"	"Window"	"P-value CCM"
129	0.000110874196227906	4739	0.00192106680326731
4986	0.00011343027055266 1	2192	0.0032841856545478
2225	0.000123986114126451	1630	0.00352376924683095
5784	0.000129450030879719	1321	0.00465613439361845
871	0.000141877856293937	4847	0.00471203141365995
1189	0.000149561267793308	4204	0.00479930828540953
3101	0.000149736481382279	2952	0.004826732467899
2195	0.000183857860640235	508	0.00495076607465683
1108	0.000200158615176083	1707	0.00496337757485787
5773	0.000200831401523605	2193	0.00572825733223425

Window	P-value TOW
251	0.000118549
5773	0.000126155
2785	0.000129112
872	0.000131648
1570	0.000138966
5662	0.000140556
3399	0.000143583
2507	0.000144919
642	0.00014917
5784	0.00015146

Window	P-value iSKAT	Window	P-value MinP
32	6.30E-05	234	0.000104629
2196	0.001177747	5452	0.000190207
4006	0.001707146	299	0.000218275
5248	0.001757833	3423	0.000229582
3805	0.002247517	5665	0.000236041
2693	0.002446083	235	0.000249251
5215	0.002464846	4991	0.000272212
5791	0.002493754	5160	0.000318506
4653	0.002695548	1707	0.000404151
2194	0.003138554	5404	0.000422441

Table 5: Range of SNPs Inside the Potentially Significant Windows for the Age Evironmental Variable

Window	Starting SNP loc	Ending SNP loc	Significant	Signaled by
105	3_2576692_G	3_2606480_C	yes	CCM
1136	3_32652332_G	3_32666083_A	yes	Min P
146	3_3473240_A	3_3497291_A	yes	ISD
1890	3_60995600_C	3_61023120_C	yes	Min P
1899	3_61226363_A	3_61256364_T	yes	iSKAT
1900	3_61256686_A	3_61297908_C	yes	TOW-SE
1912	3_61625274_C	3_61656548_G	yes	TOW-SE
2603	3_84413353_A	3_84427677_A	no	iSKAT
2607	3_84526588_C	3_84558280_T	no	iSKAT
2612	3_84695126_G	3_84726805_A	no	iSKAT
2613	3_84727898_C	3_84756552_T	no	iSKAT
2623	3_84992877_T	3_85031666_G	no	CCM
2635	3_85459212_T	3_85492002_A	no	TOW-SE
2640	3_85588494_C	3_85600771_G	no	iSKAT
2669	3_86547035_C	3_86576946_G	no	ISD
2670	3_86577017_C	3_86605162_A	no	CCM
2691	3_87295368_T	3_87339402_A	no	Min P
2762	3_89549919_C	3_89595252_T	no	CCM
2789	3_94053165_A	3_94078802_G	no	Min P
283	3_6787821_G	3_6805408_A	yes	Min P
2886	3_97152615_A	3_97191431_T	no	Min P
2918	3_98096781_T	3_98120608_A	no	Min P
3103	3_104235682_G	3_104255789_G	no	iSKAT
3150	3_105681971_C	3_105693714_C	no	iSKAT
3421	3_115221939_T	3_115260435_G	yes	CCM
3491	3_117453184_T	3_117492046_C	yes	iSKAT
3552	3_119399641_A	3_119421703_A	yes	ISD
358	3_8718008_A	3_8730289_G	no	TOW-SE
3711	3_125129489_T	3_125174636_C	yes	CCM
4004	3_134649566_A	3_134673394_T	yes	Min P
4049	3_136773875_C	3_136800244_T	yes	Min P
435	3_11062062_T	3_11099571_A	yes	TOW-SE
4353	3_147386728_C	3_147417839_T	yes	ISD
4452	3_150510227_A	3_150554959_T	yes	TOW-SE
4458	3_150726830_T	3_150774782_A	yes	ISD
448	3_11591200_G	3_11629420_G	yes	ISD
4576	3_154915593_G	3_154955403_T	yes	CCM
4796	3_162611646_T	3_162632550_A	yes	CCM
491	3_13045939_T	3_13075350_A	yes	TOW-SE
4927	3_166715641_G	3_166750007_C	yes	CCM
5257	3_178567095_C	3_178601948_G	yes	iSKAT
5452	3_186259233_G	3_186288565_A	yes	CCM
5536	3_189058427_C	3_189085393_T	yes	TOW-SE
5721	3_193972501_A	3_194006998_T	yes	TOW-SE
941	3_27088978_C	3_27102630_A	yes	Min P

Table 6: Range of SNPs Inside the Potentially Significant Windows for the Smoker Status Environmental Variable

Window	Starting SNP loc	Ending SNP loc	Significant	Signaled by
32	3_991242_C	3_1017394_A	no	iSKAT
1108	3_31929486_A	3_31957631_A	yes	ISD
1189	3_34549511_C	3_34589833_C	yes	ISD
129	3_3093955_C	3_3123326_G	yes	ISD
1321	3_39240559_T	3_39275010_A	yes	CCM
1570	**3_48623124_*A***	**3_48721040_*A***	yes	CCM
1630	3_52659079_C	3_52697566_C	yes	CCM
1707	3_55302446_G	3_55333080_C	yes	CCM
2192	3_70021965_C	3_70069611_A	yes	CCM
2193	3_70069621_A	3_70114499_C	yes	CCM
2194	3_70115600_T	3_70172654_C	yes	iSKAT
2195	3_70172923_T	3_70198308_C	yes	ISD
2196	3_70198406_G	3_70232626_T	yes	iSKAT
2225	3_71384708_T	3_71421152_T	yes	ISD
234	3_5729514_C	3_5755612_G	yes	Min P
235	3_5755763_G	3_5785973_G	yes	Min P
2507	3_80479403_A	3_80515770_T	yes	TOW-SE
251	3_6051816_A	3_6087578_A	yes	TOW-SE
2693	3_87354731_A	3_87369511_G	no	iSKAT
2785	3_93827350_A	3_93936708_G	no	TOW-SE
2952	3_99083010_T	3_99108384_T	no	CCM
299	3_7230063_C	3_7266362_G	yes	Min P
3101	3_104164622_A	3_104202009_G	no	ISD
3399	3_114110629_T	3_114144714_A	yes	TOW-SE
3423	3_115297011_A	3_115327361_A	yes	Min P
3805	3_127832554_G	3_127881613_A	yes	iSKAT
4204	3_142703435_T	3_142750908_G	yes	CCM
4653	3_157729011_A	3_157788588_T	yes	iSKAT
4739	3_161151867_T	3_161175643_T	yes	CCM
4847	3_164105512_C	3_164125983_C	yes	CCM
4986	3_169121931_A	3_169148465_T	yes	ISD
4991	3_169262775_C	3_169288729_T	yes	Min P
508	3_13522177_A	3_13557469_T	yes	CCM
5160	3_175211459_G	3_175246745_C	yes	Min P
5215	3_177007214_C	3_177067087_G	yes	iSKAT
5248	3_178291260_A	3_178330596_A	yes	iSKAT
5404	3_184409373_A	3_184441278_A	yes	Min P
5452	3_186259233_G	3_186288565_A	yes	Min P
5662	3_192287194_T	3_192315287_G	yes	TOW-SE
5665	3_192367902_G	3_192395378_C	yes	Min P
5773	3_195238993_G	3_195282609_C	yes	ISD
5784	3_195546145_G	3_195563248_A	yes	ISD
5791	3_195763241_C	3_195789428_C	yes	iSKAT
642	3_18228007_A	3_18278139_A	yes	TOW-SE
871	3_25062752_C	3_25084111_C	no	ISD
872	3_25084308_G	3_25114861_G	no	TOW-SE

5 References

1. Laird N,Lange C: **The Fundamentals of Statistical Genetics**. New York, Springer, 2011.

2. Lin X, Seunggeun L, C. Wu M, Wang C, Chen H, Li Z, Lin X. 2016 **Test for Rare Variants by Environment Interactions in Sequencing Association Studies**. Biometrics 72(1):156-64.

3. National Center for Biotechnology Information. Bethesda, Web, 2016.

4. National Human Genome Research Institute. **An Overview of the Human Genome Project**. Web. 2015.

5. Sha Q, Zhang Z, Zhang S. 2011. **An improved score test for genetic association studies**. Genet Epidemiol 35:350–359.

6. Wang X, Zhao X, Zhou J. **Testing Rare Variants for Hypertension Using Family-Based Tests with Different Weighting Schemes**. TS. 2015.