OPEN ACCESS Freely available online

PLoS one

Multiple Regression Methods Show Great Potential for Rare Variant Association Tests

ChangJiang Xu[1,2], Martin Ladouceur[1,3], Zari Dastani[1,2], J. Brent Richards[1,2,4,5], Antonio Ciampi[2☯], Celia M. T. Greenwood[1,2,6*☯]

1 Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada, **2** Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada, **3** Department of Human Genetics, McGill University, Montreal, Quebec, Canada, **4** Department of Medicine, Jewish General Hospital, McGill University, Montreal, Quebec, Canada, **5** Twin Research and Genetic Epidemiology, Kings College London, London, United Kingdom, **6** Department of Oncology, McGill University, Montreal, Quebec, Canada

Abstract

The investigation of associations between rare genetic variants and diseases or phenotypes has two goals. Firstly, the identification of which genes or genomic regions are associated, and secondly, discrimination of associated variants from background noise within each region. Over the last few years, many new methods have been developed which associate genomic regions with phenotypes. However, classical methods for high-dimensional data have received little attention. Here we investigate whether several classical statistical methods for high-dimensional data: ridge regression (RR), principal components regression (PCR), partial least squares regression (PLS), a sparse version of PLS (SPLS), and the LASSO are able to detect associations with rare genetic variants. These approaches have been extensively used in statistics to identify the true associations in data sets containing many predictor variables. Using genetic variants identified in three genes that were Sanger sequenced in 1998 individuals, we simulated continuous phenotypes under several different models, and we show that these feature selection and feature extraction methods can substantially outperform several popular methods for rare variant analysis. Furthermore, these approaches can identify which variants are contributing most to the model fit, and therefore both goals of rare variant analysis can be achieved simultaneously with the use of regression regularization methods. These methods are briefly illustrated with an analysis of adiponectin levels and variants in the ADIPOQ gene.

Citation: Xu C, Ladouceur M, Dastani Z, Richards JB, Ciampi A, et al. (2012) Multiple Regression Methods Show Great Potential for Rare Variant Association Tests. PLoS ONE 7(8): e41694. doi:10.1371/journal.pone.0041694

Editor: Zhaoxia Yu, University of California, Irvine, United States of Ameirca

Received March 16, 2012; **Accepted** June 25, 2012; **Published** August 8, 2012

Funding: This work was supported by the Canadian Institutes for Health Research grant number MOP-115110. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: celia.greenwood@mcgill.ca

☯ These authors contributed equally to this work.

Introduction

New methods for the analysis of rare genetic variants are appearing rapidly. Resequencing efforts are identifying numerous new variants but the majority of the new variants are seen only in a very small number of individuals [1]. Hence, the new methods for rare variants, in general, look for association between phenotypes and the collection of all rare variants in a defined set, such as all variants in or near a gene [2].

Hoffman [3], and Lin and Tang [4] framed the goal of rare variant statistical analysis as a problem of distinguishing which (if any) of a set of genetic variants are associated with the phenotype. Let x_{ij} be a genotype coding for the j^{th} variant in individual i, where $j=1,\ldots p$, and $i=1,\ldots n$. For example, $x_{ij}\in(0,1,2)$ for additive allele coding. Suppose that a phenotype y is related to a set of genetic variants by

$$g(\mu_i)=\alpha_0+\sum_{j=1}^{p}\beta_j x_{ij} \qquad (1)$$

for an appropriate function $g(.)$, where $\mu_i=E[y_i]$ is the mean of y_i. The parameters β_j reflect the effect of variant j on the phenotype. In this framework, therefore, rare variant analysis is used to answer two questions: (1) Are any of the parameters β_j nonzero? (2) If some parameters are nonzero, which ones?

Usually there are very few individuals carrying the minor allele at the majority of the identified variants, and therefore it is extremely challenging to estimate the parameters β_j using single marker tests. Joint analysis of a set of genetic variants has therefore been proposed as an alternative strategy to get around this issue of very sparse data. The many proposed methods encompass a wide variety of approaches [5–7]. Some approaches assume a "burden" hypothesis where the count of rare variants is associated with increased risk [4,8,9]. Others methods assume an increased variance of the phenotype or in the risk distribution in the presence of one or more causal rare variants [9–11]. A third group examines genotypic or haplotypic similarities between individuals [9,12].

Conceptually, the problem of how best to model the relationship between a phenotype and a large set of rare genetic variants is a problem of variable selection (or feature selection) and/or dimension reduction (or feature extraction) in a sparse covariate space. There are many well-studied statistical methods for feature

selection and extraction when the number of predictor variables is large. However, for the analysis of rare genetic variation, such approaches have only recently been explored. There were several groups at the GAW17 workshop in 2010 who implemented feature extraction or penalization methods, using a wide variety of different approaches [13–18], and a few other publications have appeared recently using such methods (e.g. [19–21]). Some groups first collapsed the rare variants, and then implemented a LASSO or PLS model using the common variants and the collapsed rare variants [14,16,21]. Others addressed the question of simultaneous modelling across multiple regions or genes, combining methods such as LASSO or PLS first at the gene level, and then across genes [13,15]. A few publications described innovative approaches specifically developed for the sequencing context: Ayers et al. [17] built a LASSO with three custom penalties encouraging different aspects of shrinkage; Luo et al. [20] combined LASSO with local linear embedding. Each of these papers featured a different multiple regression method.

In this paper, we explore whether several classic approaches for feature selection or extraction (ridge regression (RR) [22], LASSO [23], principal components regression (PCR) [24], partial least squares (PLS) regression [25,26], or sparse PLS (SPLS) [27]) can effectively identify associations between a genetic region and a continuous trait. Features of the five chosen methods are shown in Figure 1 and Table 1, respectively. All penalized regression methods minimize a penalized log-likelihood, so that the regression coefficients are shrunk toward zero. However, methods differ in which penalty functions are used. RR uses the L_2-norm penalty, which minimizes the sum of squares of deviations, while LASSO uses L_1-norm penalty, minimizing the absolute value of the deviations. Since the L_1 LASSO shrinks some of the coefficients to be exactly zero, it can be considered as a variable selection method (Figure 1). PCR and PLS reduce the dimension of the variable space by constructing linear combinations of the original variables, but the methods differ in how the linear combinations are constructed. In PCR, the transformed variables

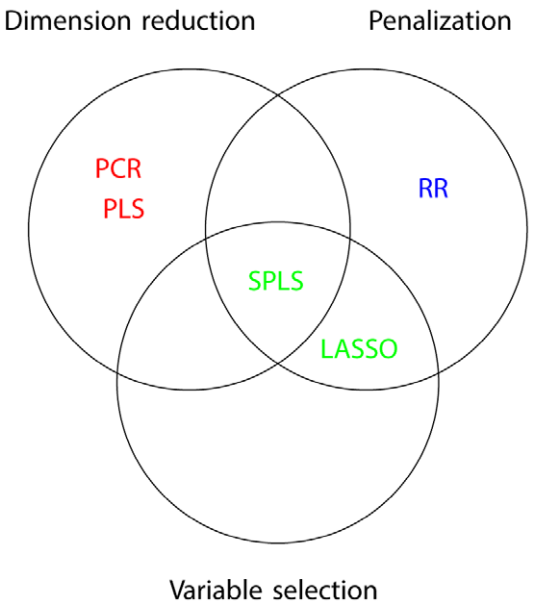

Figure 1. Characteristics of the regression regularization methods compared. The methods are characterized by whether there is variable selection, penalization of parameter estimates, or dimension reduction.
doi:10.1371/journal.pone.0041694.g001

are chosen to explain as much variance as possible in the predictor variable space. In contrast, PLS features are chosen to have high correlation with the response variable. In both PCR and PLS, the number of transformed variables or features included in the regression model must be chosen. Therefore, these methods can be considered as feature selection methods; the feature selection occurs in the transformed variable space. A sparse PLS model was proposed by adding an L_1 penalty to PLS regression [27]. In SPLS, there is variable selection in the original variable space due to the penalization of the log-likelihood and feature selection in the transformed variable space when choosing the number of components in the regression model. Details about and comparisons of the various regularization methods may be found in [28,29] and in Methods.

Since the extreme rarity of most resequencing variants could lead to computational and inferential challenges with feature selection and extraction methods, we also implemented and investigated adaptations of these methods specifically for rare variant analysis. Each method is implemented using two different model choice criteria, both with and without our rare variant adaptation. Using genetic variants identified by Sanger sequencing on three genes in 1998 individuals, we simulated phenotypes under a range of models, and then compared the ability to identify the causal variants using these regression regularization methods. We have also compared performance with three popular methods recently developed for rare variant analysis: the weighted count of Madsen and Browning (WE) [8], the variable threshold method (VT) [30], and the sequence kernel association test (SKAT) [9]. Many methods have been developed for rare variant analysis; we chose these methods for comparison since they represent both the burden methods and the variance-based methods, and have been shown to have good power [7].

In fact, we show that RR, PLS, LASSO and sparse PLS usually outperform WE, VT and SKAT as long as the causal variants are not singletons or extremely rare. Our comparison is timely, since there is great interest in methods for rare variant detection. One additional advantage of feature selection methods is that they can not only identify associations, but can also point towards which variants are likely the truly-associated ones.

Results

Commonly-used methods for rare variants often pool rare alleles and fit simple regression models relating the phenotype to rare allele counts. However, the choice of threshold below which a variant is pooled or collapsed for rare-variant analysis is, of course, arbitrary. Although a 1% threshold is the traditional standard for

Table 1. Characteristics of the five regression regularization methods.

Method	Dimension reduction	Penalization	Variable selection
PCR	$\surd$	–	$\surd$ (on transformed variables)
PLS	$\surd$	–	$\surd$ (on transformed variables)
SPLS	$\surd$	L_1	$\surd$
LASSO	–	L_1	$\surd$
RR	–	L_2	×

doi:10.1371/journal.pone.0041694.t001

differentiating between a polymorphism and a mutation [31], this may not be the optimal threshold for rare variant analysis.

In contrast, we are using multiple regression methods to look for rare variant associations. The five statistical models for feature selection and extraction were fit using well-known R packages [32] (see Methods). However, in each of these models, consideration must be given to model size. For the penalty methods, this is achieved by choosing the penalty parameter λ. For the feature extraction methods, we chose the number of features to enter the model by using three well-known approaches for model selection or choice, AIC [33], BIC [34] and GIC [35]. (See Methods for details.)

To combine our chosen multiple regression methods with the concepts of pooling and collapsing, we propose an approach motivated by the variable threshold idea [30]. We defined a set of thresholds for defining rarity, starting at 5% and including all minor allele frequency observed (MAF) values smaller than this. For each threshold, we created a new variable that contained the unweighted count of minor alleles for all variants with MAF below the threshold, and we then added the entire set of new variables to the set of variables being analyzed. Hence, we have combined the feature selection methods with a generalized pooling strategy, and we have evaluated the performance of these hybrid approaches for detection of rare genetic variants.

For our evaluations, we used genotype data on three genes where the exons and flanking regions were Sanger sequenced in 1,998 individuals (courtesy of GlaxoSmithKline (GSK)) [36,37]. We then simulated phenotypes following six simulation scenarios based on these genotypes. For simplicity, the missing values were imputed independently at each variant by randomly generating the missing genotype using the computed MAF. The three genes sequenced (anonymized data, called genes A, B and C) had respectively 98, 28 and 122 variant sites.

Continuous phenotypes were generated assuming a normal distribution $N(0,1)$ among individuals not carrying any causal genetic variants. From each gene, some rare variants (and possibly some common variants) were selected to be associated with the phenotype, and for carriers of these variants, the normal distributions were shifted. In our first set of simulations (Scenario set I), the shift is independent of allele frequency; however in a second set of simulations (Scenario set II), the size of the effect of the causal variants depends inversely on the MAF. The parameters used in the simulations are described in Table 2, and more details about the simulation design are given in Methods.

When fitting the models, a single measure of model fit was chosen for each method, after choosing all the parameters of the model (see Methods). Empirical power was calculated by comparing this test statistic to its distribution under 1000 permutations. In the analysis of permuted data, the parameters controlling model size and complexity were chosen independently within each permutation.

Labelling and Nomenclature

Each of our five chosen methods was used to analyze 1000 simulated data sets, and the results are used to calculate empirical power at significance level $\alpha = 0.01$. Permutation was used to assess significance for all methods, since the feature selection inherent in each method will lead to biased estimates of significance using asymptotic techniques. Using QQ-plots, the distribution of the empirical p-values under the null hypothesis is demonstrated in Figure S6 for most of the methods. Variability is within the expected error bounds. Power comparisons across the different methods are illustrated in Figure 2 for scenario set I and gene A, and in Figure 3 for scenario set II and gene C. Additional results (for genes B and C from Scenario set I, and for genes A and B from Scenario set II) are in Figures S1, S2, S3, and S4. Results in numeric form are also given in Table S1 and Table S2.

Methods are colour coded and labelled across the top of each figure. Several different options were used for fitting each of the multiple regression methods. For PCR and PLS, models with one component are denoted "Comp1". The label "Compk" denotes models with k components, where k represents the number of features that explained 80% of the variance in the response. For RR, $RR.0$ represents ridge regression with $\lambda = 0$, or equivalently an ordinarily linear regression. $RR.10$ implies RR with a penalty parameter of 10. For the LASSO and SPLS methods, the labels AIC, BIC, or GIC indicate the method used for selecting penalty parameters. Finally, if the label terminates with ".p", as in "Comp1.p" or "AIC.p", then the pooled rare variant set was added to the set of predictor variables.

Power Comparisons

With a few exceptions, any of the 5 multiple regression methods had better power than the three approaches developed specifically for rare variant analysis (WE, VT, SKAT), and furthermore, performance was often very similar across different variants of the multiple regression methods.

Consider first two situations where causal variants had clear large effects (I.1:Large10 and I.5:Bidirectional in Figure 2 for gene A, and Figure S2 for gene C). In these scenarios, the best powers of each of the 5 regularization methods were very comparable. A few specific choices for feature selection performed poorly: notably PCR with only one component showed poor power, as did PLS with one component (particularly when the pooled rare variants were included in the predictor space). LASSO, RR and SPLS showed very similar powers, and neither the variable selection technique nor the addition of pooled predictor variables altered power in these cases. When the effect size depends on MAF (Figure 3 and Figure S3), there is comparable performance for all multiple regression methods (apart from PCR and PLS with one component), and more power than VT, WE or SKAT.

When the causal variants had smaller effects (I.3:Modest10, I.4:Modest20 in Figure 2, similar models in Figure 3), there is slightly more variability between the different multiple regression methods. The LASSO, in particular, seems to have better power than other approaches. When 20% of the rare variants were causal, WE or VT sometimes had good power too.

With a mixture of rare and common variants, the LASSO again had better power than most other multiple regresion methods. Comparing this scenario across the three genes for scenario set I (Figures 2, S1 and S2), WE, VT and PCR had the best power for gene A but not for genes B and C. It seems that the common causal variants are aligned with the rare causals such that the first principal component captures the association identified by the burden methods.

When the effect size depends on the MAF (Scenario set II), the relative performances of the multiple regression methods were similar to Scenario set I; we saw very little alteration in the relative performances of the various algorithms. The multiple regression methods continued to perform well in comparison with VT, WE and SKAT. However, all methods had excellent power for the models with a mixture of rare and common variants (I.2 and II.2). This is due to the definition of the effect sizes in this scenario, where the average effect size was defined across all causal variants (see Table 2 and Methods).

In scenario I.6 and II.6, only variants with frequency less than 1/1000 were selected as causal. Power is substantially lower in this situation for all methods and the patterns of performance differ. In

Table 2. Parameters for Simulation Scenarios.

Scenario	Causal rare variant threshold	Direction of effect	Percentage of variants that are causal	Average effect size in standard deviations
Scenario set I: Effect size constant across MAF				
I.1 Large10	MAF $\leq$ 0.01	deleterious	10% rare	1.64
I.2 Rare/Common	MAF $\leq$ 0.01	deleterious	4 rare, 4 common	1.64, 0.07
I.3 Modest10	MAF $\leq$ 0.01	deleterious	10% rare	1
I.4 Modest20	MAF $\leq$ 0.01	deleterious	20% rare	1
I.5 Birectional	MAF $\leq$ 0.01	7.5% deleterious, 7.5% protective	15%	1.64, 1.64
I.6 VeryRare	MAF $\leq$ 0.001	deleterious	20% rare	1.64
Scenario set II: Effect size dependent (inversely) on MAF				
II.1 Large10	MAF $\leq$ 0.01	deleterious	10% rare	1.64
II.2 Rare/Common	MAF $\leq$ 0.01	deleterious	4 rare, 4 common	1.64, 0.07
II.3 Modest10	MAF $\leq$ 0.01	deleterious	10% rare	1
II.4 Modest20	MAF $\leq$ 0.01	deleterious	20% rare	1
II.5 Birectional	MAF $\leq$ 0.01	7.5% deleterious, 7.5% protective	15%	1.64, 1.64
II.6 VeryRare	MAF $\leq$ 0.001	deleterious	20% rare	1.64

doi:10.1371/journal.pone.0041694.t002

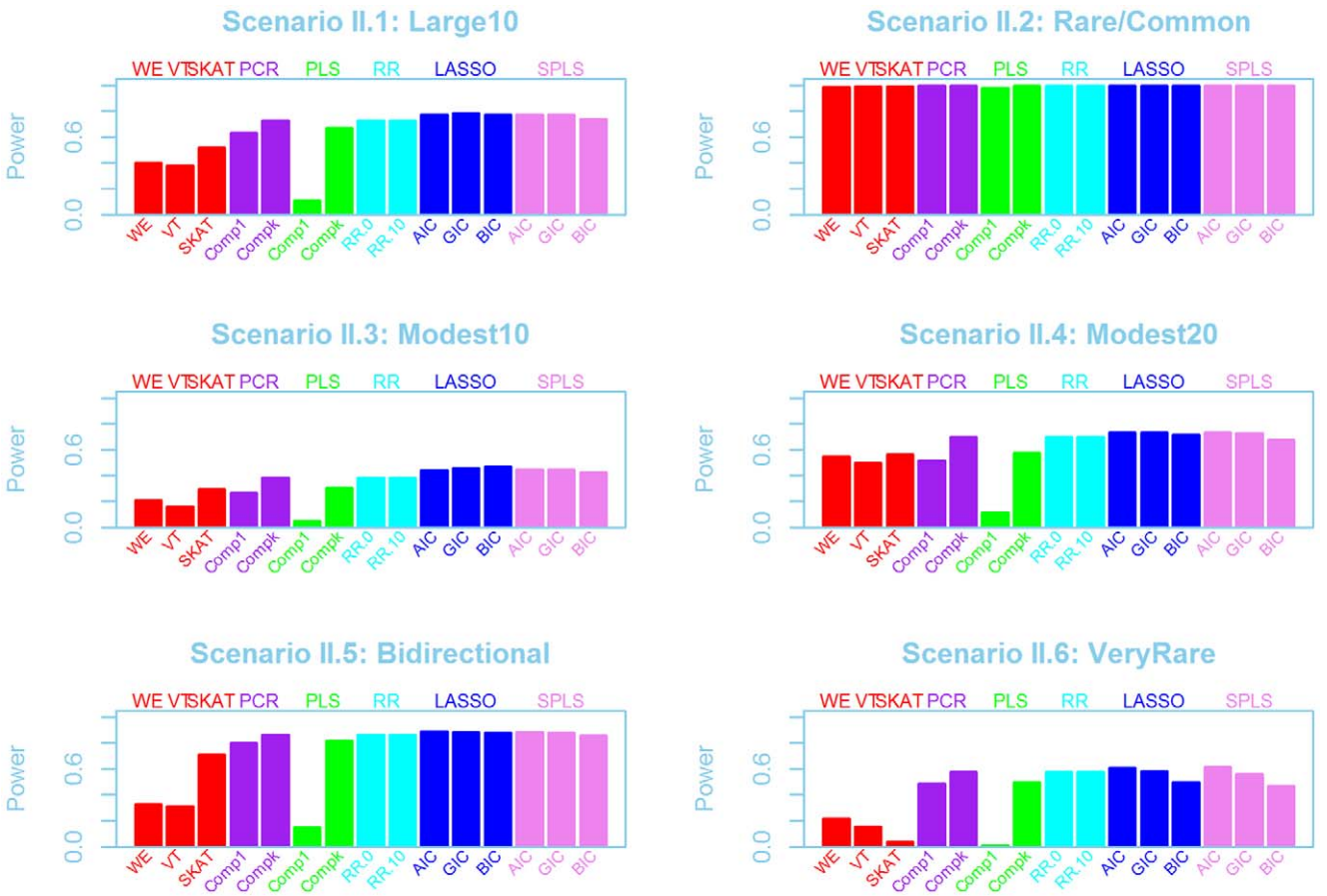

Figure 2. Test power for gene A with 98 variant sites under simulation scenarios I.1 to I.6. Power is shown for several different methods, including several options within each of the regularization methods. WE, VT and SKAT are shown in red, PCR in purple, PLS in green, RR in turquoise, LASSO in royal blue and SPLS in pink. Simulation scenarios are shown in Table 2.
doi:10.1371/journal.pone.0041694.g002

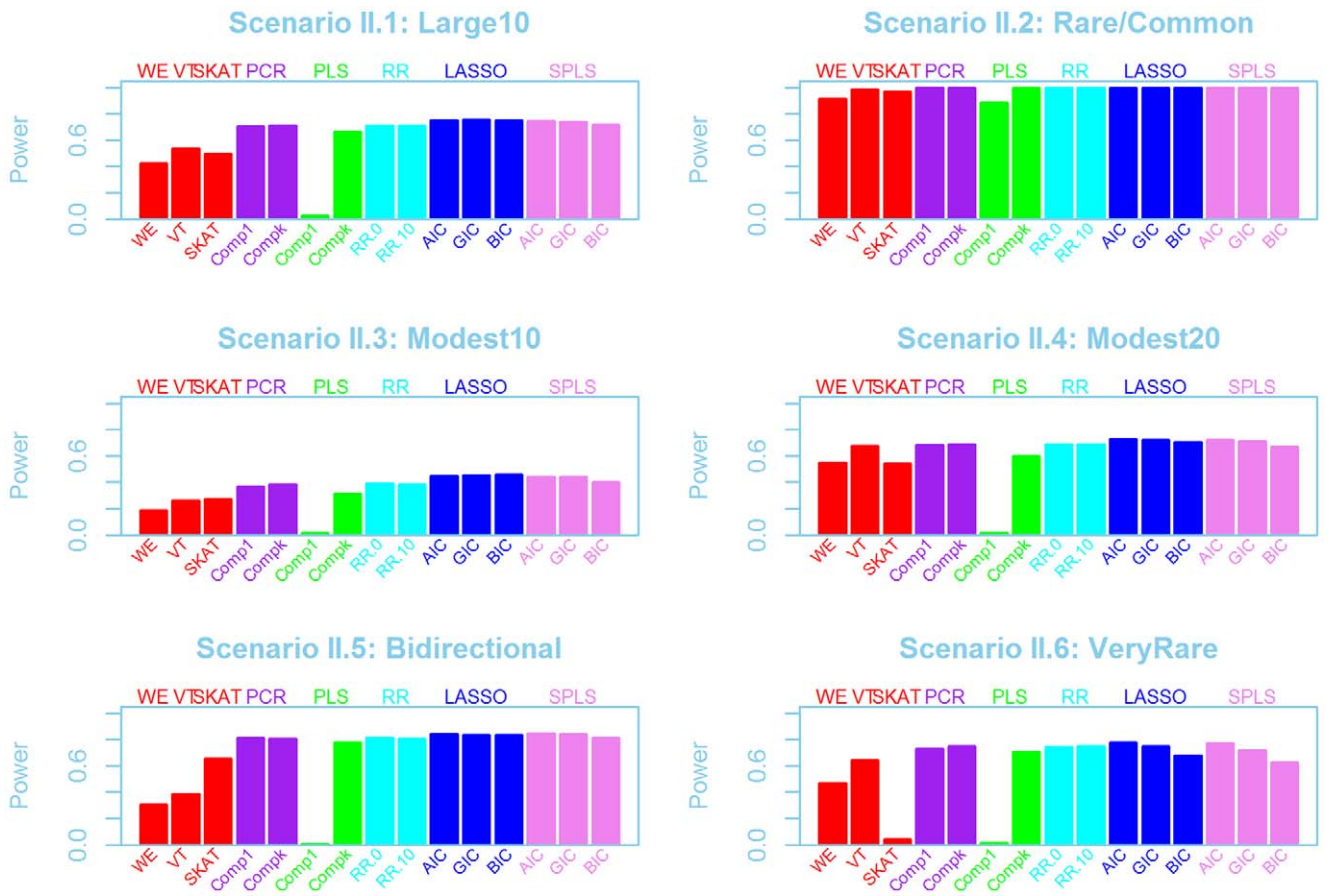

Figure 3. Test power for gene C with 122 variant sites under simulation scenarios II.1 to II.6. Power is shown for several different methods, including several options within each of the regularization methods. WE, VT and SKAT are shown in red, PCR in purple, PLS in green, RR in turquoise, LASSO in royal blue and SPLS in pink. Simulation scenarios are shown in Table 2.
doi:10.1371/journal.pone.0041694.g003

Scenario set I (Figures 2, S1 and S2), PCR (with one component and pooled predictors), PLS or RR have better power than LASSO and SPLS. In addition, VT performs well in this context for genes B and C. In contrast, when the magnitude of the effect is inversely dependent on MAF (scenario set II), all the multiple regression methods perform comparably to or better than VT; in particular, the LASSO or SPLS with AIC show good performance. It is interesting to note that SKAT performed very poorly in this scenario.

The addition of pooled variables to the predictor space did not seem to alter power in the simulations using Scenario set I. However as previously noted, PCR with one component and pooled predictor variables performed better than other approaches in scenario I.6:VeryRare, and is presumable capturing the causal rare variants through one or more of the pooled variables. Since in most cases, the pooled predictors made no difference, the figures for Scenario set II do not include pooled predictors.

When effects could act in both directions (I.4) VT and WE did very poorly, but these approaches are known to look only for variants acting in one direction [7,11]. In contrast, the power of SKAT is much better than WE and VT in the bidirectional situation, since this test looks for changes in variances rather than means. Nevertheless, the regression-based approaches all have greater power than SKAT. As the percentage of variants that are causal increases, it has been shown that SKAT loses power relative to VT and WE [7], but this parameter has a less important effect on power than the effect size or the causal MAF distribution.

Gene B contains only 28 variant sites, and as a result the power for detecting association is low for all methods (Supplemental Figures S2 and S3). Our multiple regression implementions perform just as well (or just as poorly) as WE, VT or SKAT for this gene. Power is slighly better when the model includes a mixture of both rare and common variants (I.2 and II.2), and then the multiple regression methods perform better than the rare variant methods.

Finally, we did not see consistent changes in performance when comparing variable selection using AIC, BIC or GIC for the LASSO and SPLS. There are a couple of situations where power seemed better when using AIC, and other situation were BIC or GIC appeared to be the best. Given that these power estimates are based on 1000 simulations, the standard error of the power estimates is 1.6% or less, depending on the magnitude of the power.

Causal Variable Identification

After identifying whether a gene is associated with a phenotype, there is interest in finding which variants are strongly associated. Pooling and collapsing rare variant methods do not provide this kind of information. However, parameter estimates from the LASSO and SPLS methods can be helpful for this inference, since the final models can be examined to see whether the true causal

variants were selected and retained. Figure 4 demonstrates whether the truly-associated variants were captured by LASSO or SPLS in the simulations using Scenario set I, averaged over the three genes. This figure shows three different aspects of variable selection. Firstly, the left (pink) bar in each set shows the the average numbers of variants selected across the 1000 simulations. In the centre (green) bar of each set is the average number of causal variants selected by the LASSO and SPLS methods. Finally the third (purple) bar shows the number of the pooled variables included in the final models.

AIC variable selection included more variables than BIC or GIC, for both LASSO and SPLS, and included a large proportion of the causal variants in each generating model. However, these AIC methods also selected a large number of non-associated variants. The number of non-associated variants was much smaller for BIC or GIC, but as a consequence, many truly-associated variants were missed. Power does depend on the degree of penalization, and when there is insufficient penalization or too many variants in the regression models, the power tends to be lower (Figure S5). Usually only one or fewer pooled variables were retained in the models, but occasionally there could be 2 different pooled variables (with different thresholds) kept in the same model (Figure 4).

Adiponectin and ADIPOQ

Adiponectin levels are controlled by the ADIPOQ gene, and genetic variants in this gene are known to influence adiponectin levels [38]. Therefore, we compared the results for our tests of association between rare genetic variants at ADIPOQ and adiponectin in two data sets. The first data set included individuals from Twins UK data [39,40] who were selected from the extremes of a pain phenotype; 175 out of 500 samples have available adiponectin values and had undergone exome sequencing. The exomes were captured by Nimblegen technologies and were resequenced by Beijing Genomics Institute (BGI) (unpublished data); 5 rare variants were identified in the gene and were analyzed together. The second analysis included 1375 individuals, again from the Twins UK data who were explicitly genotyped at two rare variants within ADIPOQ, and the analysis included only these 2 variants (unpublished data). These two SNPs were genotyped by Taqman in KBioscience, UK; 113 individuals were in both analyses.

Table 3 shows the results of the analyses of these two data sets. For the small data set of 175 individuals, several methods give rise to estimates of significance near 0.05, but among these, PCR and RR with a penalty of 10 show the smallest p-values. The WE method shows no relationship with phenotype in this context. Three of the 5 SNPs showed some univariate association with adiponectin. When using SPLS with AIC, all three of these markers were included in the chosen model. For LASSO with AIC, two of the three markers were selected. When using GIC or BIC, the marker with the strongest univariate significance was always included. In the analysis of the larger data set, all of the methods showed strong significance with the minimum possible p-value for 10^6 permutations, and the p-values obtained are smaller than the p-values for the comparison methods WE, VT or SKAT. Both variants were included by the LASSO and SPLS methods.

Discussion

Resequencing efforts identify many extremely rare or private genetic variants, often of unknown function. Analysis of association of such variants is difficult due to the sparsity of the data. Although any statistical inference about an event seen only once is impossible, we hypothesized that modern multiple regression methods might be able to find some associations between high-dimensional sparse data and phenotypes, and we demonstrated that this is, in fact, the case using Sanger sequencing data on almost 2000 individuals at three genes. Our analysis of adiponectin and the ADIPOQ locus confirmed this potential for increased power using multiple regression methods. Furthermore, we showed that we have the ability to identify a substantial proportion of the causal variants within each gene.

Performance of the multiple regression methods was not as good in a few situations. Dimension reduction methods (PCR, PLS) with only one component tended to perform poorly. In fact, this could probably be expected. These methods are normally used together with a rule for determining the best number of components, and so choosing only one component is not the normal implementation. Multiple regression methods also performed poorly in a simulation where the causal variants were very rare; in that situation there was a large discrepancy between the frequency of the causal variants (0.001) and the largest threshold we used for pooling rare alleles (0.05). Therefore, refining our combined regression and pooling approach may improve performance in this case. For example, we could reduce the number of MAF thresholds that we included, or use counts weighted by the inverse of MAF.

It is difficult to understand why some models would include multiple pooled variables with different thresholds, as was seen in Figure 4. Due to the rarity of the variants, many of the pooled variables may be highly collinear with the original data as well as with other pooled variables. We believe this is why the performance of the models with the pooled predictors was so similar to the performance without this set of variables. Furthermore, the selection of more than one pooled variable may also suggest that pooled variables at distinct allele frequency intervals could be useful.

Since permutations are required, the computational overhead of some of these regression approaches can be quite high. Supplemental Table S3 shows average run times on a 2.8 GHz blade, and demonstrates that PLS, LASSO and SPLS are much slower than the other methods. Taking the largest of our 3 genes, the LASSO would need 108 days of computer time to do an exome-wide analysis of 20,000 genes, with 1000 permutations. This would require the use of a processor with multiple compute nodes. In contrast, for RR, only 27 hours of CPU time would be needed to use this method with our code. Computation times could be substantially improved by using an incremental number of permutations, with more precision for smaller p-values.

It is interesting to speculate on why these methods work. Some of the rare causal variants in our simulations could have been observed in up to 20 individuals out of the 2000 in the sample, and hence there would be adequate power to identify these variants. It is also possible that the presence of long-range correlation between variants could provide information to PCR or PLS methods. Long-range correlation between common and rare variants can occur due to patterns of ancestral recombinations and co-occurrence of rare variants in some parts of a gene genealogy [41]. Power was lower across all methods for the smallest gene, but this gene contained only 2 variants with MAF over 1%, and so our strategy may have been unable to capture additional signals through linkage disequilibrium.

Nevertheless, these multiple regression methods were not designed for data as sparse as data from resequencing studies. We therefore proposed augmenting the variable space with new variables defined by pooling rare variants together. Instead of choosing a single threshold, we augmented the predictor space by

Figure 4. Number of variants selected by the LASSO and SPLS methods for Scenario set I. Numbers of variants selected are averaged across the three genes and across the simulations for AIC, GIC and BIC variable selection techniques. Within each set of three bars, the left (pink) bar shows the total number of variants selected, the middle (green) bar shows the number of causal variants selected, and the right bar (purple) shows the number of pooled variables selected.
doi:10.1371/journal.pone.0041694.g004

a set of pooled counts based on a range of thresholds. These variables would, of course, be highly correlated with each other. We did not find an improvement in power in most situations when we added the pooled variables, and we believe that this is due to the increase in the parameter space, so that the models had more difficulty identifying the best predictors.

Any division of genetic variants into "common" and "rare" is arbitrary. Our choice of 5% for the upper threshold of rarity for pooling may also have an impact on the performance of the models including pooled variable sets. It would be interesting to investigate whether considerations of population history could be used to set more appropriate thresholds distinguishing rare variants from common ones.

Simulations can be designed to favour one analytic strategy over another. When we modelled a constant effect of the causal variants (Scenario set I), as expected the VT method usually outperformed the WE method. In contrast, in Scenario set II where the effect depended on MAF, performance of WE was improved. Our simulations did not explicitly select correlated genetic variants to be causal, and therefore the dimension reduction approaches of PCR and PLS would not necessarily be expected to outperform other approaches. However, the idea that a small percentage of the rare genetic variants in a gene are causal underlies most of the methods developed for rare variant analysis, and therefore the better performance of multiple regression methods is a pleasant surprise.

All simulations were based on the genotype data from three Sanger sequenced genes. The variants chosen as causal were randomly selected for each simulated data set and therefore there was variability across the simulations in the rarity of the causal variants. However, evaluation of performance on a larger set of genotypes may provide additional insight into performance of these methods.

Predictions of changes in amino acids in proteins, or predictions of sequence conservation have been used with success to distinguish potentially causal variation from variation that is unlikely to be causal [30]. Any of the methods evaluated here could be combined with such predictions to improve the variable selections, by implementing an appropriate weighting of the variants. Similarly, covariates could easily be included into these

Table 3. P-values for association tests between Adiponectin levels and the ADIPOQ gene.

Method	175 samples, 5 SNPs	1376 samples, 2 SNPs
WE	0.7419	0.0130
VT	0.0736	0.0006
SKAT	0.1596	8×10^{-6}
PCR	0.0158	10^{-6}
PLS	0.0290	10^{-6}
RR $\lambda=0$	0.0290	10^{-6}
RR $\lambda=10$	0.0780	10^{-6}
LASSO	0.0846	10^{-6}
SPLS	0.0892	10^{-6}

Note: 5×10^3 and 10^6 permutations were used for the two datasets, respectively, to obtain empirical significance levels. PCR and PLS were fitted using only one component. When multiple components were used, the p-values were very similar. AIC was used to select model size for LASSO and SPLS.
doi:10.1371/journal.pone.0041694.t003

multiple regression models. However, our current implementation measures the strength of the genetic association through a global model fit statistic; this would therefore need alteration so that the summary statistic excludes the effects of covariates and focuses only on the genetic variants.

Theoretically, such methods could simultaneously model all the genetic variants in more than one gene, such as all variants in a pathway, or conceivably all exome-identified variants. Genome-wide simultaneous modeling has been suggested by several authors, in particular using Bayesian methods [42]. We have not attempted such models using these multiple regression methods, however, we anticipate that the penalty needed as a result of the substantial model selection steps would overwhelm power.

We obtained excellent power over a variety of simulation scenarios with many of our implementations of these multiple regression methods. Therefore, the choice of best method may be partially the preference of the data analyst. PLS or SPLS methods may be beneficial for modelling jointly covariates and genotypes. We like ridge regression since it is computationally fast, but where computation time is less of a concern or the variable selection aspect is of importance, the LASSO could be an excellent choice since the power is often slightly better.

Conclusions

Methods developed for high-dimensional data may outperform other approaches for rare variant analysis. These methods will simultaneously model the effects of all genetic variants in a gene, common or rare. Furthermore, unlike collapsing, counting, or variance-based methods for rare variant association analysis, some of these regression methods can identify the most likely causal variants.

Methods

Ethical Statement

The genotype data used for our simulations represents a re-use of data and no new human interventions were conducted. No additional IRB approvals were sought for the simulation studies. The Committee on Ethics in Clinical Research, CHUV, Lausanne University, Lausanne, Switzerland approved the original protocols for sample collection for the genotype data used in simulations. All participants in Twins UK provided informed written consent, and the research protocol was approved by institutional ethics review committees at Kings College London. Again, the data used for our analyses represents a re-use of data that has been previously analyzed and no further IRB approvals were sought.

Model Details

Suppose we have a sample of n independent individuals who have been sequenced to identify genetic variation in at least one candidate gene, and measured for a continuous trait. Assume that equation (1) describes the true relationship between the phenotype and genotypes, where the set of genotypes includes all identified locations that vary between individuals in the sample. We fit several multiple regression methods including variable selection or feature extraction methods. These methods have been previously compared, but not for analysis of rare genetic variants [43]. An R package (RVtests) containing the implementation of the tests described below is available from the authors or at www.mcgill.ca/statisticalgenetics/.

Feature Extraction Methods

In PCR [24], the original predictor variables $x_j=(x_{1j},\ldots,x_{nj})'$, $j=1,\ldots p$, are transformed to principal components that explain variance in the predictor space, without considering the relationship to the response variable. Specifically, let UDV' be the singular value decomposition of $X=\{x_{ij}\}$. Then the fitted response for PCR with k components is $\hat{y}=\sum_{j=1}^{k} U_j U_{j'} y$ [29], where U_j is the j^{th} column of U, and $y=(y_1,\ldots,y_n)'$.

In PLS [25,26], orthogonal scores are created that have both high variance and high correlation with the response, y. Let T_j, $j=1,\ldots,k$, be the orthogonal scores in a PLS model with k components. Then the fitted response can be written $\hat{y}=\hat{\alpha}_0+T_1\hat{\alpha}_1+\ldots+T_k\hat{\alpha}_k$, where $\hat{\alpha}_j=T_{j'}y/T_{j'}T_j$.

The fitted values for PCR and PLS depend on the number of components, k, which is also referred to as the size of the model. Although k is often chosen by the use of cross-validation, in the rare variant context, we wanted to identify algorithms that are computationally efficient, and therefore, we compared performance for two values of k, $k=1$, and k^*, where k^* is chosen so that 80% of variance in the response is explained by the model. We also fit models with a large value of $K(K=30)$, but results were not as good and are not shown. All our simulations were implemented with the R statistical programming language [32]. We used the R package 'pls' [44] for getting the PLS scores, T_j, and R function 'svd' for calculating the PCR scores, U_j.

L_2 Penalization Method (Ridge Regression)

Parameters in RR models [22] are shrunk towards zero by adding to the regression model a penalty parameter which is a function of the squared regression coefficients, i.e., L_2 norm. Following the notation used above, where D is a matrix containing the singular values of X, and U_j are the singular vectors, the RR fitted response is $\hat{y}=\sum_{j=1}^{p} U_j U_{j'} y\{d_j^2/(d_j^2+\lambda)\}$ [29], where d_j is the j^{th} diagonal element of D. The penalty parameter λ, where $\lambda\geq0$, controls the degree of shrinkage; for large values of λ all parameters become close to zero and the effective dimension of the model is reduced. In contrast, when $\lambda=0$, RR reduces to an ordinary linear regression model. Results are shown for $\lambda=0$ and $\lambda=10$. We also completed simulations with $\lambda=100$ but performance was comparable to $\lambda=10$ and results are not shown. We used the correlation between the observed y and the fitted values $\hat{y}$, $r=cor(y,\hat{y})$, as our measure of goodness of fit.

Methods Using L_1 Penalization

The penalty parameter in LASSO [23] and SPLS [27] can be chosen by classic model selection criteria [45,46]. Here we used AIC [33], BIC [34], and GIC [35] to choose this parameter. For SPLS, a series of models was fit, varying the number of hidden components k between 1 and 5, as well as the thresholding parameter $\eta \in (0.5,0.6,0.7,0.8,0.9)$. The best model choice over the two-way grid of parameter values was chosen by AIC, BIC or GIC.

To evaluate model performance for LASSO and SPLS, we used the selected final model F-test p-value as the score measuring model performance. The R-package 'glmnet' was used for LASSO [47] and the package 'spls' for SPLS [27].

Pooling Rare Variants

Let MAF_j, $j=1,\ldots,p$, be the MAF of the j-th variant. For a chosen threshold MAF_0, the set of rare variants can be defined as the variants j with $\mathrm{MAF}_j \le \mathrm{MAF}_0$, and a pooled rare variant count is $v_{\mathrm{MAF}_0} = \sum_{\mathrm{MAF}_j \le \mathrm{MAF}_0} x_j$, where x_j is the number of minor alleles at variant j. To combine the regression methods with the concepts of pooling and collapsing, we propose an approach motivated by the variable threshold idea [30], and so we defined a set of thresholds for defining rarity, starting at 5% and including all MAF values smaller than this. Let $\{v_T\}$ be a new set of variables $\{v_T = \sum_{\mathrm{MAF}_j \le T} x_j, T = \min(\mathrm{MAF}_j),\ldots,0.05\}$ that pool rare variants for a series of possible thresholds, T, on the minor allele frequency. These new sets of variables were added as possible predictor variables in the regression models. Since we did not identify any benefit to including these pooled variables in the set of predictor variables in Scenario set I, the simulaton results for Scenario set II are presented without the inclusion of these additional predictors.

Study Sample

The subjects used in this paper are a subset of the CoLaus study, a population-based study of 6,188 Lausanne residents aged 35 to 75 years [37].

Sanger Sequencing Data

Sanger sequence data for the exons and flanking regions of three genes from 1,998 individuals were provided by GlaxoSmith-Kline (GSK) [36]. Missing values of each rare variant were imputed independently from others based on the computed MAF, as in [7]. All non-polymorphic base-pair markers were removed from the sequence data. The three genes used in our simulations contained, after removal of monomorphic variants, 98, 28 and 122 variant sites, respectively. Of these, 85, 26 and 99 variants, respectively, were seen at allele frequencies less than 1%. Coding lengths for these genes were 4094, 1239 and 1500 base pairs.

Phenotype Simulation

Within each simulation, a proportion of the rare variants was randomly selected to be causal, depending on the simulation scenario (Table 2). The threshold for "rare" is given in the second column of Table 2; all variants with MAF below the threshold could be chosen to be causal in any simulation. The phenotypes were generated from a $N(0,1)$ distribution for individuals not carrying any rare variants. In Scenario set I, for carriers of one or more rare variants, the phenotype was assumed to be distributed as $N(-\mu,0.2)$ where the values of μ are given in the last column of Table 2. For Scenario set II, we define the effect of each variant as $b_j = \frac{K}{\sqrt{MAF_j(1-MAF_j)}}$ for a chosen constant K. This constant was chosen so that the average effect of all J causal variants, $\mu = \frac{1}{J}\sum_j b_j$, has the values shown in Table 2. For individuals carrying more than one causal rare variant, the phenotype was drawn from the normal distribution with mean corresponding to the most rare causal variant carried by that individual. For scenarios I.5 or II.5, where the effects could be bidirectional, the mean μ could be either positive or negative. Similar simulation parameters were used by [7].

Permutations

For each simulated data set, the phenotype data was permuted relative to all the genotype data and the analysis was repeated. For each permutation, the analysis included all model fitting steps, so that variable selection or identification of the best model as a function of AIC or BIC was repeated for each permutation step. Using the chosen measure of model fit for each method (described above), we then compared this statistic between the permuted data sets and the original simulated data set, and counted the number of permutations where the model fit statistic was more extreme than in the original data.

Software

An R package, RVtests, that uses these approaches to test for rare variant associations, is available from the authors or from cran-r.project.org.

Supporting Information

Figure S1 Power for Scenario set I for gene B.
(PDF)

Figure S2 Power for Scenario set I for gene C.
(PDF)

Figure S3 Power for Scenario set II for gene A.
(PDF)

Figure S4 Power for Scenario set II for gene B.
(PDF)

Figure S5 Power for the LASSO method as a function of the penalty parameter.
(PDF)

Figure S6 QQ plot of empirical p-values for gene C under the null hypothesis.
(PDF)

Table S1 Test power of WE, VT, SKAT, PCR, PLS, RR, LASSO, and SPLS for three genes and for scenario set I where variant effects do not vary with the minor allele frequency.
(PDF)

Table S2 Test power of WE, VT, SKAT, PCR, PLS, RR, LASSO, and SPLS for three genes and for scenario set II where variant effects vary with the minor allele frequency.
(PDF)

Table S3 Simulation run time for three genes and six scenarios.
(PDF)

Acknowledgements

The authors are indebtted to GlaxoSmithKline for the provision of genotype data used in simulations. We thank GlaxoSmithKline and the co-

PIs of the CoLaus study, Gerard Waeber and Peter Vollenweider, for the use of this anonymized resequencing data, and Drs. Matthew R. Nelson and Margaret G. Ehm for their helpful suggestions as well. This work was supported by the Canadian Institutes for Health Research grant number MOP-115110.

Author Contributions

Conceived and designed the experiments: CG AC CX ML. Performed the experiments: CX. Analyzed the data: CX. Contributed reagents/materials/analysis tools: ML ZD JBR. Wrote the paper: CX ML ZD JBR AC CG.

References

1. Zhu Q, Ge D, Maia JM, Zhu M, Petrovski S, et al. (2011) A genome-wide comparison of the functional properties of rare and common genetic variants in humans. American Journal of Human Genetics 88: 458–468.
2. Asimit J, Zeggini E (2010) Rare variant association analysis methods for complex traits. Annual Review of Genetics 44: 293–308.
3. Hoffmann TJ, Marini NJ, Witte JS (2010) Comprehensive approach to analyzing rare genetic variants. PLoS ONE 5: e13584.
4. Lin DY, Tang ZZ (2011) A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89: 354–367.
5. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 11: 773–785.
6. Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biology 12: 227.
7. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB (2012) The empirical power of rare variant association methods: Results from sanger sequencing in 1,998 individuals. PloS Genetics 8: e1002496.
8. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genetics 5: 1–11.
9. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics 89: 82–93.
10. Liu DJ, Leal SM (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 Genetics 6: e1001156.
11. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. PLoS Genetics 7: e1001322.
12. Gusev A, Kenny EE, Lowe JK, Salit J, Saxena R, et al. (2011) Dash: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am J Hum Genet 88: 706–717.
13. Turkmen A, Lin S (2011) Gene-based partial least-squares approaches for detecting rare variant associations with complex traits. BMC Proceedings 5: S19.
14. Scholz M, Kirsten H (2011) Comparison of scoring methods for the detection of causal genes with or without rare variants. BMC Proceedings 5: S49.
15. Brennan J, He Y, Calixte R, Nyirabahizi E, Jiang Y, et al. (2011) A lasso-based approach to analyzing rare variants in genetic association studies. BMC Proceedings 5: S100.
16. Chen H, Hendricks AE, Cheng Y, Cupples AL, Dupuis J, et al. (2011) Comparison of statistical approaches to rare variant analysis for quantitative traits. BMC Proceedings 5: S113.
17. Ayers K, Mamsoula C, Cordell H (2011) Penalized-regression-based multi-marker genotype analysis of genetic analysis workshop 17 data. BMC Proceedings 5: S92.
18. Kazma R, Hoffmann T, Witte J (2011) Use of principal components to aggregate rare variants in case-control and family-based association studies in the presence of multiple covariates. BMC Proceedings 5: S29.
19. Biswas S, Lin S (2011) Logistic bayesian lasso for identifying association with rare haplotypes and application to age-related macular degeneration. Biometrics: doi: 10.1111/j.1541-0420.2011.01680.x.
20. Luo L, Boerwinkle E, Xiong M (2011) Association studies for next-generation sequencing. Genome Research 21: 1099–1108.
21. Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26: 2375–2382.
22. Hoerl AE, Kennard R (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12: 55–67.
23. Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58: 267–288.
24. Massy WF (1965) Principal components regression in exploratory statistical research. Journal of the American Statistical Association 60: 234–256.
25. Wold H (1975) Soft modeling by latent variables: the nonlinear iterative partial least squares (nipals) approach. Perspectives in Probability and Statistics: In Honor of MS Bartlett on the Occasion of his Sixty-fifth Birthday : 117–144.
26. Wold H (1985) Partial least squares. Encyclopedia of Statistical Sciences 6: 581–591.
27. Chun H, Keles S (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B 72: 3–25.
28. Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35: 109–135.
29. Hastie T, Tibshirani R, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag, 2nd edition.
30. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832–838.
31. Nussbaum R, McInnes R, Willard H, Thompson M (2007) Thompson & Thompson genetics in medicine. Philadelphia: Saunders/Elsevier.
32. R Development Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. Accessed 2012 June 18, ISBN 3-900051-07-0.
33. Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716–723.
34. Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6: 461–464.
35. Shao J (1997) An asymptotic theory for linear model selection. Statistica Sinica 7: 221–262.
36. Song KNM, Aponte J, Manas ES, Bacanu SA, Yuan X, et al. (2011) Sequencing of Lp-PLA2-encoding PLA2G7 gene in 2000 europeans reveals several rare loss-of-function mutations. Pharmacogenomics.
37. Firmann M, Mayor V, Vida PM, Bochud M, Pecoud A, et al. (2008) The Colaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. BMC cardiovascular disorders 8: doi:10.1186/1471-2261-8-6.
38. Dastani Z, Hivert MF, Timpson N, Perry JRB, Yuan X, et al. (2012) Novel loci for adiponectin levels and their inuence on type 2 diabetes and metabolic traits: A multi-ethnic meta-analysis of 45,891 individuals. PLoS Genetics : In Press.
39. Andrew T, Hart DJ, Snieder H, Spector TD, MacGregor AJ (2001) Are twins and singletons comparable? a study of disease-related and lifestyle characteristics in adult women. Twin Research 4: 464–477.
40. Richards JB, Rivadeneira R, Inouye M, Pastinen TM, Soranzo N, et al. (2008) Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. Lancet 371: 1505–1512.
41. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biology 8: e1000294.
42. Carbonetto P, Stephens M (2011) Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Analysis 6: 1–42.
43. Frank IE, Friedman JH (1993) A statistical view of some chemometrics regression tools. Technometrics 35: 109–135.
44. Mevik BH, Wehrens R (2007) The pls package: principal component and partial least squares regression in r. Journal of Statistical Software 18: 1–24.
45. Linhart H, Zucchini W (1986) Model Selection. New York: John Wiley & Sons.
46. McQuarrie ADR, Tsai CL (1998) Regression and Time Series Model Selection. Singapore: World Scientific Publishing Company.
47. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33: 1–22.