BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 16 2014, pages 2317–2323
doi:10.1093/bioinformatics/btu207

Genetics and population analysis

Advance Access publication April 22, 2014

A power set-based statistical selection procedure to locate susceptible rare variants associated with complex traits with sequencing data

Hokeun Sun[1] and Shuang Wang[2,*]

[1]Department of Statistics, Pusan National University, Pusan 609-735, Korea and [2]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Existing association methods for rare variants from sequencing data have focused on aggregating variants in a gene or a genetic region because of the fact that analysing individual rare variants is underpowered. However, these existing rare variant detection methods are not able to identify which rare variants in a gene or a genetic region of all variants are associated with the complex diseases or traits. Once phenotypic associations of a gene or a genetic region are identified, the natural next step in the association study with sequencing data is to locate the susceptible rare variants within the gene or the genetic region.

Results: In this article, we propose a power set-based statistical selection procedure that is able to identify the locations of the potentially susceptible rare variants within a disease-related gene or a genetic region. The selection performance of the proposed selection procedure was evaluated through simulation studies, where we demonstrated the feasibility and superior power over several comparable existing methods. In particular, the proposed method is able to handle the mixed effects when both risk and protective variants are present in a gene or a genetic region. The proposed selection procedure was also applied to the sequence data on the ANGPTL gene family from the Dallas Heart Study to identify potentially susceptible rare variants within the trait-related genes.

Availability and implementation: An R package 'rvsel' can be downloaded from http://www.columbia.edu/∼sw2206/ and http://statsun.pusan.ac.kr.

Contact: sw2206@columbia.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 14, 2013; revised on April 2, 2014; accepted on April 15, 2014

1 INTRODUCTION

The fundamental problem with rare variants [with minor allele frequency (MAF) $< 1\%$] is their low frequency, i.e. the limited number of observed carriers lowers the statistical power to detect phenotypic association with any single rare variant. Thus, almost all existing statistical methods to detect disease/trait-associated rare variants follow the framework of aggregating and testing all rare variants in a gene or a candidate genomic region thereby boosting the association signal (Bhatia *et al*., 2010; Chen *et al*., 2012; Cheung *et al*., 2012; Ionita-Laza *et al*., 2011; Lee *et al*., 2012; Lin and Tang, 2011; Liu and Leal, 2008, 2010; Madsen and Browning, 2009; Neale *et al*., 2011; Price *et al*., 2010; Wu *et al*., 2011). The existing methods try to improve this basic idea of aggregating signals in two aspects: first, by more potent extraction of signals from individual rare variants; and second, by better aggregation signals extracted from multiple rare variants in a gene or a genetic region of interest. Improvements of the first kind include upweighing variants as they become rarer (Madsen and Browning, 2009) along with flexibility of the threshold for a rare variant (Price *et al*., 2010) and accommodation of both risk and protective variants in a genetic region of interest (Ionita-Laza *et al*., 2011). Methods for more powerfully aggregating statistical signals include kernel-based adaptive clustering, which assigns weights to multi- rather than single-site genotypes (Liu and Leal, 2010), and the C_α test statistic, which contrasts the observed versus expected variance of binomially distributed allele counts (Neale *et al*., 2011), and regression-based models that extract and aggregate signals (Lee *et al*., 2012; Lin and Tang, 2011; Wu *et al*., 2011).

However, these existing rare variant detection methods are not able to identify which rare variants in a gene or a genetic region out of all variants are associated with the complex diseases or traits. Once phenotypic associations of a gene or a genetic region are identified, the natural next step in the association study with sequencing data is to locate the susceptible rare variants within the gene or the genetic region. There have been a few testing procedures based on the subset selection of rare variants such as the variable thresholding (VT) (Price *et al*., 2010) and RARECOVER (Rcover) (Bhatia *et al*., 2010) methods. However, VT is not designed to select potentially causal variants within a gene or a genetic region. Rcover collapses multiple rare variants within a gene or a genetic region using the combined multivariate and collapsing test (CMC) proposed by Liu and Leal (2008). It has low power to identify causal variants when both risk and protective variants are present within a gene or a genetic region. Moreover, Rcover applies the Pearson's χ^2 statistic in the testing procedure, which may not be optimal for rare variants and is limited to case-control designs only. Recently, Bayesian hierarchical models have been proposed to estimate the effects of rare variants under a regression framework (Capanu and Begg, 2011; Yi *et al*., 2011), where Bayesian credible regions for regression coefficients were provided to assess the

*To whom correspondence should be addressed.

association of each individual rare variant. However, credible regions for variants with low frequencies are too wide, indicating the uncertainty for estimation. In addition, estimation results in Bayesian analysis often rely on settings of unknown hyper-parameters, and intensive computing is required all the time.

Zhou *et al.* (2010) considered the problem of locating susceptible variants as a statistical variable selection problem, and attempted to select both common and rare variants associated with a disease outcome using a penalized regression model, where regression coefficients were regularised to select variants with relatively high impacts on the outcome. In penalised regression models, the total number of non-zero coefficients is tuned by a regularization parameter, so the choice of an optimal regularization parameter is a crucial part of the l_1-norm regularization procedure. However, Zhou *et al.* (2010) did not try to select an optimal regularization parameter, as common methods such as cross-validation and resampling might not be applicable to rare variants because of extreme sparsity. Alternatively, the authors ranked the potentially disease-related variants with their impact on the outcome. Thus, the method is essentially unable to separate potentially causal variants from non-causal variants within a gene or a genetic region.

In this article, we propose a statistical selection procedure based on the linear combinations of the power set of a set of rare variants in a gene or a genetic region that is able to identify the locations of the potentially susceptible rare variants within a disease-related gene or genetic region. A stage I association test first identifies the phenotypic association of a gene or a genetic region using existing methods that aggregate multiple rare variants. The proposed method then selects the subset out of the power set of all the rare variants that has the highest impact on the outcome (either quantitative or qualitative).

The selection performance of the proposed method was evaluated through simulation studies, where different effect sizes, sample sizes, and proportions of risk variants, protective variants and non-causal variants were considered. We also applied the proposed method to the sequence data of the ANGPTL gene family from the Dallas Heart Study (DHS). Several studies have worked on the DHS data and have already identified some genes that are related to the traits (Cheung *et al.*, 2012; Liu and Leal, 2010, 2012; Price *et al.*, 2010; Wu *et al.*, 2011; Yi *et al.*, 2011). For example, associations between the ANGPTL4 gene and both triglyceride and very low density lipoprotein in European American population have been suggested by several studies (Liu and Leal, 2010, 2012). However, most of the existing rare variant association tests have not tried to locate potentially susceptible rare variants within the ANGPTL4 gene. Our proposed selection procedure was able to select several potentially causal rare variants within the ANGPTL4 gene that are associated with triglyceride and very low density lipoprotein.

2 METHODS

Assume we observe the number of mutations over p rare variants in a gene or a genetic region from a sequence data of n individuals. We denote the dataset of the ith individual as (y_i, x_i, u_i), $i=1,\dots,n$, where $x_i=(x_{i1},\dots,x_{ip})^{\mathrm{T}}$ is the p dimensional counting vector with $x_{ij}\in\{0,1,2\}$, $j=1,\dots,p$ and $u_i=(u_{i1},\dots,u_{im})^{\mathrm{T}}$ is the m dimensional covariate vector such as age and gender. The phenotypic outcome y_i can be either quantitative or binary for case-control status, where $y_i=1$ for cases and $y_i=0$ for controls.

To generate $K=2^p-1$ subsets of the power set of the p rare variants without the empty set, we introduce K weighting vectors for the p variants denoted as $\xi_k=(\xi_{k1},\dots,\xi_{kp})^{\mathrm{T}}$, $k=1,\dots,K$. We then code 0 as the exclusion of a rare variant and 1 as the inclusion of a rare variant in the weighting vector. For example, when $p=4$, we have a total of 15 subsets in the power set without the empty set such as $\{x_{i1}\}$, $\{x_{i2}\}$, $\{x_{i3}\}$, $\{x_{i4}\}$, $\{x_{i1},x_{i2}\}$, $\{x_{i1},x_{i3}\}$,..., $\{x_{i2},x_{i3},x_{i4}\}$, $\{x_{i1},x_{i2},x_{i3},x_{i4}\}$. The corresponding weighting vectors are then $\xi_1=(1,0,0,0)^{\mathrm{T}}$, $\xi_2=(0,1,0,0)^{\mathrm{T}}$, $\xi_3=(0,0,1,0)^{\mathrm{T}}$, $\xi_4=(0,0,0,1)^{\mathrm{T}}$, $\xi_5=(1,1,0,0)^{\mathrm{T}}$, $\xi_6=(1,0,1,0)^{\mathrm{T}},\dots$, $\xi_{14}=(0,1,1,1)^{\mathrm{T}}$, $\xi_{15}=(1,1,1,1)^{\mathrm{T}}$, respectively. As we exclude the empty set from the subsets of the power set, the K weighting vectors all have at least one '1' out of the p indicators.

We then define the new feature of the subset k for the ith individual as

$$z_{ik}=\sum_{j=1}^{p}\xi_{kj}\omega_j x_{ij}^{*},$$

where

$$x_{ij}^{*}=\begin{cases}1-x_{ij} & \text{if variant } j \text{ is potentially protective;}\\ x_{ij} & \text{otherwise.}\end{cases}$$

for $k=1,\dots,K$. The new feature z_{ik} can be viewed as a weighted linear combination of the subset k, where variants included in the subset k are combined. It is known that the association test on the weighted sum of multiple rare variants could be underpowered if both risk and protective variants are present in a gene or a genetic region (Pan, 2009). Han and Pan (2010) proposed a data-adaptive procedure to take into account this problem. In their procedure, potentially protective variants are first identified via marginal association tests. That is, each variant is tested one at a time. If the regression coefficients from the marginal association tests are negative and the *P*-value of the regression coefficients $<\alpha$, where $\alpha=0.1$, was recommended by Han and Pan (2010) based on their simulation experiments, then the variants are considered potentially protective. We adopted this data-adaptive procedure to identify potentially protective variants x_{ij}^{*}. We then flip the coding of x_{ij} in the new feature z_{ik} if variant j is potentially protective. In addition to a binary weighting $\xi_{kj}\in\{0,1\}$, we allow each variant to have a different weight $\omega_j>0$. For instance, Madsen and Browning (2009) proposed to weight variants inversely proportional to the SD of estimated allele frequencies, so that rarer variants are up-weighted. Price *et al.* (2010) used the thresholds of the allele frequencies. Recently, Lin and Tang (2011) showed that the estimated regression coefficient of an individual variant is an optimal weight with relatively large sample.

Because each feature consists of a different combination of multiple rare variants from the p rare variants in a gene or a genetic region, our goal here is to find the most outcome-related feature $z_{i\hat{k}}$ among the K new features. Ideally, the subset $\hat{k}$ contains only causal variants, i.e. $\xi_{\hat{k}j}=1$ for causal variants and $\xi_{\hat{k}j}=0$ for non-causal variants. To select the most outcome-related feature $z_{i\hat{k}}$ among the K new features, we first adjust the quantitative or binary phenotype outcomes y_i for the covariates u_i by fitting a linear regression or a logistic regression, respectively. Let us denote the ith residual from the fitted regression by $\tilde{y}_i$. We then individually test each feature z_{ik} for association with the residual $\tilde{y}_i$ and choose the one that has the maximum test statistic, i.e.

$$\hat{k}=\underset{1\le k\le K}{\arg\max}\ T(\tilde{y},z_k),$$

where $\tilde{y}=(\tilde{y}_1,\dots,\tilde{y}_n)^{\mathrm{T}}$, $z_k=(z_{1k},\dots,z_{nk})^{\mathrm{T}}$, and $T(a,b)$ is a test statistic for association between a and b.

Several association test statistics such as a marginal test, a score statistic (Lin and Tang, 2011) and a z-score (Price *et al.*, 2010) can be applied for $T(\cdot,\cdot)$ to measure the strength of association between $\tilde{y}$ and z_k. Here we simply use a sample correlation between $\tilde{y}$ and z_k to find the most outcome-related feature for computational speed-up. The feature having the highest correlation with the outcome equivalently has the maximum test statistic for testing a slope of linear regression of the outcome on the feature.

One limitation of the proposed method is computational intensity for a gene or a genetic region with a large number of rare variants. For example, we need to compute and compare $K > 10^6$ number of test statistics when $P = 20$. In the algorithm we developed, which we implemented in a statistical software R combined with Fortran codes, it took only 1.38 s to get the maximum statistic for both quantitative and binary outcomes with $P = 17$ ($K > 10^5$) and 500 subjects, ~12 s with $P = 20$ ($K > 10^6$), and ~3.5 min with $P = 24$ ($K > 10^7$). When the analysis is limited to a handful of genes, i.e. for disease-related genes that have already been detected by a stage I association test, the proposed method is computationally feasible to locate susceptible rare variants within the genes or the genetic regions. For genes with extremely large number of rare variants, one potential solution is to cut the gene or the genetic region with a large number of variants into smaller segments first, then apply a stage I association test on those segments and perform the proposed power set-based selection method to the significant segments only to select potentially causal rare variants. This way, the computational burden can be much reduced.

3 RESULTS

3.1 Simulation studies

We conducted simulation studies to evaluate the selection performance of the proposed method for both quantitative and case-control binary traits. We first generated MAF of P variants from the Wright's distribution (Ionita-Laza *et al.*, 2011; Madsen and Browning, 2009)

$$f(q) = cq^{\delta-1}(1-q)^{\delta/3-1}e^{12(1-q)},$$

where $\delta = 0.001$ and c is a normalizing constant. We focused on rare variants with $0.001 \leq q \leq 0.01$. Given the MAFs of the P variants, genotype data $(x_{i1}, \ldots, x_{ip})$ were generated under Hardy–Weinberg equilibrium. The sample size was set at $n = 1000$, 2000 and 5000. We considered equal numbers of cases and controls in a case-control setting.

Similar to Wu *et al.* (2011) and Lee *et al.* (2012), we set the regression coefficients β_j, $j = 1, \ldots, p$ for p variants to have different values to separate causal (risk or protective) and noncausal variants where

$$\beta_j = \begin{cases} \gamma|\log_{10}\text{MAF}_j|, & \text{if variant } j \text{ is risk;} \\ -\gamma|\log_{10}\text{MAF}_j|, & \text{if variant } j \text{ is protective;} \\ 0, & \text{if variant } j \text{ is noncausal.} \end{cases}$$

This setting assumes rarer variants to have larger effects, and $\gamma > 0$ controls the effect sizes of causal variants. We set a sequence of γ from 0.2 to 0.7 increased by 0.1. The average effect sizes of the risk variants are $\sim\text{avg}(|\beta_j|) = 0.5$ for $\gamma = 0.2$ and $\text{avg}(|\beta_j|) = 1.75$ for $\gamma = 0.7$. Accordingly, the average odds ratios of the risk variants with case-control outcomes are around 1.65 and 5.75, respectively.

We simulated the quantitative outcome of the ith individual from the following regression model

$$y_i = 0.5u_{i1} + 0.5u_{i2} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon,$$

where covariate u_{i1} follows a Bernoulli distribution with a probability 0.5 and covariate u_{i2} follows a standard normal distribution. An error term ε was simulated from a standard normal distribution. For case-control outcomes, we first simulated quantitative outcomes with twice of the required sample size, and then set cases ($y_i = 1$) to be the top 25% of the quantitative outcomes and controls ($y_i = 0$) to be the bottom 25% of the outcomes.

The total number of variants p within a gene was fixed as 15, each of which could be a risk (R), protective (P) or non-causal (N) variant. We considered three variant-mix scenarios: (i) 5R/0P/10N, (ii) 3R/2P/10N and (iii) 5R/5P/5N. All simulation results were summarized based on 1000 simulation replicates. For each simulation replicate, we first applied the SKAT association test (Wu *et al.*, 2011) to assess the significance of a gene. If the association of a gene was identified (i.e. the P-value of SKAT is <0.05), we then applied the proposed selection procedure to locate potentially causal variants within the gene. We considered three versions of the proposed method: the power set-based selection procedure that ignores the mixed risk and protective variants (Pset), the data-adaptive power set-based procedure that adopts Han and Pan's method to identify potentially protective variants (aPset) and the weighted data-adaptive power set-based procedure that gives different weights to different variants (wPset), where we used the weight proposed by Madsen and Browning (2009). For comparison purposes, we also applied Rcover (Bhatia *et al.*, 2010) where the χ^2 test statistic was replaced by a sample correlation with residuals, as the original Rcover is limited to case-control outcomes and also cannot handle with covariates.

In the first simulation study, the averaged selection proportions (ASP) of risk, protective and non-causal rare variants were computed for Rcover, Pset, aPset and wPset procedures with three variant-mix scenarios, different effect sizes and different sample sizes. In each simulation, the selection proportion of risk variants is defined as the number of selected risk variants divided by the total number of true risk variants. The selection proportions of protective and non-causal variants are defined similarly. ASP is then defined as the averaged selection proportion over 1000 simulations for each type of variants. The larger the ASP of causal (risk and protective) variants and the smaller the ASP of non-causal variants, the better the selection performance.

Figure 1 shows the ASPs for causal and non-causal variants of the three variant-mix scenarios when sample sizes $n = 1000$, 2000 and 5000, the trait is quantitative and the effect size is relatively small, i.e. $\text{avg}(|\beta_j|) = 0.5$. The power of SKAT for a group of 15 variants is included in each figure on the top right. We can see that the power of SKAT increases as the number of causal (risk and protective) variants, or the sample size increases. The ASPs of the four selection procedures are similar when all causal variants are risk (**A**, **D** and **G**). However, when causal variants are either risk or protective, difference among the ASPs of the four selection procedures were noticeable. Specifically, Rcover and Pset have low ASPs when the number of protective variants is high (**C**, **F** and **J**). They selected <50% causal variants even when the sample size is large because Rcover and Pset are not able to identify causal variants correctly in the presence of both risk and protective variants in a gene or a genetic region.

In contrast, aPset and wPset selected most true causal variants together with a few non-causal variants when both risk and

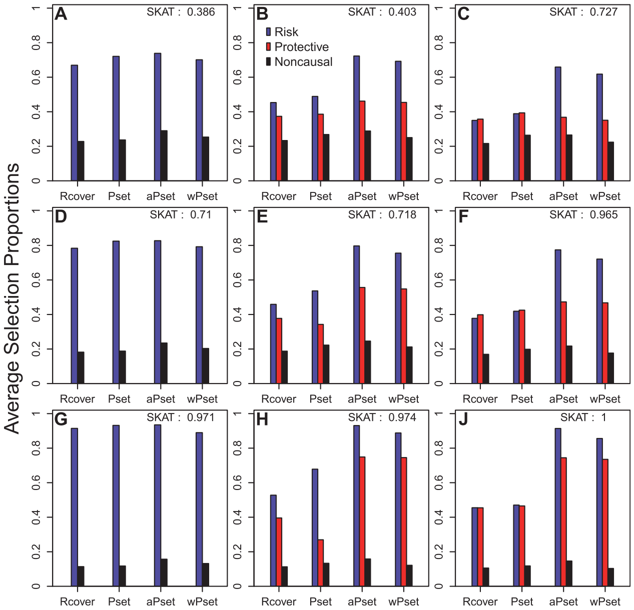

Fig. 1. Averaged selection proportions of risk (R), protective (P) and non-causal (N) rare variants for RARECOVER (Rcover), power-set-based (Pset), data-adaptive Pset (aPset) and weighted data-adaptive Pset (wPset) procedures when the outcome is quantitative and the effect size is avg($|\beta_j|$) = 0.5. The sample size is n = 1000 for **A–C**, n = 2000 for **D–F**, and n = 5000 for **G–J**. There are 5R/0P/10N variants in A, D and G, 3R/2P/10N in B, E and H, and 5R/5P/5N in C,F and J

protective variants are present. We noticed that ASPs of different types of variants for aPset and wPset were hardly affected by the proportions of causal and non-causal rare variants. For example, the ASPs of the non-causal variants of aPset are <25% when n = 2000 and <15% when n = 5000 in all variant-mix scenarios while those of wPset are lowered by 3–5%. Because aPset and wPset use the data-adaptive procedure (Han and Pan, 2010) for the initial screening of potentially protective variants, they may not perform well for small sample sizes and effect sizes, as the data-adaptive procedure is basically based on individual tests. Supplementary Figures S1 and S2 in the Supplementary Materials show ASP results when the effect sizes are avg($|\beta_j|$) = 0.75, and 1.0, respectively. In these settings, when sample size and effect size are large, the ASPs of protective variants are almost the same as those of risk variants for both aPset and wPset and the ASPs of the non-causal variants are lowered by ~3%. This might be because of the better performance of the data-adaptive procedure for the initial screening of potentially protective variants with individual tests when sample size and effect size are large. Similar patterns for case-control binary outcomes for the three effect size settings were observed in Supplementary Figures S3–S5 in the Supplementary Materials.

We also considered an additional simulation study where there is a big gene with 50 variants consisting of 10 risk, 5 protective and 35 non-causal variants. All other settings such as the sample sizes and the effect sizes remain the same as in the first simulation study. We first applied the SKAT stage I association test for the entire gene with 50 variants. Once a significant association signal is detected, we divided the gene into five segments so that each segment has 10 variants. We then reapplied the SKAT association test for each segment, and further applied Rcover, Pset, aPset and wPset for the segments with SKAT P-values <0.05. For segments that were not detected by the SKAT method, we did not apply the selection procedure and considered variants within the segments non-causal. Average selection proportions of risk, protective and non-causal variants were summarized in Supplementary Figures S6 and S7 in the Supplementary Materials for quantitative outcomes and binary outcomes, respectively. Similar to what we observed in the smaller gene setting, the proposed aPset and wPset have the best selection performance in all simulation settings.

In the second simulation study, we computed the selection power of each selection procedure, where the selection power is defined as the proportion that a procedure selects exactly all of the causal variants among 1000 simulations. The probability of randomly selecting only all of causal variants is $1/(2^{15}-1) \approx 0.3 \times 10^{-4}$ in each simulation replicate when the exact number of the causal variants is unknown. Figure 2 displays the selection power of the four procedures as the effect size avg($|\beta_j|$) increases from 0.5 to 1.75 when the traits are case-control binary outcomes. Two variant-mix scenarios: (i) 5R/0P/10N and (ii) 3R/2P/10N were considered along with different sample sizes (n = 1000, 2000 and 5000). The corresponding selection power for quantitative outcomes is summarized in Supplementary Figure S8 in the Supplementary Materials.

As expected, the power of both Rcover and Pset does not increase at all as the effect size increases when both risk and protective variants are present in a gene or a genetic region (**D–F**). In fact, they never identified exactly all of the causal variants in the variant-mix scenario (ii) 3R/2P/10N. When all causal variants are risk variants, Rcover and Pset have slightly higher power than aPset, as aPset could misidentify potentially protective variants (**A–C**). In all scenarios, wPset has the highest selection power as expected, as the weight (Madsen and Browning, 2009) was proposed for case-control outcomes to boost association signals. For quantitative outcomes, the selection powers of aPset and wPset are hardly different. Thus, the development of a different optimal weight for quantitative outcomes may improve the selection power of the proposed selection procedure for quantitative traits.

3.2 Analysis of the DHA

We applied the proposed power set-based selection procedure to the DHS data (Romeo *et al.*, 2007, 2009). Coding regions of four genes ANGTPL3, ANGTPL4, ANGTPL5 and ANGTPL6 were sequenced to detect the association with nine energy metabolism traits, namely triglyceride (TG), low-density lipoprotein (LDL), very low-density lipoprotein (VLDL), high-density lipoprotein, cholesterol, glucose, body mass index, systolic (SysBP) and diastolic blood pressure. A total of 348 nucleotide sites of sequence variations were discovered in these four genes, where the majority of them are rare (MAF <5%). We focused on the European American population to identify trait-related genes, and then to identify potentially susceptible rare variants within the trait-related genes.

We analysed the nine traits in two ways, either as the original quantitative traits or as the binary case-control outcomes, where

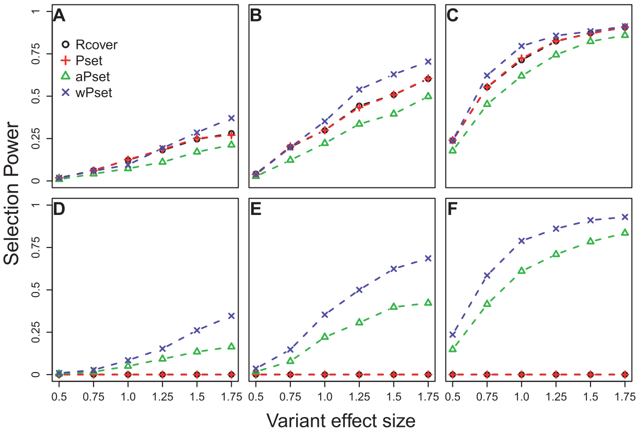

Fig. 2. Selection power of RARECOVER (Rcover), power set-based (Pset), data-adaptive Pset (aPset) and weighted data-adaptive Pset (wPset) procedures are displayed as the effect size avg($|\beta_j|$) increases when the outcome is binary (case-control status). There are 5 Risk/0 Protective/10 Non-causal variants in **A–C**, and 3 Risk/2 Protective/10 Non-causal variants in **D–F**. The sample size is n = 1000 for A and D, n = 2000 for B and E, and n = 5000 for C and F

Table 1. VT and SKAT association analysis results with *P*-values for the genes ANGPTL3, ANGPTL4, ANGPTL5 and ANGPTL6 from the Dallas Heart Study sequence data with both quantitative and case-control outcomes for the European American population

Genes	Traits	VT	SKAT	Total variants
Quantitative outcomes				
ANGPTL4	TG	0.0040	0.0022	17
ANGPTL4	VLDL	0.0070	0.0021	17
Case-control outcomes				
ANGPTL4	TG	0.0010	0.0021	9
ANGPTL4	VLDL	0.0060	0.0021	9

the case and control groups are defined as the samples in the top 25th percentile and in the bottom 25th percentile, respectively. Gender and logarithm transformed age were included in the model as covariates. We also limited the analysis to the non-synonymous rare variants with MAF $\leq 3\%$ (Liu and Leal, 2010, 2012). For the stage I association test to identify trait-related genes, we applied VT (Price *et al.*, 2010) and SKAT (Wu *et al.*, 2011) methods for the four genes. We then applied the proposed power set-based selection procedure on the identified genes and the associated traits to locate potentially susceptible rare variants within these genes.

Because for each of the quantitative or binary case-control outcome we applied VT and SKAT to four genes to detect association signals, a Bonferroni-adjusted significance level $\alpha = 0.05/8 = 0.0063$ was used to claim any significant association signals. Two traits, TG and VLDL, are associated with the gene ANGPTL4 at the multiple comparison adjusted significance level (Table 1). This association analysis result is consistent with the result in a recent application of the DHS data (Liu and Leal, 2012) where CMC, weighted sum statistic test (Madsen and Browning, 2009) and VT were applied but only to the quantitative traits. We then applied the four selection procedures (Rcover, Pset, aPset and wPset) to identify potentially causal variants within the gene ANGPTL4. Fewer number of variants of the gene ANGPTL4 were included in the analysis with the binary case-control outcome than the quantitative outcome. This is because we included variants with at least one mutation in both analyses. As fewer samples were included in the binary case-control analysis, there were more variants with no mutations in the binary case-control analysis.

The lists of the selected rare variants in the gene ANGPTL4 for the binary case-control outcomes are summarized in Table 2. The lists of the selected rare variants in the gene ANGPTL4 for the quantitative outcomes are included in Supplementary Table S1 in the Supplementary Materials. As the selection result of Pset is exactly the same as that of Rcover in all analyses, we omitted Pset in Table 2 and Supplementary Table S1. This is because most of the rare variants in the gene ANGPTL4 are singletons, thus summing p variants as in the proposed selection procedure is not much different from collapsing p variants as in Rcover. However, aPset and wPset show quite different selection results compared with Rcover. In the analyses of binary outcomes, aPset selected three rare variants for TG and four rare variants for VLDL whereas Rcover selected most of the variants in the gene ANGPTL4. Consistent with the simulation results, wPset selected the same or fewer number of rare variants than aPset did.

All four selection procedures selected the variant X1313_E40K in the gene ANGPTL4 for the traits TG and VLDL. The variant X1313_E40K has the largest difference in the number of mutations observed between cases and controls for both traits. This variant could be a candidate causal variant for future validation study. aPset also selected the variants X6113_E190Q and X8020_P251T for both traits and the variant X8262_S302fs for the trait VLDL. For the three variants selected, there is only one mutation in cases and no mutation in controls. Further investigation is necessary for these variant. We also noticed that for the variant X8364_R336C, although there are two mutations in controls and no mutations in cases for the binary trait TG (and three mutations in controls and no mutations in cases for the binary trait VLDL), it was not selected by either aPset or wPset. This may be because this variant was not identified as a potentially protective variant by the data-adaptive procedure (Han and Pan, 2010) in the screening step. The variant X1313_E40K was identified as a potentially protective variant in the screening step and its genotype coding was then flipped. We observed the similar results for the quantitative traits. Another pattern we noticed is, Rcover selected all variants with more mutations in controls (potentially protective variants) but did not select any variants with more mutations in cases (potentially risk variants), which may be owing to the strong effect of the potentially protective variant X1313_E40K and the fact that Rcover does not handle mixed effects.

Additionally, we also applied the proposed selection procedure to the Hispanic population of the DHS sequencing data. Cheung *et al.* (2012) have identified an association of the gene ANGPTL5 with the SysBP in the Hispanic group although the

Table 2. Susceptible rare variants in the ANGPTL4 gene that is associated with the binary TG and VLDL traits for the European American population

Variants	MAF	Cases	ctrls	Rcover	aPset	wPset
ANGPTL4—binary TG						
X1313_E40K	0.0150	1	11	×	×	×
X3145_E167K	0.0023	0	1	×		
X6113_E190Q	0.0023	1	0		×	
X7936_G223R	0.0023	0	1	×		
X8003_K245fs	0.0023	0	1	×		
X8020_P251T	0.0023	1	0		×	
X8262_S302fs	0.0023	0	1	×		
X8364_R336C	0.0035	0	2	×		
X10621_G361S	0.0023	0	1	×		
ANGPTL4—binary VLDL						
X1313_E40K	0.0152	1	12	×	×	×
X3145_E167K	0.0022	0	1	×		
X6113_E190Q	0.0022	1	0		×	×
X7936_G223R	0.0022	0	1	×		
X8003_K245fs	0.0022	1	0		×	×
X8020_P251T	0.0022	1	0		×	×
X8262_S302fs	0.0022	0	1	×		
X8364_R336C	0.0043	0	3	×		
X10621_G361S	0.0022	0	1	×		

Note: × indicates a selected variant.
'cases' and 'ctrls' indicate the number of mutations in cases and controls, respectively.

association *P*-value is not extremely small. Our repeated analysis using SKAT stage I association test has a *P*-values of 0.0367 and 0.0370 for the quantitative and binary SysBP traits, respectively. We then applied the four selection procedures to the gene ANGPTL5 to separate potentially causal variants from noncausal variants within the gene. The selection result is shown in Supplementary Table S2 in the Supplementary Materials. It is noticeable that in the analysis of binary SysBP trait, the proposed selection procedures identified both risk and potentially protective variants S93N, L98P and T268M, whereas Rcover only identified two potentially protective variants L98P and T268M. Moreover, Rcover selected two completely different lists of variants for the quantitative and binary SysBP traits, whereas aPset and wPset performed consistent selections in both analyses with the quantitative and binary traits.

4 DISCUSSION

In this article, we proposed a power set-based selection procedure to identify the locations of susceptible rare variants associated with complex diseases with sequencing data. The proposed selection procedure not only aggregates multiple rare variants to boost association signals, but also locates potentially causal and noncausal variants within a disease-related gene or a genetic region. Using the idea of the power set of all rare variants in a gene or a genetic region, the most outcome-related subset of rare variants among the power set is identified. Our simulation studies suggested that the proposed power set-based selection procedures (aPset and wPset) are able to locate potentially causal variants relatively accurately when both risk and protective variants are present in a gene or a genetic region. In the analysis of the DHS sequencing data, the best subset of variants selected by the proposed method are promising and could be potential trait-related variants for future validation.

Selecting susceptible rare variants is statistically challenging because of sparsity. To improve the selection performance of the proposed power set-based selection procedure, there are three potential extensions. First, we used the data-adaptive procedure (Han and Pan, 2010) to screen potentially protective variants. However, the procedure may not perform well when sample size and effect size are small, as the procedure is based on a marginal association test. We noticed in the simulation studies that ASPs for both risk and protective variants could be improved by a maximum of 10% (data not shown) if we knew which variants are protective. Therefore, it is essential to develop an improved procedure to identify potentially protective variants. Second, we used allele frequency weights (Madsen and Browning, 2009) for wPset and observed a similar performance of wPset and aPset in quantitative analysis. One possible explanation is that this weight is designed for binary case-control outcomes. Another possible explanation is that the weight is assigned on the individual variant level while the selection is on the feature level. Thus, the variant-level weights may not influence much on selecting the most outcome-related feature. The development of optimal weights for each feature rather than each variant may improve the selection performance of the proposed method. Third, we combined variants within a subset using a weighted linear combination. A recent publication by Chen *et al*. (2012) has suggested that an exponential combination might be more powerful than a linear combination to detect rare variant associations when the number of causal variants within a gene or a genetic region is small. In addition to the improvement in the detection power, the exponential combination procedure (Chen *et al*., 2012) combines both risk and protective variants without the need of initial screening of protective variants, as it is not affected by the directions of variants. However, the numerical values of the exponential combinations depend on the number of variants when the test statistics of individual variants are exponentially combined, i.e. the more the variants that are combined, the larger statistic it has. Thus, the proposed power set idea cannot be directly applied to the exponential combinations. Some modification of the exponential combination may be preceded to extend the proposed power set method. To improve in these three areas are our future work.

Our proposed selection procedure can be easily applied to different types of phenotypic outcomes. We limited an outcome to be either quantitative or binary in this article. However, because our procedure is based on an association with regression residuals, it is readily extended to other types of outcomes by simply applying the generalized regression model.

ACKNOWLEDGEMENT

The authors thank the Yale High Performance Computing center for providing the computing facility and Prof. Jonathan Cohen for providing the DHS data. The authors also thank Dr. Iuliana Ionita-Laza for helpful discussions.

Funding: Dr.Hokeun Sun was supported by a 2-Year Research Grant of Pusan National University.

Conflict of Interest: none declared.

REFERENCES

Bhatia,G. *et al.* (2010) A covering method for detecting genetic associations between rare variants and common phenotypes. *PLoS Comput. Biol.*, **6**, e1000954.

Capanu,M. and Begg,C.B. (2011) Hierarchical modeling for estimating relative risks of rare genetic variants: properties of the pseudo-likelihood method. *Biometrics*, **67**, 371–380.

Chen,L.S. *et al.* (2012) An exponential combination procedure for set-based association tests in sequencing studies. *Am. J. Hum. Genet.*, **91**, 977–986.

Cheung,Y. *et al.* (2012) A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. *Genet. Epidemiol.*, **36**, 675–685.

Han,F. and Pan,W. (2010) A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.*, **70**, 42–54.

Ionita-Laza,I. *et al.* (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.*, **7**, e1001289.

Lee,S. *et al.* (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, **91**, 224–237.

Lin,D.Y. and Tang,Z.Z. (2011) A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.*, **89**, 354–367.

Liu,D. and Leal,S. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Liu,D. and Leal,S. (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.*, **6**, e1001156.

Liu,D. and Leal,S. (2012) Estimating genetic effects and quantifying missing heritabaility explained by identified rare-variant associations. *Am. J. Hum. Genet.*, **91**, 585–596.

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

Neale,B. *et al.* (2011) Testing for an unusual distribution of rare variants. *PLoS Genet.*, **7**, e100132.

Pan,W. (2009) Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genet. Epidemiol.*, **33**, 497–507.

Price,A.L. *et al.* (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, **86**, 832–838.

Romeo,S. *et al.* (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, 513–516.

Romeo,S. *et al.* (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J. Clin. Invest.*, **119**, 70–79.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Yi,N. *et al.* (2011) Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. *PLoS Genet.*, **7**, e1002382.

Zhou,H. *et al.* (2010) Association screening of common and rare genetic variannts by penalized regression. *Bioinformatics*, **26**, 2375–2382.