A variational Bayes discrete mixture test for rare variant association

Benjamin A. Logsdon[1,2,*], **James Y. Dai**[1,3], **Paul L. Auer**[1,20], **Jill M. Johnsen**[4,19], **Santhi K. Ganesh**[5], **Nicholas L. Smith**[6,7,8], **James G. Wilson**[9], **Russell P. Tracy**[10], **Leslie A. Lange**[11], **Shuo Jiao**[12], **Stephen S. Rich**[13], **Guillaume Lettre**[14], **Christopher S. Carlson**[1], **Rebecca D. Jackson**[15], **Christopher J. O'Donnell**[16], **Mark M. Wurfel**[17], **Deborah A. Nickerson**[2], **Hua Tang**[18], **Alexander P. Reiner**[1,6], and **Charles Kooperberg**[1] **on behalf of the NHLBI GO Exome Sequencing Project**[21]

[1]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

[2]Department of Genome Science, University of Washington, Seattle, WA 98195

[3]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

[4]Research Institute, Puget Sound Blood Center, Seattle, WA 98109

[5]Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48108

[6]Department of Epidemiology, University of Washington, Seattle, WA 98195

[7]Group Health Research Institute, Group Health Cooperative, Seattle, WA 98101

[8]Seattle Epidemiologic Research and Information Center, VA Office of Research and Development, Seattle, WA 98108

[9]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216

[10]Department of Pathology and Biochemistry, University of Vermont College of Medicine, Burlington, VT 05405

[11]Departments of Epidemiology, Genetics and Biostatistics, Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC, 27599

[12]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109

[13]Center for Public Health Genomics, Department of Public Health Sciences, University of Virgnia, Charlottesville, VA 22908

[14]Montreal Heart Institute and Départment de Médecine, Université de Montréal, Montréal, QC H15 1C8, Canada

---

[*]To whom correspondence should be addressed: Box 352350, University of Washington, Seattle, WA 98195, phone: 206-616-6025, blogsdon@uw.edu.
[21]Full authorship banner included in the supplementary materials

[15]Division of Endocrinology, Diabetes and Metabolism, Ohio State University, Columbus, OH 43210

[16]National Heart, Lung, and Blood Institute Center for Population Studies, The Framingham Heart Study, Framingham, MA 01702

[17]Division of Pulmonary and Critical Care Medicine, University of Washington, Seattle, WA 98104

[18]Department of Statistics and Department of Genetics, Stanford University, Stanford, CA

[19]Department of Medicine, University of Washington, Seattle, WA 98195

[20]School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53201

Abstract

Recently, many statistical methods have been proposed to test for associations between rare genetic variants and complex traits. Most of these methods test for association by aggregating genetic variations within a predefined region, such as a gene. Although there is evidence that "aggregate" tests are more powerful than the single marker test, these tests generally ignore neutral variants and therefore are unable to identify specific variants driving the association with phenotype. We propose a novel aggregate rare-variant test that explicitly models a fraction of variants as neutral, tests associations at the gene-level, and infers the rare-variants driving the association. Simulations show that in the practical scenario where there are many variants within a given region of the genome with only a fraction causal our approach has greater power compared to other popular tests such as the Sequence Kernel Association Test (SKAT), the Weighted Sum Statistic (WSS), and the collapsing method of Morris and Zeggini (MZ). Our algorithm leverages a fast variational Bayes approximate inference methodology to scale to exome-wide analyses, a significant computational advantage over exact inference model selection methodologies. To demonstrate the efficacy of our methodology we test for associations between von Willebrand Factor (VWF) levels and *VWF* missense rare-variants imputed from the National Heart, Lung, and Blood Institute's Exome Sequencing project into 2,487 African Americans within the *VWF* gene. Our method suggests that a relatively small fraction (~10%) of the imputed rare missense variants within *VWF* are strongly associated with lower VWF levels in African Americans.

Keywords

Exome sequencing study; approximate inference; von Willebrand Factor genetics

Introduction

Genome-wide association studies (GWAS) have discovered thousands of common single nucleotide polymorphisms (SNPs) associated with complex phenotypes [1, 2], yet these common variants generally explain a small fraction (<10%) of trait heritability [3, 2]. Rare coding polymorphisms, those with minor allele frequency (MAF) less than 1–5%, may in part explain this missing heritability [3, 4]. With advances in sequencing technologies, it has become affordable to assay genome-wide rare variation in large studies and identify novel rare variant associations[5, 6].

While rare variants may have larger effects than common variants [7], detecting such variants by a single variant test is typically underpowered due to the low frequencies of the variants and the large number of rare variants in the human genome [5]. The standard approach has been to aggregate rare variant associations across a gene or region of the genome [8, 9, 10, 11, 12, 13, 14]. Many different aggregate rare variant tests have been proposed, largely grouped into three classes.

The first class tests whether or not the (weighted) sum of all rare variants in a gene or region of the genome is associated with a shift in the mean of a trait, so-called "burden tests". Popular examples include the weighted sum statistic (WSS) [9], the Morris-Zeggini (MZ) test [13], and the combined multivariate and collapsing test (CMC) [8]. These tests are well powered if the proportion of variants with non-zero effect is high and each variant's effect is in the same direction. However, these assumptions may be violated, e.g. if most variants have little or no effect or some variants are protective and other variants are deleterious.

To address these potential limitations, tests in the second class model heterogeneity in the effect distribution of rare variants. Essentially these methods test for a non-zero variance component in a normal random effect model. Popular examples include the c-alpha test [11] and the more general Sequence Kernel Association Test (SKAT) [14]. These tests are advantageous when the effects of rare variants are either in the opposite direction, or a mixture of neutral and non-neutral effects [14]. Under diverse genetic architectures, neither of the two aforementioned classes of tests is uniformly most powerful. This drives development of the third class of tests, that combine the mean shift test with the test for heterogeneity of effect (e.g., SKAT-O [15], aSum [12] and others [16]), aiming to achieve greater power over a range of genetic architectures.

Previous authors have noted that even among nonsynonymous variants only approximately 20% are likely functional [17, 18]. This implies that aggregate tests may lose power because the inclusion of a substantial number of neutral variants increases the "noise" in the aggregate test. In particular, both SKAT and SKAT-O assume that the effects of rare variants follow a normal distribution, yet in published simulations for demonstrating power, only a portion of variants were set to be causal [14, 15]. Model misspecification may lead to power loss.

Other authors have explored variant level model selection methods, including a step-wise selection model [19] and Bayesian hierarchical models with different variant level priors [20, 21]. In this article we propose a test based on a simple discrete mixture model for rare variant effects. Rather than lumping all SNPs together, every variant has a latent variable determining whether it has an effect on the phenotype. Unlike the random effect tests such as SKAT and SKAT-O, we assume that the variant level effects are distributed as a mixture of a point mass at zero and a point mass away from zero. The separation of neutral and causal variants is thus directly modeled, thereby improving power. This is similar in vein to rare variant model selection methods proposed by Quintana et al. [22, 23]. The two major differences of our method compared to the method of Quintana et al. [22] are 1) we propose a computationally efficient variational Bayes approximate inference algorithm that is scalable to whole-genome analysis, and 2) we propose a likelihood ratio test that can be used

to prioritize genes or regions of the genome for further investigation that has an approximately χ_1^2 asymptotic distribution. Similar to Quintana et al. [22, 23], another feature of our method is that a variant-level posterior probability of association is generated, which can be used to prioritize variants for follow-up.

Through extensive simulations, we demonstrate our proposed method has proper type I error rates and superior power performance over existing methods in a wide range of parameter settings. As an example to illustrate the test statistic and the posterior probability of association, we present results from applying our method to data on quantitative von Willebrand Factor levels and rare missense variants in the VWF gene from the National Heart, Lung, and Blood Institute (NHLBI)-Exome Sequencing Project (ESP), and the NHLBI Candidate Gene Association Resource (CARe) consortium.

Materials and Methods

Variational Bayes discrete mixture model

Suppose n subjects are sequenced at a region with J variants observed. For the ith subject, $i = 1, ..., n$, observed data contain (Y_i, X_i, G_i), where Y_i denote the disease trait, $X_i = (X_{i1}, ..., X_{iK})$ denote K covariates (that may include eigenvectors used for controlling for population stratification), and $G_i = (G_{i1}, ..., G_{iJ})$ denote the vector of 0, 1, or 2 at J loci for an additive genetic model. Our goal is to test for genetic association in the region.

For a quantitative trait, we propose the following probabilistic discrete mixture model

$$Y_i = \gamma_0 + \sum_k^K X_{ik} \gamma_k + \theta \sum_j^J Z_j G_{ij} + e_i,$$

where Z_j is a latent indicator variable that determines whether the jth variant is truly causal ($Z_j = 1$) or neutral ($Z_j = 0$), θ is the "average" effect among causal variants, γ_k is the kth covariate effect, and e_i is the error distributed as $\mathcal{N}\left(0, \sigma^2\right)$. We assume that Z_j is distributed as Bernoulli(p_{fun}) for all j, so that p_{fun} denotes the proportion of causal variants in the region.

Compared to the burden tests and the tests for heterogeneity [8, 9, 13, 14, 15], our model adds a probabilistic assessment of whether a variant is causal and should be lumped with other causal variants, thereby filtering out neutral variants based on their evidence of association. One could extend the model to allow causal variants that have different effects by assigning a normal prior distribution to θ. However for testing genetic association in the region as a whole, a simple model with a single average effect θ, after teasing out the neutral variants, offers computational simplicity and good power. Similar arguments have been made for why burden tests are more powerful than SKAT when a majority of variants are causal [24].

The null hypothesis of no genetic association can be expressed as

$$\theta=0 \quad \text{or} \quad p_{fun}=0.$$

The two parameters θ and p_{fun} are identifiable under the alternative hypothesis, but not under the null. This causes irregular large-sample behavior of the usual likelihood ratio test (LRT) statistic. For a simple finite mixture model, the asymptotic distribution of the likelihood ratio test of the mixing parameter has a complex distribution based on an autoregressive Gaussian process [25]. To remedy this problem, we take a Bayesian view and apply an informative prior distribution Beta (2, 2) for p_{fun}. The density function of Beta (2, 2) is between 0 and 1, peaking at 0.5. Use of the Beta prior discourages p_{fun} from being near 0 or 1, the boundary of the parameter space, essentially adding a penalty to the log likelihood, in a frequentist version of the model, similar to Chen et al. [25]. Such a penalty can simplify the asymptotic distribution of the LRT to a more tractable distribution.

To fit this discrete mixture model using a standard expectation-maximization algorithm [26], we would have to sum over all possible assignments of Z_j for each variant within a gene in the likelihood, which can be prohibitive for genes with many variants. Alternatively, we could derive a Bayesian Markov chain Monte Carlo solution with a Gibbs sampler [27], but this is also computationally intensive because of the number of iterations necessary for the Markov chain to converge to the posterior distribution. To overcome these computational difficulties in genome-wide analysis, we use a variational Bayes approximate inference method [28].

The variational Bayes method was first developed for problems in statistical mechanics as mean-field theory [29], and recently has found popular applications in fitting complex statistical models for high dimensional data in genomics [30, 31, 32, 33]. In particular, when the covariates under model selection are independent, the variational Bayes approximation is nearly the same as the exact likelihood calculation [34], suggesting that the method should work well for rare variant association due to typically low correlation between rare variants.

For a quantitative trait, a variational Bayes algorithm approximates the true posterior distribution $p(Z_1, \dots, Z_J, p_{fun} \mid G, X, y, \gamma, \theta, \sigma^2)$ by estimating a factorized posterior distribution $Q(Z, p_{fun}) = q_{p_{fun}}(p_{fun}) \prod_j q_j(Z_j)$ by minimizing the Kullback-Leibler divergence

$$D_{KL}(Q(Z,p_{fun})\|p(Z,p_{fun}|G,X,Y,\gamma,\theta,\sigma^2))=\sum_Z \int Q(Z,p_{fun})\log\left(\frac{Q(Z,p_{fun})}{p(Z,p_{fun}|G,X,Y,\gamma,\theta,\sigma^2)}\right)dp_{fun},$$

where Σ_Z indicates $\sum_{Z_1=0}^{1}\cdots\sum_{Z_J=0}^{1}$. Because of the factorization approximation, there exist efficient closed form coordinate updates for all $q_j(Z_j)$ and $q_{p_{fun}}(p_{fun})$ to find a local solution to the optimization problem

$$\hat{Q}\left(Z, p_{fun}\right) = \operatorname{argmin}_{Q\left(Z, p_{fun}\right)} D_{KL}\left(Q\left(Z, p_{fun}\right) \| p\left(Z, p_{fun} | G, X, Y, \gamma, \theta, \sigma^2\right)\right).$$

Further details of the derivations of these approximate posterior distributions are provided in the Supplementary Information. Fitting this model produces a lower bound

$$\mathscr{L}\left(Y | G, X, \gamma, \theta, \sigma^2\right) = \sum_Z \int \hat{Q}\left(Z, p_{fun}\right) \log\left(\frac{p\left(Z, p_{fun}, Y | G, X, \gamma, \theta, \sigma^2\right)}{\hat{Q}\left(Z, p_{fun}\right)}\right) dp_{fun}$$

for the marginal log-posterior probability of the data log ($p\left(Y / G, X, \gamma \, \theta, \sigma^2\right)$). The parameters γ, θ, and σ^2 are estimated as

$$\left(\hat{\gamma}, \hat{\sigma^2}, \hat{\theta}\right) = \operatorname{argmax}_{\gamma, \sigma^2, \theta} \mathscr{L}\left(Y | G, X, \gamma, \theta, \sigma^2\right).$$

Based on the Beta prior distribution for p_{fun}, we propose the following test statistic

$$\chi^2_{dm} = -2\mathscr{L}_0\left(Y | G, X, \hat{\gamma}, \hat{\theta}=0, \hat{\sigma^2}\right) + 2\mathscr{L}\left(Y | G, X, \hat{\gamma}, \hat{\theta}, \hat{\sigma^2}\right),$$

where $\mathscr{L}_0\left(Y | G, X, \hat{\gamma}, \hat{\theta}=0, \hat{\sigma^2}\right)$ is the lower bound of the data when fitting the null model ($\theta = 0$). Our simulation results suggest that the asymptotic approximation of this test statistic as a χ^2_1 holds fairly well for realistic data-sets. We found that different parameters in the Beta prior on p_{fun} had very little effect on the model inference (results not shown), and that fitting this model with $\theta = 0$ produces the expected approximate posterior densities for q_{Z_j} ($Z_j = 1$) = 0.5, and q_{pfun} (p_{fun}) = Beta(J/2+2, J/2+2). It would be possible to incorporate different types of priors based on evidence of functionality in each individual variant, for example from information obtained from PolyPhen-2 [35].

In addition to the gene level test statistic χ^2_{dm}, the unique feature of our method compared to competing aggregate rare variant tests is a per-variant measure of association based on the estimated probability q_j (Z_j). This is the approximate posterior probability that a given variant is associated with the phenotype. This posterior probability estimate can be used to assess which variants in a gene are driving the association with phenotype, and to rank these variants for functional follow-up or replication.

Simulation Design

Genotypes were simulated based upon the observed minor allele frequency spectrum from all imputed exonic variants from the NHLBI-ESP project with imputation $R^2 > 0.5$ and $0.001 < \mathrm{MAF} < 0.05$ (Figure S1). As other authors have noted, based on the recent hyper exponential population growth of the last 2,000 years, there is a severe shift in the site

frequency spectrum for variants with very low minor allele frequencies [5]. Because of the possible biological relevance of this excess of very rare variation we accounted for it in our simulations. We simulated genes with 10, 20, or 40 independent rare variants, with minor allele frequencies sampled from this empirical frequency spectrum. Rare variants are often effectively independent because they lie on different haplotype backgrounds [36].

Phenotypes were simulated under multiple models. We first simulated phenotypes based on the null model with 10^6 replicates. Phenotype data were simulated under the alternative hypothesis for two distinct cases for a range of genetic architectures. The first case focused on simulated genotype data and genetic architectures where the effect sizes were independent of allele frequency (i.e. the unweighted model). We simulated data with $\theta = 1$, the proportion of functional variants within a gene was either 0.1, or 0.2, the total heritability associated with a gene was between 0.005 and 0.02, the sample size was 2,000 and 1,000 replicate simulations were run. Next, we simulated a model where $\theta \sim \mathcal{N}$ $\mu = 1$, $\sigma^2 = 1/4$ to model a deviation from the assumption that each functional variant has the same effect. The significance level a for all tests was 10^{-3}.

For the second case we used the 30 imputed missense variants from *VWF* with $0.001 < \text{MAF} < 0.05$ and imputation $R^2 > 0.5$ as the genotype data, and simulated models where the variant effect magnitudes were dependent on allele frequency. We simulated phenotype data with $\theta = 1/\sqrt{(f_j(1-f_j))}$, the proportion of functional variants within a gene was either 0.1, or 0.2, the total heritability associated with a gene was between 0.005 and 0.02, and 1,000 replicate simulations were run. Next, we simulated a model where

$$\theta \sim \mathcal{N}\left(\mu = 1/\sqrt{(f_j(1-f_j))}, \sigma^2 = 1/(4f_j(1-f_j))\right)$$ to model a deviation from the assumption that each functional variant has the same effect. Age, sex, sub-study and their corresponding estimated effects were also included in all of these simulations. As before, the significance level a for all tests was 10^{-3}. Finally, for the imputed *VWF* genotypes, we also considered simulations where the proportion of non-null variants was either 0.5 or 1.0, with effect distributions as before.

NHLBI ESP, WHI SHARe, and CARe data

The NHLBI's ESP was designed to identify novel rare variant associations with common heart, lung, and blood traits in 6,823 individuals from 20 studies. Genotype data were collected at both the University of Washington and the Broad Institute via deep whole-exome sequencing on Illumina GAII or Illumina HiSeq platforms using one of four capture targets (CCDS 2008 (~26 Mb), Roche/Nimblegen SeqCap EZ Human Exome Library v1.0 (~32 Mb; Roche Nimblegen EZ Cap v1) or EZ Cap v2 (~34 Mb), and Agilent Ref-Seq2010V2, (~36 Mb)). Single nucleotide variants were called using the UMAKE pipeline at the University of Michigan, which allowed all samples to be analyzed simultaneously, both for variant calling and filtering [5]. Samples were filtered based on call-rate, heterozygosity, concordance with GWAS data, concordance between self-reported race and genetic ancestry as determined through principal component analysis, and concordance between self-reported sex and depth of coverage on the X and Y chromosomes. Duplicate and related samples were identified using PLINK [37], for each duplicate/related pair, we

retained the sample with the higher call-rate. Variants were filtered based on a support vector machine classifier [5], depth of coverage, and Hardy-Weinberg equilibrium. After filtering, there were 1,904,903 variants available for analysis. The final data-set contained > 111 fold coverage of coding sequence in the majority of samples. Further details on study design, sequencing, and variant calling are provided in Tennessen et al. [5]. Next, we briefly describe the primary studies which contain the African American samples used in this study.

Data from Womens Health Initiative (WHI) SNP Health Association Resource (WHI-SHARe) and the NHLBI Candidate Gene Association Resource (CARe) consortium (Atherosclerosis Risk in Communities (ARIC), Cardiovascular Risk in Communities (CARDIA), Jackson Heart Study (JHS), Multiethnic Study of Atherosclerosis (MESA)) were used to impute variants identified in ESP into over 13,000 African Americans. Genome-wide genotyping was performed in all participants using the Affymetrix 6.0 platform, either at Affymetrix (for the WHI-SHARe project) or at the Broad Institute (for the CARe consortium). DNA samples with a genome-wide genotyping success rate < 97%, duplicate discordance or sex mismatch, genetic ancestry outliers (as determined by cluster analysis performed using principal components), SNPs with genotyping success rate < 98%, monomorphic SNPs, SNPs with minor allele frequency (MAF) < 1%, and SNPs with a p-value testing Hardy-Weinberg equilibrium $< 10^{-4}$ were removed from the analyses. Data from 7,563 WHI AAs on 816,361 SNPs and 6,205 CARe AAs on 712,556 SNPs remained after QC. We took the intersection of SNPs passing QC in SHARe and CARe for a total of 702,205 GWA markers. The 6,205 CARe and 7,563 SHARe samples were combined into a single target set of 13,768 samples for phasing. Phasing was performed using BEAGLE v3.3.1 [38]. An imputation reference panel of 2,163 samples was constructed using individuals with both ESP and Affymetrix 6.0 data. These samples did not overlap with the 13,768 in the target set. Prior to imputation, the imputation reference panel was pre-phased across all 2,163 samples using BEAGLE v3.3.1. Using the 4,326 statistically phased haplotypes in the reference panel we imputed 375,024 bi-allelic autosomal markers into the SHARe and CARe target panels using minimac [39], similar to Auer et al. [40].

We tested for association of imputed rare variants with natural log transformed von Willebrand Factor levels in the CARe (n=2,487) study. Von Willebrand factor antigen levels were measured in ARIC, CARDIA, and MESA participants using stored EDTA plasma samples, as previously described [41]. Only exonic missense variants with an imputation R^2 > 0.5 and 0.001 < MAF < 0.05 were considered in our model. Because of the sensitivity of the aggregate rare variant tests for genes with many very rare variants, we also Winsorized the log(VWF) data at the 0.005 and 0.995 percentiles; this improved the quality of the test statistic for all tests considered. The imputed dosages are used in all statistical tests (though the dosages are rounded for visualization purposes). All analyses were run with covariates of the first two principal components from the GWAS data to correct for possible population structure, age, sex, and sub-study.

Ethics Statement

Informed consent for collection of genetic and phenotypic data was received from all participants within each of the studies that contributed to the Exome Sequencing Project, the

NHLBI Candidate Gene Association Resource, and the Womens Health Initiative SNP Health Association Resource, where each individual study received approval from all respective institutional review boards. In our analyses all data was analyzed anonymously.

Results

Simulations

First we demonstrate that the asymptotic VBDM test statistic has appropriate type-I error control. As described in the methods, we performed 10^6 replicate null simulations of genes with either 10 or 20 independent variants with allele frequencies sampled from the empirical allele frequency distribution of imputed exome variants in the WHI-SHARe and CARe studies (as shown in Figure S1). In Table 1 we show the type-I error rates for both the weighted and unweighted version of the VBDM test statistic for different sig-nificance thresholds. The weighted statistic assumes the effect of a variant is proportional to its minor allele frequency, while the unweighted version does not make that assumption. The type-I error is well controlled for moderate significance cutoffs, and becomes slightly inflated for more stringent cutoffs, e.g. $\alpha < 10^{-3}$. We therefore recommend running permutations on any result that is significant based on this likelihood ratio test, to ensure that the type-I error rate is exactly controlled. The inflation of the test statistic in the tails likely arises because the χ^2 reference distribution for our test statistic only holds approximately.

For the power simulations we compared our approach to other popular aggregate rare variant tests including WSS, MZ, SKAT, and SKAT-O (the "w" in front of names of methods refers to the variation on the method where each variant is re-weighted based upon the standard deviation of each locus, WSS is essentially a weighted version of MZ). As shown in Figure 1, we see that VBDM has greater power when compared to the other methods when there are 40 variants within the simulated gene. Because the variant effects were simulated independently of allele frequency for the simulations shown in Figure 1, we see that the unweighted versions of each statistic tend to do better than their weighted counterparts (e.g. SKAT vs. wSKAT). The performance gain for VBDM is present across a range of simulations, including varying sparsity, heritability, and effect distributions at a significance cutoff of $\alpha = 10^{-3}$. For genes with 10 and 20 variants a similar pattern is observed (Figures S2 and S3), though the performance gain is attenuated. Adaptive permutations were performed to ensure the type-I error rate of VBDM was exactly controlled. Specifically, our approach has much greater power than the naive collapsing methods of WSS and MZ when the true distribution of effects is sparse because our mixture model is directly accounting for neutral variants in the effect distribution.

We also compared our approach to another model selection approach, the 'step-up' approach of Hoffmann et al. [19], for a smaller set of simulations (simulated genes with 10 variants, 20% of variants as causal, and effects sampled from a N(1,1/4) distribution). The results of the simulations are shown in Figure S4, where the 'step-up' method performed worse than both our method as well as the simple collapsing methods. We used the implementation of the 'step-up' method available in the 'thgenetics' R package. We also found that the 'step-up' method was more computationally intensive than our approach, where on average our algorithm took 1.0 seconds to fit a single model (with 1000 permutations), whereas it took

the method of Hoffmann et al. 3.0 seconds to fit a single model (with 1000 permutations). We didn't include the method of Hoffmann et al. in all of the simulations studies because it became computationally infeasible for genes with greater numbers of variants.

In addition, for the simulations shown in Figure 1C where the proportion of causal variants is 20% and the effect size is fixed, we also inspected the behavior of the per variant posterior probability $p_j = q_j(Z_j)$. The densities of the empirical p_j statistics for these simulations are shown in Figure S5. For the low heritability simulations in Figure S5A ($h^2 = 0.005$), there is not enough signal in the data to over-come the weak prior that each variant has a probability of 0.5 of being associated with outcome. As the heritability increases (Figure S5B-5D), the p_j statistics for casual variants shift to the right, and the p_j statistics for neutral variants shift to the left as the p_j statistics become more informative with respect to which variants are driving the association. For the highest heritability simulations ($h^2 = 0.02$, Figure S5D) any effect of the weak prior distribution on the posterior probabilities p_j has disappeared and the p_j statistics are highly informative of which variants are neutral and which variants are causal.

Next we considered simulations using the 30 imputed rare missense variants within *VWF* with MAF < 5%. In this case the effects were sampled in a way that was dependent on allele frequency, and hence the weighted versions of all the test statistics tend to do better. We see in Figure 2 that now wVBDM outperforms the alternative methods when only 10–20% of variants are functional within the gene. In contrast when 50%–100% of variants are functional our method is still competitive, but is not the most powerful (as shown in Figure S6). Our method outperforms both SKAT approaches when the number of variants in a gene is large and the signal is sparse, and has similar performance for genes with fewer variants. The SKAT tests have more power than the naive collapsing methods for the sparse models because of the heterogeneity in the simulated effect distribution, even though the assumed normal effect distribution is misspecified. The SKAT model performs better when the simulated model is sparse, whereas SKAT-O performs better when the true distribution of effects is not sparse. Yet neither SKAT approach has the additional advantage of ranking variants within a gene for follow-up based on each variant's estimated posterior probability of association.

Analyses of von Willebrand Factor Levels in African Americans

Von Willebrand Factor (VWF) is a multimeric plasma glycoprotein that plays an important role in hemostasis and thrombosis [42]. VWF binds to platelets at sites of vascular injury, stabilizes coagulation factor VIII (FVIII). High levels of VWF are associated with increased risk of various cardiovascular outcomes [43, 44]. Decreased VWF levels are associated with the common bleeding disorder von Willebrand Disease (VWD) [45]. The VWF gene on chromosome 12 contains a large number of non-synonymous coding variants, some of which have been identified in patients with VWD. Using data from ESP AA participants, Johnsen et al., recently showed that common and lower frequency variants (including Ser1486Leu and Arg2287Trp) VWF missense variants are associated with VWF levels within the African American population [41]. Therefore, we performed a gene-wide

association analysis of plasma VWF levels using imputed genotype data on low-frequency VWF missense variants from the ESP AA to demonstrate the efficacy our approach.

We analyzed log transformed von Willebrand factor levels of 2,487 AA participants from the CARe consortium and 30 imputed missense *VWF* variants derived from exome sequencing. The 30 imputed missense variants, their allele frequencies, and imputation quality scores are shown in Table S1. The results of the aggregate rare variant analyses for each method are shown in Figure 3, with significance cutoffs of 0.05 and 5×10^{-7} shown in black and red respectively. Our wVBDM approach found a significant association between rare-variants in the *VWF* gene and lower VWF levels, with an approximate likelihood ratio test p-value of 1.59×10^{-9}, and a permutation p-value of $< 1 \times 10^{-7}$. Both SKAT and SKAT-O appear to detect some signal, but neither are as significant as wVBDM. Moreover, SKAT is unable to localize which variants are driving the association. Interestingly, the unweighted version of our method is unable to detect the association signal. We speculate that by not weighting variants based on allele frequency, our model was dominated by weak positive effects from the slightly more common imputed variants. This suggests that even in cases when the rare variant signal is strong, it can be hidden if there is additional weak signal in the opposite direction. Under these circumstances models that allow for effects in both directions may be advantageous. Our results support this assertion, where we see in Figure 3 that the SKAT methods have increased power to detect the *VWF* association as compared to the simple burden tests methods (MZ and WSS). Luckily, in this particular case the weighted VBDM test prioritized the rarer variants, and was then able to identify the small subset that were driving this strong negative *VWF* association.

Figure 4 shows the burden plot for this association. The burden plot depicts the distribution of the log VWF levels, the distribution of alleles for each variant as a function of phenotype, and the mean effect of each variant on phenotype. The posterior probabilities from the VBDM algorithm are depicted to indicate which variants contribute the most to the association. It appears that the four low-frequency variants driving the strong association with phenotype are rs61750625 (Arg2287Trp), rs149424724 (Ser1486Leu), rs150077670 (Val1439Met), and rs141211612 (Ala1377Val). Based on the VBDM analysis, the estimated additive effect of the burden of rare variants is −39.1 IU/dL lower VWF levels.

Discussion

We have demonstrated several advantages of our discrete mixture test. Compared to other popular rare variant aggregate tests such as the SKAT-O test, our VBDM test had greater power, particularly as the number of variants increases (e.g., for testing large genes or multiple genes within a pathway). In addition, the exact inference implementation of this model proposed in Quintana et al. [22, 23] relies on computationally inefficient MCMC methodologies that can take multiple hours to run for a single gene with 134 variants [23]. In contrast, our method takes a fraction of a second to fit a single gene, and can therefore be used efficiently for genome-wide analyses. As the size of sequencing studies grows, the number of rare variants in any given region of the genome will also increase dramatically, while the likelihood that any given variant in that region will be functional for a given phenotype will be low. Having a test that performs well for this circumstance will be

invaluable to association mapping practitioners. Not only is this test well powered to detect rare variant associations when there is significant heterogeneity in the effect distribution (e.g., the association is driven by a small proportion of the overall number of variants tested), it can also provide a variant level measure of association with phenotype. The variant level association works by determining the level of evidence that a particular variant has an effect θ on the phenotype. In cases where there is weak or no association ($\theta = 0$) the Beta(2,2) prior on the p_{fun} parameter will cause p_j to have a null expectation of 0.5 (Figure S5A). Alternatively, in cases where there is ample evidence of $\theta \ 0$, the p_j statistics are highly informative with respect to which variants have an effect θ (Figure S5D).

We showed an example of the utility of the wVBDM statistic in identifying rare variant associations with lower von Willebrand Factor levels. Our test statistic captures a more realistic effect distribution of a mixture of null and non-null variants as compared to the popular SKAT random effect test. It can also detect both heterogeneity in the effects distribution and a mean shift, making it competitive with the SKAT-O type tests. Extensions of the VBDM statistic that allow different effects for some variants in one direction and for other variants in other directions will be considered in future work. These extensions will hopefully allow for a more natural accommodation of the phenomena we observed for *VWF*, where there appeared to be many slightly more common variants with positive effects, and only a few rare variants with large negative effects. An additional extension would be to apply VBDM to dichotomous outcomes. The variational algorithm applied to the logistic regression models is non-trivial [34], therefore we also plan to extend the VBDM framework to logistic models in future work. Finally, another extension of the VBDM model would be to apply it to testing for associations genome-wide. This would require a very careful selection of the hyperparameters α and β to prevent model over-fitting, since the number of variants would exceed the observed sample size. Of course one could also apply VBDM to each gene across the genome in a genome-wide search.

In our data set of >2,400 African Americans, the association of the 30 *VWF* imputed variants reached an exome-wide significance level (5×10^{-7}) with our wVBDM test. To the authors knowledge, the application of aggregate rare-variant tests to imputed rare variants is novel, and could be used to identify sparse aggregate associations in general GWAS studies where imputation has been performed for rare variation. The association of *VWF* with VWF levels was completely undetectable by the naive collapsing methods (WSS and MZ) that don't take into account sparsity or heterogeneity in the effect distribution. Notably, our wVBDM model provides additional insight into the possible genetic architecture underlying *VWF*'s putative association with VWF levels. Specifically, four rare variants with MAF < 1% appeared to be driving this strong association with lower VWF levels.

Our results from the aggregate VDBM test are consistent with a prior single-variant variant analysis [41], in which 3 of these four variants (Arg2287Trp,Ser1486Leu, and Val1439Met) were significantly associated with lower plasma VWF levels in ESP AA. The fourth variant (Ala1377Val) had an effect size of similar magnitude in the single-variant analysis [41], but did not reach statistical significance (perhaps due to power/sample size limitations). Additional evidence supports the functional importance of these 4 VWF missense variants. Arg2287Trp has been reported in a patient with VWD and low VWF levels [46, 47].

Arg2287Trp has also been shown to be associated with abnormal expression in multiple studies [48, 47] Arg2287Trp and Ala1377Val are predicted to be functionally deleterious based on sequence conservation analyses [41], while Ser1486Leu disrupts a highly conserved O-linked glycosylation site that modulates VWF function [49]. As mentioned previously, another natural extension of the model would be to incorporate these functional predictions into the prior distribution over individual variants. This generalization would provide a test of association that integrates multiple sources of prior belief about variant effect distributions and probability of functionality.

Studies of the distribution of genetic variation within human populations have observed a severe excess of very rare genetic polymorphism [5, 50] because of recent hyper exponential population growth. Being able to account for this severe excess of rare variation in association models will be necessary to identify the source of missing heritability for complex disease and complex phenotypes. Our method provides a novel approach, that exploits a biologically plausible assumption of the effect distribution, is well powered to identify genes or regions of the genome with an aggregation of rare variation associated with phenotype, and classifies which variants in that region are most likely to be contributing to phenotype. Software for the VBDM algorithm is freely available in the "vbdm" package from CRAN (http://cran.r-project.org).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9:356–369. [PubMed: 18398418]

2. Visscher P, Brown M, McCarthy M, Yang J. Five years of gwas discovery. Am J Hum Genet. 2012; 90:7–15. [PubMed: 22243964]

3. Manolio T, Collins F, Cox N, Goldstein D, Hindorff L, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

4. Eichler E, Flint J, Gibson G, Kong A, Leal S, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11:446–450. [PubMed: 20479774]

5. Tennessen J, Bigham A, O'Connor T, Fu W, Kenny E, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012; 337:64–69. [PubMed: 22604720]

6. Emond M, Louie T, Emerson J, Zhao W, Mathias R, et al. Exome sequencing of extreme phenotypes identifies dctn4 as a modifier of chronic pseudomonas aeruginosa infection in cystic fibrosis. Nat Genet. 2012; 44:887.

7. McClellan J, King M. Genetic heterogeneity in human disease. Cell. 2010; 141:210–217. [PubMed: 20403315]

8. Li B, Leal S. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–321. [PubMed: 18691683]

9. Madsen B, Browning S. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]

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

11. Price A, Kryukov G, De Bakker P, Purcell S, Staples J, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86:832–838. [PubMed: 20471002]

12. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70:42–54. [PubMed: 20413981]

13. Morris A, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010; 34:188–193. [PubMed: 19810025]

14. Wu M, Lee S, Cai T, Li Y, Boehnke M, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]

15. Lee S, Wu M, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012; 13:762–775. [PubMed: 22699862]

16. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. Genet Epidemiol. 2013 In Revision.

17. Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov A, et al. Prediction of deleterious human alleles. Hum Mol Genet. 2001; 10:591–597. [PubMed: 11230178]

18. Ng P, Levy S, Huang J, Stockwell T, Walenz B, et al. Genetic variation in an individual human exome. PLoS Genet. 2008; 4:e1000160. [PubMed: 18704161]

19. Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PLoS One. 2010; 5:e13584. [PubMed: 21072163]

20. Capanu M, Begg CB. Hierarchical modeling for estimating relative risks of rare genetic variants: Properties of the pseudo-likelihood method. Biometrics. 2011; 67:371–380. [PubMed: 20707869]

21. Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. Genetic epidemiology. 2011; 35:57–69. [PubMed: 21181897]

22. Quintana MA, Berstein JL, Thomas DC, Conti DV. Incorporating model uncertainty in detecting rare variants: the bayesian risk index. Genetic epidemiology. 2011; 35:638–649. [PubMed: 22009789]

23. Quintana MA, Schumacher FR, Casey G, Bernstein JL, Li L, et al. Incorporating prior biologic information for high-dimensional rare variant association studies. Human heredity. 2013; 74:184–195. [PubMed: 23594496]

24. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genetic Epidemiology. 2011; 35:606–619. [PubMed: 21769936]

25. Chen H, Chen J, Kalbfleisch J. A modified likelihood ratio test for homogeneity in finite mixture models. J Roy Stat Soc B Met. 2001; 63:19–29.

26. Dempster A, Laird N, Rubin D. Maximum likelihood from incomplete data via the em algorithm. J Roy Stat Soc B Met. 1977:1–38.

27. Smith A, Roberts G. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. J Roy Stat Soc B Met. 1993; 55:3–23.

28. Beal, M. PhD thesis. Gatsby Computational Neuroscience Unit, University College London; 2003. Variational algorithms for approximate Bayesian inference.

29. Parisi, G. Statistical field theory. Addison-Wesley; 1988.

30. Teschendorff A, Wang Y, Barbosa-Morais N, Brenton J, Caldas C. A variational bayesian mixture modelling framework for cluster analysis of gene-expression data. Bioinformatics. 2005; 21:3025–3033. [PubMed: 15860564]

31. Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes C. Genosnp: a variational bayes within-sample snp genotyping algorithm that does not require a reference population. Bioinformatics. 2008; 24:2209–2214. [PubMed: 18653518]

32. Logsdon B, Hoffman G, Mezey J. A variational bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinformatics. 2010; 11:58. [PubMed: 20105321]

33. Li Z, Sillanpää M. Estimation of quantitative trait locus effects with epistasis by variational bayes algorithms. Genetics. 2012; 190:231–249. [PubMed: 22042575]

34. Carbonetto P, Stephens M. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian analysis. 2012; 7:73–108.

35. Adzhubei I, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]

36. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease–common variant or not? Human molecular genetics. 2002; 11:2417–2423. [PubMed: 12351577]

37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

38. Browning S, Browning B. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007; 81:1084–1097. [PubMed: 17924348]

39. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis G. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012; 44:955–959. [PubMed: 22820512]

40. Auer P, Johnsen J, Johnson A, Logsdon B, Lange L, et al. 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. 2012; 91:794–808. [PubMed: 23103231]

41. Johnsen J, Auer P, Morrison A, Jiao S, Wei P, et al. Common and rare vwf coding variants, von willebrand factor levels, and factor viii levels in african americans: the nhlbi exome sequencing project. Blood. 2013 In Press.

42. Denis CV, Lenting PJ. von willebrand factor: at the crossroads of bleeding and thrombosis. International journal of hematology. 2012; 95:353–361. [PubMed: 22477538]

43. Martinelli, I. Seminars in hematology. Vol. 42. Elsevier; 2005. von willebrand factor and factor viii as risk factors for arterial and venous thrombosis; p. 49-55.

44. van Schie M, van Loon J, De Maat M, Leebeek F. Genetic determinants of von willebrand factor levels and activity in relation to the risk of cardiovascular disease: a review. Journal of Thrombosis and Haemostasis. 2011; 9:899–908. [PubMed: 21342431]

45. Johnsen, J.; Ginsburg, D. von willebrand disease. In: Kaushansky, K.; Licthman, M.; Beutler, E.; Kipps, T.; Seligsohn, U., et al., editors. Williams Hematology. New York: McGraw Hill, Inc; 2010. p. 2069-2087.

46. Goodeve A, Eikenboom J, Castaman G, Rodeghiero F, Federici AB, et al. Phenotype and genotype of a cohort of families historically diagnosed with type 1 von willebrand disease in the european study, molecular and clinical markers for the diagnosis and management of type 1 von willebrand disease (mcmdm-1vwd). Blood. 2007; 109:112–121. [PubMed: 16985174]

47. Ahmad F, Jan R, Kannan M, Obser T, Hassan MI, et al. Characterisation of mutations and molecular studies of type 2 von willebrand disease. Thrombosis and haemostasis. 2013; 109:39–46. [PubMed: 23179108]

48. Eikenboom J, Hilbert L, Ribba A, Hommais A, Habart D, et al. Expression of 14 von willebrand factor mutations identified in patients with type 1 von willebrand disease from the mcmdm-1vwd study. Journal of Thrombosis and Haemostasis. 2009; 7:1304–1312. [PubMed: 19566550]

49. Badirou I, Kurdi M, Legendre P, Rayes J, Bryckaert M, et al. In vivo analysis of the role of o-glycosylations of von willebrand factor. PLoS One. 2012; 7:e37508. [PubMed: 22616016]

50. Nelson M, Wegmann D, Ehm M, Kessner D, Jean P, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012; 337:100–104. [PubMed: 22604722]
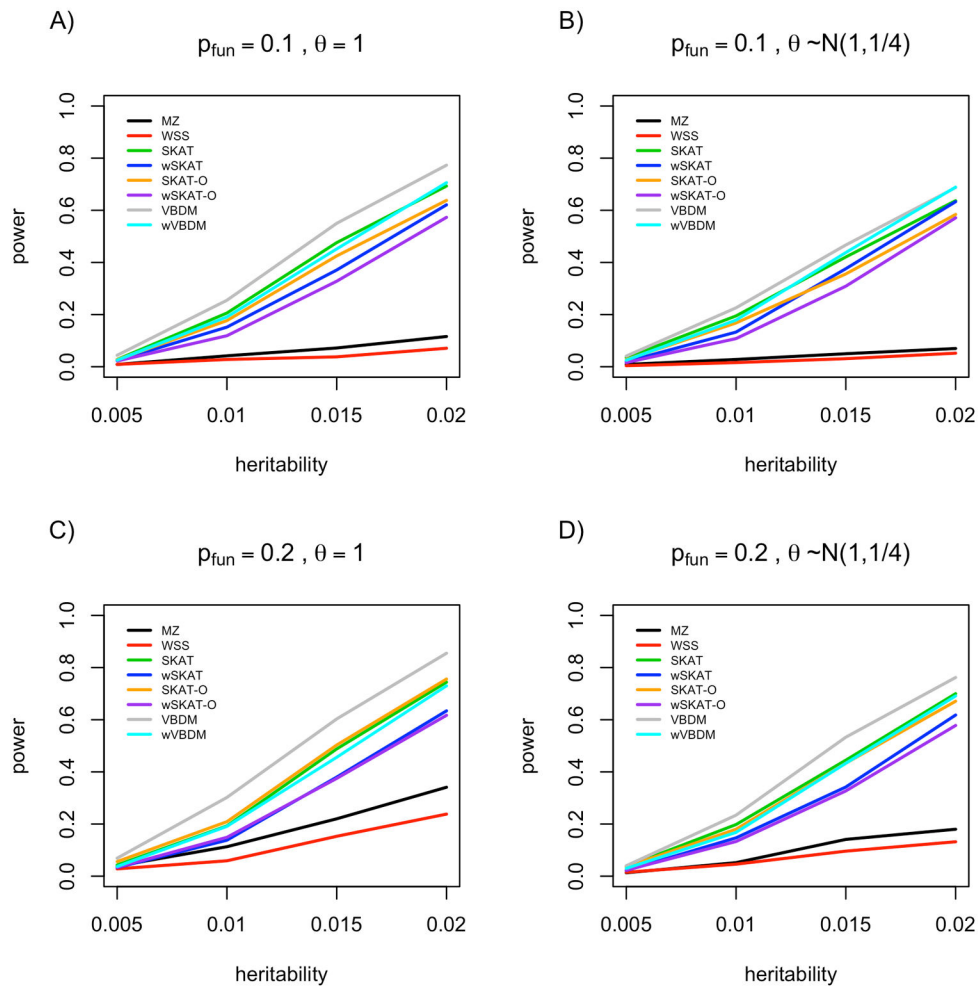
Figure 1.
Power results from simulations of genes with 40 rare variants. A) 10% of variants within a gene have an effect on phenotype, with a fixed effect size. B) 10% of variants have an effect on phenotype, with effects sampled from a N(1,1/4) distribution. C) 20% of variants within a gene have an effect on phenotype, with a fixed effect size. D) 20% of variants within a gene have an effect on phenotype, with effects sampled from a N(1,1/4) distribution.
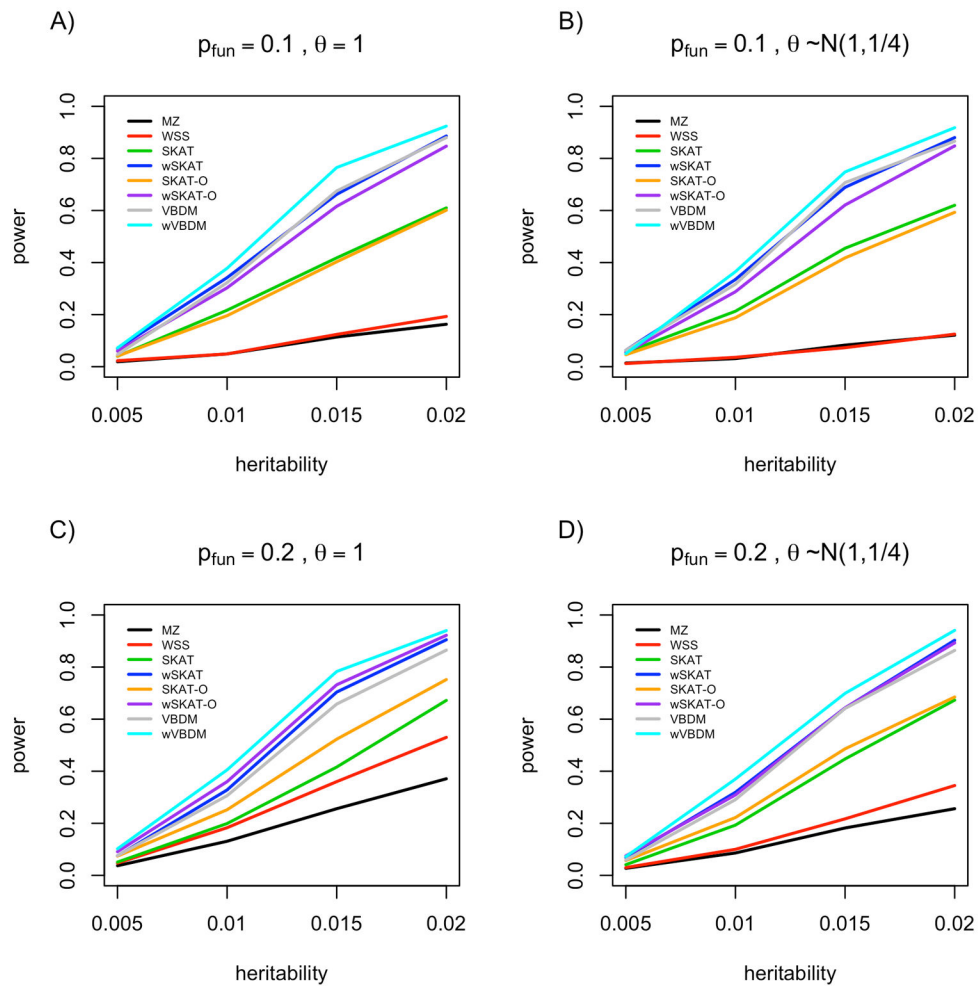
Figure 2.
Power results from phenotypic simulations based on 30 imputed VWF missense variants with MAF < 5%. A) 10% of missense variants within VWF have an effect on phenotype, with a fixed effect size. B) 10% of variants have an effect on phenotype, with effects sampled from a N(1,1/4) distribution. C) 20% of variants within VWF have an effect on phenotype, with a fixed effect size. D) 20% of variants within a gene have an effect on phenotype, with effects sampled from a N(1,1/4) distribution.
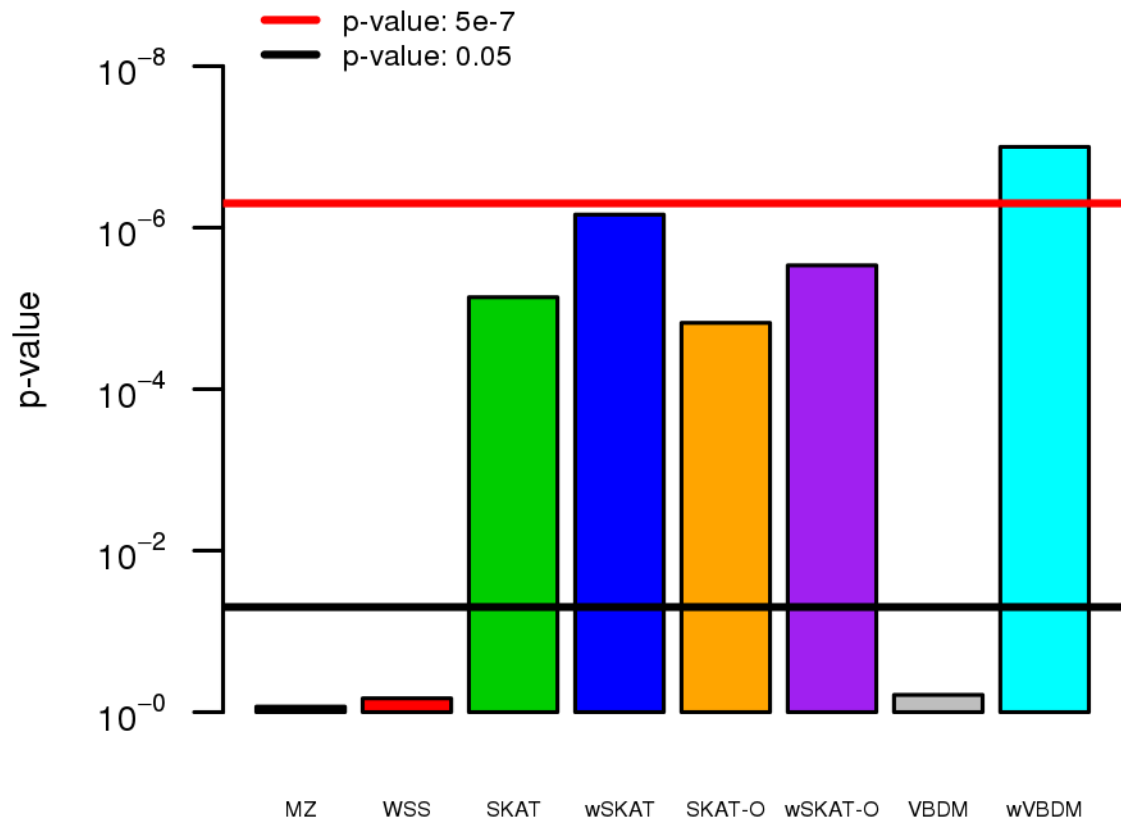
Figure 3.
Aggregate rare variant test results for the association of imputed missense variants within VWF with VWF levels in African Americans in the CARe consortium.
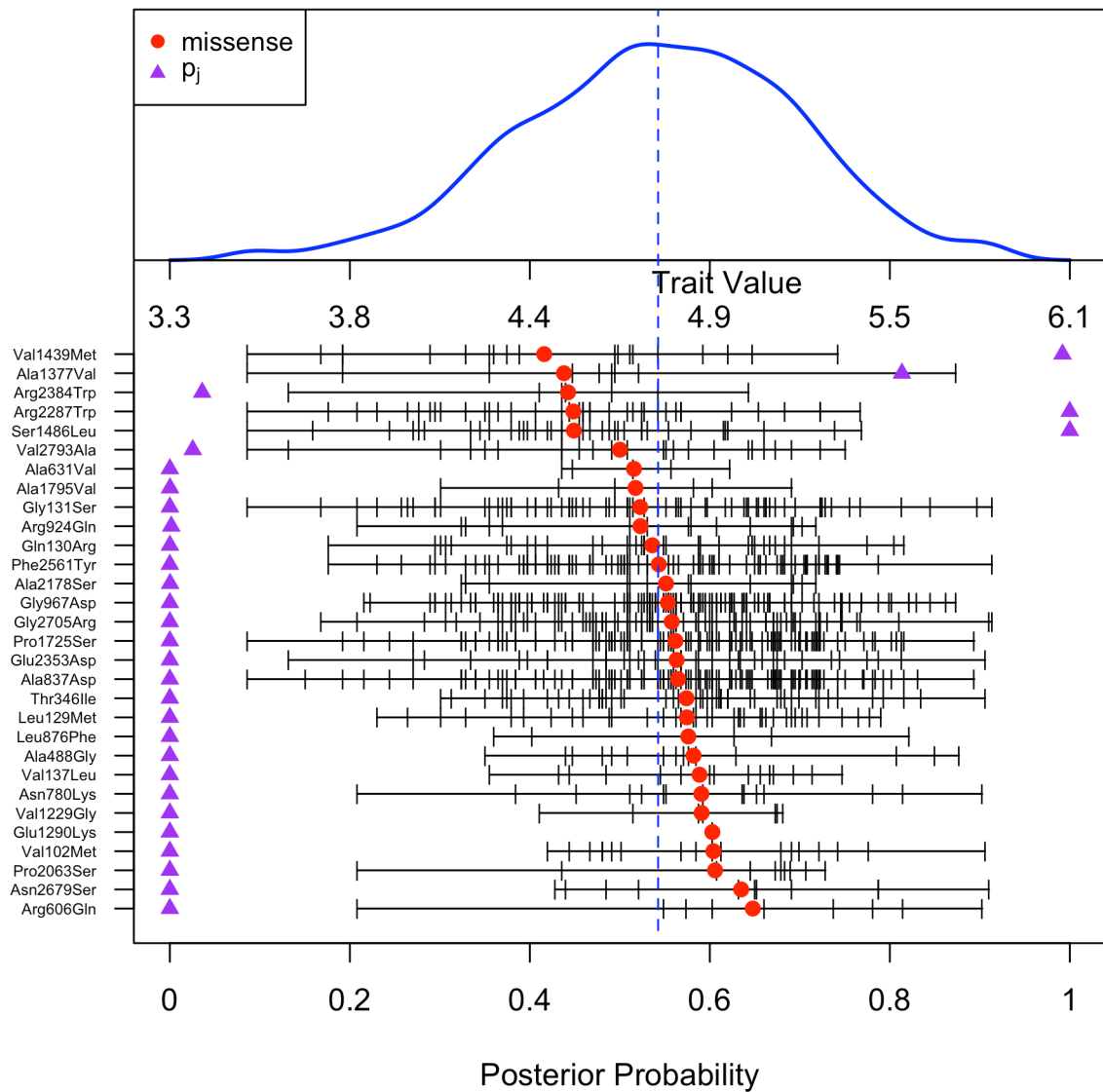
Figure 4.
Burden plot of VWF levels for the *VWF* gene from the CARe consortium. Histogram of Winsorized log transformed VWF levels is shown on top. The thirty imputed missense variants are shown on bottom, ordered by mean log VWF levels of the imputed non reference allele. Each tick mark for a given variant represents the phenotypic value of an individual estimated to have the non-reference allele for that variant based on the imputation. The solid circles indicate the mean of all individuals with imputed non-reference alleles. The dotted line indicates the overall mean of VWF levels. The purple triangles indicate the posterior probability of association of a given variant with VWF levels.