Original Paper

Human Heredity

Hum Hered 2012;74:51–60
DOI: 10.1159/000343797

Received: April 30, 2012
Accepted after revision: September 30, 2012
Published online: November 13, 2012

An Optimum Projection and Noise Reduction Approach for Detecting Rare and Common Variants Associated with Complex Diseases

Asuman Turkmen[a, b] Shili Lin[a]

[a]Department of Statistics, The Ohio State University, Columbus, Ohio, and [b]The Ohio State University at Newark, Newark, Ohio, USA

Key Words
Missing heritability · Noise reduction · Partial least squares · Rare variants · Regularization/LASSO · Super variant

Abstract
Background: Despite the thrilling advances in identifying gene variants that influence common diseases, most of the heritable risk for many common diseases still remains unidentified. One of the possible reasons for this missing heritability is that the genome-wide association study (GWAS) approaches have been focusing on common rather than rare single nucleotide variants (SNVs). Consequently, there is currently a great deal of interest in developing methods that can interrogate rare variants for association with diseases. ***Methods:*** We propose a two-step method (termed rPLS) to reveal possible genetic effects related to rare as well as common variants. The procedure starts with removing irrelevant variants using penalized regression (regularization) which is followed by partial least squares (PLS) on the surviving SNVs to find an optimal linear combination of rare and common SNVs within a genomic region that is tested for its association with the trait of interest. ***Results:*** Simulation settings based on the 1000 Genomes sequencing data and reflecting real situations demonstrated that rPLS performs well compared to existing methods especially when there are a large number of noncausal variants (both rare and common) present in the gene and when causal SNVs have different effect sizes and directions.

Introduction

Many methods for testing association with common variants have been extensively developed due to the long-held 'common disease common variants' (CDCV) assumption [1, 2] suggesting that common variants may hold the secrets to many disease susceptibilities. As a result of this, genome-wide association studies (GWASs) have led to the identification of well over one thousand common single nucleotide variants (SNVs) associated with more than 200 complex traits. However, most of these associated common variants have very small effect size and it becomes clear that the CDCV model cannot explain the 'missing heritability' of complex traits [3]. One popular paradigm for explaining this missing genetic component is the 'common disease rare variants' (CDRV) hypothesis, in which the complex traits are as-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/12/0741–0051$38.00/0

Accessible online at:
www.karger.com/hhe

Asuman Turkmen, PhD
The Ohio State University at Newark
1179 University Drive
Newark, OH 43055 (USA)
E-Mail turkmen@stat.osu.edu

sumed to be caused by multiple rare variants with moderate to large effects in addition to common ones, and it has shown promise in explaining disease etiology in multiple studies. For instance, it has been reported that rare variants in three genes (SLC12A3, SLC12A1 and KCNJ1) contribute to the reduction in blood pressure and protection from hypertension [4]. Cohen et al. [5] found that multiple rare variants in three genes (ABCA1, APOA1 and LCAT) significantly contribute to low plasma HDL cholesterol level.

With improvements in the efficiency of next-generation sequencing technologies, assessing the role of rare variants in complex diseases is becoming increasingly cost-effective and feasible. This technological development leads to a necessity for effective statistical methods to detect effect of rare variants in a gene on disease susceptibility. Although methods developed for analysis of common variants, such as single-marker tests, can be easily extended to rare variants, they suffer from reduced power due to low frequency of rare SNVs even in very large samples. Therefore, a number of statistical methods have been developed and proven to be powerful for detecting associations with rare SNVs [6].

Collapsing methods constitute a broad class in the current literature. The basic idea behind collapsing methods is aggregating low-frequency variants in the same genomic region (such as a gene) so that they may be common enough to account for variation in common traits. The cohort allelic sums test (CAST) [7] is one of the existing collapsing strategies; it is a simple grouping method that compares the number of individuals having at least one of the rare variants between cases and controls by using the x^2 or Fisher exact test. Li and Leal [8] modified the CAST to improve its performance with both rare and common variants. In their combined multivariate and collapsing (CMC) method, rare variants are first divided into two subgroups on the basis of predefined criteria (e.g. allele frequencies = 0.01 as the threshold); the variants within the rare subgroup are collapsed into a single variant while keeping each common variant without collapsing. Then, a multivariate test (e.g. Hotelling's T^2 test or Fisher product method) is applied to analyze the collapsed rare and common variants. Obviously, this method is sensitive to the threshold for grouping and may lead to decreased power. Madsen and Browning [9] introduced a nonparametric weighted sum test in which each variant within the functional unit is assigned a weight to give rare, highly penetrant, mutations greater influence on the test statistic. The incorporation of weights improves the power of the test unless common variants are functionally relevant. Since both CAST and weighted sum methods emphasize rare variants by either excluding or down-weighting common variants, CMC outperforms them when a genomic region has both rare and common causal variants. Nevertheless, all of these collapsing methods implicitly assume that all the rare variants affecting the trait act in the same direction. Therefore, they experience a loss in power when both protective and risk variants are present in the region under analysis. Another important pooling method is the sum test (SUM) that was originally designed for the analysis of multiple SNVs (not necessarily rare ones) within a genomic region and was demonstrated to work well for rare variants under certain conditions [10–12]. The sum test summarizes information across multiple SNVs with only one degree of freedom (DF) by creating a super-variant that is simply the sum of all SNVs and needs no multiple test adjustment since it tests the association between the super-variant and the trait. On the other hand, this test is based on an unrealistic assumption that not only SNVs affect the trait in the same direction but also with a common effect size. Although this approach would be reasonable to aggregate rare SNVs within a genomic region, it yields a low power when SNVs have different effect sizes.

Pan [11] proposed sum of squared score (SSU) test which can be considered as a modified score test feasible for high-dimensional data. The SSU test is shown to be equivalent to the permutation-based version of Goeman's test [13]. Since Goeman's test is a variance component score test for a random-effects regression model, SSU can be regarded as a variance component test. SSU test is known to be sensitive to the presence of noncausal common variants. The weighted version of the SSU (denoted by SSUw) [11] is proposed to remedy this drawback of SSU. SSUw employs weights that are inversely proportional to minor allele frequencies (MAFs), i.e. SNVs with lower MAFs are given higher weights, as in the weighted sum test, hence it is more resistant to the noncausal common variants than SSU. On the other hand, SSUw will suffer from power loss when there are causal common variants since it weights the variants based on their overall MAFs and a causal variant may have a higher MAF in cases but a lower MAF in controls resulting in a higher overall MAF across both cases and controls. Basu and Pan [14] proposed another weighted version of SSU to overcome the power loss issue of the SSUw (and SSU) in the presence of both causal rare SNVs and common SNVs in which the weights of the SNVs are determined using the MAFs of control samples only. The resulting method, denoted by wSSU-P, uses permutation to calculate p values.

In this study, we introduce a partial least squares (PLS) approach for identifying rare and common causal variants within a gene or genomic region in association studies. Although our method can deal with various phenotypes, we demonstrate its performance with a binary disease trait in population-based case-control studies so that comparisons can be easily made with existing methods discussed above. The methodology, termed rPLS, starts with a screening procedure based on penalized regression (regularization) that selects a subset of SNVs within the gene to reduce noise caused by the inclusion of noncausal variants. We then obtain an optimum linear combination of the SNVs surviving the first step. More specifically, the linear combination, termed the super-variant, is the first PLS component when we apply PLS on the trait and the surviving SNVs. This yields a variant that not only explains the variation in the SNVs but also maximizes the correlation with the trait. This allows varying signs and varying magnitudes of the contributing SNVs driven by the data themselves. Once we have constructed the super-variant, its association with the trait of interest is tested. Our approach, therefore, has the advantage of low DF as the collapsing based test while does not depend on the strict assumption that all SNVs have common effect size and sign, nor the need to use a rare variant threshold. Furthermore, the initial screening helps to reduce contamination of noncausal variants so that the noise will be lessened in the super-variant. Another important feature of the rPLS is its ability to identify the most important SNVs within a gene or genomic region, a unique feature not available in the other methods. We used extensive simulations, partially based on the Genetic Analysis Workshop 17 (GAW17) 1000 Genomes sequencing data [15], to evaluate the proposed method and to compare its performance with existing methods. Our results show that rPLS is competitive in all, and outperforms existing approaches in most of the simulations scenarios.

Materials and Methods

We first introduce notation and the hypothesis of interest. Suppose that a population-based association study consists of n unrelated individuals, with n_0 controls and n_1 cases, with their phenotypes for a binary disease trait denoted by $Y = [Y_1, Y_2, ..., Y_n]'$ where $Y_i = 1$ if i is a diseased sample and $Y_i = 0$ otherwise. Genotypes for p SNVs in a gene or a genomic region are available with $x_{ij} = 0$, 1 or 2 for $i = 1, 2, ..., n$; $j = 1, 2, ..., p$, coding for the number for minor alleles at locus j for individual i. We would like to test whether the SNVs within the candidate gene are associated with the trait, Y. In what follows, we explain our method that conducts the selection and test in a two-step procedure.

Proposed Method: rPLS

The key idea in the proposed method is to find an optimum linear combination of the SNVs within a gene so that one can detect associations that are too weak to be detected for individual variants. Typically, only a subset of SNVs within a gene are associated with disease outcome. Hence, using all SNVs to summarize information from a gene without a screening step can result in reduced test power due to inclusion of SNVs irrelevant to the disease. Thus, it is highly desirable to conduct an initial SNV selection step before combining information in the gene.

Here, we employ a penalized regression method to select a subset of important SNVs. LASSO [16] and Ridge [17] penalties have been applied to a variety of phenotype prediction tasks using genomic data. For low-dimensional settings, when there are high correlations between predictors, it has been empirically observed that the prediction performance of the LASSO is dominated by Ridge regression. Therefore, Zou and Hastie [18] proposed the Elastic Net (EN), with its penalty function being a weighted combination of L_1 and L_2 norms thus leading to LASSO and Ridge penalties as its two special cases. In our application, performances of both LASSO and EN will be investigated.

We consider the following logistic regression model between the trait and genotypes:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + x_i'\beta, \tag{1}$$

where $p_i = P(Y_i = 1|x_i)$ and $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]'$ for $1 \le i \le n$. The negative log-likelihood function using the EN penalty is given as:

$$L(\beta;\lambda,\alpha) = -\sum_{i=1}^{n}\left\{Y_i\log(p_i) + (1-Y_i)\log(1-p_i)\right\} + \lambda\sum_{i=1}^{p}\left\{\alpha|\beta_j| + (1-\alpha)\beta_j^2\right\}, \tag{2}$$

where $\lambda \ge 0$ is the regularization parameter (degree of penalty) and $0 \le \alpha \le 1$ is the weighting proportion associated with the L_1 norm. Since LASSO is a special case of EN, our discussion below focuses on that. Specifically, the EN estimator of β is

$$\hat{\beta}_{EN}(\lambda,\alpha) = \underset{\beta}{\operatorname{argmin}} L(\beta;\lambda,\alpha) \tag{3}$$

Once we have $\hat{\beta}_{EN}$, we can define the set of indices corresponding to the initial subset of important SNVs as

$$I_c = \{j : \hat{\beta}_{EN(j)} \neq 0\} \tag{4}$$

with cardinality $||I_c|| = k$ where $k \le p$. Using the I_c, a subset of genome matrix can be defined as $X^* = \{x_{ij} : 1 \le i \le n; j \in I_c\}$ which will be used in the subsequent analysis.

In the second step of the analysis, the purpose is to construct a super-variant that is a linear combination of the SNVs in X^*. The proposed method utilizes PLS to construct the super-variant. PLS is a member of the nonlinear iterative least squares (NILES) procedures developed by Wold [19]. The main idea of PLS is to summarize explanatory variables (i.e., X^*) into a smaller set of uncorrelated, so called latent, variables which have the 'optimal' predictive power for the response (i.e., Y). In general, the direction for the hth PLS component for X^* and Y is

$$r_h = \underset{r}{\operatorname{argmax}}((r'X^{*\prime}Y)^2) \quad \text{subject to } r_h' X^{*\prime}X^* r_j = 0 \text{ for } 1 \le j < h. \tag{5}$$

Table 1. Description of the first simulation setting

Gene	Rare variants			Common variants	
	causal %	noncausal %	direction of causal	causal n	noncausal n
1	100	0	same	–	–
2	100	0	different	–	–
3	30	70	same	–	–
4	100	0	same	1	–
5	30	70	same	1	–
6[a]	r_c	r_{nc}	same	1	c_{nc}

[a] In the 'Rare variants' section, the numbers given in the table are the percentages except for gene 6, where r_c and r_{nc} are the number of causal and noncausal rSNVs, respectively. In the 'Common variants' section, the values given are number of cSNVs where c_{nc} denotes the number of noncausal cSNVs, and '–' denotes the type of cSNVs was not considered. The distribution of causal and noncausal SNVs for gene 6 is such that $r_c + 1$ causal SNVs constitute 30% of all SNVs (i.e. $(r_c + 1)/(r_c + 1 + r_{nc} + c_{nc}) = 0.3$) and the remaining 70% (i.e., $r_{nc} + c_{nc}$) are noncausal SNVs.

The first PLS component is taken as the super-variant, Z, which is the projection of X^* on the $k \times 1$ vector r_1, that is, $Z = X^* r_1$.

Boulesteix [20] has shown that the components of r_1 satisfy an interesting property with respect to variable ranking. In the context of the current problem, this property is interpreted as the larger the component weight is, the larger the association between the SNV and Y is. Therefore, in addition to summarizing the SNV information within the gene using Z, we can also use the component weights in r_1 to rank the k SNVs in terms of their relative prediction power. This is a feature unique to our proposed method which is critical for prioritizing the SNVs in follow-up evaluation. The details of the algorithm to solve the optimization problem in (5) is provided in the Appendix. We have also shown in the Appendix that the SSU test is directly related to the first PLS component.

Without loss of generality, if we assume $I_c = \{j : 1 \le j \le k\}$, then the ith row of Z can be written in terms of the original genome matrix as

$$Z_i = \sum_{j=1}^{p} w_j x_{ij} \tag{6}$$

where $w = [w_1, w_2, ..., w_p]' = [r_1, 0]' \in R^p$ and $i = 1, 2, ..., n$. Then, we can test the association between the trait and the Z through the logistic regression model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_{C0} + \sum_{j=1}^{p}\left(w_j x_{ij}\right)\beta_C \tag{7}$$

and test $H_0 : \beta_C = 0$ from which a p value for rPLS can be calculated. This demonstrates the relationship between the proposed method and the sum test where all weights are taken as 1. Obviously, rPLS is also related to the existing model-selection-based methods that are proposed as alternative to sum test where the weights can be –1, 0 or 1 [21–23]. However, although model-selection-based methods allow one to take into consideration opposite association directions; all of these methods assume that the effect sizes of SNVs are equal, which is obviously not realistic. In contrast, rPLS selects weights in such a way that SNVs having stronger association with the trait will have larger weights with the relationship directions reflected as the sign of the weights. In addition, initial screening step assigns a 0 weight to the SNVs that are irrelevant to the trait. It is important to note that both models in (1) and (7) can be easily extended to model other types of traits and can accommodate non-SNV covariates.

Simulation Study

We carried out a simulation to compare the performance of the rPLS with performances of aforementioned methods: SSU, wSSU-P, CMC and the SUM. The notations CMC1 and CMC2 correspond to CMC with MAF cutoffs of 0.01 and 0.05, respectively, and throughout the paper CMC denotes the method with a default cutoff of 0.05. Similarly, rPLS1 and rPLS2 correspond to LASSO (with $\alpha = 1$) and EN estimators (with $\alpha = 0.5$) where rPLS denotes the method using LASSO. A total of three settings were considered. In the first two settings, we used the genetic variants from the GAW17 data, which contain sequencing data for 24,487 SNVs in 3,205 genes for 697 unrelated subjects from the 1000 Genomes Project. We kept these genotypes fixed and generated model-dependent phenotypes in these simulations for 697 individuals. Among all the 24,487 SNVs, 91 and 88% have MAFs less than 0.1 and 0.05, respectively. In the third setting, we generated simulated data as in Basu and Pan [14]. For the rest of paper, rSNV (cSNV) are used to denote rare (common) SNVs.

In the first setting, we randomly selected 6 causal genes among 590 genes (out of 3,205) that contain 10 or more SNVs with at least 50% rSNVs and at least 2 cSNVs. One hundred control genes were randomly selected from the remaining set excluding the 6 causal genes. This design is to realistically reflect the GWAS setting. We modeled the relationship between these 6 genes and the trait as in table 1 to account for various association scenarios. Note that we only considered rSNVs in genes 1 to 3, that is, the cSNVs in these three genes were excluded in the simulation and the analysis.

Once the causal genes were constructed, the continuous trait was generated using the linear regression model. The coefficients of causal rSNVs were generated from $U(2.5, 3.5)$ while coefficients of causal cSNVs were generated from $U(0.65, 0.85)$, with the only exception for gene

2. For this gene, half of the generated normal values were set to be negative to create different association directions. The quantitative trait was calculated for each individual and the top 30% of the distribution was declared affected. The above procedure was repeated 500 times to produce 500 data sets.

To further investigate the impacts of 'noise levels' and 'effect sizes', in the second setting, we considered 740 genes containing at least 10 SNVs and selected one causal and one noncausal gene randomly from these genes. We took 0, 25, 50, and 75% of the SNVs in the causal gene as noise, while the remaining SNVs were taken as causal SNVs. The coefficients of causal rSNVs were generated from either $U(2.5, 3.5)$ or $U(3.5, 4.5)$ while the coefficients for cSNVs were generated from either $U(0.65, 0.85)$ or $U(0.9, 1.15)$ and the trait is generated as in the first setting. The causal gene was used to calculate power, whereas noncausal ones were used for type 1 error calculations.

For the third setting, we generated genotypes for 500 cases and 500 controls based on a latent multivariate normal model as in Basu and Pan [14]. The binary trait was generated using the logistic regression model. We considered eight causal rSNVs with 4 different scenarios of noncausal SNVs: 8 noncausal rSNVs together with 8, 16, 32 or 64 noncausal cSNVs within the gene. The odds ratios (ORs) for the 8 causal rSNVs were [3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2]′. Also, instead of taking 8 causal rSNVs, we took 6 rare and 2 common causal variants with ORs [3, 1/3, 2, 2, 1/2, 1/2]′ and [1.15, (1/1.15)]′, respectively, but kept the 4 noise scenarios the same. For this setting with both rare and common causal SNVs, we also considered the noise scenario with 8 noncausal cSNVs together with 8, 16, 32 or 64 noncausal rSNVs within the gene. The performance of the SNV selection in the first step of the rPLS algorithm is investigated in this simulation setting where 2 causal cSNVs, 6 causal rSNVs and 16 (8 rare, 8 common) noncausal SNVs are present within the gene. Along with the aforementioned, we consider ORs [3, 1/3, 4, 4, 1/4, 1/4]′ for causal rSNVs to see how SNV selection step is influenced by the effect size. In each of the 1,000 simulations, we determined how many causal and noncausal SNVs are retained after the first step.

We used the R packages 'glmnet' and 'plsgenomics' for the implementation of EN and PLS, respectively. The optimal value of λ in (3) was obtained by 10-fold cross validation and α was set to either 0 (LASSO) or 0.5. R codes provided on the website (http://www.biostat.umn.edu/~weip/prog.html) were used for implementing CMC, SSU, wSSU-P and the SUM.

Results

Setting 1

In this setting, we conducted tests for each of the 106 (6 causal and 100 noncausal) genes. We used the criterion of p value $<0.05/106$ to declare a gene as significant based on the Bonferroni correction. The 6 causal genes were used to compute the power (true positive), while the 100 control genes were used to gauge the false positive. The boxplots for the power and type 1 error calculated over 500 replications are given in figure 1. As can be seen from the figure, rPLS and wSSU-P yielded higher median power compared to the other methods and all have comparable type 1 error rates. Figure 2 illustrates the power comparison of the methods for identifying each of the 6 causal genes. Specifically, figure 2a represents the genes with only causal rSNVs (i.e. genes 1 to 3) whereas figure 2b represents the genes with both rare and common causal variants within the same gene (i.e. genes 4 to 6). Pooled association tests (SUM and CMC), as expected, were the most powerful when association directions are the same and when there are no or few noncausal SNVs (e.g., gene 1). On the other hand, rPLS and wSSU-P are comparable, and they overpower CMC, SSU, and SUM when association directions differ (gene 2) and when there are noncausal rSNVs (gene 3). In general, inclusion of causal cSNVs increases the power for all methods (fig. 2b). CMC seems to be more resistant to the presence of noncausal cSNVs than the SUM test does. Although SSU has good performance for gene 5, it lost its competitiveness as soon as the noncausal cSNVs were introduced into the model. Overall, rPLS outperforms SSU, CMC and SUM for genes 2, 3, 5 and 6 while still being competitive for the other two genes (gene 1 and 4) in which there are only causal variants with same directions, settings that CMC was designed for. The performances of rPLS and wSSU-P were quite comparable in most settings, wSSU-P giving slightly higher power. For the rest of the simulation, we only considered CMC, SSU, wSSU-P, and the rP LS settings due to comparability of the SUM test with CMC.

Setting 2

Figure 3a and b illustrate the accuracy of the CMC, SSU, wSSU-P and rPLS methods with respect to effect size and noise levels. We used two versions of rPLS, rPLS1 (LASSO) and rPLS2 (EN) to evaluate the relative merits of the two penalties. Both versions of rPLS have the highest accuracy among the methods, where accuracy is defined as the rate of correct inference (both positive and negative). We can see that, as the noise level increases, the

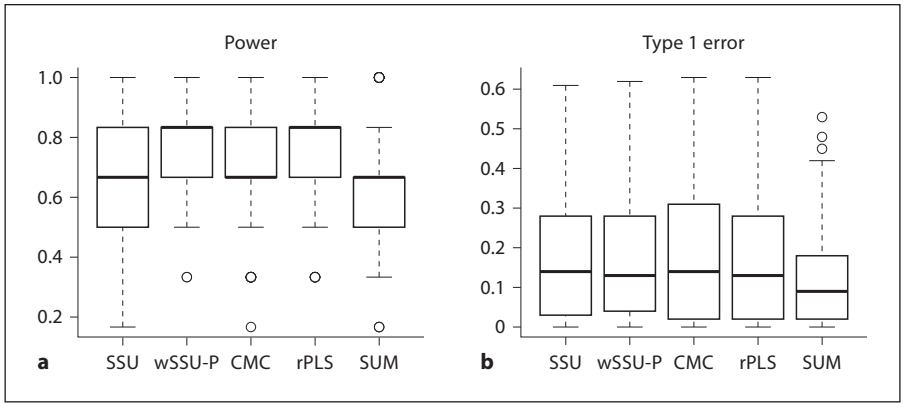


Fig. 1. Overall results from setting 1. Boxplots of the true positive and false positive rates for SSU, wSSU-P, CMC (with α = 0.05), rPLS (LASSO) and SUM test based on 500 iterations are given in **a** and **b**, respectively.

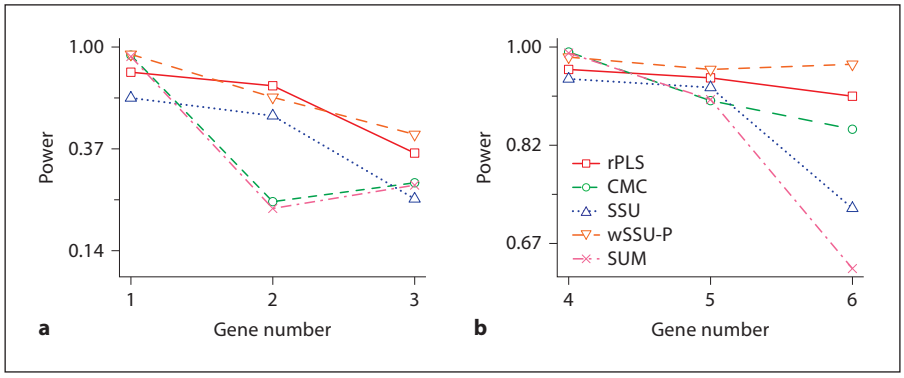


Fig. 2. Individual gene-based results for setting 1: **a** Comparison of power for genes 1 to 3 (i.e. only causal rSNVs); **b** comparison of power for genes 4 to 6 (i.e. rare and common causal SNVs). The powers are calculated at the 0.05 significance level.

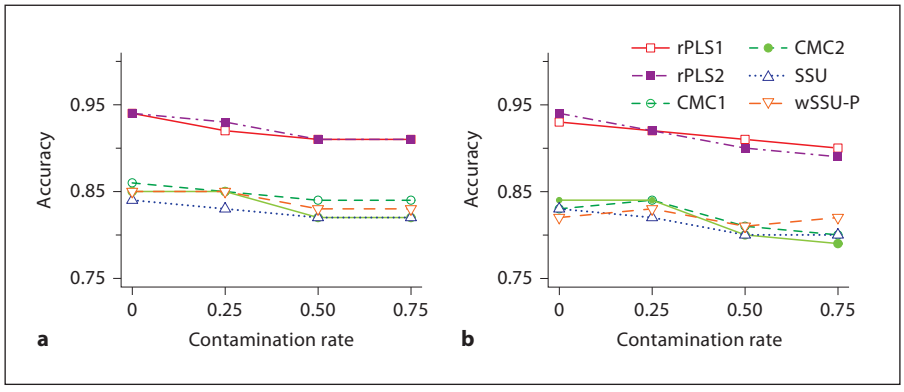


Fig. 3. Results from setting 2. The accuracy values at 0.05 significance level for two levels of effect sizes. **a** Small effect sizes: coefficients of the rare and common causal SNVs are from $U(0.65, 0.85)$ and $U(2.5, 3.5)$; **b** Large effect sizes: coefficients of the rare and common causal SNVs are from $U(0.95, 1.15)$ and $U(3.5, 4.5)$.

accuracy decreases for all methods, as expected. Nevertheless, LASSO seems to work slightly better than the EN when the noise level is high. The two versions of the CMC (corresponding to 0.05 and 0.01 cutoffs) performed comparably. When power and actual type 1 error rates are compared separately (see table 2 for the large effect size case), one can see that the power are all high and comparable, although the power of rPLS is slightly lower. In contrast, the type 1 error rates for all methods are inflated. However, rPLS has much less elevation in its type 1 errors compared to those for the other methods, which range from 400–700% of the nominal type 1 error rate.

Setting 3

Results for the more controlled setting based entirely on simulated data for $\alpha = 0.01$ are summarized in table 3 and figure 4. Figure 4 provides a comparison of accuracy of the methods. Regardless of the underlying setting of

Table 2. Power and type 1 error results for the second setting with large effect sizes at $\alpha = 0.05$

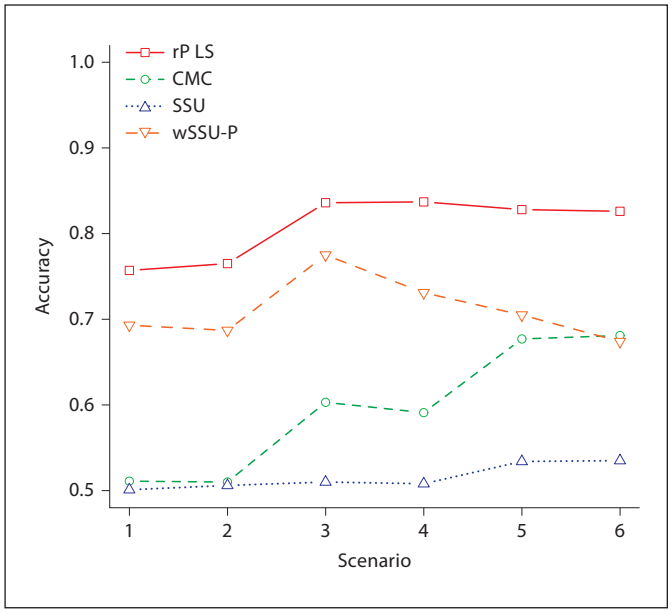
Contamination	Power				Type 1 error			
	0%	25%	50%	75%	0%	25%	50%	75%
SSU	0.973	0.941	0.907	0.880	0.307	0.303	0.299	0.274
wSSU-P	0.991	0.976	0.946	0.909	0.345	0.319	0.324	0.259
SUM	1.000	0.982	0.934	0.851	0.239	0.236	0.224	0.207
CMC1	1.000	0.986	0.938	0.885	0.331	0.303	0.312	0.280
CMC2	1.000	0.982	0.928	0.865	0.316	0.307	0.330	0.276
rPLS1	0.962	0.926	0.884	0.844	0.100	0.092	0.080	0.060
rPLS2	0.970	0.927	0.878	0.838	0.096	0.093	0.081	0.062

Table 3. Power and type 1 error results for the third setting at $\alpha = 0.01$

Power						
Causal	8R		6R, 2C[a]			
Noncausal	8R, 8C	8R, 16C	8R, 32C	8R, 64C	32R, 8C	64R, 8C
SSU	0.010	0.014	0.032	0.022	0.084	0.078
wSSU-P	0.393	0.389	0.561	0.473	0.425	0.354
CMC	0.029	0.028	0.220	0.208	0.369	0.366
rPLS	0.646	0.637	0.748	0.715	0.734	0.700
Type 1 error						
Causal	8R		6R, 2C			
Noncausal	8R, 8C	8R, 16C	8R, 32C	8R, 64C	32R, 8C	64R, 8C
SSU	0.007	0.002	0.012	0.006	0.015	0.007
wSSU-P	0.007	0.015	0.011	0.012	0.015	0.006
CMC	0.007	0.008	0.015	0.027	0.014	0.004
rPLS	0.132	0.107	0.077	0.042	0.078	0.048

[a] R and C notations are used to denote rare and common SNVs, respectively.

causal variants (combination of rare and common) and noncausal variants (also combination of rare and common), it is clear that rPLS outperforms the other methods by a large margin. Considering the power and type 1 error separately (results given in table 3), one can see that rPLS has much higher power. As in previous settings, rPLS has elevated type 1 errors as expected due to the initial SNV selection step. Nevertheless, the increase in type 1 error is once again modest, and it gets smaller when there are more noncausal variants in the gene, a more realistic reflection of real data. On the other hand, many of the type 1 errors for SSU, wSSU-P, and CMC are smaller than the nominal, signaling a more conservative result, the exact opposite of the results in setting 2, where the type 1 errors for these methods are severely inflated. The performance of the SNV selection step are summarized in table 4. Here, the mean, median, first and third quartiles for the number of causal and noncausal SNVs retained are provided. The number in the tables are for the setting where rSNV ORs are [3, 1/3, 4, 4, 1/4, 1/4]′ whereas those in parentheses correspond to the setting where rSNV ORs are [3, 1/3, 2, 2, 1/2, 1/2]′. As expected, LASSO ($\alpha = 1$) can detect causal cSNVs better than the causal rSNVs, and rSNVs with larger effect sizes are more likely to be retained in the model. LASSO seems to be effective in removing both rare and common noncausal variants



Color version available online

Fig. 4. Results from setting 3. Accuracy plots for a variety of scenarios: (1) 8 rare causal, 8 rare and 8 common noncausal; (2) 8 rare causal, 8 rare and 16 common noncausal; (3) 6 rare and 2 common causal, 8 rare and 32 common noncausal; (4) 6 rare and 2 common causal, 8 rare and 64 common noncausal variants within a gene; (5) 6 rare and 2 common causal, 32 rare and 8 common noncausal; (6) 6 rare and 2 common causal, 64 rare and 8 common noncausal variants within a gene.

Table 4. Performance of the SNV selection step in rPLS

SNV	Original number of SNVs	Number of SNVs retained after step 1[a]			
		mean	1st quartile	median	3rd quartile
Lasso: $\alpha = 1$					
Causal rSNV	6	4.1 (1.8)[b]	3 (0)	4 (2)	5 (3)
Causal cSNV	2	1.4 (1.1)	1 (0)	2 (1)	2 (2)
Noncausal rSNV	8	0.7 (0.5)	0 (0)	0 (0)	1 (1)
Noncausal cSNV	8	0.7 (0.5)	0 (0)	0 (0)	1 (1)
EN: $\alpha = 0.5$					
Causal rSNV	6	4.2 (1.9)	3 (0.8)	4 (2)	5 (3)
Causal cSNV	2	1.5 (1.2)	1 (0.8)	2 (1)	2 (2)
Noncausal rSNV	8	0.9 (0.6)	0 (0)	1 (0)	1 (1)
Noncausal cSNV	8	0.9 (0.6)	0 (0)	1 (0)	1 (1)

[a] Larger for causal SNVs (ideally 6 and 2 for rare and common SNVs, respectively) and smaller for noncausal variants (ideally 0 for both rare and common SNVs) are better. [b] The numbers given in the table correspond to SNV selection results for large effect sizes, i.e. ORs = [3, 1/3, 4, 4, 1/4, 1/4]′, while the numbers in parentheses correspond to SNV selection results for moderate effect sizes, i.e. ORs = [3, 1/3, 2, 2, 1/2, 1/2]′.

while EN with $\alpha = 0.5$ yields slightly better performance for detecting causal SNVs yet slightly worse performance for removing noncausal variants. Overall, the results from these two selection procedures are fairly similar.

Discussion

We have proposed here a two-step methodology, rPLS, to detect rare and common variants associated with a complex trait. The approach has several advantages over previously proposed test statistics. Its primary advantage is that it has better performance than existing approaches in the presence of both deleterious and protective variants in a gene or genomic region. Unlike the current collapsing methods that assign uniform weights to SNVs within a genomic region, the rPLS searches for the optimal projection direction based on the association strength with the trait of interest allowing different signs and magnitudes of the SNV contributions. Secondly, the method utilizes an initial SNV selection step to weed out the irrelevant SNVs and to avoid introducing the effect of noise into the projected variant. Therefore, rPLS is robust to the presence of noncausal SNVs. Another important advantage of the rPLS is that the optimal projection vector can also be used to determine which SNVs within a gene are more important by quick evaluation of magnitudes of the vector components. The proposed method does not require an MAF cutoff as in CMC.

We can apply the rPLS on any trait (quantitative or categorical) while having flexibility of including non-SNV covariates and adjusting for population stratification as extra terms in the (generalized) linear model framework. Though, our simulation study in this paper only considers binary trait without covariates nor adjusting for population stratification to facilitate comparisons with other methods.

Results from the three different simulation settings clearly demonstrate the above-discussed advantages of rPLS. These settings were chosen to provide an extensive study of the performance of rPLS and to compare with performances from other promising methods in the literature. The settings are realistic, including the use of the 1000 Genomes sequencing data. Various combinations of rare/common causal and rare/common noncausal variants were considered to reflect a variety of realistic scenarios throughout the genome and plausible hypotheses of association with common diseases. In particular, the noise reduction feature of rPLS is clearly seen across all three settings and highlighted in setting 2 where the issue

was specially investigated. This is a significant step forward as it is expected that there are noncausal SNVs within a gene. As such, rPLS performs well when both power and type 1 error are taken into account. In other words, when the type 1 error rates are controlled to be the same, rPLS provides the highest power in most of the cases studied by examining the receiver operating characteristic (ROC) curves (results not shown).

From a pure power comparison standpoint, there are clearly cases in which existing methods outperform rPLS. For instance, rPLS cannot beat CMC nor SUM test when most of the variants are causal and they influence the trait in the same direction. When there are only rare variants in the gene being considered, SSU performs better than rPLS. We also observed that when there are only causal rSNVs, wSSU-P can have higher power than rPLS. However, wSSU-P is computationally more expensive than rPLS since p values are calculated using permutations.

As expected and shown in all our results, the initial variant screening step can lead to an elevated type 1 error, but we also observe that increase in the type 1 error reduces when more noncausal variants are present in the gene. More importantly, while rPLS has a moderately inflated type 1 error rate consistently across all settings (type 1 error across settings 2 and 3 ranges from 0.048–0.132), the direction and size of the actual type 1 errors for the other methods are much less consistent and thus less predictable. Methods such as CMC, SSU, wSSU-P can either be extremely liberal or conservative depending on settings (type 1 error across settings 2 and 3 ranges from 0.002–0.345). As such, even though other methods may have higher power than rPLS in some settings, it is difficult to gauge whether a resulting positive is a true or a false positive. On the other hand, rPLS has higher accuracy in all settings when both power and type 1 error are taken into account. More importantly, the consistent nature of rPLS makes it a worthy competitor when selecting a test statistic given the underlying setting is unknown in a real data analysis.

Finally, we would like to note that the number of PLS components can be tuned by cross validation, but we do not pursue it because there is no obvious power advantage to the proposed one. We also investigated sparse version of the PLS to simultaneously reduce the noise and the collapse of the SNVs; however, we found it computationally expensive. Adaptive versions of LASSO can also be employed in the first step but are not investigated in this paper.

Appendix

Relationships between SSU and PLS

NIPALS [19] and SIMPLS [24] are the most popular algorithms to solve the nonlinear optimization problem in (5) and they differ by the deflation theme required for the orthogonality of derived components. Although PLS was originally designed for problems with quantitative response, it has started to be used frequently as a dimension-reduction tool for classification problems where the response variable is qualitative. There are mainly two approaches when PLS is employed as a dimension-reduction method for binary responses. One approach is to utilize the NIPALS algorithm to determine components. However, since the NIPALS algorithm consists of regression steps, it seems to be unappealing to use the NIPALS algorithm designed to handle continuous response models that do not suffer from heteroscedasticity. The other most commonly used approach is to determine PLS components for classification problem applying the original SIMPLS algorithm. Nguyen and Rocke [25, 26] and Boulesteix [20] proposed the use of SIMPLS for dimension reduction based on SIMPLS as a preliminary step to classification problems. In this study, the SIMPLS algorithm is considered due to its computational advantages over the NIPALS algorithm. The SIMPLS algorithm can be summarized as follows:

Step 1: Compute cross-product matrix: $S^0_{xy} = X'Y$

Step 2: Repeat steps 2.1–2.4 for $h = 1, 2, ..., m$:

Step 2.1: Compute first left singular vector of S^{h-1}_{xy} as hth PLS weight vector r_h,

Step 2.2: Compute hth score, $t_h = Xr_h$, normalize $t_h =: t_h \|t_h\|$, and update

$$r_h := r_h / \left(\sqrt{r_h^\top X^\top X r_h}\right),$$

Step 2.3: Compute hth x-loading by regressing X on t_h: $p_h = X't_h$,

Step 2.4: Store vectors r_h, t_h, and p_h into matrices $R_h = [r_1, r_2, ..., r_h]$, $T_h = [t_1, t_2, ..., t_h]$, and $P_h = [p_1, p_2, ..., p_h]$, respectively.

Step 2.5: $h =: h + 1$ and $S^{h-1}_{xy} = (I_p - V_{h-1}V'_{h-1})X'Y$ where columns of V_{h-1} form an orthonormal base for P_{h-1}.

The first PLS-weight vector, r_1, is proportional to the dominant eigenvector of $S_{xy}S'_{xy} = S_{xy}S_{yx}$ or equivalently first left eigenvector of S_{xy}, i.e.

$$r_1 = \frac{X'Y}{\sqrt{r_1^\top X^\top X r_1}}$$

Since SSU test statistic is defined as $U'U$ where $U = X'(Y - \bar{Y})$, U is proportional to the first PLS weight vector when applied to non-centered X and centered Y. Therefore, if we denote the obtained first PLS weight vector by $\tilde{r_1}$, the SSU test statistic is equal to the $(\tilde{r_1}'X'X\tilde{r_1})\|\tilde{r_1}\|^2$.

References

1 Reich DE, Lander ES: On the allelic spectrum of human disease. Trends Genet 2001; 17:502–510.

2 Iyengar SK, Elston RC: The genetic basis of complex traits: rare variants or 'common gene, common disease'? Methods Mol Biol 2007;376:71–84.

3 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Wittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM: Finding the missing heritability of complex diseases. Nature 2009;461:747–753.

4 Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, Newton-Cheh C, State MW, Levy D, Lifton RP: Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 2008;40:592–599.

5 Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH: Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 2004;305:869–872.

6 Asimit J, Zeggini E: Rare variant association analysis methods for complex traits. Annu Rev Genet 2010;44:293–308.

7 Morgenthaler S, Thilly WG: A strategy to discover genes that carry multiallelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mutat Res 2007; 615:2856.

8 Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 2008;83:31121.

9 Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 2009; 5:e1000384.

10 Chapman JM, Whittaker J: Analysis of multiple SNPs in a candidate gene or region. Genet Epidemiol 2008;32:560566.

11 Pan W: Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol 2009;33:497507.

12 Wang T, Elston RC: Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet 2007;80: 353–360.

13 Goeman JJ, van de Geer S, van Houwelingen HC: Testing against a high dimensional alternative. J R Stat Soc B 2006;68:477493.

14 Basu S, Pan W: Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 2011;35:606–619.

15 Almasy LA, Dyer TD, Peralta JM, Kent JW Jr, Charlesworth JC, Curran JE, Blangero J: Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc 2011;5(suppl 9):S2.

16 Tibshirani R: Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B 1996; 58:267288.

17 Hoerl AE, Kennard RW: Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12:5567.

18 Zou H, Hastie T: Regularization and variable selection via the elastic net. J R Stat Soc Ser B 2005;67:301301.

19 Wold H: Estimation of principal components and related models by iterative least squares. Multivariate Analysis, Academic Press, New York, 1966, pp 391–420.

20 Boulesteix AL: PLS dimension reduction for classification with high-dimensional microarray data. Stat Appl Genet Mol Biol 2004;3: 33.

21 Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K: A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput Biol 2010;6:e1000954.

22 Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 2010a;70:4254.

23 Hoffmann TJ, Marini NJ, Witte JS: Comprehensive approach to analyzing rare genetic variants. PLoS One 2010;5:e13584.

24 De Jong S: SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 1993;18:251–263.

25 Nguyen DV, Rocke DM: Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002a;18:39–50.

26 Nguyen DV, Rocke DM: Multi-class cancer classification via partial least squares using gene expression profiles. Bioinformatics 2002b;18:1216–1226.

27 Basu S, Pan W, Shen X, Oetting WS: Multilocus association testing with penalized regression. Genet Epidemiol 2011;35:755–765.