RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Identification of Grouped Rare and Common Variants via Penalized Logistic Regression

Kristin L. Ayers* and Heather J. Cordell

Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne NE1 3BZ, United Kingdom

Received 20 December 2012; Revised 24 May 2013; accepted revised manuscript 24 May 2013.
Published online 8 July 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21746

ABSTRACT: **In spite of the success of genome-wide association studies in finding many common variants associated with disease, these variants seem to explain only a small proportion of the estimated heritability. Data collection has turned toward exome and whole genome sequencing, but it is well known that single marker methods frequently used for common variants have low power to detect rare variants associated with disease, even with very large sample sizes. In response, a variety of methods have been developed that attempt to cluster rare variants so that they may gather strength from one another under the premise that there may be multiple causal variants within a gene. Most of these methods group variants by gene or proximity, and test one gene or marker window at a time. We propose a penalized regression method (PeRC) that analyzes all genes at once, allowing grouping of all (rare and common) variants within a gene, along with subgrouping of the rare variants, thus borrowing strength from both rare and common variants within the same gene. The method can incorporate either a burden-based weighting of the rare variants or one in which the weights are data driven. In simulations, our method performs favorably when compared to many previously proposed approaches, including its predecessor, the sparse group lasso [Friedman et al., 2010].**
Genet Epidemiol 37:592–602, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: **penalized likelihood; lasso; elastic net; association analysis; rare variants**

Introduction

Although genome-wide association studies have been successful at identifying many common variants as contributors to complex diseases, most of these variants seem to have very small estimated effect sizes and explain only a small proportion of the heritability of complex diseases [McCarthy and Hirschhorn, 2008]. Attention has turned to the analysis of rare variants, where the suggestion by previous studies that multiple rare variants within the same gene can contribute to largely monogenic disorders (for a summary, see Bansal et al. [2010]) has led to the development of a variety of methods that group or collapse variants within a region, gene, or gene pathway. Testing individual rare variants within a gene is likely to be highly underpowered unless the effect sizes are huge. Collapsing or grouping variants together capitalizes on the fact that the gene is the relevant functional biological unit that may be expected to have some relationship with phenotype. Burden tests (such as CAST [Morgenthaler and Thilly, 2007], GRANVIL [Morris and Zeggini, 2009], and the variable threshold (VT) method [Price et al., 2010]) collapse the rare variants into a single variable, such as an indicator or count, for analysis, ignoring the effects from the common variants that may contain additional information. These methods require the use of a minor allele frequency (MAF) threshold cutoff to define what constitutes a rare variant and difficulties arise when the number of rare variants in the region is so small or so large that either none or all of the individuals within a phenotype group (e.g. cases or controls) carry rare variants. The combined multivariate and collapsing (CMC) method allows rare variants to be simultaneously analyzed with common variants in a multivariate test [Li and Leal, 2008]. The problem of defining and separating common and rare variants was avoided with the introduction of weighting methods that compute a weighted sum statistic (WSS), such as that proposed by Madsen and Browning [2009]. Most methods use weights inversely related to the MAF (resulting in rarer variants having higher weights). All of these methods suffer when there are protective variants in addition to risk variants, as they sum over variables with effects in potentially opposite directions. To overcome this problem, methods such as C-alpha were introduced that compare the expected variances of the distribution of the allele frequencies within the cases and controls to the actual variance [Neale et al., 2011]. SKAT, the sequence kernel association test, is a generalized version of C-alpha that allows for weights [Wu et al., 2011]. Since burden tests have been shown to be more powerful when most variants in a region are causal and have effects in the same direction, Lee et al. [2012] have developed the method SKAT-O, an extension of the SKAT test, which optimally combines a burden test and the nonburden SKAT test. Han and Pan [2009] developed the

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Kristin Ayers, Institute of Genetic Medicine, Newcastle University, Central Parkway, Newcastle upon Tyne NE1 3BZ, United Kingdom. E-mail: kayers@ucla.edu

adaptive sum method (aSum) that determines the direction of the variant weights from the data and incorporates them into the burden test. The VW-TOW method, which estimates weights from the data, places large weights on variants that have strong associations with the trait and on rare variants [Sha et al., 2012]. To allow larger weights on common variants, the authors propose dividing the variants into common and rare variants, applying their score test to each group separately, and finding the optimal combination of the two test statistics.

All of the methods above operate on a single region or gene at a time, and, since multiple genes can contribute to disease risk, we propose instead to analyze all genes simultaneously in a regression framework. Analyzing variables together in a regression model allows one to consider the impact of one variable on another, the hope being that a weak effect may become more visible when other causal effects are already accounted for. Previous studies have shown that joint modeling may improve power in certain situations for both quantitative and qualitative traits [Ayers and Cordell, 2010; Clayton, 2012; Hoggart et al., 2008; Pirinen et al., 2012]. Currently, many sequencing studies are underway, resulting in enormous amounts of detected variants. However, sample sizes remain limited to several hundred up to a thousand, and consequently, we have many more predictors than the number of test subjects, overwhelming standard regression methods. In genetic studies, we expect that only a handful of our genes will have true effects on our trait. Penalized regression methods can be used on these underdetermined problems, shrinking the size of the coefficients, pushing the coefficients of variants with little or no apparent effect on a trait down toward zero and performing model selection. With the aim of finding the subset of genes most associated with the disease, we propose PeRC (**Pe**nalized regression of **R**are and **C**ommon variants), a method that groups SNPs by genes, and collapses the rare variants in the gene into a single variable where the rare variants are allowed to contribute different effects. This approach capitalizes on the recent success of genome-wide association studies (GWAS) that shows there will generally be adequate power to detect/select common variants associated with phenotype.

Methods

Penalized Regression Approach

Regression methods can be used to analyze both qualitative and quantitative traits. Logistic regression is often used to analyze binary phenotypes such as case/control status. Given a phenotype vector y of 0's and 1's for m observations and a matrix of SNP genotypes X, if we let $p = P(y = 1|X = x)$, our logistic regression equation for individual i may be written as:

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta^T X_{i.} \quad (1)$$

where β is our vector of regression coefficients. The likelihood may be formulated as a product over all individuals i:

$$L = \prod_{i=1}^{m} p_i^{y_i}(1-p_i)^{1-y_i}.$$

With some rearranging, the log likelihood may be written as a sum over the m individuals:

$$\log L = \sum_{i=1}^{m} y_i\eta_i - \log(1+\exp(\eta_i)). \quad (2)$$

For a quantitative trait, we maximize the negative sum of squares of differences (RSS) between observed and predicted trait values, rather than maximizing the above likelihood:

$$\log L = -RSS(\beta|X, Y) = -\sum_{i=1}^{m}(y_i - \eta_i(\beta|X))^2. \quad (3)$$

With current genotyping and sequencing studies, the number of markers is typically on the order of hundreds of thousands to millions, while sample size is on the order of hundreds to thousands, leading to underdetermined problems where standard regression methods cannot produce a unique interpretable model. Penalized likelihood methods can be applied to these high dimensional regression problems to perform model selection. We maximize the log likelihood subject to a penalty that is dependent on the magnitude of the estimated parameters. A penalty on the log likelihood will penalize models that have a large number of large regression coefficients more heavily, and thus the penalized likelihood will be optimized with a sparser model. In genetics, we suspect that there are only a modest number of underlying causal variants compared to the total number of variants, and our ideal penalty would quickly exclude variables with little effect, retaining only the most relevant variables in the model. Thus, we choose to maximize the penalized log likelihood:

$$\log L(X, Y, \beta) - f(\beta, \lambda)$$

where the penalty f is a function of the regression coefficients and penalty parameters. Many different penalty functions have been proposed such as the L_1 norm (or Lasso) [Tibshirani, 1996], the L_2 norm (or ridge) [Hoerl and Kennard, 1970; Le Cessie and van Houwelingen, 1992], and the combination of these two norms, the elastic net [Zou and Hastie, 2005]. The elastic net penalty may be written as:

$$f(\lambda, \beta) = \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$$

where $\|\beta\|_1 = \sum_j |\beta_j|$ and $\|\beta\|_2^2 = \sum_j \beta_j^2$ are the L_1 and L_2 norm, respectively (with j indexing variables) and λ_1 and λ_2 are fixed parameters controlling the penalty strengths. The elastic net penalty above is reduced to the Lasso if we let $\lambda_2 = 0$ and to the ridge if $\lambda_1 = 0$. The L_1 norm imposes heavy shrinkage and drives the coefficient of many variables to zero and generally includes only one of a group of highly correlated variables [Bondell and Reich, 2008]. Ridge regression in contrast results in similar coefficients for highly correlated variables. The elastic net is somewhere in the middle,

encouraging correlated variables to enter the model together. The penalty functions also have Bayesian interpretations: the lasso penalty corresponds to a double exponential or Laplace prior on β, the ridge penalty corresponds to a zero mean Gaussian or normal prior, and the elastic net is a mixture of Gaussian and Laplace priors. As most of the mass of these priors is around zero, most of the coefficient estimates will be near zero. Penalized regression and Bayesian selection methods have previously been applied to a variety of problems in human genetics [Ayers and Cordell, 2010; Hoggart et al., 2008; Li et al., 2010; Malo et al., 2008; Yi and Zhi, 2011] and in animal and plant genetics [Mutshinda and Sillanpää, 2010, 2011; Sun et al., 2010; Xu, 2010; Yi and Xu, 2008]. The focus in the animal and plant literature has been in prediction of phenotype or genetic breeding value, rather than in variable selection per se.

In disease association studies, if we suspect that there may be several genes causing a disease, and that there may be more than one causal variant within a gene, we can take advantage of multiple signals within a gene by analyzing our variables in groups. This is consistent with the idea that the gene is the functional biological unit, and so evidence for the existence of effects at some variants within the gene should effectively upweight the prior for other variants within the same gene. To force variables to be grouped by (a) encouraging variables within a group to enter a model together, and (b) encouraging sparsity between groups, we can use the group lasso or the sparse group lasso [Friedman et al., 2010a; Meier et al., 2008; Yuan and Lin, 2006]. The sparse group lasso encourages sparsity between and within groups, and has been previously applied to GWAS for variants with frequencies > 1% [Zhou et al., 2010]. If g indexes the G groups, this penalty function may be written as:

$$f(\lambda, \beta) = \sum_{g=1}^{G} \left[\lambda_1 \left(\sum_{j \in g} \beta_j^2 \right)^{1/2} + \lambda_2 \sum_{j \in g} |\beta_j| \right]$$

where λ_1 is a parameter that controls the strength of the group penalty and λ_2 is a parameter that controls the strength of the sparsity penalty. If one variable within a group enters the model, then this penalty does not strongly discourage another variable within that group from also entering the model. Zhou et al. [2010] recommend setting $\lambda_1 = \lambda_2$ as it performed well in simulations. The group lasso penalty corresponds to a multivariate p_g dimensional, multi-Laplacian prior over each group, where p_g is the number of variables in group g.

In PeRC, we choose to use a combination of these two penalties to group both rare and common variants within a region, such as a sliding window, gene, or gene network. We propose to first collapse/cluster the RVs within a group into a single variable to model a common effect. However, we allow rare variants to contribute differently to this effect by allowing the weights of the rare variants (denoted as α_r below) to be estimated as we optimize. We can replace η_i in our likelihood with:

$$\eta_i = \beta_0 + \sum_g \left\{ \sum_{c \in g_c} x_{ic} \beta_c + \gamma_g \sum_{r \in g_r} x_{ir} \alpha_r \right\}.$$

Here $\gamma_g > 0$ is the coefficient for the linear combination of the rare variants in group g, α_r are the estimated weights for each rare variant in g_r (the set of rare variants with MAF $\leq \tau$), and g_c is the set of common variants (MAF $> \tau$) in g. If we force all α_r to be constant, we are performing a procedure similar to a burden test. Otherwise, we have an unidentifiable model unless restrictions are placed on the weights α_r. We can do this via a penalty, restricting the rare variants to have weights in the approximate range of (–1,1) by using a penalty that penalizes variables little inside this range, and heavily outside this range. This prevents the influence of any particular rare variant from becoming too large and keeps the grouped rare variants coefficient γ_g on the same scale as the coefficients β_c for the common variants. These properties can be achieved with a fourth order polynomial penalty, which is a form of bridge regression (equivalent to a prior from the exponential power family) that does not provide sparse solutions and is similar to a uniform prior in the range (-1,1). This idea is similar to the hierarchical prior used by Yi et al. [2011], who place an informative prior on α_r and a weakly informative prior on the γ_g, and model all variables in the likelihood as we model our rare variables. The method is used to group synonymous and nonsynonymous rare variants and common variants into four separate groups within a gene. Here, we encourage rare and common variants in the same gene or window to be in the model together via a group penalty. The penalty structure imposed in our approach allows an individual regression coefficient to be estimated for each common variant, effectively allowing individual common variants to be selected, while our grouping penalty allows a borrowing of strength between common and rare variants within the same gene.

Our generalized penalty function can be written as:

$$\begin{aligned} f(\lambda, \beta) = \sum_{g=1}^{G} \Bigg[& \lambda_1 s_g \left(\sum_{c \in g_c} w_c \beta_c^2 + r_g \gamma_g^2 \right)^{1/2} \\ & + \lambda_2 \left(\sum_{c \in g_c} w_c |\beta_c| + r_g |\gamma_g| \right) \\ & + \lambda_3 \left(\sum_{c \in g_c} w_c \beta_c^2 + r_g \gamma_g^2 \right) + \lambda_4 \sum_{r \in g_r} d_r \alpha_r^4 \Bigg]. \end{aligned}$$

The first term groups the rare and common variants within our region of interest, the second and third terms correspond to the elastic net and promote sparsity of the individual common variants and the groups of rare variants, while the final term prevents the coefficients for the rare variants from becoming too large and promotes a small amount of sparsity in the rare variants. If $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, then when $\tau = 1$, $\lambda = (\lambda_1, \lambda_2, 0, 0)$ corresponds to a sparse group lasso, and $(0, \lambda_2, \lambda_3, 0)$ corresponds to the elastic net. We realize that using a ridge penalty with a group penalty may be slightly redundant as the sparse group lasso gives an elastic net fit within each nonzero group, but this addition makes PeRC more flexible in terms of the analyses one can perform. To perform a burden-based procedure, we simply set $\lambda_4 = 0$, and

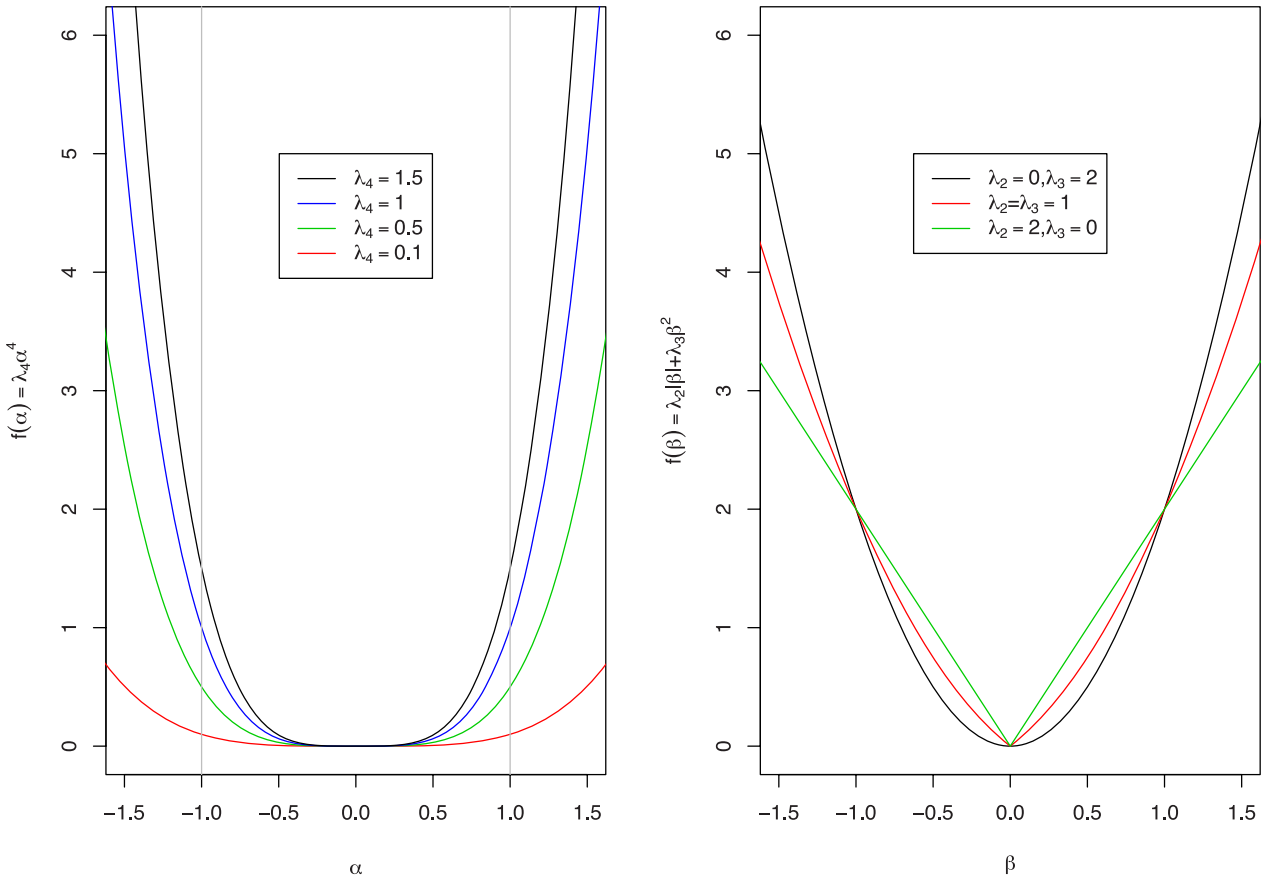

Figure 1. The first plot is the penalty function for the weights on the rare variants which applies little penalty between 0 and 1 and a large penalty elsewhere. The second plot is the elastic net penalty function vs. β.

force $\alpha_r = 1$ for all r. We could also force $\alpha_r = 2/|g_r|$ (to give a proportion of rare variants on a scale of 0–2 rather than a count), but we found that this did not perform as well in practice. We will refer to the weighted procedure as PeRC_W and the burden procedure as PeRC_B. For our weighted analysis, we keep λ_4 constant at 0.5 to maintain the shape of that part of the penalty function (although we could choose instead to make it slightly higher to encourage more sparsity). We place a weight s_g on each group dependent on its size, for example, $\sqrt{(l_g/\max(l_g))}$, where l_g is the total number of common variants in the group plus one to account for the rare group coefficient. This prevents the preferential selection of large groups solely for their ability to explain a larger proportion of phenotype variance due to increased degrees of freedom. Additionally, we can assign individual weights to the penalty terms for each variable. For instance, we may choose to penalize the common variants based on their MAF and set w_c equal to $2\sqrt{MAF_c(1-MAF_c)}$, which results in $w_c = 1$ when $MAF_c = 0.5$, as implemented in the software Mendel [Zhou et al., 2010, 2011]. This downweights the penalty of less common variants relative to more common variants. We also place a weight r_g on the rare group coefficient of similar form, where the MAF is replaced by the average MAF of the variants in the rare group, or for the case of the burden procedure, the MAF of the collapsed locus. With weights $r_g = w_c = 1$ and unstandardized genotypes, the method preferentially selects mostly common variants. For each rare variant, we place weight $d_r = \sqrt{MAF_r(1-MAF_r)}/\sqrt{\tau(1-\tau)}$ to allow little penalty to be placed on very rare variants with frequencies much smaller than τ. After some experimentation, we have currently set $(\lambda_1, \lambda_2, \lambda_3) = \kappa(1, 1, 1)$ for the burden procedure, and $(\lambda_1, \lambda_2, \lambda_3) = \kappa(4, 1, 1)$ for the weighted procedure, where κ is a penalty strength to be determined by permutation testing for both the weighted and burden tests. This choice of penalty parameters allows for strong grouping, leading to heavy group sparsity and intermediate sparsity within a selected gene, yet has the ability to select a wide range of causal gene configurations. The log likelihood is maximized using cyclic coordinate ascent. See Figure 1 for a pictorial representation of the resulting penalty function shapes.

Cyclic Coordinate Ascent

The log likelihood and the negative penalty functions are concave, and since the sum of concave functions is

Table 1. The distribution of the causal variants for each causal gene

Causal gene #	Total Variants (causal)	Minor allele frequency range						Causal Variant MAF	
		> 0.08 Count	0.03 – 0.08 Count	0.01 – 0.03 Count	.005 – 0.01 Count	0.005 – 0.001 Count	< 0.001 Count	Sum	Average
1	35 (11)	0	1	0	0	7	3	0.063	0.0058
2	40 (13)	1	0	2	1	5	4	0.175	0.0135
3	10 (2)	1	1	0	0	0	0	0.255	0.1276
4	86 (18)	0	0	3	1	4	10	0.066	0.0037
5	50 (13)	0	0	1	2	1	9	0.033	0.0045
6	80 (5)	0	0	1	0	2	2	0.023	0.0025
7	168 (36)	0	1	0	2	8	25	0.107	0.0029
8	5 (3)	0	0	1	0	0	2	0.012	0.0040
9	290 (24)	0	1	0	4	8	11	0.078	0.0032
10	20 (7)	0	1	1	0	3	2	0.080	0.0114

concave, we can use the CLG algorithm [Genkin et al., 2005] for optimization. We maximize the penalized log likelihood via Newton's method and cyclic coordinate descent [Friedman et al., 2007; Friedman et al., 2010b; Wu and Lange, 2008; Wu et al., 2009]. The coefficient update is:

$$\beta_j^{n+1} = \beta_j^n - \frac{O'(\beta^n)}{O''(\beta^n)}$$

where n is the iteration number. The derivative of the penalty function is not continuous nor differentiable at zero. When the current coefficient estimate β_j^n is zero, special steps must be taken. The elastic net penalty has continuous first and second derivatives, but the derivative is discontinuous at zero due to the absolute value term. Additionally, the derivative of the group penalty has a singularity when all coefficients in a group are zero. We attempt a move in the direction that improves the penalized log likelihood given the other penalty parts, but this move is not accepted if the derivative of the objective function changes sign (we pass the local maximum). We do not allow coefficients to take large steps or to change signs in one iteration. If our Newton update is $\Delta\beta = \beta_j^{n+1} - \beta_j^n$, then let

$$\beta_j^{n+1} - \beta_j^n = \begin{cases} -\delta & \text{if } \Delta\beta < -\delta \\ \Delta\beta & \text{if } -\delta \le \Delta\beta \le \delta \\ \delta & \text{if } \Delta\beta > \delta \end{cases}$$

where δ is currently set at 0.1. This is the proposed new value for β_j^{n+1}. If this value does not improve the objective function, we halve δ and reattempt. For the weighted procedure, the coefficient γ_g is restricted to be nonnegative. One potential problem occurs when the group coefficient γ_g is zero. At this point, changing the value of $\alpha_r \forall r \in g_r$ cannot improve the likelihood, it can only change the penalty, thus we would expect all α_r to be driven to zero, causing γ_g to remain at zero. To overcome this, we perturb γ_g, allow the α_r and subsequently γ_g be reestimated , and cycle through again, making sure the surrogate is improved. Similar issues occur with the fused lasso [Friedman et al., 2007].

Simulation Study

Using FREGENE [Hoggart et al., 2007], we simulated five chromosomes consisting of 50 Mb of sequence data for approximately 20K individuals. Gene regions were simulated for each of the chromosomes as follows, with each chromosome having a different gene density of between 250 and 1,150 genes per chromosome, for a total of 3,600 genes. First, the number of variants within a gene was simulated from a log normal density and restricted to between 3 and 300 variants per gene. A region of SNPs of this size was then randomly selected to be a gene. If the selected region had a length greater than 50 kb, another region was selected. This was repeated until we had 3,600 genes. Second, we chose two causal genes per chromosome for a total of 10 causal genes. For each causal gene, a fixed number of causal variants from given frequency ranges were randomly selected. Table 1 shows the distribution of causal variant MAFs for each gene along with the gene size, the sum of the causal variant MAFs, and the average causal variant frequency of that gene. Although there are a large number of causal variants in some genes, many are quite rare and thus unlikely to occur in a given population sample.

Case/control status was simulated 100 times for the entire population using $\text{logit}(P(y_i = 1)) = \alpha_0 + \beta_1 x_{ij_1} + \cdots + \beta_n x_{ij_p}$, where j indexes the causal variants, and p is the number of causal variants. The regression coefficient β_j for each causal variant was set to $\ln\frac{5}{4}|\log(\text{MAF}_j)|$, as implemented in the simulations performed by Wu et al. [2011] when evaluating the software SKAT. α_0 was adjusted to give a population prevalance around 11%. For each of the 100 population replicates, we randomly selected 1,000 cases and 1,000 controls to analyze. In Scenario 1, all causal variants were risk variants. In Scenario 2, 1/3 of the causal variants were selected to be protective. Although we are only considering two different scenarios of 100 replicates each, in actuality, this design generates 20 different causal variant structures, for a total of 2,000 tests on causal genes, and approximately 72K tests on null genes, the results of which can be used to assess true and false detection rates.

Analysis was performedusing our penalized regression approach PeRC with rare variant threshold $\tau = .01$. Analysis was also performed in the R package SKAT [Wu et al., 2011] and in the Score-seq software [Lin and Tang, 2011]. In SKAT, we performed four analyses with different prechosen weights based on MAFs drawn from the $Beta(\text{MAF}; a_1, a_2)$ distribution: (1) $a_1 = 1$ and $a_2 = 25$, the SKAT default, (2) $a_1 = 1$

and $a_2 = 1$, equivalent to C-alpha, (3) $a_1 = 0.5$ and $a_2 = 0.5$, equivalent to the weights used in Madsen and Browning [2009], which we will refer to as MB, and (4) SKAT-O with defaults. In Score-seq, we considered several different analyses: (1) the VT test, (2) T1 with rare variants < 1%, (3) T5 with rare variants < 5%, and (4) the Fp test (which also uses MB weights). We attempted to use the Score-seq EREC test, which is a permutation test, but found it to be too slow for this size of analysis. We also compared our results to those obtained using GRANVIL [Morris and Zeggini, 2009] (with default rare variant threshold $\tau = .05$), single marker (SM) analysis in PLINK [Purcell et al., 2007] with an additive model, the sparse group lasso as implemented in Mendel (ML) [Zhou et al., 2010, 2011] with nonuniform weights based on the MAFs, and VW-TOW with $\tau = .01$ and 10,000 permutations as recommended by the authors [Sha et al., 2012]. Although some of these methods use very similar statistics, they are implemented in slightly different ways.

Results

The different methods have their own strengths and weaknesses, as they have been tailored to perform well in different situations. For example, by design, C-alpha should be more likely to detect common causal variants than SKAT, and SKAT should be better at detecting rare variants than C-alpha. Thus, we expect the best performing method to depend on the type of data simulated.

We implemented each method over a range of *P*-value cutoff thresholds and penalty parameters and counted the number of true and false detections of genes at each cutoff. To our knowledge, Mendel v12.0 does not allow selection of the penalty parameter, but instead allows one to set the number of desired predictors, which we will use as a defined cutoff point. We calculated average per gene empirical true detection rates (power) and false positive detection rates (type I error) by summing the number of true and false positives for a given *P*-value/penalty parameter over all replicates. For all methods and each gene, we summed the number of gene detections at a given cutoff or penalty parameter value over all replicates to get a gene detection count. For the single marker test SM, we counted a gene detection for any gene containing a marker below the *P*-value threshold. To obtain power, we added the detection counts for all 10 causal genes, and divided this by the total number of possible true detections (1,000), i.e. the number of true positives (10) times the number of replicates (100). For type I error, we added detection counts for all noncausal genes, and divided this sum by the number of true negatives (3,590) times the number of replicates, i.e. by the total of 359,000 possible true negative detections (null genes). Figures 3 and 2 show the resulting power vs. false positive (FP) rates per gene for all methods considered, for Scenarios 1 and 2, respectively.

PeRC appears to perform well on this data set. In Figure 3, the methods in Score-seq (T1, VT, Fp, and T5), seem to outperform the other methods. In Figure 2, when a third of the causal variants are protective, most of the methods perform worse than in Scenario 1 (with Fp taking a huge loss of power), but PeRC_W seems to perform similarly to the case with all risk causal variants. The Score-seq methods take the biggest hit compared to Scenario 1, having similar true detection rates to SKAT, MB, and Calpha. GRANVIL and PeRC_B also lose power, as expected. Single marker (SM) analysis and the sparse group lasso (GL) perform poorly in both cases. Table 2 shows gene detection power and false positive rates for: (1) the single gene based methods (using the the Bonferonni corrected *P*-value of $.05/3,600$ to declare a gene as significant) (2) the SM method (using the Bonferroni corrected *P*-value of $.05/120,608$ to declare a SNP as significant, with a gene declared as significant if any SNPs within it are significant), and (3) the results for PeRC when using a permutation-based test to find the penalty parameter that gives the desired experimentwise 5% false positive rate. In PeRC, to find the κ that yields the desired false positive rate for our simulations, we assumed each replicate should have the same κ (since they were drawn from the same population on the same set of SNPs and from the same simulation model), in order to reduce computational costs. For each scenario, we permuted case/control status once for each replicate and recorded the resulting number of variables in the model over a range of κ. We selected the κ that gave approximately five false positives over the 100 replicates, which gave us our desired per gene false positive rate. For a real data set, we could permute the case/control status 100 times and select the value of κ that results in five false positives over all 100 replicates, or we could use a similar procedure, for example, to that described in Ayers and Cordell [2010]. As Mendel v12.0 did not allow selection of the penalty parameter and had little power in Figures 3 and 2, we did not include the sparse group lasso in this table. Table 2 shows that SM, GRANVIL, PeRC, VW-TOW, and the Score-seq methods control type I error at the nominal level while SKAT and SKAT-O do not; although SKAT and SKAT-O had the highest power, they had higher false positive rates than most of the other methods.

Next, we examined that genes each method preferentially selected to see why our method was performing so much better than the other methods in Scenario 2. The results are shown in Figures 4 and 5. These figures give the detection power of the causal genes over the 100 replicates for each method at several different average false positive rates. On further inspection, it appears as if our method preferentially selects the longest causal gene with many causal variants, even though the penalty has been adjusted to penalize longer genes more heavily. Most of the methods preferentially select genes with a higher sum of causal variant MAFs. The sparse group lasso appears to only detect the small causal gene with 2 common causal variants. SKAT-O and VW-TOW select the widest range of genes at low false positive rates. All methods struggled to detect genes 3, 4, 5, 6, and 8 in both Scenarios. Gene 3 contained common variants, while genes 4, 5, 6, and 8 had a low sum of causal variants minor allele frequencies.

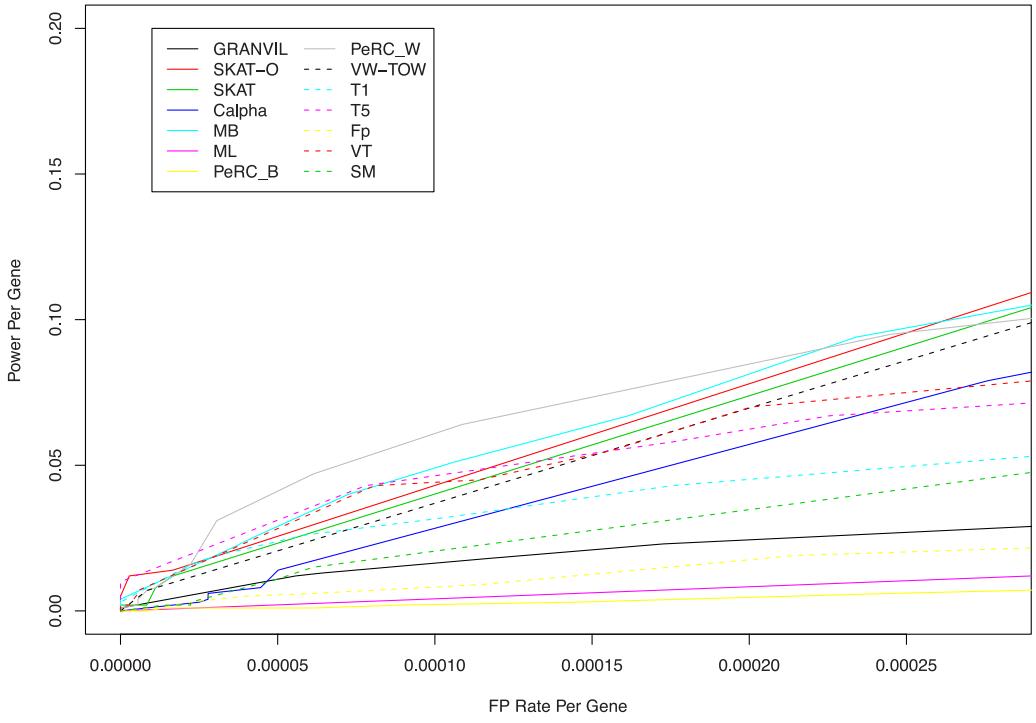

Figure 2. Power vs. false positive rates for the case where two-thirds of the causal variants are risk variants and the rest are protective (Scenario 2). The *P*-value threshold or penalty parameter rate was varied to obtain points for the curves, which are an average over all replicates at each point.

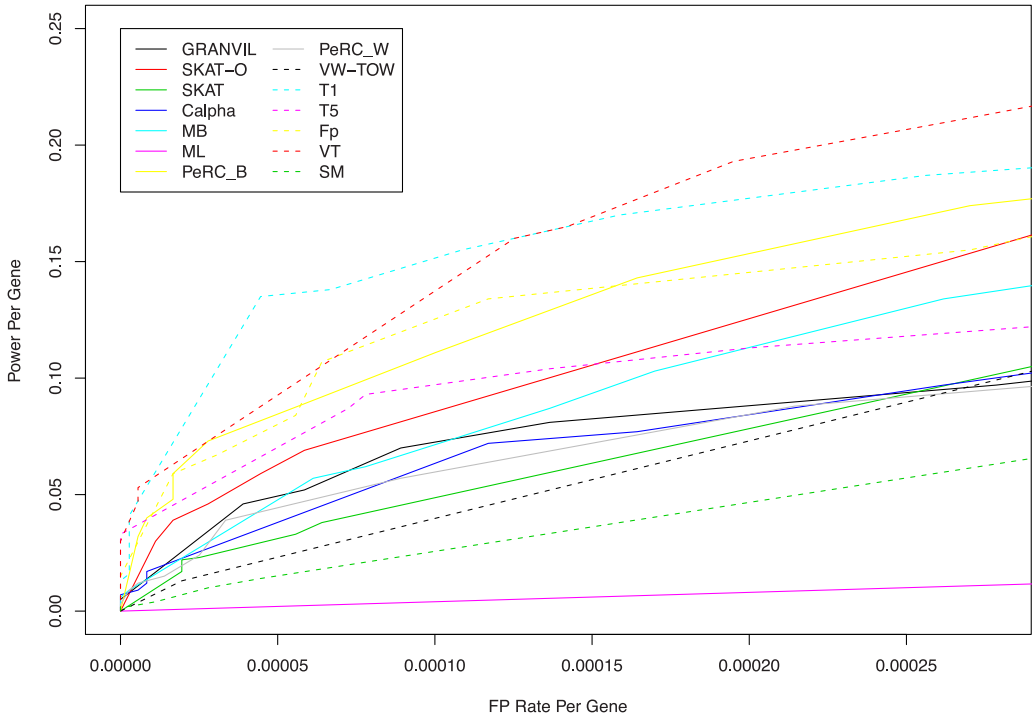

Figure 3. Power vs. false positive rates for the case where all causal variants are risk variants (Scenario 1). The *P*-value threshold or penalty parameter rate was varied to obtain points for the curves, which are an average over all replicates at each point.

Table 2. Estimated average power and false positive rates (FPR) and their standard errors (SE) for the simulated data

		Method											
		GRANVIL	SKAT-O	SKAT	Calpha	MB	PeRC_B	PeRC_W	VW-TOW	T1	T5	Fp	VT	SM
Scenario 1														
Power	Est	0.046	0.247	0.171	0.072	0.057	0.073	0.057	0.013	0.135	0.087	0.059	0.160	0.010
	SE	(0.006)	(0.011)	(0.011)	(0.008)	(0.006)	(0.007)	(0.006)	(0.008)	(0.007)	(0.007)	(0.011)	(0.003)	(0.004)
FPR	Est	3.9e-05	5.0e-04	5.1e-04	1.2e-04	6.1e-05	2.8e-05	8.9e-05	1.9e-05	4.5e-05	7.2e-05	1.7e-05	1.3e-04	2.8e-05
	SE	(1.0e-05)	(4.5e-05)	(4.3e-05)	(2.2e-05)	(1.3e-05)	(8.4e-06)	(1.7e-05)	(1.3e-05)	(1.4e-05)	(6.6e-06)	(1.7e-05)	(1.9e-05)	(8.2e-06)
Scenario 2														
Power	Est	0.012	0.120	0.122	0.079	0.022	0.001	0.047	0.007	0.016	0.043	0.001	0.043	0.011
	SE	(0.004)	(0.010)	(0.010)	(0.007)	(0.005)	(0.002)	(0.006)	(0.004)	(0.006)	(0.001)	(0.006)	(0.003)	(0.003)
FPR	Est	5.6e-05	3.2e-04	3.4e-04	2.8e-04	3.6e-05	3.6e-05	6.1e-05	8.4e-06	2.2e-05	7.8e-05	1.1e-05	8.1e-05	5.0e-05
	SE	(1.3e-05)	(4.2e-05)	(4.1e-05)	(3.0e-05)	(1.0e-05)	(1.1e-05)	(1.2e-05)	(8.6e-06)	(1.5e-05)	(5.5e-06)	(1.6e-05)	(2.1e-05)	(4.8e-06)

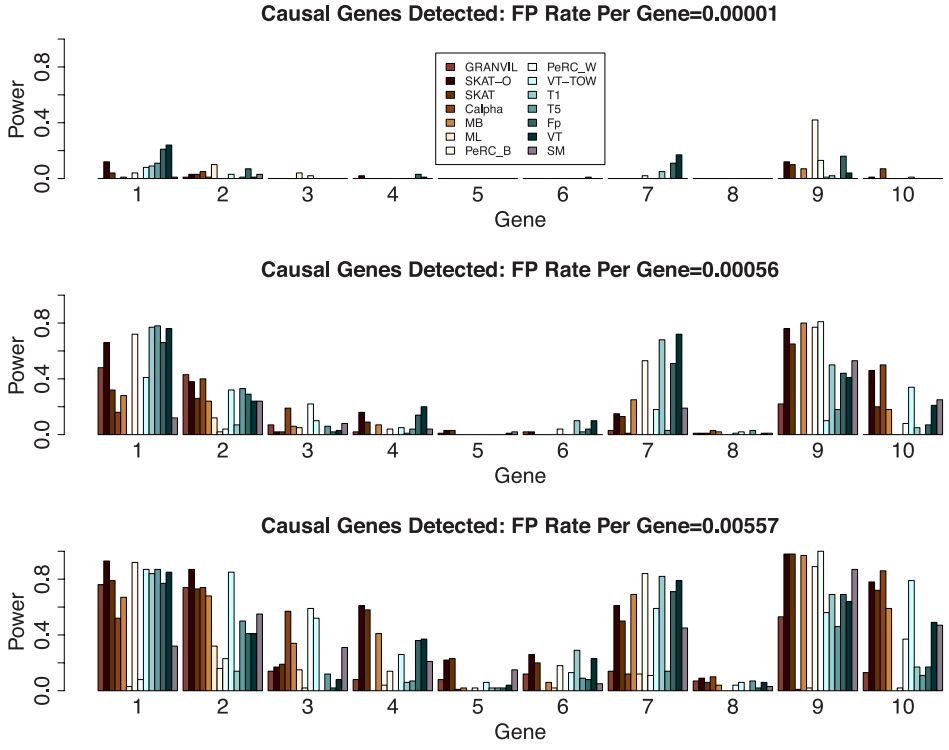

Figure 4. Average gene power over all replicates for each causal gene at three different false positive rates for Scenario 1.

To investigate whether PeRC_W selects genes solely due to their length, we permuted the genotypes for the longest gene (9) to break the genotype phenotype correlation and re-ran the analysis (Supplementary Figures S1 and S2). At low false detection rates (top two rows of Supplementary Figures S1 and S2), our method did not detect gene 9. However, at a much higher false detection rate (bottom row of Supplementary Figures S1 and S2), gene 9 was selected as a false detection for a large proportion of the replicates. PeRC_B did not detect gene 9 after permutation of the genotypes for either Scenario (data not shown). We then removed all of the SNPs in gene 9 from the data set and re-ran both PeRC and Mendel. We recomputed the power and false detection rates for all of the methods excluding gene 9, (see Supplementary Figures S3 and S4). In this scenario, PeRC_W performed poorly, although it did maintain some of its power when a proportion of the causal variants were protective. Currently, we control for gene length bias via the s_g term in the penalty function. Silver and Montana [2012] suggest using weights determined from permutations with a null response to control this bias, which may be a beneficial addition to PeRC_W.

We investigated the power of PeRC_B and PeRC_W over a small selection of different choices for the values of the penalty parameters (Figures 6 and 7). Although the lasso-like penalty (with $\lambda_1 = \lambda_3 = 0$) performed best for both methods

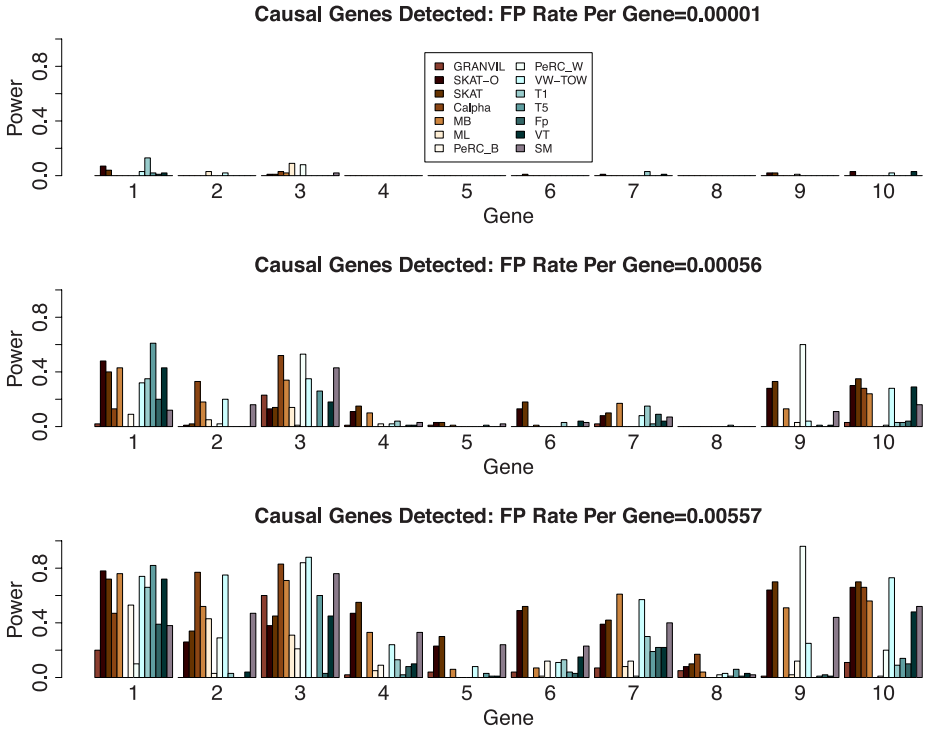

Figure 5. Average gene power over all replicates for each causal gene at three different false positive rates for Scenario 2.

at low false positive rates, it performed poorly at larger false positive rates. Thus, in Figures 3– 5, we focused on parameters that lead to higher power at higher false positive rates and had the widest range of gene detections (shown in Supplementary Figures S5– S8).

Conclusions

Many recent methods developed for rare and common variant analysis have been geared toward determining the optimal weights for the different variants. Here, we have proposed a flexible regression framework to test for association between a dichotomous phenotype or a quantitative trait with rare and common grouped variants where the weights for the rare variants can be determined by the data. With weights based on minor allele frequencies for the group and common coefficients, we find that we may lose power to detect common variants as they are more heavily penalized. The large number of possible weights and penalty parameters creates an enormous search space of parameters of which we have only touched the surface. Here, we have selected the penalty parameters $(\lambda_1, \lambda_2, \lambda_3) = \kappa(4, 1, 1)$ for PeRC_W, which encourages heavy grouping. For PeRC_B, the lasso $(\lambda_1 = \lambda_3 = 0)$ seems to be optimal for Scenario 1, but very strong grouping, $(\lambda_1, \lambda_2, \lambda_3) = \kappa(8, 1, 1)$, is optimal for Scenario 2, perhaps because we are drawing more information from the common variants, searching for hints of association from any common variants that might be in association with a causal rare variant (Figures 6 and 7). A more detailed investigation of the choice of penalty parameter is beyond the scope of this manuscript, but would be an interesting topic for further work. An interesting exploratory approach would be to investigate the number and positions of selected predictors at a variety of penalty parameter ratios and over a range of κ values.

We recently became aware of a new hierarchical method CHARM [Cardin et al., 2012] that shares the same spirit as PeRC_W, allowing each variant to have a different effect size but without model selection. CHARM uses a prior distribution of effect sizes centered at zero with a hierarchical parameter controlling the degree to which the effect sizes vary from zero. Like PeRC, CHARM is able to distinguish between multiple signals and linkage disequilibrium. Unlike PeRC, CHARM only analyzes one gene at a time. The authors of CHARM state that it requires approximately one minute of computation time per SNP, so we did not attempt to run CHARM on our 120K SNP data set.

The advantage of our model is that we can include all genes simultaneously, considering the impact of one variable on another, so that, in principal, a weak effect may become more visible when other causal effects are already accounted for. Penalty parameters can be adjusted to change the weights on the rare variants, common variants, and groups of variants through various weighting options to tailor the method to the required analysis. The user can also input their own weights. PeRC also allows for penalized and unpenalized covariates, and can perform the lasso, sparse group lasso, elastic

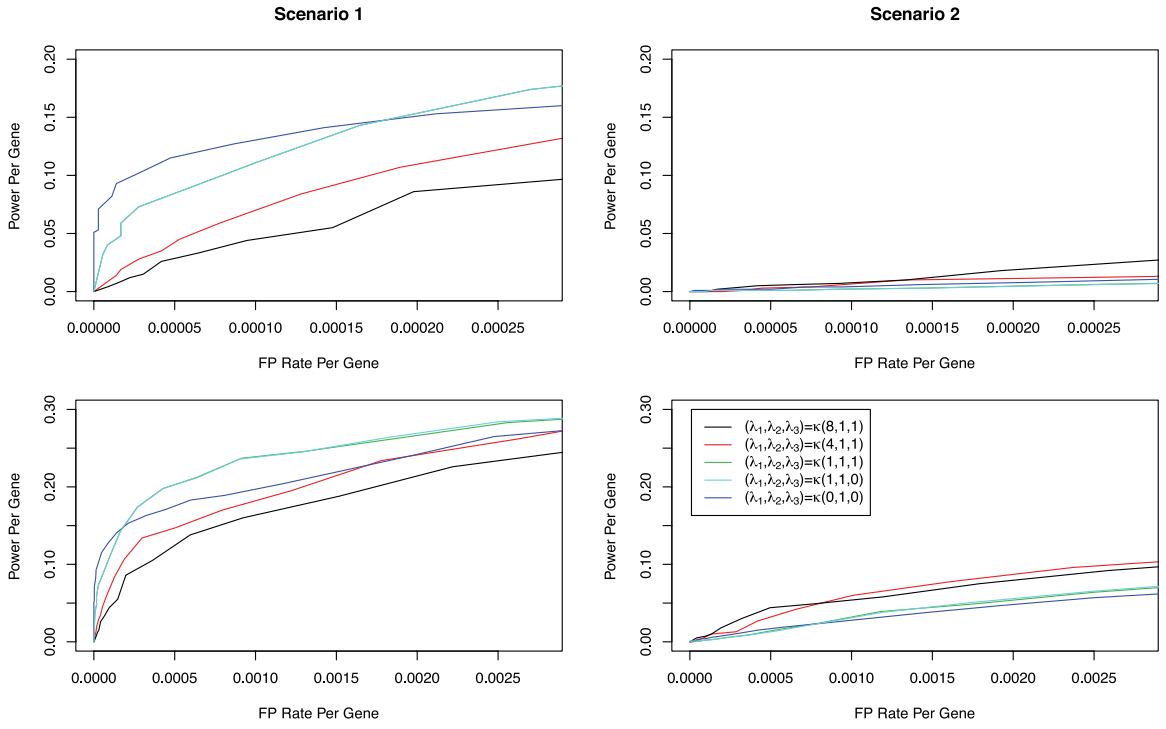

Figure 6. Power vs. false positive rates for PeRC_B over a variety of penalty parameters. The two plots on the left represent Scenario 1, where the bottom plot is a larger range of false positive rates than the top plot. The two plots on the right correspond to Scenario 2.

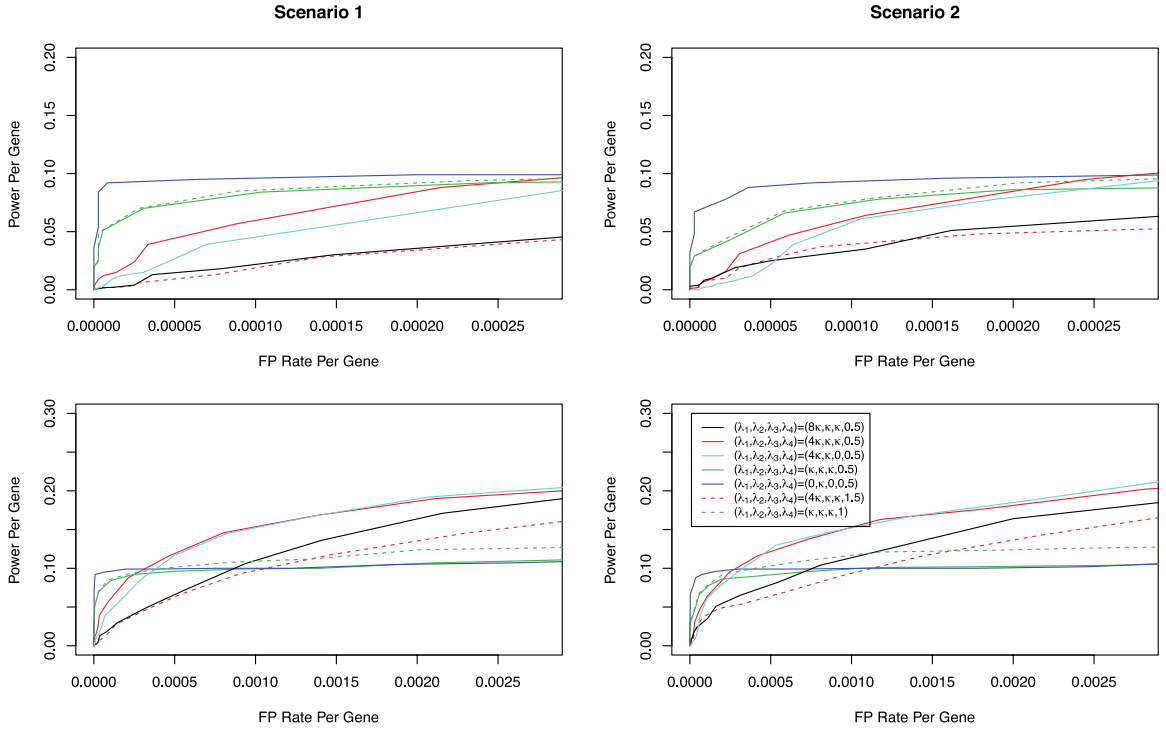

Figure 7. Power vs. false positive rates for PeRC_W over a variety of penalty parameters. The two plots on the left represent Scenario 1, where the bottom plot is a larger range of false positive rates than the top plot. The two plots on the right correspond to Scenario 2.

net, and ridge by controlling the penalty parameters appropriately. Currently, PeRC does not handle missing genotypes, so they must be imputed beforehand and input to PeRC as dosage data. For 20 values of the penalty parameter κ on approximately 120K SNPs, PeRC_W required around 4 hr for to run, while PeRC_B required around an hour. Although PeRC does not always outperform other methods, its improved performance over the sparse group lasso is encouraging for rare variant penalized regression. Additionally, we may be able to improve performance by using an optimal combination of the burden and weighted models, similarly to SKAT-O, to take advantage of the strengths of each model within the different scenarios of risk variants (risk vs. protective). We have shown that our method is able to detect long causal genes with many very rare variants with the current selected penalty parameters. The coefficients for the rare variants are estimated from the data, allowing variants to have both risk and protective effects. Thus, PeRC_W is fairly insensitive to whether or not the causal variants are risk or protective. PeRC can also be used with sequence data for a set of genes or pathways that may have been shown or suspected to be associated with a trait in attempt to determine which genes and/or SNPs are contributing most to the trait variance.

Acknowledgments

This work was supported by the Wellcome Trust (grant reference 087436).

References

Ayers KL, Cordell HJ. 2010. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol* 34: 879–891.

Bansal V, Libiger O, Torkamani A, Schork NJ. 2010. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11: 773–785.

Bondell HD, Reich BJ. 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics* 64: 115–123.

Cardin NJ, Mefford JA, Witte JS. 2012. Joint association testing of common and rare genetic variants using heirarchical modeling. *Genet Epidemiol* 36: 642–651.

Clayton D. 2012. Link functions in multi-locus genetic models: implications for testing, prediction, and interpretation. *Genet Epidemiol* 36: 409–418.

Friedman J, Hastie T, Hofling H, Tibshirani R. 2007. Pathwise coordinate optimization. *Ann Appl Statist* 1: 302–32.

Friedman J, Hastie T, Tibshirani R. 2010a. A note on the group lasso and sparse group lasso. Technical report, Department of Statistics, Stanford University.

Friedman J, Hastie T, Tibshirani R. 2010b. Regularization paths for generalized linear models via coordinate descent. *J Statist Software* 33: 1–22.

Genkin A, Lewis DD, Madigan D. 2005. Sparse logistic regression for text categorization. DIMACS Working Group on Monitoring Message Streams Project Report, April 2005.

Han F, Pan W. 2009. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 70: 42–54.

Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12: 55–67.

Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, Iorio MD, Balding DJ. 2007. Sequence-level population simulations over large genomic regions. *Genetics* 177: 1725–1731.

Hoggart CJ, Whittaker JC, Iorio MD, Balding DJ. 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing studies. *PLoS Genet* 4(7):e1000130.

Le Cessie S, van Houwelingen JC. 1992. Ridge estimators in logistic regression. *Appl Statist* 41: 191–201.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Mickerson DA, NHLBI GO Exome Sequencing Project-ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91: 224–237.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83: 311–321.

Li J, Das K, Fu G, Li R, Wu R. 2010. The Bayesian lasso for genome-wide association studies. *Bioinformatics* 27: 516–523.

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.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5(2):e1000384.

Malo N, Libiger O, Schork NJ. 2008. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet* 82: 375–385.

McCarthy MI, Hirschhorn JN. 2008. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet* 17: R156–R165.

Meier L, van de Geer S, Bühlmann P. 2008. The group lasso for logistic regression. *J R Statis Soc Ser* B 70: 53–71.

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutat Res* 615: 28–56.

Morris AP, Zeggini E. 2009. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34: 188–193.

Mutshinda CM, Sillanpää MJ. 2010. Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics* 186: 1067–1075.

Mutshinda CM, Sillanpää MJ. 2011. Bayesian shinkage analysis of qtls under shape-adaptive shrinkage priors, and accurate re-estimation of genetic effects. *Heredity* 107: 405–412.

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. 2011. Testing for an unusual distribution of rare variants. *PLoS Genet* 7: e1001322.

Pirinen M, Donnelly P, Spencer CC. 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat Genet* 44: 848–851.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86: 832–838.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, de Bakker P, Daly M, Sham P. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575.

Sha Q, Wang X, Wang X, Zhang S. 2012. Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genet Epidemiol* 36: 567–571.

Silver M, Montana G, Alzheimer's Disease Neuroimaging Initiative. 2012. Fast identification of biological pathways associated with a quantitative trait using group lasso with overlaps. *Stat Appl Genet Mol Biol* 11: 1–43.

Sun W, Ibrahim JG, Zou F. 2010. Genome-wide mulitple loci mapping in experimental crosses by the iterative adaptive penalized regression. *Genetics* 185: 349–359.

Tibshirani R. 1996. Regression shrinkage via the lasso. *J R Statis Soc* 58: 267–88.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89: 82–93.

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. 2009. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* 25: 714–21.

Wu TT, Lange K. 2008. Coordinate descent algorithms for lasso penalized regression. *Ann Appl Statist* 2: 224–44.

Xu S. 2010. An expectation maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity* 105: 483–494.

Yi N, Liu N, Zhi D, Li J. 2011. Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. *PLoS Genet* 7(12):e1002382.

Yi N, Xu S. 2008. Bayesian LASSO for quantitative trait loci mapping. *Genetics* 179: 1045–1055.

Yi N, Zhi D. 2011. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol* 35: 57–69.

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J R Statis Soc Ser B* 68: 49–67.

Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel EM, Lange K. 2011. Penalized regression for genome-wide association screening of sequence data. *Pac Symp Biocomput* 2011: 106–117.

Zhou H, Sehl ME, Sinsheimer JS, Lange K. 2010. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26(19):2375–2382.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J R Statis Soc Ser B* 67: 301–320.