RESEARCH ARTICLE

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

The Value of Statistical or Bioinformatics Annotation for Rare Variant Association With Quantitative Trait

Andrea E. Byrnes,[1] Michael C. Wu,[1] Fred A. Wright,[1] Mingyao Li,[2] and Yun Li[1,3,4]*

[1]*Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina;* [2]*Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania;* [3]*Department of Genetics, University of North Carolina, Chapel Hill, North Carolina;* [4]*Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina*

Received 25 March 2013; Revised 20 May 2013; accepted revised manuscript 3 June 2013.
Published online 8 July 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21747

ABSTRACT: In the past few years, a plethora of methods for rare variant association with phenotype have been proposed. These methods aggregate information from multiple rare variants across genomic region(s), but there is little consensus as to which method is most effective. The weighting scheme adopted when aggregating information across variants is one of the primary determinants of effectiveness. Here we present a systematic evaluation of multiple weighting schemes through a series of simulations intended to mimic large sequencing studies of a quantitative trait. We evaluate existing phenotype-independent and phenotype-dependent methods, as well as weights estimated by penalized regression approaches including Lasso, Elastic Net, and SCAD. We find that the difference in power between phenotype-dependent schemes is negligible when high-quality functional annotations are available. When functional annotations are unavailable or incomplete, all methods suffer from power loss; however, the variable selection methods outperform the others at the cost of increased computational time. Therefore, in the absence of good annotation, we recommend variable selection methods (which can be viewed as "statistical annotation") on top of regions implicated by a phenotype-independent weighting scheme. Further, once a region is implicated, variable selection can help to identify potential causal single nucleotide polymorphisms for biological validation. These findings are supported by an analysis of a high coverage targeted sequencing study of 1,898 individuals.
Genet Epidemiol 37:666–674, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: rare variants; association; weighting; variable selection; variant annotation

Introduction

Recent studies have shown that rare variants may be important to the underlying etiology of complex traits [Cohen et al., 2004; Dickson et al., 2010; Gorlov et al., 2008; Haase et al., 2012; Nelson et al., 2012; Zawistowski et al., 2010] and that they may account for part of the "missing heritability" [Eichler et al., 2010; Gibson, 2010; Maher, 2008; Manolio et al., 2009] left by genome-wide association studies (GWAS). Conventional association analysis methods, which evaluate each variant independently of all others, lack the statistical power to evaluate rare variants given the sample size of sequencing data currently available. However, there is increasing evidence that the combined effects of rare variants in the same exon, gene, region, or biological pathway can be used to elucidate complex phenotypes [Cohen et al., 2004; Nejentsev et al., 2009; Sanna et al., 2011]. Where the effect size of a single variant may not be large enough to detect with the sample sizes available, a collection of variants with small effect size, taken together, may be detectable. In order to explore the potential effects of rare variants in present-day genomic data, a large number of methods [Bacanu et al., 2011; Cheung et al., 2012; Lee et al., 2012; Li and Leal, 2008; Li et al., 2010a; Madsen and Browning, 2009; Mao et al., 2013; Neale et al., 2011; Price et al., 2010; Tzeng et al., 2011; Wu et al., 2011; Xu et al., 2012; Yi et al., 2011] for aggregating information across variants have emerged. However there is little consensus on which method is most effective. The weighting scheme adopted when aggregating across variants is an important consideration, as is the use of functional or bioinformatics information when available.

We present an evaluation of multiple weighting schemes through a series of simulations. We evaluate several existing phenotype-independent [Cohen et al., 2004; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007] and phenotype-dependent weighting schemes [Wu et al., 2011; Xu et al., 2012], as well as weighting schemes determined by linear regression, penalized regression, and variable selection methods, including Lasso [Tibshirani, 1996], Elastic Net (EN) [Zou and Hastie, 2005], and SCAD [Xie and Huang, 2009]. We conduct simulations under a variety of scenarios with different numbers of true causal variants, mixtures of direction of effect, and availability of functional information, mimicking sequencing studies of a quantitative trait. We then apply each of these methods to a set of high coverage targeted sequencing data [Nelson et al., 2012] of 1,898 individuals from the CoLaus population-based cohort [Firmann et al., 2008].

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

*Correspondence to: Yun Li, Department of Genetics, Campus Box 7264, University of North Carolina, Chapel Hill, NC 27599. E-mail: yunli@med.unc.edu

© 2013 WILEY PERIODICALS, INC.

Materials and Methods

Over the last few years, numerous sensible weighting schemes have been proposed. In most of these methods a genomic region or variant set is assigned a weighted sum over the variants meant to describe the burden of potentially influential variants carried by each individual. We call this weighted sum S_i. Further, we assume there are N individuals under study, indexed by i, and for each individual we have M variants in the region or variant set, indexed by j.

Phenotype-Independent Weighting Schemes

First, we examine three approaches that are independent of the observed phenotype. The first of these is a simple indicator of whether or not rare variants (minor allele frequency, MAF < 0.01) are present in the region [Cohen et al., 2004]. That is,

$$S_i = I\left(\sum_{j=1}^{M} I(\hat{q}_i < Q)x_{ij} > 0\right),$$

where x_{ij} is the number of minor alleles observed for individual i at variant j. $\hat{q}_i = \frac{\sum_{i=1}^{N} x_{ij}+1}{2N+2}$ is the estimated MAF of variant j in the data with pseudocounts and Q is the MAF threshold. In this work, we consider $Q = 0.05$.

Second, we examine a count approach, which assigns a higher score to individuals carrying a larger number of rare alleles [Morgenthaler and Thilly, 2007];

$$S_i = \sum_{j=1}^{M} I(\hat{q}_j < Q)x_{ij}$$

with x_{ij} being the count of rare alleles for individual i at variant j and $\hat{q}_j$ being the estimated MAF, as defined above.

We also consider the approach proposed by Madsen and Browning [2009] where the weight for variant j is a function of the MAF:

$$S_i = \sum_{j=1}^{M} \xi_j x_{ij}, \text{ where } \xi_j = \frac{1}{\sqrt{N \times \hat{q} \times (1-\hat{q})}}$$

with x_{ij} and $\hat{q}_j$ as above. In the original Madsen and Browning framework for case-control studies, MAFs are estimated using controls only. However, in this paper, the outcome of interest is quantitative and we estimate MAF using the entire sample, which makes the method phenotype-independent in this context.

Phenotype-Dependent Weighting Schemes

We also consider phenotype-dependent regression-based methods. First, we examine the performance of marginal regression coefficients. That is, we fit the simple linear regression model $Y = x_j\beta + \varepsilon$ for each variant j separately and independently and then take the fitted values $\tilde{\beta}_j$ to be our weights.

$$S_i = \sum_{j=1}^{M} \xi_j x_{ij}, \text{ where } \xi_j = \tilde{\beta}$$

the MLE of β for the model above.

Though imperfect, this weighting scheme allows investigators to test for associations with multiple rare variants in cases where $N < M$ and begin to follow up on individual variants that may potentially be of interest.

Second, we consider weights from ordinary multiple regression, modeling all of the M variants simultaneously. That is, we fit the model $Y = X\beta + \varepsilon$, where the (i, j)th element of the matrix $X = x_{ij}$, the minor allele count for individual i at variant j. We then take S_i to be as above, with the fitted values from this multiple regression, $\hat{\beta}_j = \xi_j$ [Lin and Tang, 2011; Xu et al., 2012].

We also consider weights from several variable selection methods. Such methods are appealing because we expect the majority of rare variants not to influence the quantitative trait of interest. Use of penalized regression is therefore expected to reduce the number of nonzero weights. Similar strategies were recently proposed in the context of rare variant association testing [Turkmen and Lin, 2012; Zhou et al., 2010]. In penalized regression, we solve for the $\hat{\beta}'s$, which best fit the data, subject to some constraint(s) or penalty. That is, instead of minimizing the sum of squared error, $(Y - \beta X)'(Y - \beta X)$, we aim to minimize the sum of squared errors and an additional penalty term, $(Y - \beta X)'(Y - \beta X) + P(\lambda, \beta)$. In general, the greater the number of parameters included in the model, the greater the penalty. A number of penalty functions have been proposed and extensively studied in the recent statistical literature [Heckman and Ramsay, 2000; Hesterberg et al., 2008; Kyung et al., 2010; Wu and Lange, 2008]. Of these, we chose three: the Lasso, which imposes a linear penalty [Tibshirani, 1996], EN, which imposes a quadratic penalty [Zou and Hastie, 2005], and SCAD, which is designed to penalize smaller coefficients more heavily than larger coefficients [Xie and Huang, 2009].

For Lasso and SCAD, only one tuning parameter, λ, is required. We used the R packages *lars* [Efron et al., 2004] and *ncvreg* [Breheny and Huang, 2011] with default parameter values, which is to choose the optimal λ among a grid of 100 possible values equally spaced on the log-scale. For EN, there are two tuning parameters, one for the linear component and one for the quadratic component. The linear term, λ_1, is chosen in the same way as the λ parameter for the Lasso and SCAD methods, discussed above. The quadratic parameter, λ_2, was set to 1 in all simulations and for the real data. We used the R package *elasticnet* to fit the EN models [Zou and Hastie, 2005]. After model fitting, we then use estimated coefficients from each of these variable selection methods as weights. The number of nonzero coefficients included is upper-bounded by 100 for each of these schemes throughout this work.

Under each weighting scheme examined, we determine the significance of a genomic region using a score test of the

following form:

$$U = \sum_{i=1}^{N} (Y_i - \bar{Y}) S_i,$$

where $S_i = \sum_{j=1}^{M} \xi_j x_{ij}$, in which N is the number of individuals under study, and Y_i is the quantitative trait value for the ith individual. S_i is the genetic score for the ith individual, a weighted sum across multiple variants. Specifically, x_{ij} is the number of minor alleles observed for individual i at variant j, where x_{ij} are not normalized. M is the number of variants in the region under study (discovered through sequencing in our context) and ξ_j is the weight of variant j under one of the above weighting schemes. The analytical distribution for this statistic is not generally known in this context, so significance must be assessed empirically by permutation.

Additionally, we apply the similarity-based method SKAT [Wu et al., 2011] to each of our simulated data sets and the real data set for comparison. We use weights based on the default beta distribution implemented in the SKAT package, version 0.79. We will comment in the Discussion section on the conceptual differences between the weighting schemes we consider in this work and the SKAT methodology.

Simulation Setup

We simulate 45,000 chromosomes for a series of hundred 50 kb regions with a coalescent model [Schaffner et al., 2005] that mimics linkage disequilibrium (LD) in real data, accounts for variations in local recombination rates, and models population history consistent with the CEU samples. We then randomly select 2,000 simulated chromosomes (forming 1,000 diploid individuals) to mimic a large sequencing study. For each region, we simulate one single pool of 45,000 chromosomes instead of multiple pools of 2,000 chromosomes so that the causal variants in each region can be determined by population MAFs (MAFs calculated using the entire population of 45,000 chromosomes) and thus retained across replicates from the same region. We assume only rare variants (0.001 < population MAF < 0.05) influence the value of the quantitative trait and we randomly select m variants that truly influence the quantitative trait value. For each variant, we independently assign the direction of influence according to r, the probability that a causal variant will increase the trait value. Following Wu et al. [2011], we then simulate quantitative traits under the null model:

$$y_i = 0.5E_{1i} + 0.5E_{2i} + \varepsilon_i \qquad \text{(null model)},$$

where E_{1i}, E_{2i}, and ε_i are independent with $E_{1i} \sim Bernoulli(0.5)$ to mimic a binary covariate, $E_{2i} \sim Normal(0,1)$ to mimic a continuous covariate, and $\varepsilon_i \sim Normal(0,1)$. We also simulate quantitative traits under an alternative model:

$$y_i = 0.5E_{1i} + 0.5E_{2i} + \sum_{j=1}^{m} \beta x_{ij}^{C} + \varepsilon_i \quad \text{(alternative model)},$$

where $\beta_j = r_j\,|k \times F(\mathrm{MAF}_j)|$ and $r_j = 1$ with probability r and $r_j = -1$ with probability $(1 - r)$. E_{1i}, E_{2i}, and ε_i are as before, j indexes the truly causal variants, and x_{ij}^{C} is the number of minor alleles individual i has at causal variant j. The link function F takes one of the following forms:

$$F_{\log}(q) = k \times \log(q),\ F_{\log}(q) = k \times \log\left(\frac{q}{1-q}\right),$$

$$F_{MB}(q) = k \times \frac{1}{\sqrt{q(1-q)}},$$

where N is the number of individuals sequenced. We call the first link function log, the second logit, and the third Madsen-Browning (MB). In addition, we also consider $F_{random}(q)$, a random value chosen from the *exponential*(1) distribution, independent of q and multiplied by k. The constant k is a scaling factor to control the magnitude of the change in quantitative trait due to truly causal genetic variants. In our simulations k is set to 0.2, which keeps the heritability h^2, between 0.1% and 2.5%. Complex human quantitative traits are thought to have heritability estimates in this range [Manolio et al., 2009]. In the Results section, we report the results for the logit link function; results for all four link functions are given in the supplementary materials.

To assess significance in each simulated setting, score test statistic from each weighting scheme is compared to the empirical distribution of the test statistic obtained under the null simulations. We assess the significance of each test at the $\alpha = 0.01$ level using the empirical null distribution, which we approximate using 100,000 data sets simulated under the null hypothesis of no variant contributing to the quantitative trait.

Simulation of Data Sets Under the Null Hypothesis

For each of the 100 regions we simulate, we randomly select 100 samples of 2,000 chromosomes (forming 1,000 diploid individuals). We then assign quantitative trait values under the null model specified above. Using these 100 × 100 = 10,000 data sets simulated under the null hypothesis, we obtain the empirical null distribution of the test statistics for each method.

Simulation of Data Sets Under Different Alternative Hypotheses

For each choice of r, m, and $F(\cdot)$, we select 2,000 chromosomes from the population of 45,000 chromosomes again via simple random sampling. Again, we randomly pair these chromosomes to form diploid individuals and replicate 100 times for each region. For each replicate, we randomly select m rare variants to be causal. Each causal variant is assigned a direction in which to exert its effect (positive with probability r and negative with probability $1 - r$).

Simulation of "Good" Functional Annotation

In each simulated data set, we annotate variants as "functional" or "non-functional." We assume that we have a

reasonably good bioinformatics tool such that a true causal variant has 90% probability to be annotated as "functional." Even a perfect bioinformatics tool can only predict functionality, not causality or association with a particular trait of interest. Because of this, we annotate an additional random number of W noncausal variants as "functional." Kryukov et al. [2009] have estimated that approximately one-third of de novo missense mutations (that would be predicted as functional by a sensible bioinformatics tool) have no effect on phenotypic traits. We therefore used 1/3 as the lower bound for the fraction of noncausal variants annotated and simulated $W \sim N(25, 5)$, rounded to the nearest integer. We evaluate the performance of each of these weighting schemes both using all variants without the help of the bioinformatics tool, and using only the "functional" variants annotated. Under the null distribution, W variants are selected at random.

Simulation of GWAS Data Sets

We use the same choice of causal variants in each region as in the simulated sequencing data. Consequently, the direction of association and true effect size of each of these are unchanged. In order to simulate GWAS SNPs, we select 1,000 chromosomes from the total 45,000 to mimic the 1000 Genomes [Abecasis et al., 2012] sample. The simulated 1000 Genomes sample is used to define LD, based on which GWAS SNPs are selected. For each region, we choose 75 GWAS SNPs consisting of the first 70 tagSNPs (SNPs with the highest number of LD buddies where an LD buddy is an SNP for which the $r^2 > 0.8$) and five SNPs at random from the remaining set of SNPs, mimicking the Illumina Omni5 or Affymetrix Axiom high-density SNP genotyping platforms.

Results

In the Absence of a Bioinformatics Tool

Throughout our simulations, we observe several consistent patterns. First, when we apply these methods in the absence of a bioinformatics tool (thus, all variants are included in analysis), variable selection schemes (most noticeably Lasso and EN) outperform other methods, including SKAT, in nearly all situations (notable exceptions are discussed below). For example, under the simulated setting of 10 causal variants, among which five are expected to increase quantitative trait value, the power is 80.0% and 83.7% for Lasso and EN, and is 0.4%, 7.3%, 7.6%, 43.2%, 25.3%, 60.5%, 41.3%, and 46.6% for Indicator, Count, MB, Marginal Regression, Multiple Regression, SCAD, SKAT (all variants), and SKAT (rare variants only), respectively (Fig. 1a). Under the simulated setting of 50 causal variants among which 40 are expected to increase quantitative trait value, power is 100% for both Lasso and EN, and is 0.03%, 0.19%, 0.07%, 99.63%, 100%, 100%, 96.9%, and 98.5% for Indicator, Count, MB, Marginal Regression, Multiple Regression, SCAD, SKAT (all variants), and SKAT (rare variants only), respectively (Fig. 1b).

In the Presence of a Good Bioinformatics Tool

In the presence of a good bioinformatics tool (as introduced in the Materials and Methods section) the power increases

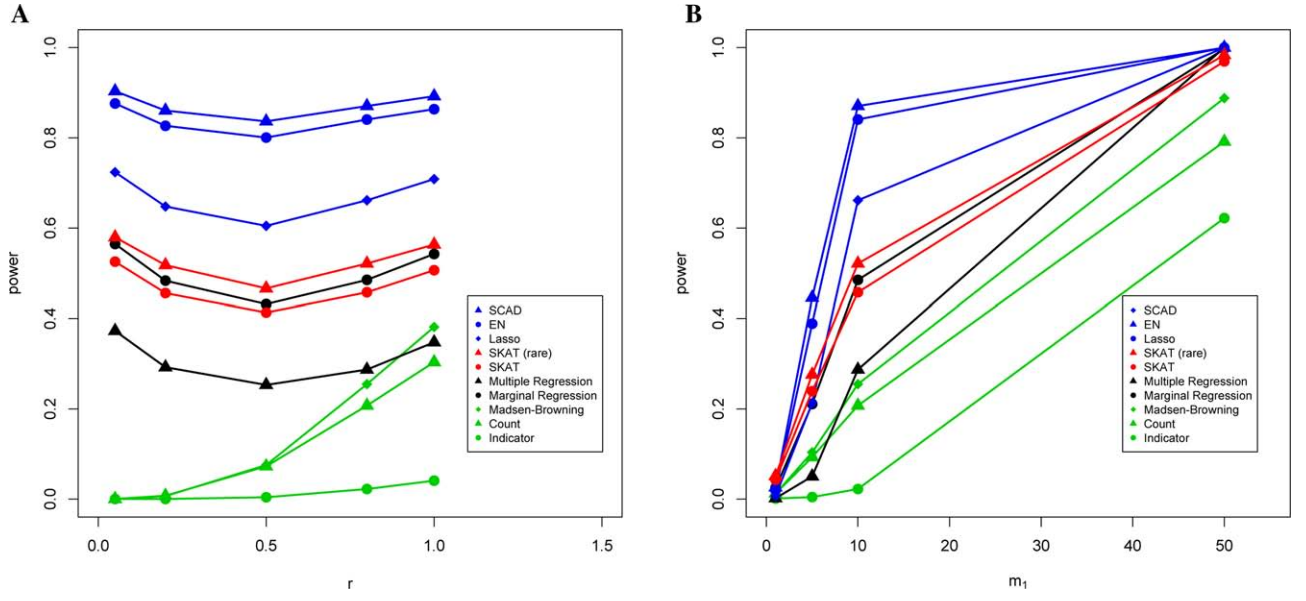

Figure 1. Power comparison in the absence of a bioinformatics tool. The figure shows the power (*Y*-axis) of the different methods across a wide spectrum of m (the number of true causal variants) and r (the proportion of variants that contribute to our quantitative trait in a positive direction) in the absence of a bioinformatics tool. In (a), we fix m at 10 and show power comparisons across the entire spectrum of r (*X*-axis). In (b) we show how power changes as a function of m (*X*-axis) with r fixed at 0.8. Here we use the logit link function.

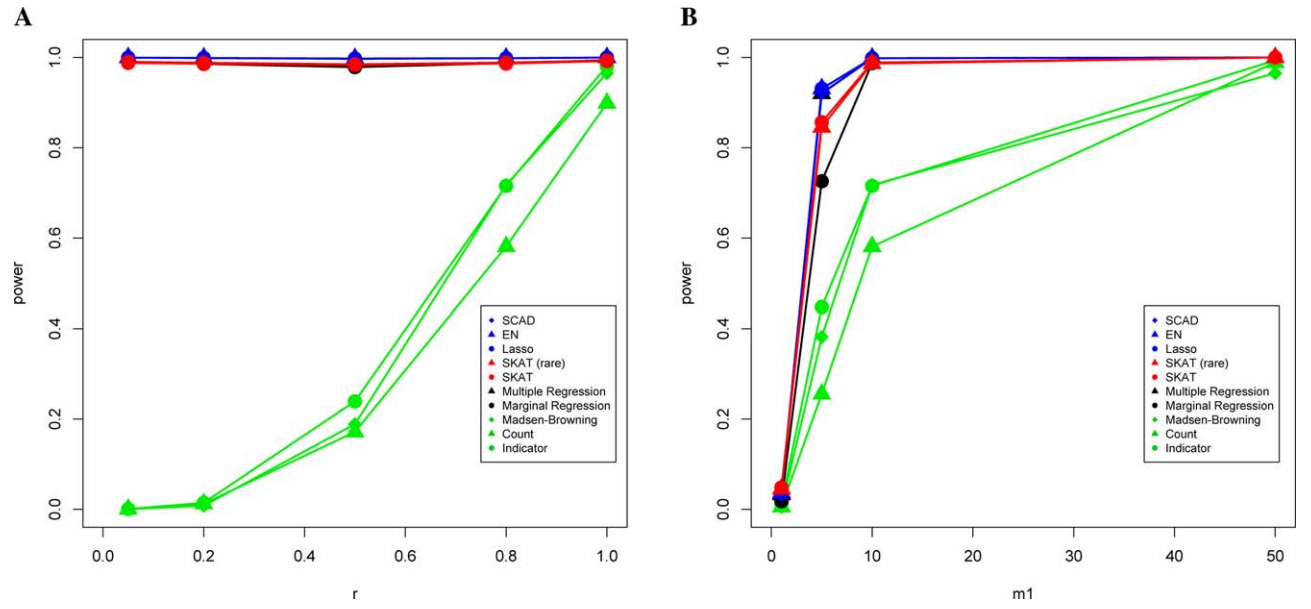

Figure 2. Power comparison in the presence of the good bioinformatics tool. The figure shows the power (*Y*-axis) of the different methods across a wide spectrum of *m* (the number of true causal variants) and *r* (the proportion of variants that contribute to our quantitative trait in a positive direction) in the presence of the good bioinformatics tool described in the Materials and Method section. Like in Figure 1A, we fix *m* at 10 and show power comparisons across the entire spectrum of *r* (*X*-axis) in (A). Similarly, in (B) we show how power of the methods changes as a function of *m* (*X*-axis) with *r* fixed at 0.8. Again the logit link function is used.

for each of the methods previously discussed. Most notably, the phenotype-independent methods show a substantial gain in power once the bioinformatics tool is applied. For example, under the simulated setting of 10 causal variants, among which five are expected to increase the quantitative trait value, the power is 99.83% and 99.80% for Lasso and EN, and is 23.91%, 17.15%, 18.85%, 97.87%, 99.73%, 99.76%, 98.49%, and 98.34% for Indicator, Count, MB, Marginal Regression, Multiple Regression, SCAD, SKAT (all variants), and SKAT (rare variants only), respectively (Fig. 2a). Under the simulated setting of 50 causal variants, among which 40 increase quantitative trait value, power is 100% for both Lasso and EN, and is 99.38%, 98.89%, 96.51%, 100%, 100%, 100%, 100%, and 100% for Indicator, Count, MB, Marginal Regression, Multiple Regression, SCAD, SKAT (all variants), and SKAT (rare variants only), respectively (Fig. 2b). Although power increases for all methods, the relative performance of the methods changes little from that under the absence of a bioinformatics tool.

Effect of *m* (the Number of Causal Variants) and *r* (Percent of Positive Causal Variants)

As the number of true causal variants (m) increases, so does power for all methods. This is to be expected because adding more causal variants increases the signal-to-noise ratio. When the number of true causal variants is very small, none of the methods have adequate power. Interestingly, it is in these situations where m is very small that SKAT manifests its advantage over other methods examined. As r gets smaller (i.e., the probability that a causal variant will contribute positively to the quantitative trait values gets smaller), the power of the phenotype-independent methods decreases. For example, the phenotype-independent methods have close to 0 power when $r = 0.05$; while the phenotype-dependent methods are relatively unaffected by changing values of r (Figs. 1a and 2a). We also observe a slight dip in power in all of the phenotype-dependent schemes when $r = 0.5$ and no bioinformatics information is used (Fig. 1a), which is to be expected because the signals from different directions are canceling one another. Similar trends are seen in all simulations with all four link functions (shown in supplementary materials).

Weight Estimation Accuracy for Individual Variants

Table 1 shows the correlation between the true and estimated values of the weights for each method under the

Table 1. Average Pearson correlation of true and estimated weights ($m = 10$ and $r = 0.8$)

Method	All markers	Limited to functional markers
Indicator	–	–
Count	0.0126	0.2386
Madsen-Browning	0.0591	0.1225
Marginal Regression	0.1588	0.6490
Multiple Regression	0.0883	0.6537
Lasso	0.2852	0.7436
EN	0.3555	0.7787
SCAD	0.2301	0.7344
SKAT (all)	–	–
SKAT (rare only)	–	–

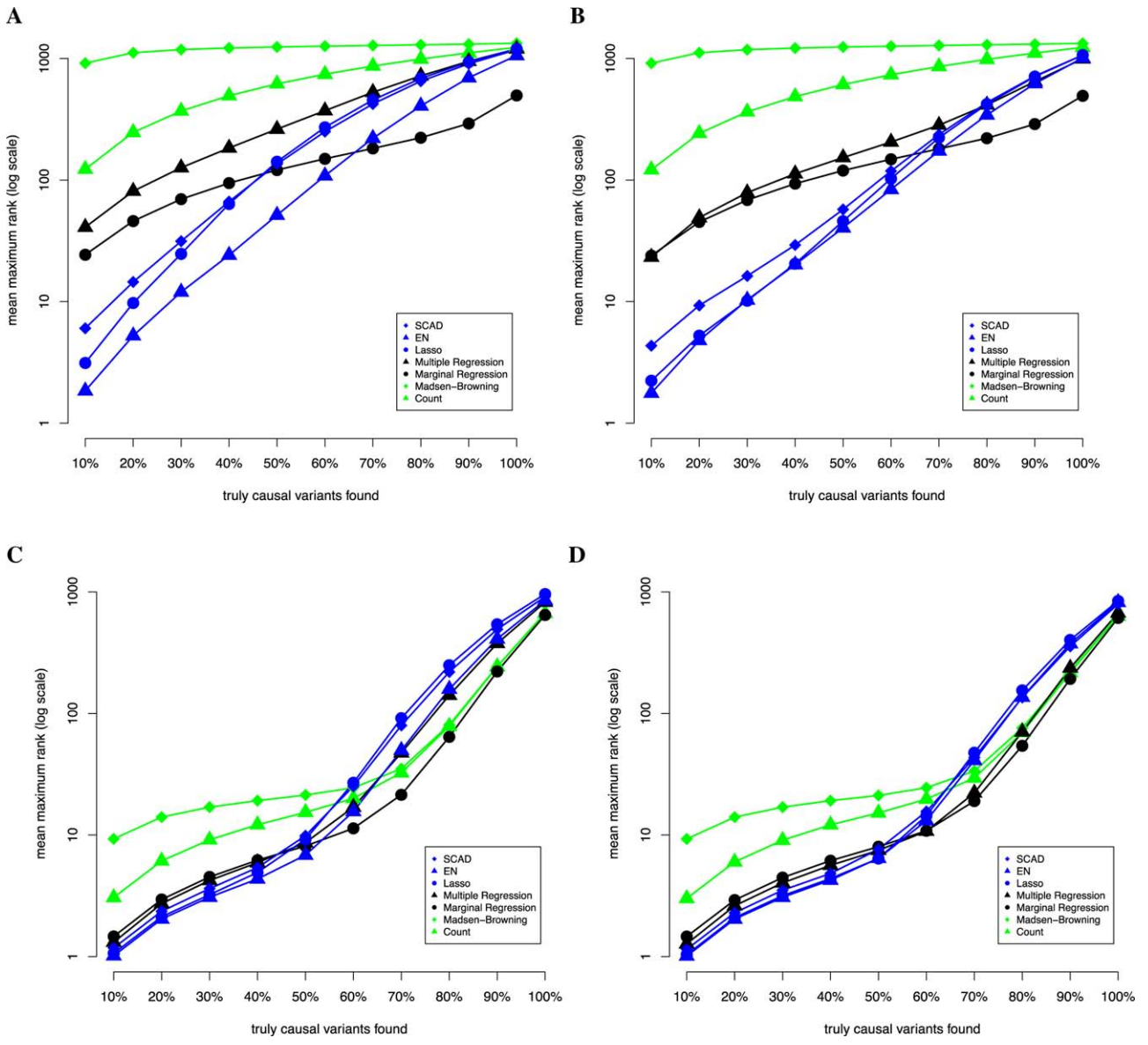

Figure 3. How far down the ranked list are the truly causal variants when all variants are included? (A) The number of variants that must be considered (*Y*-axis) in order to catch the top 10%, 20% ... 100% of truly causal variants (*X*-axis) in simulation when all variants are considered. We assume that the variants are ranked in order of significance. These plots aggregate true and estimated weights from all 10,000 replicates of the experiment and once again, we fix *r* at 0.8, *m* at 10 and use the logit link function. (B)LD buddies (variants with $r^2 > 0.8$ with causal variant) are taken into consideration. (C) The results are restricted from (A) to functional variants only using a good bioinformatics tool. (D) Functional variants are restricted only and takes LD buddies into account.

simulation settings in which the number of truly causal variants, *m*, is 10 and the proportion of variants contributing in the positive direction, *r*, is 80%. Of note, the correlation between true and estimated weights increases for all methods with the addition of bioinformatics filtering. The EN and Lasso yield the highest correlations between estimated and true weights, both in situations where we restrict to variants that are likely to be functional (Pearson correlations of 0.285 and 0.355), and when we do not (Pearson correlations of 0.744 and 778).

Identification of Individual Causal Variants

When using variable selection schemes, we have the opportunity to identify individual causal variants within the region or variant set under study. Figure 3 illustrates the accuracy with which the causal variant(s) can be identified by each weighting scheme. Note that the causal variant(s) are not always 100% identified, but in many cases, the causal variant, or a variant in high LD ($r^2 > 0.8$), has estimated nonzero weights. For example, if we fix $m = 10$, $r = 0.8$, and the logit

link function, without considering LD buddies, we need to consider the top 696 (109 and 12) variants in order to detect 90% (60%, 30%) of the causal variants using EN (Fig. 3a); taking LD buddies into consideration, the numbers decrease to 378 (14 and 4; Fig. 3b). While considering functional information we consider fewer variants and narrow the field to include a higher proportion of truly causal variants. In this case, we need to consider the top 408 (16 and 4) variants in order to detect 90% (60%, 30%) of the causal variants (Fig. 3c) without considering LD buddies; with LD buddies taken into consideration, the numbers decrease to 374 (13 and 3; Fig. 3d).

Results With GWAS Data Sets

Studies that sequence a portion or the entirety of the genome are becoming increasingly common, but still much more GWAS data exist than sequencing data. Imputation has been shown to accurately predict genotypes at untyped variants from GWAS data in a variety of circumstances [Auer et al., 2012; de Bakker et al., 2008; Li et al., 2009b; Li et al. 2010a; Liu et al., 2012; Marchini and Howie, 2010]. Using our simulated GWAS data and simulated reference, we observe that variable selection can improve power for GWAS data as well. However, the power is consistently lower than that under the sequencing setting due to the imperfect rescue of information through imputation (comparing Fig. 1 with supplementary Fig. S3). In our simulations, the imputation accuracy is 99.66% for all variants and 99.98% for rare variants, but most of the inaccuracies are due to missed rare variants. In fact, among variants with MAF < 0.001 nearly all inaccuracies are due to failure to identify the minor allele. Specifically, the squared Pearson correlation between the imputed genotypes (continuous, ranging from 0 to 2) and the true underlying genotypes (coded as 0, 1, and 2) is only 0.2397 for variants with MAF < 0.001. Supplementary Figure S3 shows the relative power of these weighting schemes over a range of r (supplementary Fig. S3a) and m (supplementary Fig. S3b).

Results With Real Data Set

Of the over 6,000 individuals in the CoLaus cohort [Firmann et al., 2008], 1,898 had recorded total cholesterol and targeted sequence data in 202 drug target genes [Nelson et al., 2012]. Sequencing was done at moderately high coverage (with median coverage 27X) and genotype calls were obtained using *SOAP-SNP* [Li et al., 2009a]. Sporadic missing genotypes were imputed with *MaCH* [Li et al., 2010b]. One gene previously known to be associated with total cholesterol in these data is used as a positive control. We test each of the 172 autosomal genes with and without removing nonfunctional variants using *ANNOVAR* [Wang et al., 2010]. For each method, we estimate weights in association with total cholesterol and, for the methods that accommodate covariates, we adjust for age, age^2, sex, and the first five principal components. For the phenotype-independent methods, no covariate adjustment is performed and significance is assessed by permutation of the Y_i's. For methods allowing covariates (marginal and multiple regression, Lasso, EN, and SCAD), permutation of outcomes alone is not appropriate. For these methods, we fit a regression model, $Y_i \sim Z_i$, where Z is the matrix of covariates and then obtain residuals, ε_i. The ε_i's are then randomly permuted to obtain a set of ε_i^*'s, the permuted residuals. For each permutation, we fit the model $\varepsilon_i^* \tilde{X}_i$ in order to re-estimate the weights ξ_j and scores S_i as in Davidson and Hinkley, [1997]. We do 10,000 such permutations and, from these, obtain a null distribution of statistics with which to assess significance. Because SKAT produces analytical P-values shown to preserve type I error [Wu et al., 2011], we use the SKAT analytical P-values without permutation.

Table 2. Permuted *P*-values[a] on the positive control gene in the real data set

Method	All variants (491)	Limited to functional variants (13)
Indicator	0.208	0.00057
Count	0.068	***0.00017***[b]
Madsen-Browning	0.090	**0.00041**[c]
Marginal Regression	0.166	0.00420
Multiple Regression	0.136	0.00395
Lasso	**0.017**[c]	0.00053
EN	***0.008***[b]	0.00059
SCAD	0.111	0.00078
SKAT (all)	0.329	0.00142
SKAT (rare only)	0.348	0.00142

[a] Except for SKAT(all) and SKAT(rare only).
[b] Most significant P-value under each column is in bold, italicized.
[c] Second most significant P-value under each column is in bold.

When all variants regardless of bioinformatics prediction are included, the variable selection methods Lasso and EN yield the smallest P-values compared to other methods for the previously implicated gene. However, the previously implicated gene is not the most significant among the 172 genes tested. Using *ANNOVAR* annotations [Wang et al., 2010], we restrict to nonsynonymous variants in coding regions of the genome only. When considering only these functional variants, most weighting schemes identify the correct gene with highly significant P-values (Table 2 and supplementary Fig. S4).

Discussion

In summary, through extensive simulation studies with varying number, model, and direction of causal variant(s) contributing to a quantitative trait, we find that functional annotations derived from good set of bioinformatics tools can substantially boost power for rare variant association testing. In the absence of good bioinformatics tools, "statistical" annotation based on phenotype-dependent weighting of the variants, particularly through variable selection based methods to both select potentially causal/associated variants and estimate their effect sizes, manifests advantages. This observation holds for both sequencing-based studies or studies based on a combination of genotyping, sequencing, and imputation. We also find supporting evidence from application to a real sequencing-based data set.

The price one has to pay for adopting phenotype-dependent methods is the necessity of permutation, which can be easily performed through permuting of residuals for the analysis of quantitative traits [Davidson and Hinkley, 1997; Lin, 2005] or using the BiasedUrn method [Epstein et al., 2012] recently proposed for binary traits. This, in turn, increases computational costs. Therefore, we recommend primarily using phenotype-dependent weighting for refining the level of significance. That is, we recommend applying phenotype-dependent weighting only to genomic regions or variant sets that have strong evidence of association (but not necessarily reaching genome-wide significance) from methods that do not require permutation (e.g., SKAT [Wu et al., 2011]).

We note that testing over a region by aggregating information across variants is a different task from estimating effect sizes of individual variant (as measured by the variant weights in our work). Perfection in the latter (i.e., being able to estimate weights for each individual variants accurately) leads to perfection in the former (i.e., maximal testing power over the region harboring those variants), but not vice versa. Based on our simulations where we know the true contribution (effect size) of each individual variant, we find that individual effect sizes cannot be well estimated (Pearson correlation between true and estimated effect sizes <0.5 even for the best variable selection based methods). However, these methods can still increase power of region or variant set association analysis without accurate estimation of individual variant effect sizes. In addition, these methods are able to identify the vast majority of the causal variants, particularly when LD buddies are considered.

In this paper, we mainly consider aggregation of information at the genotype level (where we first obtain a regional genotype score via a weighted sum of genotype scores for individual variants and then assess the association between the regional genotype score and the phenotype of interest), which underlies the largest number of rare variant association methods published. In contrast, there are methods that aggregate information at the effect size level (e.g., SKAT [Wu et al., 2011] where the final regional score test statistic is a weighted sum of the test statistics for individual variants) or at the *P*-value level, for example, in Cheung et al. [2012]. Our comparisons with SKAT suggest that the same conclusions apply to aggregation methods at levels other than genotype.

Lastly, although one could potentially argue that the phenotype-dependent methods require an undesirable computing-power trade-off in the presence of good bioinformatics tools, in practice, we rarely (if ever) get perfect bioinformatics tools. In addition, even perfect bioinformatics tools can only predict functionality but not causality or association with particular phenotypic trait(s) of interest. Therefore, we view that the application of "statistical annotation" through phenotype-dependent weighting, particularly using variable selection based methods, to top regions or variant sets implicated by computationally efficient phenotype-independent methods, is valuable.

Acknowledgments

We thank GlaxoSmithKline, especially Drs. Margaret G. Ehm, Matthew R. Nelson, Li Li, and Liling Warren for sharing the targeted sequencing data. We also thank our CoLaus collaborators for providing the phenotypic data. The research is supported by R01HG006292, R01HG006703 (awarded to Y.L.), and R01HG004517, R01HG005854 (to M.L.).

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.

Auer PL, Johnsen JM, Johnson AD, Logsdon BA, Lange LA, Nalls MA, Zhang G, Franceschini N, Fox K, Lange EM and others. 2012. Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am J Hum Genet* 91(5):794–808.

Bacanu SA, Nelson MR, Whittaker JC. 2011. Comparison of methods and sampling designs to test for association between rare variants and quantitative traits. *Genet Epidemiol* 35(4):226–235.

Cheung YH, Wang G, Leal SM, Wang S. 2012. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. *Genet Epidemiol* 36(7):675–685.

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305(5685):869–872.

Davidson AC, Hinkley DV. 1997. *Bootstrap Methods and Their Applications.* New York: Cambridge University Press.

de Bakker PIW, Ferreira MAR, Jia XM, Neale BM, Raychaudhuri S, Voight BF. 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 17:R122–R128.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. *PLoS Biol* 8(1):e1000294.

Efron B, Hastie T, Johnstone I, Tibshirani R. 2004. Least Angle Regression. *Ann Stat* 32(2):407–409.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11(6):446–450.

Epstein MP, Duncan R, Jiang Y, Conneely KN, Allen AS, Satten GA. 2012. A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *Am J Hum Genet* 91(2):215–223.

Firmann M, Mayor V, Vidal PM, Bochud M, Pecoud A, Hayoz D, Paccaud F, Preisig M, Song KS, Yuan X and others. 2008. The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc Disord* 8:6.

Gibson G. 2010. Hints of hidden heritability in GWAS. *Nat Genet* 42(7):558–560.

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 82(1):100–112.

Haase CL, Frikke-Schmidt R, Nordestgaard BG, Tybjaerg-Hansen A. 2012. Population-based resequencing of APOA1 in 10,330 individuals: spectrum of genetic variation, phenotype, and comparison with extreme phenotype approach. *PLoS Genet* 8(11):e1003063.

Heckman NE, Ramsay JO. 2000. Penalized regression with model-based penalties. *Can J Stat* 28(2):241–258.

Hesterberg T, Choi NH, Meier L, Fraley C. 2008. Least angle and $\ell 1$ penalized regression: a review. *Statist Surv* 2:61–93.

Kryukov GV, Shpunt Stamatoyannopoulos JA, Sunyaev SR. 2009. Power of deep all-exon resequencing for discovery of human trait genes. *P Natl Acad Sci USA* 106(10):3871–3876.

Kyung M, Gill J, Ghosh M, Casella G. 2010. Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Anal* 5(2):369–411.

Lee S, Wu MC, Lin X. 2012. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 89:82–93.

Li BS, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83(3):311–321.

Li RQ, Li YR, Fang XD, Yang HM, Wang J, Kristiansen K. 2009a. SNP detection for massively parallel whole-genome resequencing. *Genome Res* 19(6):1124–1132.

Li Y, Willer C, Sanna S, Abecasis G. 2009b. Genotype imputation. *Annu Rev Genomics Hum Genet* 10:387–406.

Li Y, Byrnes AE, Li M. 2010a. To identify associations with rare variants, just WHaIT: weighted haplotype and imputation-based tests. *Am J Hum Genet* 87(5): 728–735.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010b. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 34(8):816–834.

Lin DY. 2005. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 21(6):781–787.

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.

Liu EY, Buyske S, Aragaki AK, Peters U, Boerwinkle E, Carlson C, Carty C, Crawford DC, Haessler J, Hindorff LA and others. 2012. Genotype imputation of metabochip SNPs using a study-specific reference panel of ∼4,000 haplotypes in African Americans from the women's health initiative. *Genet Epidemiol* 36(2): 107–117.

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

Maher B. 2008. Personal genomes: the case of the missing heritability. *Nature* 456(7218):18–21.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A. 2009. Finding the missing heritability of complex diseases. *Nature* 461(7265):747–753.

Mao X, Li Y, Liu Y, Lange L, Li M. 2013. Testing genetic association with rare variants in admixed populations. *Genet Epidemiol* 37(1):38–47.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11(7):499–511.

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(1–2):28–56.

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(3):e1001322.

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. 2009. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324(5925):387–389.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337(6090):100–114.

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(6):832–838.

Sanna S, Li B, Mulas A, Sidore C, Kang HM, Jackson AU, Piras MG, Usala G, Maninchedda G, Sassu A and others. 2011. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet* 7(7):e1002198.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 15(11):1576–1583.

Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. *J Ro Stat Soc Ser B* 58:267–288.

Turkmen A, Lin S. 2012. An optimum projection and noise reduction approach for detecting rare and common variants associated with complex diseases. *Hum Hered* 74(1):51–60.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet* 89(2):277–288.

Wang K, Li M, Harkonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nuc Acids Res* 38(16):e164.

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(1):82–93.

Wu TT, Lange K. 2008. Coordinate descent algorithms for lasso penalized regression. *Ann Appl Stat* 2(1):224–244.

Xie HL, Huang J. 2009. SCAD-penalized regression in high-dimensional partially linear models. *Ann Stat* 37(2):673–696.

Xu C, Ladouceur M, Dastani Z, Richards JB, Ciampi A, Greenwood CM. 2012. Multiple regression methods show great potential for rare variant association tests. *PLoS One* 7(8):e41694.

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.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet* 87(5):604–617.

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 Stat Soc Ser B* 67:301–320.