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

Hokeun Sun¹ and Shuang Wang²,*

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

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.


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; Leal et al., 2011; 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 Cₜ 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., 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 Cₜ 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 χ² 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 (Capanna et al., 2011; Yi et al., 2011), where Bayesian credible regions for regression coefficients were provided to assess the
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 hyperparameters, 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 regularisation parameter, so the choice of an optimal regularisation parameter is a crucial part of the 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 ANGPTL4 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 $i$th individual as $(y_i, x_i, u_i)$, $i = 1, \ldots, n$, where $x_i = (x_{i1}, \ldots, x_{ip})$ is the $p$ dimensional counting vector with $x_{ij} \in \{0, 1, 2\}$, $j = 1, \ldots, p$ and $u_i = (u_{i1}, \ldots, u_{im})$ 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 $\xi_k = (\xi_{k1}, \ldots, \xi_k),$ $k = 1, \ldots, K$. 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 $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_1, x_4\}$, $\{x_2, x_3, x_4\}$, $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$, $\{x_2, x_3, x_4\}$. The corresponding weighting vectors are then $\xi_1 = (0, 0, 0, 0)^T$, $\xi_2 = (0, 0, 0, 1)^T$, $\xi_3 = (0, 0, 1, 0)^T$, $\xi_4 = (0, 0, 1, 1)^T$, $\xi_5 = (1, 0, 0, 0)^T$, $\xi_6 = (1, 0, 0, 1)^T$, $\xi_7 = (1, 0, 1, 0)^T$, $\xi_8 = (1, 0, 1, 1)^T$, $\xi_9 = (1, 1, 0, 1)^T$, $\xi_{10} = (1, 1, 1, 0)^T$, $\xi_{11} = (1, 1, 1, 1)^T$, $\xi_{12} = (0, 0, 0, 0)^T$.

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 ANGPTL4 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 $i$th individual as $(y_i, x_i, u_i)$, $i = 1, \ldots, n$, where $x_i = (x_{i1}, \ldots, x_{ip})$ is the $p$ dimensional counting vector with $x_{ij} \in \{0, 1, 2\}$, $j = 1, \ldots, p$ and $u_i = (u_{i1}, \ldots, u_{im})$ 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 $\xi_k = (\xi_{k1}, \ldots, \xi_k),$ $k = 1, \ldots, K$. 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 $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_1, x_4\}$, $\{x_2, x_3, x_4\}$, $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$, $\{x_2, x_3, x_4\}$. The corresponding weighting vectors are then $\xi_1 = (0, 0, 0, 0)^T$, $\xi_2 = (0, 0, 0, 1)^T$, $\xi_3 = (0, 0, 1, 0)^T$, $\xi_4 = (0, 0, 1, 1)^T$, $\xi_5 = (1, 0, 0, 0)^T$, $\xi_6 = (1, 0, 0, 1)^T$, $\xi_7 = (1, 0, 1, 0)^T$, $\xi_8 = (1, 0, 1, 1)^T$, $\xi_9 = (1, 1, 0, 1)^T$, $\xi_{10} = (1, 1, 1, 0)^T$, $\xi_{11} = (1, 1, 1, 1)^T$, $\xi_{12} = (0, 0, 0, 0)^T$.

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 ANGPTL4 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.
outcomes and controls (y = 0) to be the bottom 25% of the outcomes.

The total number of variants p within a gene was fixed at 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 χ² 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. avg(|β|) = 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

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)

where δ = 0.001 and c is a normalizing constant. We focused on rare variants with 0.001 ≤ q ≤ 0.01. Given the MAFs of the P variants, genotype data (x₁, ..., xₚ) 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 βⱼ, j = 1, ..., p, for p variants to have different values to separate causal (risk or protective) and non-causal variants where

This setting assumes rarer variants to have larger effects, and γ > 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 |avg(βⱼ)| = 0.5 for γ = 0.2 and |avg(βⱼ)| = 1.75 for γ = 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 i-th individual from the following regression model

where covariate u₁ follows a Bernoulli distribution with a probability 0.5 and covariate u₂ 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 (ŷ = 1) to be the top 25% of the quantitative outcomes and controls (ŷ = 0) to be the bottom 25% of the outcomes.

The total number of variants p within a gene was fixed at 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 χ² 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.
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 $\text{avg}(\beta_j) = 0.5$. 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 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 $1/(2^n - 1) \approx 0.3 \times 10^{-4}$ 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 $\text{avg}(\beta_j)$ 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 triacylglyceride (TG), low-density lipoprotein (LDL), very low-density lipoprotein (VLDL), high-density lipoprotein, cholesterol, glucose, body mass index, systolic 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...
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 ≤ 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 α = 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

<table>
<thead>
<tr>
<th>Variants</th>
<th>MAF</th>
<th>Cases</th>
<th>ctrls</th>
<th>Rcover</th>
<th>aPset</th>
<th>wPset</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANGPTL4—binary TG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X1313_E40K</td>
<td>0.0150</td>
<td>1</td>
<td>11</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>X3145_E167K</td>
<td>0.0023</td>
<td>0</td>
<td>1</td>
<td>×</td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>X6113_E190Q</td>
<td>0.0023</td>
<td>1</td>
<td>0</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X7936_G223R</td>
<td>0.0023</td>
<td>0</td>
<td>1</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X8003_K245fs</td>
<td>0.0023</td>
<td>0</td>
<td>1</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X8020_P251T</td>
<td>0.0023</td>
<td>1</td>
<td>0</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X8262_S302fs</td>
<td>0.0023</td>
<td>0</td>
<td>1</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X8364_R336C</td>
<td>0.0035</td>
<td>0</td>
<td>2</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X10621_G361S</td>
<td>0.0023</td>
<td>0</td>
<td>1</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANGPTL4—binary VLDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X1313_E40K</td>
<td>0.0152</td>
<td>1</td>
<td>12</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>X3145_E167K</td>
<td>0.0022</td>
<td>0</td>
<td>1</td>
<td>×</td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>X6113_E190Q</td>
<td>0.0022</td>
<td>1</td>
<td>0</td>
<td>×</td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>X7936_G223R</td>
<td>0.0022</td>
<td>1</td>
<td>0</td>
<td>×</td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>X8003_K245fs</td>
<td>0.0022</td>
<td>1</td>
<td>0</td>
<td>×</td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>X8020_P251T</td>
<td>0.0022</td>
<td>1</td>
<td>0</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X8262_S302fs</td>
<td>0.0022</td>
<td>0</td>
<td>1</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X8364_R336C</td>
<td>0.0043</td>
<td>0</td>
<td>3</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X10621_G361S</td>
<td>0.0022</td>
<td>0</td>
<td>1</td>
<td>×</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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**


