Published in final edited form as:
Genet Epidemiol. 2014 February ; 38(2): 104–113. doi:10.1002/gepi.21783.

Regularized Rare Variant Enrichment Analysis for Case-Control Exome Sequencing Data

Nicholas B. Larson[1,§] and **Daniel J. Schaid**[1]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract

Rare variants have recently garnered an immense amount of attention in genetic association analysis. However, unlike methods traditionally used for single marker analysis in GWAS, rare variant analysis often requires some method of aggregation, since single marker approaches are poorly powered for typical sequencing study sample sizes. Advancements in sequencing technologies have rendered next-generation sequencing platforms a realistic alternative to traditional genotyping arrays. Exome sequencing in particular not only provides base-level resolution of genetic coding regions, but also a natural paradigm for aggregation via genes and exons. Here we propose the use of penalized regression in combination with variant aggregation measures to identify rare variant enrichment in exome sequencing data. In contrast to marginal gene-level testing, we simultaneously evaluate the effects of rare variants in multiple genes, focusing on gene-based LASSO and exon-based sparse group LASSO models. By using gene membership as a grouping variable, the sparse group LASSO can be used as a gene-centric analysis of rare variants while also providing a penalized approach toward identifying specific regions of interest. We apply extensive simulations to evaluate the performance of these approaches with respect to specificity and sensitivity, comparing these results to multiple competing marginal testing methods. Finally, we discuss our findings and outline future research.

Keywords

rare variants; lasso; regularization; exome sequencing; association analysis

INTRODUCTION

Genome-wide association studies (GWAS) have been conducted for a large number of phenotypes on the principle that complex diseases are the result of multiple common single nucleotide polymorphisms (SNPs), known as the common disease common variant hypothesis (CD-CV) [Reich and Lander 2001]. While much success has been achieved in identifying a large number of common risk-associated loci for various complex diseases [Hindorff, et al. 2009], the reported SNP associations are generally of low effect sizes and

[§]Corresponding author Nicholas B. Larson, Ph.D. Mayo Clinic, 200 First Street SW, Rochester, MN 55905, Larson.nicholas@mayo.edu, (507) – 293 – 1700 (phone), (507) – 284 – 1516 (fax).

The authors declare no conflict of interest.

explain a modest proportion of the total estimated heritability [Manolio, et al. 2009]. This has prompted investigators to extend association analysis beyond the scope of CD-CV and consider other sources of genetic variation, notably with respect to rare variants [Pritchard 2001]. While traditional genotyping microarray chips are generally not designed to capture rare variation, recent advancements in next-generation sequencing technologies, such as whole-exome sequencing (WES), have made large-scale interrogation of rare variation logistically plausible. These assays have resulted in an invigorated evaluation of rare variant analysis for complex traits, with multiple associations identified [Holm, et al. 2011; Nejentsev, et al. 2009; Rivas, et al. 2011].

Unlike common polymorphisms, traditional single marker analyses for rare variants are often underpowered for typical sequencing study sample sizes due to low minor allele frequencies (MAFs) [Li and Leal 2008] and specialized statistical methods are required for association analysis. It has been noted that multiple causal variants may aggregate in biologically relevant functional domains, such as genes [Bansal, et al. 2010]. This has led to multi-marker analysis strategies that evaluate the aggregate effect of multiple variants within a given region of interest (ROI), rather than at the individual variant level. These modes of analysis amount to identifying regions of rare variant enrichment, indicating incidence disparity of rare variation in the ROI across affected status. There have been recent developments of multiple statistical methods to detect such clustering of rare variation. Burden testing, such as the Madsen-Browning weighted sum statistic (WSS) [Madsen and Browning 2009], collapses all of the variants in the ROI into a single variable of interest,. Collapsing methods will suffer if there is a presence of both risk and protective variants in the ROI, since the effects will be canceled out in the collapsing measure. To address this, variance component tests such as C-alpha test [Neale, et al. 2011] and the sequence kernel association test (SKAT) [Kwee, et al. 2008; Wu, et al. 2011], a generalized form of the C-alpha, were developed to evaluate the association of a genomic ROI with a trait. Other methods include scan-based clustering approaches, which use a sliding window to localize variant clustering over a much larger genomic segment [Fier, et al. 2012; Ionita-Laza, et al. 2012; Schaid, et al. 2013].

Given that disease risk may be the result of contributions from many genes, there is interest in jointly modeling the effects of rare variants across multiple genes simultaneously. Such a model may result in more powerful detection of associations in contrast to marginal ROI analyses [Ayers and Cordell 2013; Clayton 2012]. However, regression models attempting to simultaneously assess all potential rare variant risk associations will be grossly underdetermined for any typical sequencing study sample size, requiring some form of model selection. Exhaustive combinatorial searches over the model space in conjunction with a model selection criterion would be computationally infeasible for most applications. An alternative strategy is to apply regularized regression methods, such as the least absolute shrinkage and selection operator (LASSO) [Tibshirani 1996], to simultaneously perform model selection and parameter estimation. Regularized regression approaches are not uncommon in genomic data analysis due to the vast feature sets, with applications proposed for microarray [Gui and Li 2005], GWAS [Liu, et al. 2011], and methylation data [Sun and Wang 2012]. Basu et al. [Basu, et al. 2011] explored global testing using penalized regression models and permutation for common variation. Recently, a number of methods

have been proposed [Ayers and Cordell 2013; Zhou, et al. 2011; Zhou, et al. 2010] using group penalties on individual rare (and common) variation based upon prior biological information (i.e., gene structure) for large-scale applications on multiple genes.

Due to rapid population expansion and weak purifying selection, rare variation is actually quite common with respect to the number of identified polymorphic sites in a given study. Recent work by Nelson et al. [Nelson, et al. 2012] revealed that variant sites occurred every 21 bp on average across 12,514 subjects of European ancestry, with a vast majority observed in one or two subjects. For a ~30 Mb exome capture, this variant density would yield over 1 million variants for simultaneous consideration, which could be computationally taxing for complex penalties. Alternatively, gene structures can be used to inform the formulation of collapsing measures, corresponding to weighted sums of minor allele counts over predefined regions such as genes or exons. This approach also defines the maximal model dimensionality, since the quantity of unique ROIs is fixed *a priori*.

Examination of the latest build (hg19) of the consensus coding sequence (CCDS) regions [Pruitt, et al. 2009] reveals the human exome consists of 18,305 protein coding genes corresponding to ~180,000 unique exons in the human genome. While genes tend to be favored for ROI definition as a base biologically functional unit, WES also provides a natural alternative through using multi-marker collapsing measures of variants with respect to individual exons. Collapsing rare variation at the exon-level strikes a compromise between gene-level aggregations and no collapsing altogether, resulting in a fine partitioning of the exome. Examination of the CCDS data reveals that, with rare exception, exon length ≤ 1000 bp and has a median value equal to 126 bp. Figure 1 illustrates the distributions of exon lengths and number of exons per gene across the genome. There is also strong evidence that exons are correlated with functional domains in proteins for complex organisms [Kolkman and Stemmer 2001; Liu and Grigoriev 2004], supported by the “exon shuffling” theory [Gilbert 1978; Gilbert 1985]. Thus, it is possible that rare variants may not only cluster in genes, but that a given complex disease may be the result of specific alteration in protein activity by disrupting functional domains through localized exon-specific rare variation. A notable example of such a phenomenon is exon 11 in *BRCA1*, which harbors multiple variants associated with increased risk of breast and ovarian cancers [Risch, et al. 2001]. Biologically, exon 11 in *BRCA1* encodes two putative nuclear localization signals [Chen, et al. 1996] and contains a domain that interacts with the DNA damage repair gene *RAD51* [Zhang and Powell 2005]. Another example of acute localization was found in the rare variant analysis of *DISC1* exon 11 for schizoaffective disorders [Green, et al. 2011], which identified multiple case-only missense RVs.

Given that exons naturally group to form genes, it may be reasonable to expect the existence of an effect for a given exon to be potentially related to other exons within its respective gene, a hierarchical relationship which can be exploited in regularized modeling procedures. It is also possible that only certain exons within a gene may be enriched for RVs. Use of the sparse group LASSO [Friedman, et al. 2010a] can impose sparsity both across groups and within groups, which accommodates rare variant enrichment localized to specific domains within a gene. This strategy is of particular relevance to sequencing studies, and naturally accommodates WES data analysis.

In this paper, we explore the use of the penalized regression with variant collapsing measures to assess rare variant enrichment for case-control WES studies, which we respectively refer to as gene-based LASSO (GB-L) and exon-based sparse group LASSO (EB-SGL). We analyze their performance under a variety of disease model scenarios via simulation study, characterizing sensitivity and specificity and comparing gene-level performance against existing methods. We conclude with a discussion of the benefits and shortcomings of these approaches, as well as outline future research directions.

MATERIALS AND METHODS

Aggregation Measures

Consider an exome sequencing study dataset involving J genes on $N = N_C + N_D$ subjects, where N_C (N_D) is the number of controls (cases) and each gene consists of K_j exons (j = 1, …, J) for a total of $p=\sum_{j=1}^{J} K_j$ exons. Define a set of genetic positions within a given exon designated to be rare by some minor allele frequency (MAF) threshold criterion (e.g., MAF ≤ 0.05), whereby the vector $\boldsymbol{g}_{ijk} = (g_{ijk1}, g_{ijk2}, \ldots)'$ represents the minor allele count of these variants for subject i, gene j, exon k. This vector can in turn be summarized using a scalar collapsing measure, z_{ijk}, which is a univariate summary of $\boldsymbol{g}_{ijk}$ and can be considered a "super-variant." Although there are many options for z_{ijk} as a function of $\boldsymbol{g}_{ijk}$ [Dering, et al. 2011a; Dering, et al. 2011b], we consider the Madsen and Browning weighted-sum (WS) genetic score [Madsen and Browning 2009], which up-weights positions based upon the empirical rarity in the control cohort. The aggregation measure is then defined such that

$$z_{ijk}=z(\boldsymbol{g}_{ijk})=\frac{1}{2L_{jk}}\sum_{l}^{L_{jk}} w_{jkl}g_{ijkl} \text{ where } w_{jkl}=\frac{1}{\sqrt{Nq_{jkl}(1-q_{jkl})}} \text{ and } q_{jkl}=\frac{\sum g_{ijkl}^{C}+1}{2N_C+2},$$

where L_{jk} is total number of variant positions in $\boldsymbol{g}_{ijk}$ and $\sum g_{ijkl}^{C}$ is the sum of control subject alternative alleles at position l. The discriminative value of z_{ijk} is thus a function of the underlying MAF distribution, sample size, and number of true risk RVs for the given exon. We can similarly define a gene-based measure where exon membership is ignored and collapsing is conducted at the gene level, resulting in $z_{ij}^{*}=z(\boldsymbol{g}_{ij})$, where $\boldsymbol{g}_{ij} = (\boldsymbol{g}_{ij1}, \ldots, \boldsymbol{g}_{ijK_j})'$. Alternative weighting schemes could be used that take advantage of any additional annotation, including functional prediction tools such as SIFT [Kumar, et al. 2009] or Polyphen2 [Adzhubei, et al. 2010].

Penalized Models

Let $\boldsymbol{y}$ be a vector of binary disease status indicators, such that y_i is 1 if the i^{th} subject is a case and 0 if a control, and define $\boldsymbol{X}$ to be an $N \times m$ matrix of additional confounding covariate data, such as age or sex. For EB-SGL, let $\boldsymbol{Z}$ be an $N \times p$ matrix of aggregation measures, such that the i^{th} row of $\boldsymbol{Z}$ is $\boldsymbol{z}_i = (z_{i11}, z_{i12}, \ldots)'$. We similarly define the $N \times J$ matrix $\boldsymbol{Z}^{*}$ with i^{th} row $\boldsymbol{z}_i^{*}=(z_{i1}^{*}, z_{i2}^{*}, \ldots)'$ for GB-L. We use logistic regression on these data in order to detect differences in the collapsing measures across case-control status. Under

this model, we apply a logit link on the probability of affected status for subject i, $\Pr(y_i = 1 | \boldsymbol{x}_i, \boldsymbol{z}_i) = \mu_i$, such that for EB-SGL

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\theta}$$

where $\boldsymbol{x}_i$ is the i^{th} column of $\boldsymbol{X}$, $\boldsymbol{\beta}$ the respective model parameters corresponding to $\boldsymbol{X}$, and $\boldsymbol{\theta}$ the corresponding parameters to $\boldsymbol{Z}$. Model fitting can then be conducted by maximizing the log-likelihood, which can be written as

$$\ln L(\boldsymbol{\Theta}; \boldsymbol{y}) = \ln\left(\prod_{i=1}^{N} \mu_i^{y_i}(1-\mu_i)^{1-y_i}\right) = \sum_{i=1}^{N} y_i\eta_i - \ln(1+\exp(\eta_i))$$

where $\boldsymbol{\Theta} = (\beta_0, \boldsymbol{\beta}, \boldsymbol{\theta})$ is the full set of model parameters. In order to induce sparsity at both the gene and exon level for EB-SGL, we apply the sparse group LASSO penalty, such that the criterion for the penalized model fit is given as

$$\min_{\boldsymbol{\theta} \in \mathscr{R}^p}\left(-\frac{1}{N}\ln L(\boldsymbol{\Theta}; \boldsymbol{y}) + \lambda_1 \sum_{j=1}^{J} \sqrt{K_j}\|\boldsymbol{\theta}_j\|_2 + \lambda_2\|\boldsymbol{\theta}\|_1\right)$$

where $\sqrt{K_j}$ is correction factor accounting for the group size of gene j, $\boldsymbol{\theta_j} = \theta_{j1}, \theta_{j2}, \ldots)'$ is a vector of exon-level parameters for a given gene, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_J')'$ is the complete p-length exon regression coefficient vector. In this formulation β_0 and $\boldsymbol{\beta}$ are unpenalized parameters, but $\boldsymbol{\beta}$ could be also included in the LASSO penalty if desired.

For GB-L, the regression model is very similar to EB-SGL, except that no group penalty is included and the model fitting criterion is given as

$$\min_{\boldsymbol{\theta} \in \mathscr{R}^p}\left(-\frac{1}{N}\ln L(\boldsymbol{\Theta}^*; \boldsymbol{y}) + \lambda\|\boldsymbol{\theta}^*\|_1\right)$$

where $L(\boldsymbol{\Theta}^*, \boldsymbol{y})$ is the model likelihood regressing affected status on $\boldsymbol{Z}^*$ rather than $\boldsymbol{Z}$ and $\boldsymbol{\theta}^*$ is the J-length vector of regression coefficients corresponding to the gene-level collapsing measures.

Model fitting for G-BL was conducted in R using the *glmnet* package [Friedman, et al. 2010b], which employs cyclical coordinate descent algorithms. To implement model fitting for EB-SGL, we utilized a blockwise descent (BD) algorithm put forth by Simon et al. [Simon, et al. 2013] provided in the R package *SGL*. The BD algorithm takes advantage of the group-level separability of the penalty function by cyclically iterating over the covariate

groups, performing generalized gradient steps for each group with at least one non-zero regression coefficient.

Tuning Parameter Selection

Specific to EB-SGL, previous work by Zhou et al. [Zhou, et al. 2010] applying mixtures of group and traditional LASSO penalties for association analysis indicated that framing the relationship $\frac{\lambda_2}{\lambda_1+\lambda_2}=c$ is an effective strategy for identifying disease associations. This also reduces the path space for tuning parameter selection by fixing the value of c *a priori*, such that optimization corresponds to identifying a proper value of $\lambda = \lambda_1 + \lambda_2$ where $\lambda_1 = (1 - c)\lambda$ and $\lambda_2 = c\lambda$. Based upon empirical evidence under various simulation conditions, we conducted our analyses with c fixed at 0.80.

Many strategies exist in general for selecting an appropriate value of λ in penalized regression. One option is to define a fine grid of values for λ and apply cross-validation, although this may be a very computationally demanding process. If we can assume a very sparse set of true associations, an efficient alternative approach is to employ a bracketing and bisection algorithm [Wu, et al. 2009] to identify a set of values for λ that result in small additions to the active set of covariates. Let $\boldsymbol{\theta_\lambda}$ indicate the vector of non-zero coefficients under a model fit for a given value of λ, such that these coefficients correspond to the *active set* of $n_\lambda = \|\boldsymbol{\theta_\lambda}\|_0$ covariates. If we *a priori* fix the maximum possible size of n_λ to be very small n_θ, we can iteratively fit penalized models by reducing λ until n_λ exceeds n_θ. This is a particularly appealing strategy in exploratory analyses when a small set of ROIs is desired for further investigation. This approach can be combined with a model selection criterion, such as the Bayesian Information Criterion (BIC) [Schwarz 1978], which is defined as

$$BIC(\boldsymbol{\Theta};\lambda)=-2\ln L_\lambda(\boldsymbol{\Theta};\boldsymbol{y})+d\cdot\ln N$$

where d is the model degrees of freedom and $L_\lambda(\boldsymbol{\Theta}, \mathbf{y})$ is the model likelihood specific to the value of λ. The minimization of $BIC(\boldsymbol{\Theta}; \lambda)$ can then be used to select a value of λ over the path of values in order to yield a final fit such that the size of $\boldsymbol{\theta_\lambda}$ is equal to $n_\lambda \leq n_\theta$. The calculation of d requires that the model selection procedure be taken into consideration. For the LASSO procedure, this is equivalent to the size of the active set [Zou, et al. 2007]. An approximation for EB-SGL can be shown (see Appendix) to reduce to the trace of the smoother hat matrix of the model fit, which can be written as

$$d=tr\left(\boldsymbol{W}^{\frac{1}{2}}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\boldsymbol{W}\tilde{\boldsymbol{X}}+\boldsymbol{\Omega})^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{W}^{\frac{1}{2}}\right)$$

where $\boldsymbol{W} = diag(\mu_i(1 - \mu_i))$, $\tilde{\boldsymbol{X}} = [\mathbf{1}\ \boldsymbol{X}\ \boldsymbol{Z_\lambda}]$, $\boldsymbol{Z_\lambda}$ is a subset of $\boldsymbol{Z}$ corresponding to the nonzero subset of $\boldsymbol{\theta}$, $\boldsymbol{\theta_\lambda}$, for a given value of λ, and $\boldsymbol{\Omega} = diag(\boldsymbol{\omega})$, where the first $m + 1$ elements of $\boldsymbol{\omega}$ are 0 since they correspond to unpenalized parameters (β_0, $\boldsymbol{\beta}$), and the remaining n_λ are penalty weights corresponding to $\boldsymbol{\theta_\lambda}$ and equal to $\frac{c\lambda}{|\theta_{jk}|}+\frac{(1-c)\lambda\sqrt{K_j}}{\sqrt{\sum_j\theta_{jk}^2}}$.

Model criteria based in part upon prediction performance may result in a large number of false positives, which stands in contrast to the prevailing objective of identifying a confident subset of associated genes for follow-up analyses. If we are instead interested in gene-level control of the false positive rate, we may employ the permutation-based method described in Ayers et al. [Ayers and Cordell 2010], which selects λ by repeatedly permuting case-control status and identifying values that result in the first covariate (i.e., false positive) entering the active set. For EB-SGL, translating this procedure to gene-level association is complicated by defining what constitutes a false positive result. We adopt a minor modification that is based upon admittance of covariates at the group level, such that we record smallest values of λ that allow exon-level measures into the active set from at most one gene. We focus on this mode of tuning parameter selection for the remainder of the paper, as we foresee it being of greatest utility for large-scale sequencing studies (e.g., whole exome).

Data Simulation

To base our simulations on realistic exome sequence data characteristics, we downloaded the empirical distribution of rare (MAF≤0.05) variants from the NHLBI-ESP Exome Variant Server (release ESP6500). We isolated the European American minor allele frequencies for all observed polymorphic sites on chromosome 1, which we then annotated with exon membership information from the CCDS data (including appended 10bp flanking regions for potential splice site variation). Finally, we excluded genes exhibiting minimal variation, such that for all included genes the sum of MAFs across all variants in a given gene ≥ 0.005 and the total number of variants ≥ 10. Our final variant list consisted of a total of 28,552 unique sites on 6000 exons in 597 genes, with median values of 3 variants per exon and 37 variants per gene. The corresponding MAFs ranged from 0.00012 to 0.05, with a mean value of 0.00121. We generated genotype data for 30,000 simulated individuals, assuming Hardy-Weinberg equilibrium proportions and independence between variant sites.

For our disease models, we selected five unique genes to harbor causal variants, concentrating on moderately sized genes with a variety of exon distributions and total variant counts (see Table I). While some of these genes have a relatively large number of variants, the vast majority of these variants are very rare and unlikely to be observed on an individual basis in any given simulation replicate. To simulate phenotypes based upon genotypes, we applied a logistic regression approach whereby the Bernoulli probability of affected status conditional on genotype is given as

$$\mu_i = \frac{\exp(\beta_0 + \sum_j \ln(OR_j) g_{ij})}{1 + \exp(\beta_0 + \sum_j \ln(OR_j) g_{ij})}$$

where OR_j is the odds ratio associated with variant j, g_{ij} the corresponding minor allele count for subject i, and β_0 is an intercept adjusted to result in a population prevalence of approximately 5%.

Since the success of specific rare variant association methods depends largely upon the underlying assumptions of the disease model, we examined a variety of potential scenarios

in our simulation study. We considered models where approximately 20% or 40% of the total number of variants are deleterious, such that the corresponding $OR_j > 1$. To define each OR_j, we considered two different approaches: a uniform, or fixed, effect model where $OR_j = 3.0$ for all causal variants regardless of population allele frequency, and an MAF-dependent scheme where $\ln(OR_j) = \frac{\ln(10)}{4}|\log_{10} MAF_j|$, akin to settings in Wu et al. [Wu, et al. 2011]. Deleterious variant selection was also conducted using two different approaches. We first considered causal variants to be randomly sampled uniformly from the total set of variants, regardless of exon membership. The second method selected causal variants from a subset of variants defined by falling within designated causal exons. These exons were selected to capture approximately 50% of the total number of variants, enforcing localization of causal variation to a fraction of the total exons for each gene. Table II details the causal variants for each gene by derived MAF across these scenarios. A total of 1000 cases and 1000 controls were sampled from the population of 30,000 subjects for each simulation replicate, with 500 replicates conducted per unique set of simulation conditions. In total, we evaluated eight distinct simulation scenarios to examine performance under a variety of underlying causal variant models.

Comparison with Other Methods

We have framed the application of EB-SGL and GB-L with respect to hypothesis testing via variable selection, with non-zero coefficients corresponding to significantly associated genes. While no other explicit exon-based methods exist to our knowledge, there is interest in how EB-SGL compares to current gene-based testing approaches that are agnostic to the additional exon structure information. We denote gene-level significance in the context of variable selection in EB-SGL such that gene *j* is said to be associated with the response if $||\boldsymbol{\theta}_j||_2 > 0$ for a given value of λ.

We compared gene-level power of EB-SGL and GB-L to a number of leading omnibus tests: C-alpha; the sequence kernel association test SKAT; SKAT-O [Lee, et al. 2012], an optimized combination of SKAT and burden-based testing; the Madsen-Browning WSS test; and the Cohort Allelic Sums Test (CAST) [Morgenthaler and Thilly 2007]. We applied SKAT and SKAT-O under the linear weighted kernel using the *SKAT* package in R, p-values from the WSS testing were derived with the approximation method under 100 permutations, and the C-alpha and CAST tests were performed using code from the *AssotesteR* package reporting asymptotic p-values. All remaining settings were set to the default values for the respective tests, with statistical significance for each test determined based upon p-values using a Bonferroni-corrected Type I error level of $\alpha = 0.05$ (i.e, $0.05/597 \approx 8.38\text{e-}05$).

To achieve a comparably sized testing for GB-L and EB-SGL at the gene-level, we applied the previously described permutation-based approach for *a priori* selecting the tuning parameter λ. In the interest of computational cost for our simulation study, we approximated this procedure by assuming that the same values of λ were applicable across all simulation replicates for a given scenario, since the sampled population and causal variation were identical across replicates. We conducted 500 initial simulation replicates for each scenario, permuting case-control status once prior to computing $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$, and ran EB-SGL and GB-L

on the respective permuted datasets. We selected the λ value that admitted a total of 25 false positive genes across all replicates for each unique simulation scenario, resulting in a comparable Type I error rate (25/(500*597) = 0.05/597).

Finally, for both EB-SGL and GB-L, we noted in our preliminary simulation studies that the presence of singletons often led to poor statistical power. Although the mechanism behind this issue has not yet been fully resolved, we nonetheless present simulation results for both approaches with singletons either included or filtered prior to analysis. Since previous work [Ladouceur, et al. 2013] has illustrated that removal of singletons in general leads to decreases in power in the competing gene-level tests, these methods were applied solely to the unfiltered data.

RESULTS

We first examined the Type I error rate control across the various testing procedures, presented in Table III. Surprisingly, these varied greatly across methods, but were relatively stable within methods across simulation scenarios. The C-alpha testing results yielded highly elevated Type I error rates, generally two orders of magnitude larger than the nominal level, indicating possible issues with asymptotic assumptions. SKAT, SKAT-O, and CAST generally exhibited near nominal Type I error control and tended toward being conservative, while GB-L and WSS results were slightly liberal and comparable in value. EB-SGL Type I error rates were also elevated, although moreso than GB-L and WSS and often close to an order of magnitude increase relative to the nominal level. With respect to both of the penalized approaches, the exclusion of singletons led to differing impacts on Type I error rates, with EB-SGL exhibiting larger scale effects than GB-L, although these impacts were relatively modest overall.

Simulation results with respect to overall power (Table IV) indicated that EB-SGL and GB-L performed very poorly when the variant-level effect sizes were not dependent upon the underlying MAF, especially relative to the competing gene-level methods. This is somewhat unsurprising, since the weighting scheme explicitly places more weight on rarer variation. A similar pattern was observed with respect to the WSS test, which uses the same weighting scheme, although the overall power for this approach was much higher relative to GB-L and EB-SGL. Statistical power was also in general higher for each given method when the true causal proportion was higher and when the effect sizes were MAF-dependent rather than uniform across variants. SKAT-O and WSS tended to perform the best across all scenarios, taking into account proper control of Type I errors.

Focusing on gene-specific power under the MAF-dependent effect models (Figure 2), illustrates that the performance of the penalized methods was highly dependent upon the underlying causal variation. Overall, GB-L exhibited higher power over EB-SGL in every scenario except for the results for gene 2 under the exon-based variant selection model. Interestingly, the gene-wise power for the penalized approaches were much more powerful than the competing tests in isolated instances, while in others the opposite was observed. For example, from Figure 2 we observe that GB-L performed well for genes 1, 3, and 4 across all scenarios, with the version excluding singletons often exhibiting comparable, or better,

power relative to the competing approaches. However, these competing approaches grossly outperform GB-L for gene 2. Examination of the results relative to the underlying causal variation for each scenario indicates that the penalized approaches tend to perform better when both the causal and neutral variant MAFs are relatively low (<0.005).

DISCUSSION

In this paper, we have extended existing work for penalized regression on sequencing data by combining LASSO and sparse group LASSO logistic regression with variant collapsing to identify rare variant enrichment in case-control exome sequencing studies. While collapsing methods for rare variant analysis are commonplace for omnibus testing, our approaches explore simultaneous assessment of multiple collapsing measures in a regression framework, which can also accommodate adjustment for confounding variables. We considered a gene-based method using a simple LASSO logistic regression model, GB-L, where rare variants within individual genes are summarized by a univariate weighted sum for each subject. Motivated by biological principles of the coding content of exons within genes, we also examined an exon-based sparse group LASSO approach, EB-SGL, which may provide more acute localization of rare variant associations in contrast to gene-level testing by operating on an ROI definition with finer granularity.

Overall, our simulation results indicated that the power of EB-SGL to be quite poor unless causal variation is highly localized with multiple causal variants. This may in part be attributable to the exon structure of the genes we modeled, which we observed to be often composed of multiple very small exons. Given that much of the observed variation in the ESP6500 data is very rare or private, this resulted in most exons in our simulations harboring ≤ 1 variant, rendering any collapsing measure over exons superfluous. In this context, these results agreed with Ayer et al. [Ayers and Cordell 2013], who noted similarly poor performance for variant-level sparse group LASSO. As gene-wise functional domain annotation for coding proteins improves, however, it may be worthwhile to revisit this type of sub-gene aggregation.

The simulation performance of the GB-L penalized regression approach was much more encouraging, and was the highest powered method in two of our eight simulation scenarios. Although the Type I error rate was slightly elevated relative to the nominal level, this may be potentially due to the approximation approach we employed to select the tuning parameter λ. Of greater interest was the wide variety of gene-specific performance observed across simulation scenarios, as seen in Figure 2, with some instances where the power for GB-L was dramatically greater than the competing testing procedures. Examination of these cases indicated this disparity to be dependent upon the MAF distribution of the causal variation, as we observed GB-L to be relatively more sensitive when the causal variants were very rare. It was also interesting to note the sizable increases in power for both GB-L and EB-SGL by filtering out singletons in the model fitting procedure while maintaining comparable Type I error rates. This may be due in part to issues of quasi-complete separability in the regression model for genes with minimal variation, and further evaluation of filtering mechanisms may be necessary prior to model fitting.

From a computational perspective, the source of the computational burden for both EB-SGL and GB-L largely stems from running the permutation fits necessary to select the value of λ, although this is an easily parallelizable process. Final model fits completed in < 2 minutes for either approach on a modern Linux workstation equipped with a Quad-Core AMD Opteron™ Processor and 16 Gb of RAM. Methods based upon iterative model fits, particular bracketing and bisection algorithms, require considerably more computational time. This is particularly true for EB-SGL, where the burden of such model fits is heavily dependent upon the maximum desired model size, n_θ.

One caveat of variant collapsing approaches is that they implicitly presume that the direction of the effects in the ROI is uniform (i.e., all protective or all risk), an issue left unexplored in this paper. Reductions in statistical power can arise if a mixture of both types occurs in a given ROI, resulting in a "canceling" of effects. This is clearly a pitfall for GB-L, and an issue shared by other collapsing procedures. EB-SGL does not completely avoid this problem, but it does allow for the effects of aggregation measures at specific exons to be either risk or protective oriented. A possible remedy would be to apply the sign assignment procedure used by Hoffman et al. [Hoffmann, et al. 2010], where the weighted sum collapsing measure takes into account whether a given variant is observed more often in cases or controls. Evaluation of this approach when both protective and risk variants are observed in a given gene is the focus of future work.

While we have presented these approaches with respect to detecting rare variant clustering in case-control studies, they can be extended to other response types, such as quantitative traits or censored survival responses under appropriate variant weighting schemes. We can also extend GB-L to take advantage of available gene set annotation by reintroducing group-level penalties, which can accommodate overlapping groups in the penalty function. This circumvents the common problem of pathway-based analyses which result in dependent pathway-specific test statistics if genes are shared across multiple pathways. More work is necessary to further evaluate and build upon these penalized regression variant collapsing methods.

Acknowledgments

This research was supported by the US Public Health Service, National Institutes of Health (NIH), contract Grant Number GM065450 (DJS, JPS).

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature methods. 2010; 7(4): 248–9. [PubMed: 20354512]

Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic epidemiology. 2010; 34(8):879–91. [PubMed: 21104890]

Ayers KL, Cordell HJ. Identification of Grouped Rare and Common Variants via Penalized Logistic Regression. Genetic epidemiology. 2013

Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nature reviews Genetics. 2010; 11(11):773–85.

Basu S, Pan W, Shen XT, Oetting WS. Multilocus association testing with penalized regression. Genetic epidemiology. 2011; 35(8):755–765. [PubMed: 21922539]

Chen Y, Farmer AA, Chen CF, Jones DC, Chen PL, Lee WH. BRCA1 is a 220-kDa nuclear phosphoprotein that is expressed and phosphorylated in a cell cycle-dependent manner. Cancer research. 1996; 56(14):3168–72. [PubMed: 8764100]

Clayton D. Link Functions in Multi-Locus Genetic Models: Implications for Testing, Prediction, and Interpretation. Genetic epidemiology. 2012; 36(4):409–418. [PubMed: 22508388]

Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genetic epidemiology. 2011a; 35:S12–S17. [PubMed: 22128052]

Dering C, Ziegler A, Konig IR, Hemmelmann C. Comparison of collapsing methods for the statistical analysis of rare variants. BMC proceedings. 2011b; 5(Suppl 9):S115. [PubMed: 22373249]

Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, Silverman EK, Pagano M, Mangold E, Lange C. 'Location, Location, Location': a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. Bioinformatics. 2012; 28(23):3027–33. [PubMed: 23044548]

Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. 2010a arXiv preprint arXiv:1001.0736.

Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010b; 33(1):1–22. [PubMed: 20808728]

Gilbert W. Why genes in pieces? Nature. 1978; 271(5645):501. [PubMed: 622185]

Gilbert W. Genes-in-pieces revisited. Science. 1985; 228(4701):823–4. [PubMed: 4001923]

Green EK, Grozeva D, Sims R, Raybould R, Forty L, Gordon-Smith K, Russell E, St Clair D, Young AH, Ferrier IN, et al. DISC1 Exon 11 Rare Variants Found More Commonly in Schizoaffective Spectrum Cases Than Controls. American Journal of Medical Genetics Part B-Neuropsychiatric Genetics. 2011; 156B(4):490–492.

Gui J, Li H. Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics. 2005; 21(13):3001–8. [PubMed: 15814556]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(23):9362–7. [PubMed: 19474294]

Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PloS one. 2010; 5(11):e13584. [PubMed: 21072163]

Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, Magnusson OT, Helgason A, Saemundsdottir J, Gylfason A, et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nature genetics. 2011; 43(4):316–20. [PubMed: 21378987]

Ionita-Laza I, Makarov V, Buxbaum JD. Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. American journal of human genetics. 2012; 90(6):1002–13. [PubMed: 22578327]

Kolkman JA, Stemmer WP. Directed evolution of proteins by exon shuffling. Nature biotechnology. 2001; 19(5):423–8.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009; 4(7):1073–1082.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. American journal of human genetics. 2008; 82(2):386–97. [PubMed: 18252219]

Ladouceur M, Zheng HF, Greenwood CMT, Richards JB. Empirical power of very rare variants for common traits and disease: results from sanger sequencing 1998 individuals. European Journal of Human Genetics. 2013; 21(9):1027–1030. [PubMed: 23321613]

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American journal of human genetics. 2012; 91(2):224–37. [PubMed: 22863193]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American journal of human genetics. 2008; 83(3):311–21. [PubMed: 18691683]

Liu J, Wang K, Ma S, Huang J. Regularized regression method for genome-wide association studies. BMC proceedings. 2011; 5(Suppl 9):S67. [PubMed: 22373491]

Liu M, Grigoriev A. Protein domains correlate strongly with exons in multiple eukaryotic genomes--evidence of exon shuffling? Trends in genetics: TIG. 2004; 20(9):399–403. [PubMed: 15313546]

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. Plos Genetics. 2009; 5(2):e1000384. [PubMed: 19214210]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53. [PubMed: 19812666]

Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutation research. 2007; 615(1–2):28–56. [PubMed: 17101154]

Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. Plos Genetics. 2011; 7(3):e1001322. [PubMed: 21408211]

Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009; 324(5925):387–9. [PubMed: 19264985]

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012; 337(6090):100–4. [PubMed: 22604722]

Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? American journal of human genetics. 2001; 69(1):124–37. [PubMed: 11404818]

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome research. 2009; 19(7):1316–23. [PubMed: 19498102]

Reich DE, Lander ES. On the allelic spectrum of human disease. Trends in genetics: TIG. 2001; 17(9): 502–10. [PubMed: 11525833]

Risch HA, McLaughlin JR, Cole DE, Rosen B, Bradley L, Kwan E, Jack E, Vesprini DJ, Kuperstein G, Abrahamson JL, et al. Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of 649 women with ovarian cancer. American journal of human genetics. 2001; 68(3):700–10. [PubMed: 11179017]

Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics. 2011; 43(11):1066–73. [PubMed: 21983784]

Schaid DJ, Sinnwell JP, McDonnell SK, Thibodeau SN. Detecting genomic clustering of risk variants from sequence data: cases versus controls. Human genetics. 2013

Schwarz G. Estimating Dimension of a Model. Annals of Statistics. 1978; 6(2):461–464.

Simon N, Friedman J, Hastie T, Tibshirani R. A Sparse-Group Lasso. Journal of Computational and Graphical Statistics. 2013; 22(2):231–245.

Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics. 2012; 28(10):1368–75. [PubMed: 22467913]

Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B-Methodological. 1996; 58(1):267–288.

Wu MC, Lee S, Cai TX, Li Y, Boehnke M, Lin XH. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. American journal of human genetics. 2011; 89(1):82–93. [PubMed: 21737059]

Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009; 25(6):714–21. [PubMed: 19176549]

Zhang J, Powell SN. The role of the BRCA1 tumor suppressor in DNA double-strand break repair. Molecular cancer research: MCR. 2005; 3(10):531–9. [PubMed: 16254187]

Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel EM, Lange K. Penalized regression for genome-wide association screening of sequence data. Pacific Symposium on Biocomputing. 2011:106–17. [PubMed: 21121038]

Zhou H, Sehl ME, Sinsheimer JS, Lange K. Association screening of common and rare genetic variants by penalized regression. Bioinformatics. 2010; 26(19):2375–82. [PubMed: 20693321]

Zou H, Hastie T, Tibshirani R. On the “degrees of freedom” of the lasso. Annals of Statistics. 2007; 35(5):2173–2192.

APPENDIX

Approximating the degrees of freedom for the EB-SGL model fit

An approximation for the degrees of freedom *d* for a LASSO linear regression model fit was provided by Tibshirani [Tibshirani 1996], noting that the LASSO penalty can be written as a ridge regression problem by redefining the penalty term as $\sum|\beta_i|=\sum\frac{\beta_i^2}{|\beta_i|}$. This results in the LASSO estimator being defined as ridge estimator, such that

$$\tilde{\boldsymbol{\beta}}=(\boldsymbol{X}'X+\boldsymbol{\Omega})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{\Omega}=diag\left(\frac{\lambda}{|\beta_i|}\right)$. The hat matrix is then given as $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Omega})^{-1}\boldsymbol{X}'$, and Tibshirani approximates the effective number of parameters *d* by *tr*(***H***).

Consider a traditional logistic regression model

$$logit(\boldsymbol{\mu})=\boldsymbol{\eta}=\boldsymbol{X}\boldsymbol{\beta}$$

where ***X*** is some design matrix, $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients, and $\boldsymbol{\mu}$ the vector of success probabilities associated with binary response vector ***y***. Recall that the iteratively reweighted least squares (IRWLS) estimator of regression coefficients $\boldsymbol{\beta}$ in traditional logistic regression is given as

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}^*$$

where $y_i^*=\eta_i+(y_i-\mu_i)/(\mu_i(1-\mu_i))$ is the working response and $\boldsymbol{W} = diag(\mu_i(1 - \mu_i))$. The hat matrix for $\boldsymbol{y}^*$ is then given as

$$\boldsymbol{S}=\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{\frac{1}{2}}$$

Now, consider the penalized logistic regression model posed in our EB-SGL approach, which can be written similarly as

$$logit(\boldsymbol{\mu}) = \mathbf{1}\beta_0 + \boldsymbol{X\beta} + \boldsymbol{Z\theta}$$

with a model fit criterion is given by

$$\min_{\alpha \in R^p} \left(-\frac{1}{N}\log(L(\boldsymbol{\Theta};\boldsymbol{y})) + \lambda_1 \sum_{j=1}^{J} \sqrt{K_j} \|\boldsymbol{\theta}_j\|_2 + \lambda_2 \|\boldsymbol{\theta}\|_1 \right)$$

First, note that the lasso penalty term can be rewritten in the same fashion proposed by Tibshirani, such that

$$\lambda_2 \|\boldsymbol{\theta}\|_1 = \lambda_2 \sum_{j=1}^{J} \sum_{k=1}^{K_j} \frac{\theta_{jk}^2}{|\theta_{jk}|}$$

Similarly, the group penalty term can be rewritten such that

$$\lambda_1 \sum_{j=1}^{J} \sqrt{K_j} \|\boldsymbol{\theta}_j\|_2 = \lambda_1 \sum_{j=1}^{J} \sqrt{K_j} \sqrt{\sum_{k=1}^{K_j} \theta_{jk}^2} = \lambda_1 \sum_{j=1}^{J} \sqrt{K_j} \frac{\sum_k \theta_{jk}^2}{\sqrt{\sum_k \theta_{jk}^2}} = \lambda_1 \sum_{j=1}^{J} \sum_{k=1}^{K_j} \alpha_{jk}^2 \frac{\sqrt{K_j}}{\sqrt{\sum_i \theta_{jk}^2}}$$

Thus, the full penalty can be rewritten as

$$\lambda_1 \sum_{j=1}^{J} \sum_{k=1}^{K_j} \theta_{jk}^2 \frac{\sqrt{K_j}}{\sqrt{\sum_i \theta_{jk}^2}} + \lambda_2 \sum_{j=1}^{J} \sum_{k=1}^{K_j} \frac{\theta_{jk}^2}{|\theta_{ij}|} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} \theta_{jk}^2 \left(\frac{\lambda_2}{|\theta_{jk}|} + \frac{\lambda_1 \sqrt{K_j}}{\sqrt{\sum_i \theta_{jk}^2}} \right)$$

Notice that the penalty function is now in the form of a ridge penalty with weights defined by the nature of the sparse group penalty. Thus, we can frame our calculation of the degrees of freedom of the EB-SGL model fit in the context of a ridge-type penalized model fit, where the diagonal penalty matrix $\boldsymbol{\Omega} = diag(\boldsymbol{\omega})$ for a penalty weight vector ω consisting of $m + n_\lambda + 1$ elements. The first $m + 1$ elements of ω are 0, corresponding to the unpenalized parameters β_0 and $\boldsymbol{\beta}$, while the remaining n_λ parameters correspond to $\boldsymbol{\theta_\lambda}$ and are defined as $\frac{\lambda_2}{|\theta_{jk}|} + \frac{\lambda_1 \sqrt{K_j}}{\sqrt{\sum_i \theta_{jk}^2}}$. If we assume that $\lambda = \lambda_1 + \lambda_2$ and $c = \frac{\lambda_2}{\lambda}$ for some fixed value of c, then this reduces to $\frac{c\lambda}{|\theta_{jk}|} + \frac{(1-c)\lambda \sqrt{K_j}}{\sqrt{\sum_i \theta_{jk}^2}}$. Approximating the degrees of freedom then amounts to

$$d = tr\left(\boldsymbol{W}^{\frac{1}{2}}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}'\boldsymbol{W}\tilde{\boldsymbol{X}}+\Omega)^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{W}^{\frac{1}{2}}\right)$$

such that $\tilde{\boldsymbol{X}} = [\mathbf{1}\ \boldsymbol{X}\ \boldsymbol{Z}_{\boldsymbol{\lambda}}]$, where $\boldsymbol{Z}_{\boldsymbol{\lambda}}$ is a submatrix of $\boldsymbol{Z}$ with columns corresponding to the nonzero subset of $\boldsymbol{\theta}$, $\boldsymbol{\theta}_{\boldsymbol{\lambda}}$, for a given value of λ.

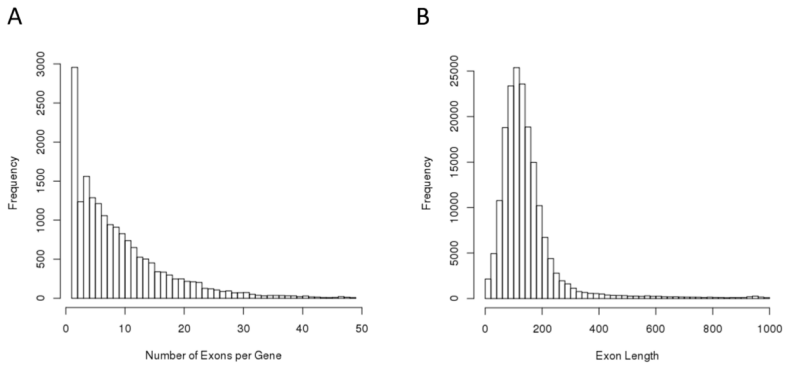

Figure 1.

Histograms depicting distributions of (A) number of exons per gene and (B) individual exon length for ~180,000 exons constituting 18,305 genes in CCDS data for the hg19 human genome build. The right tails of each figure are censored, excluding approximately with 0.8% and 1.5% of the distributions, respectively.

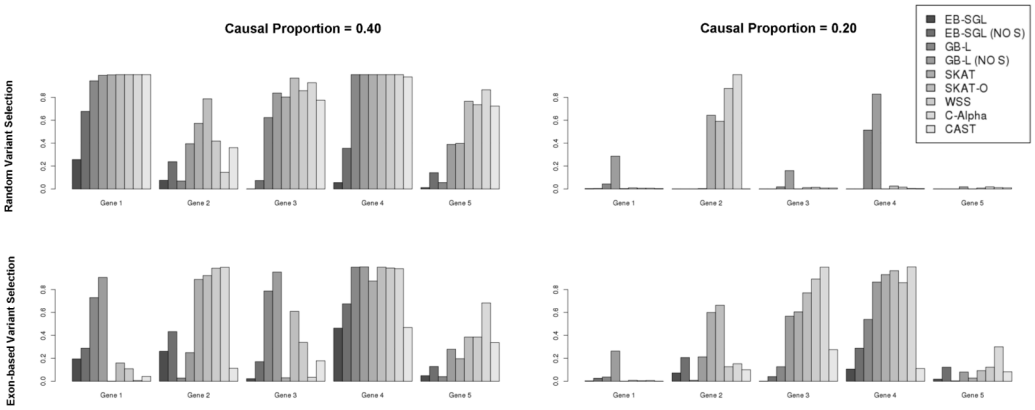

Figure 2.

Barplots indicating statistical power for detecting individual disease genes harboring causal variants OR_j across various methods and simulation conditions where the are defined using the MAF-dependent scheme. Analysis results using EB-SGL and GB-L where singletons have been removed are indicated by "NO S" in the legend.

Table I

Description of disease-related genes used in the simulation study, characterizing gene size, quantity of exons and variants, as well as MAFs of variants.

Gene	Length (bp)	Total Exons	# Var.	Mean MAF	Dist. of All Variants by MAF		
					0.05–0.01	0.01–0.001	<0.001
1	2083	11	87	0.0012	3	6	78
2	1062	6	37	0.0019	1	7	29
3	3754	17	82	0.0009	2	3	77
4	1592	7	43	0.0004	0	3	40
5	3052	17	114	0.0017	4	10	100

Table II

MAF distribution of causal rare variants by four unique variant sets.

		Distribution of Causal Variants by MAF [0.050 – 0.010] [0.010 – 0.001] [<0.001]														
Causal Variant %	**Selection Method**	**Gene 1**			**Gene 2**			**Gene 3**			**Gene 4**			**Gene 5**		
40	Random	2	2	31	0	3	12	1	0	32	0	1	16	2	4	40
	Exon	0	0	34	1	1	12	0	2	30	0	1	16	1	3	41
20	Random	0	0	17	1	0	6	0	0	16	0	0	9	0	0	23
	Exon	0	1	16	0	3	4	1	0	15	0	1	7	1	2	19

Table III

Empirical false positive rates across simulation replicates for gene-wise testing at Bonferroni corrected $\alpha = 0.05$ (0.05/597 ≈ 8.38e-05) for all simulation scenarios and methods.

Scenario			Methods								
			EB-SGL	EB-SGL (no sing)	GB-L	GB-L (no sing)	SKAT	SKAT-O	WSS	C-alpha	CAST
Causal Variant %	OR model	Causal variants									
40	Fixed	Random	5.92e-04	6.40e-04	3.00e-04	3.44e-04	5.85e-05	6.22e-05	1.06e-04	5.69e-03	4.02e-05
		Exon	6.29e-04	7.64e-04	3.07e-04	2.89e-04	2.92e-05	6.94e-05	1.39e-04	5.78e-03	5.11e-05
	Var.	Random	4.79e-04	7.17e-04	2.63e-04	1.83e-04	2.55e-05	4.75e-05	1.06e-04	5.48e-03	2.92e-05
		Exon	7.50e-04	5.78e-04	1.79e-04	2.05e-04	6.21e-05	6.94e-05	1.57e-04	5.70e-03	6.58e-05
20	Fixed	Random	6.69e-04	5.78e-04	2.82e-04	2.27e-04	6.94e-05	7.31e-05	2.05e-04	5.86e-03	8.41e-05
		Exon	5.81e-04	5.78e-04	3.55e-04	2.52e-04	6.94e-05	7.67e-05	1.54e-04	5.94e-03	4.75e-05
	Var.	Random	7.09e-04	5.74e-04	3.25e-04	3.22e-04	4.39e-05	8.04e-05	1.90e-04	6.00e-03	5.12e-05
		Exon	8.00e-04	6.00e-04	4.13e-04	3.95e-04	5.85e-05	9.51e-05	1.83e-04	5.98e-03	6.94e-05

Table IV

Empirical power across simulation replicates for gene-wise testing at Bonferroni corrected $\alpha = 0.05$ (0.05/597 ≈ 8.38e-05) for all simulation scenarios and methods. The methods with the highest power per scenario are highlighted in bold (C-alpha was ignored due to its high Type I error rate).

Simulation Scenario			Methods								
			EB-SGL	EB-SGL (no sing)	GB-L	GB-L (no sing)	SKAT	SKAT-O	WSS	C-alpha	CAST
Causal Variant %	OR model	Causal variants									
40	Fixed	Random	0.002	0.088	0.102	0.231	0.635	**0.661**	0.533	0.744	0.394
		Exon	0.012	0.081	0.065	0.135	0.349	0.369	**0.390**	0.440	0.008
	MAF	Random	0.080	0.297	0.538	0.722	0.754	**0.903**	0.802	0.787	0.767
		Exon	0.198	0.339	0.516	**0.676**	0.398	0.614	0.561	0.540	0.228
20	Fixed	Random	0.001	0.007	0.000	0.004	0.184	0.177	**0.200**	0.205	0.000
		Exon	0.001	0.023	0.006	0.029	**0.383**	0.378	0.370	0.416	0.004
	MAF	Random	0.004	0.008	0.115	**0.259**	0.129	0.130	0.186	0.206	0.004
		Exon	0.040	0.137	0.142	0.400	0.432	**0.500**	0.401	0.491	0.113