RESEARCH ARTICLE

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

VAAST 2.0: Improved Variant Classification and Disease-Gene Identification Using a Conservation-Controlled Amino Acid Substitution Matrix

Hao Hu,[1†] Chad D. Huff,[1†] Barry Moore,[2] Steven Flygare,[2] Martin G. Reese,[3] and Mark Yandell[2*]

[1]*Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas;* [2]*Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah;* [3]*Omicia, Inc., Emeryville, California*

Received 14 January 2013; Revised 9 April 2013; accepted revised manuscript 21 May 2013.
Published online 8 July 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21743

ABSTRACT: The need for improved algorithmic support for variant prioritization and disease-gene identification in personal genomes data is widely acknowledged. We previously presented the Variant Annotation, Analysis, and Search Tool (VAAST), which employs an aggregative variant association test that combines both amino acid substitution (AAS) and allele frequencies. Here we describe and benchmark VAAST 2.0, which uses a novel conservation-controlled AAS matrix (CASM), to incorporate information about phylogenetic conservation. We show that the CASM approach improves VAAST's variant prioritization accuracy compared to its previous implementation, and compared to SIFT, PolyPhen-2, and MutationTaster. We also show that VAAST 2.0 outperforms KBAC, WSS, SKAT, and variable threshold (VT) using published case-control datasets for Crohn disease (*NOD2*), hypertriglyceridemia (*LPL*), and breast cancer (*CHEK2*). VAAST 2.0 also improves search accuracy on simulated datasets across a wide range of allele frequencies, population-attributable disease risks, and allelic heterogeneity, factors that compromise the accuracies of other aggregative variant association tests. We also demonstrate that, although most aggregative variant association tests are designed for common genetic diseases, these tests can be easily adopted as rare Mendelian disease-gene finders with a simple ranking-by-statistical-significance protocol, and the performance compares very favorably to state-of-art filtering approaches. The latter, despite their popularity, have suboptimal performance especially with the increasing case sample size.
Genet Epidemiol 37:622–634, 2013. © 2013 Wiley Periodicals, Inc.

KEY WORDS: disease-gene finder; variant classifier; aggregative association test; rare Mendelian disease; complex disease

Introduction

Traditionally, genome-wide association studies (GWASs) have been used to identify disease-associated variants using sets of "tagging" single-nucleotide polymorphisms (SNPs) distributed across the genome. GWAS approaches, however, are underpowered to detect the effects of rare casual variants because they are usually in poor linkage disequilibrium with the tagging SNPs [Visscher et al., 2012]. New sequencing technologies have significantly reduced the price of human genome resequencing, and are identifying many novel rare variants. The classification and prioritization of these rare variants for disease-gene studies has thus become a significant problem.

To date, several variant prioritization tools have been developed to identify damaging alleles in personal genomes data. SIFT [Ng and Henikoff, 2006] and AlignGV-GD [Tavtigian et al., 2006], for example, use multiple alignments to assay conservation levels of novel amino acid changing variants with the underlying assumption that sequence variants, which alter highly conserved positions in protein sequences are a priori more likely to be damaging. Two more recently published algorithms, PolyPhen-2 [Adzhubei et al., 2010] and MutationTaster [Schwarz et al., 2010], improve upon this basic approach, integrating other information (e.g., protein structural changes) into the calculation, and thus significantly improving their variation prioritization accuracies compared to SIFT [Ng and Henikoff, 2006].

A major weakness of many variant prioritization tools is that they can only prioritize variants within phylogenetically conserved coding regions and thus have poor coverage across the proteome. For example, SIFT and PolyPhen can score only 60% and 81% of the human proteome, respectively [Adzhubei et al., 2010]. Another weakness of these approaches is that they make no use of allele frequency information. It has long been known that minor allele frequency (MAF) is negatively correlated with purifying selection pressure [Kryukov et al., 2007]. Thus, publicly available human-genome databases (e.g., HapMap [Thorisson et al., 2005], the 1000 Genomes Project (1KGP) [Altshuler et al., 2010], and

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

[†]These authors contributed equally to this work.

[*]Correspondence to: Prof. Mark Yandell, 15 North 2030 East, Room 2100, Salt Lake City, UT 84112. E-mail: myandell@genetics.utah.edu

dbSNP [Smigielski et al., 2000] provide valuable frequency information that can, in principle, be used for variant prioritization. VAAST [Yandell et al., 2011] is a step forward in both regards in that it uses an approach to variant classification that combines both amino acid substitution (AAS) information with variant frequency information, allowing it to score all variants no matter where they lie in the genome and with greater accuracy [Yandell et al., 2011].

The widened scope of the VAAST approach, however, comes at a cost: VAAST, in its original form, does not make any use of phylogenetic conservation data. In the present study, we describe an extension of the VAAST variant prioritization approach that makes use of a conservation-controlled AAS matrix (CASM) to overcome this shortcoming. The CASM approach allows VAAST to score every variant in the genome, and to employ phylogenetic conservation information at the same time. Our benchmark analyses presented here demonstrate that CASM approach results in the highest variant prioritization accuracies yet achieved.

Employing rare variants for disease-gene identification is another challenge. One approach is simply to search case genomes for regions having an increased density of rare variants. This is the approach taken by ANNOVAR [Wang et al., 2010], which allows users to impose a threshold on variant frequencies as observed in dbSNP or in the 1000 Genomes Project [Altshuler et al., 2010; Smigielski et al., 2000], excluding from further consideration variants with population frequencies above a user-defined threshold. A strength of the tool is that it can use third-party variant prioritization scores such as those produced by SIFT and PolyPhen to improve search accuracy; its principle weakness is that excluding variants with MAFs above a user-defined threshold renders the tool ineffective for searching datasets containing disease-causing alleles distributed across a range of population frequencies. In response, probabilistic approaches that overcome this limitation have emerged. These tests aggregate prioritization information from each variant in a gene to achieve greater statistical power, allowing them to bypass the need for large statistical corrections for multiple tests. These tools include CAST [Morgenthaler and Thilly, 2007], CMC [Li and Leal, 2008], WSS [Madsen and Browning, 2009], KBAC [Liu and Leal, 2010], VT [Price et al., 2010], SKAT [Wu et al., 2011], and VAAST [Yandell et al., 2011]. Although each algorithm approaches the problem differently, all either explicitly or implicitly use the MAF information to weight variants. In addition, VT and VAAST 2.0 can also use functional predictions from third-party variant prioritization tools such as PolyPhen and PhastCons [Yang, 1995] to weight variants [Price et al., 2010]. We refer to these approaches collectively as *aggregative variant association tests.*

To date, aggregative variant association tests have been seen as a means to identify genes and variants associated with common diseases. However, the performance characteristics of different association tests as rare disease gene finders are still largely unknown. Also largely undetermined to date is the impact of factors such as population attributable risk (PAR) and allelic and locus heterogeneity on their ability to identify genes and alleles responsible for both rare and common disease [Madsen and Browning, 2009].

Here we describe the new version of VAAST (VAAST 2.0) and the CASM approach. We employ a variety of datasets to benchmark VAAST 2.0, systematically comparing its performance to the original version of VAAST [Yandell et al., 2011] and to other published association tests, including WSS [Madsen and Browning, 2009], KBAC [Liu and Leal, 2010], SKAT [Wu et al., 2011], and VT [Price et al., 2010]. Our results demonstrate the improvements to VAAST made possible by the CASM approach; they also provide a general framework with which to investigate the performance of different aggregative variant association tests using published and simulated datasets. These results shed considerable light on the complexities involved in searching personal genomes data for disease-causing alleles as they reveal unexpected strengths and weaknesses of different approaches under different scenarios, providing a roadmap for future improvements to each method.

Materials and Methods

The CASM Approach

VAAST uses an extended composite likelihood ratio test (CLRT) to determine a severity score for genomic variants [Yandell et al., 2011]. The null model of the CLRT states that the frequency of a variant or variant group is the same in the control population (background genomes) and case population (target genomes), while the alternative model allows these two frequencies to differ. Under a binomial distribution, the likelihood for both models can be calculated based on observed allele frequencies in the control and case datasets. In VAAST 1.0 this likelihood ratio (LR) is further updated by the AAS severity parameter (a_i/h_i), where h_i is the likelihood that an AAS does not contribute to the disease and a_i is the likelihood that it does. We estimate h_i by setting it equal to the frequency of this type of amino acid change in the background population, and a_i by setting it equal to the frequency of the amino acid change among all disease-causing mutations in OMIM. VAAST 1.0 uses (a_i/h_i) to model the severity of each amino acid change. This approach, however, does not take into account phylogenetic conservation at that position of the protein, which can in theory be used to improve the accuracy of (a_i/h_i). In VAAST 2.0, we have extended this severity parameter by using an additional conservation measurement, PhastCons [Yang, 1995] scores; these scores estimate the probability that the locus is under negative selection and are calculated using multiple species nucleotide alignments.

The CASM operates as follows: Consider first, a variant occurring at a position in the genome having some PhastCons score, and changing a valine (V) to an alanine (A). To calculate the severity parameter, we first calculate the relative frequencies of V to A causing variants at any conservation level within a disease and a nondamaging variant database. In practice, this approach is hindered by the fact that the number of such variants in the disease database may be limited. To overcome this problem, we start with estimating (a_i/h_i) for each type of amino acid with PhastCons scores of 0 and 1 (the two end points), as follows. For any given type of AAS i ($i = 1, 2, \ldots, m$), suppose that there are n_i variants in the disease database and each variant j (j = 1, 2, . . . , n_i) has a PhastCons score of P_{ij}. Because P_{ij} can be interpreted as the probability that the variant is at a conserved locus [Yang, 1995], the likelihood that a variant is disease causing can be estimated by

$$a_{i1} = \left(\sum_{j=1}^{n_i} P_{ij} \right) \Big/ C_D \tag{1}$$

for variants with a PhastCons score of 1, and

$$a_{i0} = \left(\sum_{j=1}^{n_i} 1 - P_{ij} \right) \Big/ C_D \tag{2}$$

for variants with a PhastCons score of 0, where C_D is the total number of variants in the disease allele database used for training. Similarly using a database of nondamaging variants, the likelihood that a variant is not disease causing can be estimated by

$$h_{i1} = \left(\sum_{j=1}^{n_i} P_{ij} \right) \Big/ C_N \tag{3}$$

for variants with a PhastCons score of 1, and

$$h_{i0} = \left(\sum_{j=1}^{n_i} 1 - P_{ij} \right) \Big/ C_N \tag{4}$$

for variants with a PhastCons score of 0, where C_N is the total number of variants in the nondamaging allele database used for training.

Thus, the severity parameter for AAS type i with a PhastCons score of 0 and 1 is (a_{i0}/h_{i0}) and (a_{i1}/h_{i1}), respectively. For variants with other PhastCons scores (x; $0 < x < 1$), the likelihood is estimated by a linear combination of (a_{i0}/h_{i0}) and (a_{i1}/h_{i1}), namely,

$$\frac{a_{ix}}{h_{ix}} = \frac{a_{i0}}{h_{i0}} \times (1 - x) + \frac{a_{i1}}{h_{i1}} \times x, \tag{5}$$

where a_{ix}/h_{ix} are the terms in the CASM. This provides an estimate of LR of a given amino acid change being disease-causal vs. being nondamaging, controlled for the phylogenetic conservation level in the gene context. The procedures for training and testing the CASM method are detailed in the Supporting Information.

Unless otherwise noted, we calculated the severity parameter using variants from the Human Gene Mutation Database (HGMD) [Cooper et al., 1998] as disease variants and using variants from the 1000 Genomes Project (Phase I data) [Altshuler et al., 2010] with MAFs ^{3} 0.05 as the nondamaging variants. We first evaluated CASM on a small test dataset using PhastCons scores from three different genome alignments: UCSC vertebrate, mammal, and primate [Karolchik et al., 2004]. The vertebrate alignment produced the most accurate CASM scores and was used in all subsequent analyses.

Indel Support in VAAST 2.0

VAAST 2.0 also supports small insertion and deletion (indel) mutations. The VAAST Annotation Tool, a component of VAAST package [Reese et al., 2010; Yandell et al., 2011] now annotates the functional impact of indels on protein-coding genes in GVF format [Reese et al., 2010]. These annotations include (1) determination of whether or not the indel disrupts the reading frame of one or more protein-coding genes and if so which ones; and (2) whether the indel causes an AAS, insertion, or deletion. VAAST 2.0 then scores indels with the same CLRT as single-nucleotide variants (SNVs) that is, it calculates the LR of null model vs. alternative model for each indel variant based on its observed allele frequencies in background and target genomes, and then updates the LR with the severity parameter (a_i/h_i), which is estimated as follows. First, indels are classified into categories based on three properties: (1) whether it is an insertion or a deletion, (2) the affected nucleotide length, and (3) whether it disrupts the protein translation reading frame. For each category of indels, we calculate the proportion of HGMD variants falling into this category, which is our estimate of disease-causal likelihood. We also use a nondamaging variant database to determine the likelihood of being noncausal for each category. The ratio of these two likelihoods is used as (a_i/h_i) term to update the original LR. Note that rare indel variants are collapsed before being scored, as described in Yandell et al. [2011]. This is especially important for indels, because the exact boundaries of indel variants are often called imprecisely. Collapsing variants thus allows VAAST to assess the impact of multiple overlapping indels in the cases.

Results

Variant Prioritization

We compared the performance of VAAST 2.0 to other variant classifiers. Whereas tools such as SIFT, PolyPhen, and Align-GD [Adzhubei et al., 2010; Ng and Henikoff, 2006; Tavtigian et al., 2006] cannot score regions lacking multiple sequence alignment information, VAAST 2.0 suffers from no such limitation. In regions where no nucleotide or protein conservation data are available, VAAST 2.0 uses allele frequencies and global amino acid substation frequencies as the basis for variant prioritization; in regions where conservation information is available, VAAST 2.0 supplements this information with PhastCons scores [Yang, 1995], which cover 99.9% of the human proteome. For this comparison, we limited our benchmark analysis to variants that can be scored by all four algorithms (SIFT, PolyPhen-2, MutationTaster, and VAAST 2.0). It should be kept in mind, however, that in addition to these, VAAST can also score many other variants in these datasets that the other tools cannot.

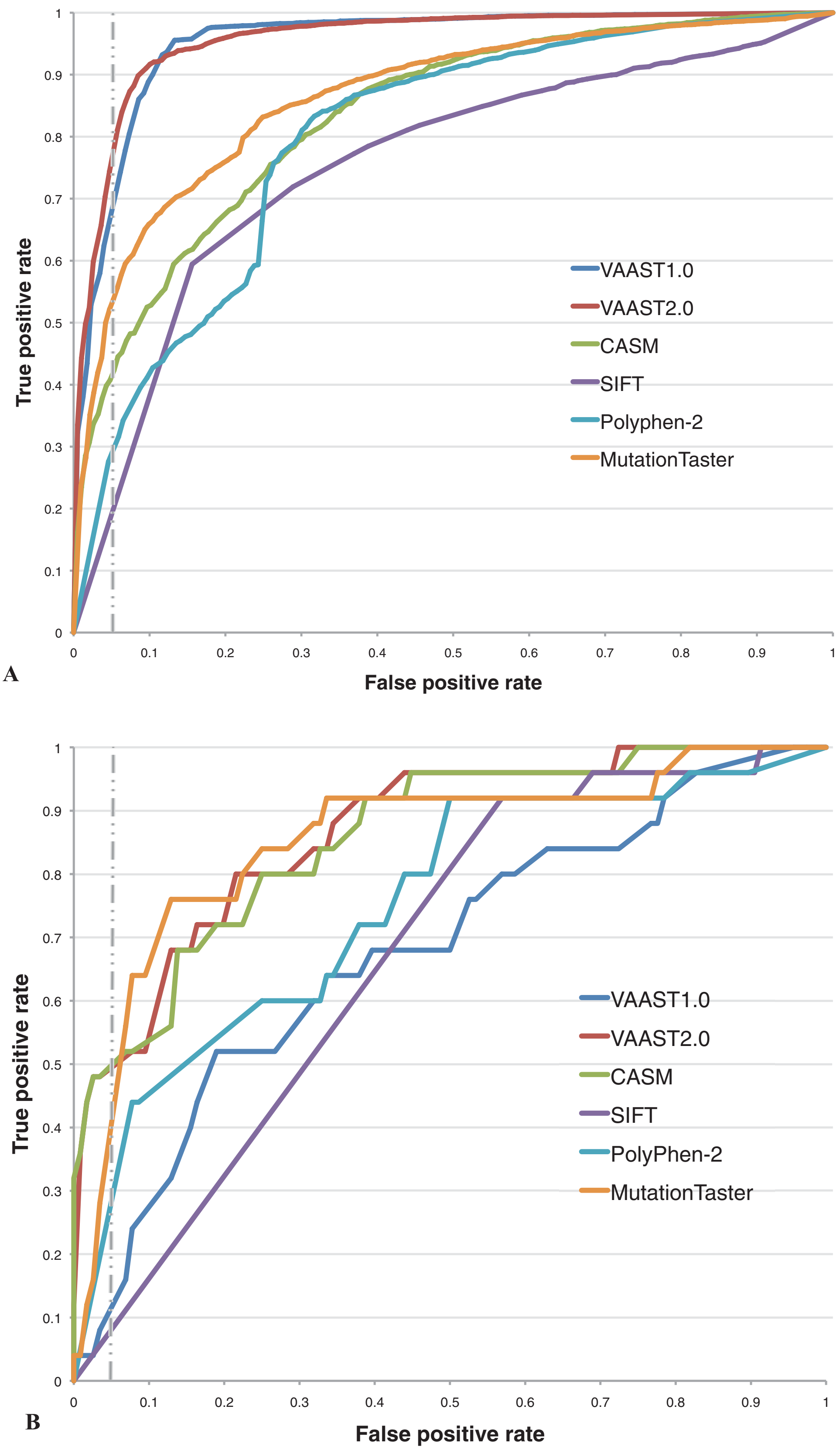

Figure 1. Receiver operator curves (ROC) for the variant prioritization tools. Shown are ROCs for VAAST 1.0, VAAST 2.0, CASM, SIFT, PolyPhen-2, and MutationTaster, using two benchmark datasets: (A) common and rare variants from HGMD and 1000 Genomes Project; (B) *BRCA1* and *BRCA2* rare variant set. *x*-axis: false-positive rate; *y*-axis: true positive rate. Dashed line denotes the false-positive rate of 0.05.

To evaluate the prioritization performance of each tool, we plotted the Receiver Operator Curve for each algorithm using a set of nondamaging variants (drawn randomly from the 1000 genomes project (1KGP) Pilot Phase [Altshuler et al., 2010]) and a set of disease-causal variants (from HGMD database) (see Supporting Information for details). Figure 1A demonstrates that the accuracy of VAAST 2.0 and 1.0 is considerably better than the other algorithms, with the true positive rate (TPR) reaching 76% for VAAST 2.0 and 68% for VAAST 1.0 when the false-positive rate (FPR) is 5%. The third best tool is MutationTaster, whose TPR is 23% lower than VAAST 2.0 at the same FPR level. VAAST 2.0 using the CASM method alone without recourse to variant frequency information ("CASM" in Fig. 1) is the fourth best performing approach, followed by PolyPhen-2 and SIFT. We also calculated the area under the curve value and the accuracy at FPR = 0.05 for each algorithm, which demonstrates the same trends (Table 1).

For a second variant prioritization benchmark, we compared the performance of each of these algorithms using a set of 143 rare missense variants in the *BRCA1* and *BRCA2* genes whose clinical significance was assessed by a third party [Easton et al., 2007]. This variant set differs from the HGMD/1KGP variants used to produce Figure 1A in that the dataset used to produce Figure 1A contains both common and rare variants for nondamaging and deleterious alleles, whereas this set Figure 1B only contains very rare variants (MAF << 1%). The results of this benchmark analysis are shown in Figure 1B and Table 1. Because majority of the variants in this set are observed only once, VAAST 2.0 cannot use the allele frequency information to improve its power, thus the performance of the full VAAST 2.0 algorithm is only marginally better than the CASM method alone in this case. Nevertheless, by a small margin, VAAST 2.0 is still the most accurate classifier. At FPR = 0.05, the accuracy of VAAST 2.0 is 4% higher than MutationTaster, the next best classifier.

The variant prioritization accuracies of VAAST 1.0 and 2.0 on the HGMD/1KGP dataset (Fig. 1A and Table 1) are very similar. This is because, on this dataset, both algorithms derive most of their power from variant MAF information in a control population. However, in cases where such information is unavailable (e.g., all variants are equally rare), the accuracy of VAAST 1.0 drops, while VAAST 2.0 still accurately predicts the severity of variants using the CASM method. This is illustrated by the *BRCA* variants benchmark dataset in Figure 1B and Table 1.

Table 2. Characteristics of the *NOD2*, *LPL*, and *CHEK2* datasets

	Average number of variants per case genome	Number of variants with odds ratio >1	Number of unique multisite genotypes	PAR[a]
NOD2	1.19	27	566	44.7%
LPL	0.10	10	14	8.4%
CHEK2	0.05	22	30	3.81%

[a] The population attributable risk (PAR) is calculated as the sum of PAR values of all susceptibility variants.

Benchmark Analyses on Multigenic Common Diseases

Next we compared the power of six aggregative variant association tests using three different published sequence-based disease-gene datasets. The three datasets used are *NOD2*, implicated in Crohn disease [Lesage et al., 2002]; *LPL*, implicated in hypertriglyceridemia [Johansen et al., 2010]; and *CHEK2*, a gene involved in breast cancer [Le Calvez-Kelm et al., 2011]. In the *NOD2* dataset, both rare and common variants are present, while only rare variants (MAF < 0.05) are present in the *LPL* and *CHEK2* dataset. In each study, either genotype or allele frequency data has been reported for diseased and control individuals. In the latter case, genotypes for case and control genomes were simulated, assuming no linkage disequilibrium between variants. Summary statistics for each of the three datasets are presented in Table 2. We calculated power using a bootstrap approach. Specifically, we sampled cases and controls with replacement, evaluating the proportion of the resampled datasets that achieved statistical significance [Yandell et al., 2011]. We used a genome-wide significance level of 2.4×10^{-6} for *NOD2* and *LPL*. For *CHEK2*, we set the significance level to 0.0005 for *CHEK2* in concordance with the original study [Le Calvez-Kelm et al., 2011].

In all three datasets VAAST 2.0 is consistently the most powerful association test (Fig. 2). For *LPL*, for example, at a sample size of 400, VAAST 2.0 has 10% more power than VAAST 1.0 (second) and 25% more power than KBAC (third); For *CHEK2*, VAAST 2.0 has 3% more power than VAAST 1.0 at its maximal sample size and 9% more than KBAC (third); for *NOD2*, the power of VAAST 2.0 is 4% better than VAAST 1.0 and 9% better than WSS (third). Each of the other algorithms seems to have a niche. KBAC, for example, performs well on the two datasets (*LPL*, *CHEK2*) where only rare variants contribute to the disease, but its performance drops significantly where both common and rare causal variants are present (e.g., *NOD2*). WSS, on the other hand, performs well under both scenarios, and outperforms KBAC, SKAT, and

Table 1. Variant prioritization performance benchmarks

	VAAST1.0	VAAST2.0	CASM	SIFT	PolyPhen-2	MutationTaster
Area under the curve (AUC)						
Dataset 1 (HGMD+1KGP)	0.95	0.96	0.83	0.76	0.8	0.87
Dataset 2 (rare *BRCA* variants)	0.68	0.87	0.86	0.73	0.76	0.85
Accuracy at FPR of 0.05						
Dataset 1 (HGMD+1KGP)	0.81	0.86	0.68	0.57	0.62	0.74
Dataset 2 (rare *BRCA* variants)	0.53	0.72	0.72	0.52	0.62	0.68

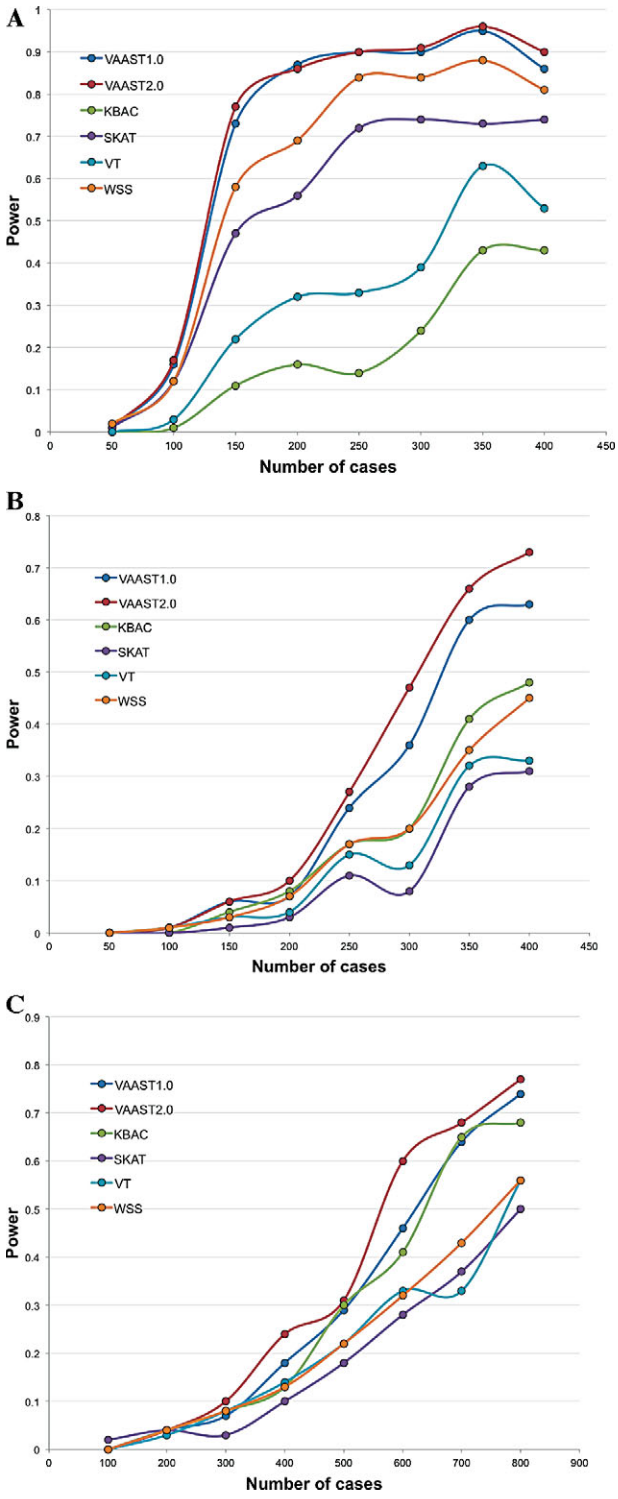

Figure 2. Power comparisons over three published common disease datasets. (A) *NOD2*, (B) *LPL*, (C) *CHEK2*. The *x*-axis shows the number of case genomes and the *y*-axis shows the statistical power. The power is calculated based on 100 bootstraps.

Table 3. Significance of associations (showing *P*-values) between low triglyceride levels and rare variants in the *ANGPTL4* gene

VAAST1.0	VAAST2.0	KBAC	SKAT	VT[a]	WSS
0.000371	0.000508	0.00402	0.00677	0.00452	0.00402

[a] VT is run with PolyPhen-2 scores.

VT when common variants are observed (e.g., the *NOD2* data).

We also benchmarked VAAST 2.0 on the Dallas Heart Study dataset [Romeo et al., 2009], in which rare variants in *ANGPTL4* gene were found to be associated with low triglyceride levels within 3,551 sequenced individuals. For this study, we tested for different distributions of rare variants in *ANGPTL4* gene between the highest quartile and lowest quartile of triglyceride levels in the 3,551 individuals. Ethnicity and gender status are matched, in accordance with the original study [Romeo et al., 2009]. For this benchmark experiment, we did not use a bootstrap approach, because the original study did not report the ethnicities and gender information for each individual and as a result we cannot re-create a balanced experimental sampling design using bootstraps. The uncorrected significance values for each test are reported in Table 3. All the tests, obtained a $P < 0.05$. Consistent with our other benchmarks, VAAST 1.0 and VAAST 2.0 obtained the lowest *P*-value.

Benchmark Analyses on Simulated Datasets

Simulated datasets provide an opportunity to investigate the performance of different approaches on datasets presenting specific challenges; for example, under various PARs or under different degrees of allelic heterogeneity, and in a controlled fashion. For these reasons, we used a previously published simulation framework [Madsen and Browning, 2009] to compare the power of six aggregative variant association tests (see Supporting Information for details).

We first benchmarked the power of these tests under different aggregated PAR [Madsen and Browning, 2009] values, which reflects the aggregated disease risk of all simulated mutations. These results are shown in Figure 3. Under a dominant model, VAAST 2.0 rapidly achieves 80% power with PARs less than 0.04, and achieves a power of 100% when PAR = 0.05. The power of VAAST 2.0 is followed by VAAST 1.0 and VT, both of which exhibit 10–15% lower power than VAAST 2.0 before reaching 80% power. In contrast, SKAT reached 80% power around PAR = 0.06 and WSS after PAR = 0.07. This trend is also seen in the recessive inherence scenario at various PARs (Fig. 3B). Note that in this experiment we assumed an equal number of causal and non-causal mutation sites, but we also explored other proportions (Fig. 4).

Both VAAST 2.0 and WSS can use user-specified inheritance models (e.g., dominant or recessive) to boost power.

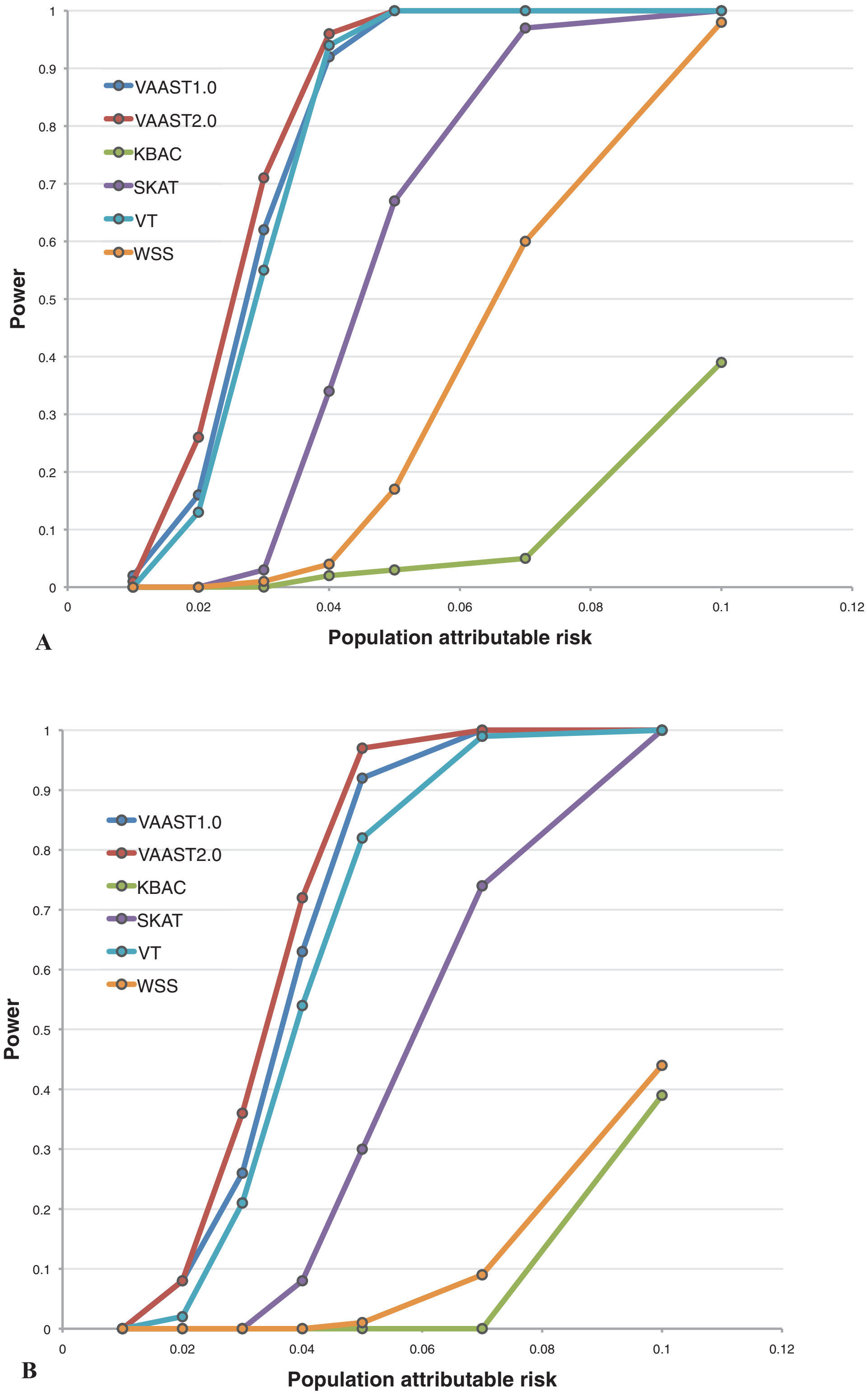

Figure 3. Impact of PAR. Shown is the power of six association tests under different total population attributable risk (PAR) levels. *x*-axis shows the total PAR values from all contributing variants; *y*-axis shows the statistical power based on 100 bootstraps. (A) Dominant model, (B) recessive model. The number of cases and control are set at 1,000, with the number of disease-causal alleles and noncausal alleles both fixed at 50.

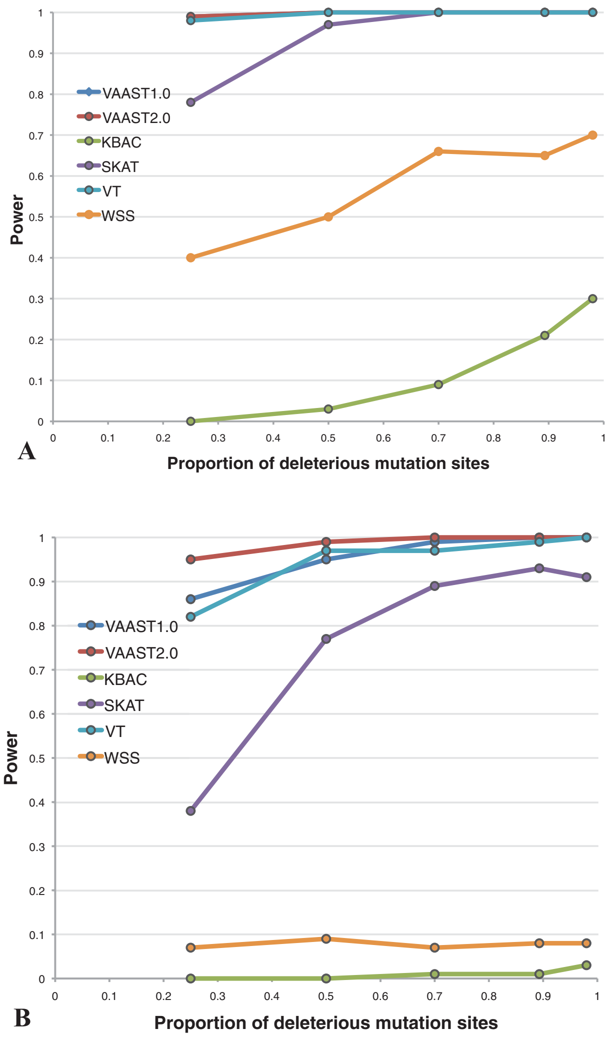

Figure 4. Impact of different proportions of deleterious mutation sites contributing to the disease risk. *x*-axis is the proportion of deleterious mutation sites among all simulated sites; *y*-axis statistical power. (A) Dominant model; (B) recessive model. Total PAR is fixed at 10%; the numbers of case/controls are set at 500; the number of casual variants is 50 with varying number of noncasual variants.

However, for the analyses presented in Figure 3, we did not invoke these options, as (1) the other tests have no such functionalities and (2) the mode of inheritance model is not always known. In the published WSS manuscript [Madsen and Browning, 2009] where genetic model information is used, WSS achieves 80% power at PAR = 0.05 under the recessive model; in contrast, even without genetic model information VAAST 2.0 has a power of 97% at PAR = 0.05.

Next we explored the effect of increasing the number of disease-causal variants (ND) while holding PAR constant in order to model the impact of allelic heterogeneity on the performance of the different algorithms. These results are shown in Figure 5. The actual number of observed causal variant sites per individual is reported in supplementary Table S1. As can be seen, as ND increases, each variant's risk contribution decreases, along with power. For example, under both dominant and recessive inheritance models, when the number of deleterious variants is 150, each individual variant will only have a PAR of 0.07%. Under this model, both VAAST 1.0 and VAAST 2.0 have greater than 80% power. VT with PolyPhen2 scores seems robust to increasing ND values until ND is greater than 100. For SKAT, the power dropped below 80% between ND of 50 and 100 under dominant model and around 50 under recessive model. KBAC and WSS are less robust to increasing ND than the other methods. We summarize the number of cases/controls required for each algorithm to achieve 80% power in Table 4 for ND = 5 and ND = 50.

WSS generally performed quite well, and in many cases outperformed KBAC. We note that the opposite behavior is reported in Liu and Leal [2010]. We believe differences in allelic heterogeneity are responsible for this discrepancy. Because KBAC calculates the sample risk for each multisite genotype, in cases where many different casual alleles or common casual alleles are present, the number of multisite genotypes grows rapidly, with a concomitant loss in power. This behavior can be seen quite clearly in Figure 5. Consistent with this hypothesis, KBAC performs well on the *CHEK2* and *LPL* datasets, but does much worse on the *NOD2* data, likely because *NOD2* contains the highest number of multisite genotypes (Table 2). We tested this hypothesis by comparing the power of WSS and KBAC under different numbers of deleterious alleles (supplementary Fig. S2, Table S2). When ND = 2 and there are less than 10 multisite genotypes, KBAC has 3–5% more power than WSS before it reaches 80% power. However, as the number of multisite genotypes increases with ND, KBAC gradually looses power, and when there are more than 40 multisite genotypes, the power of KBAC is severely compromised. This result is consistent with its performance on the *LPL*, *NOD2*, and *CHEK2* datasets, suggesting that KBAC is probably best suited for analyses of datasets where the number of distinct multisite genotypes is not large, as demonstrated in Figures 2 and 5.

Benchmark Analyses on Rare Mendelian Diseases

VAAST was designed to be a general-purpose disease-gene finder capable of identifying both rare and common alleles responsible for both rare and common diseases [Rope et al., 2011; Yandell et al., 2011]. Although the majority of aggregative variant association tests have been designed for common genetic diseases, there is no a priori reason that they cannot be applied to rare Mendelian diseases. To this end, we benchmarked the six aggregative variant association tests using the benchmarking pipeline from Yandell et al. [2011]. Briefly, this pipeline was employed to randomly select 100 Mendelian disease causal genes from the OMIM database, where each gene has at least six disease-causal variants. For each of these genes, we inserted published, disease-causing variants into from one to three healthy Caucasian genomes sequenced on the Complete Genomics platform [Drmanac et al., 2010] in order to simulate diseased individuals. For the

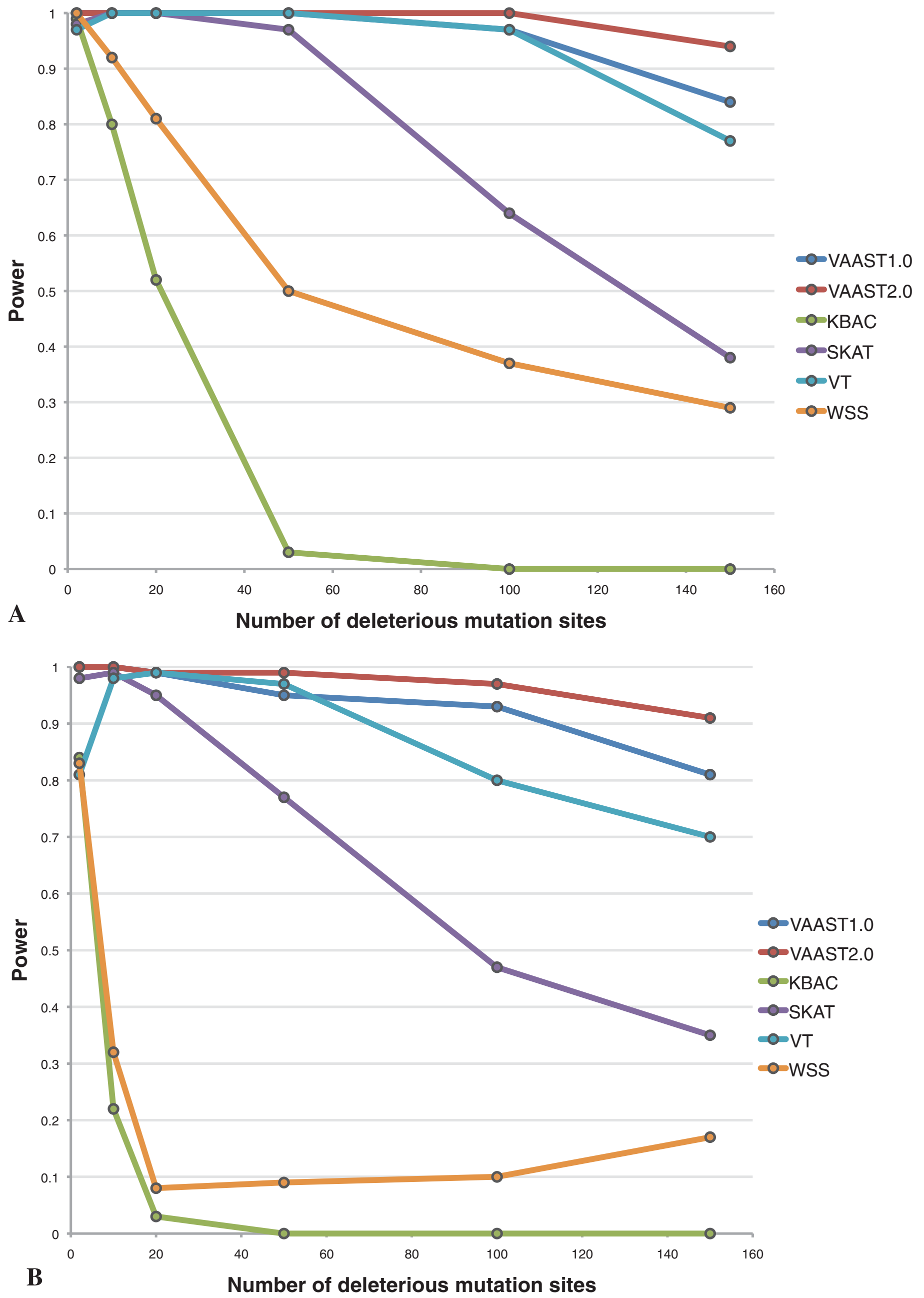

Figure 5. Impact of differing numbers of deleterious mutation sites. *x*-axis is the number of deleterious mutation sites (ND); *y*-axis shows the statistical power based on 100 bootstraps. (A) Dominant model, (B) recessive model. The number of cases and control are set at 500, and the total PAR value is set at 10%.

dominant scenario, we inserted a different single allele into each case genome; for the recessive scenario, we inserted two different alleles into each case genome.

All protein-coding genes are ranked according to the significance of associations between genotypes and dichotomous disease phenotypes. To our knowledge this is the first time that a benchmark of aggregative variant association tests has been conducted on rare Mendelian diseases.

The results are shown in Figure 6. Figure 6 reports the proportion of the 100 OMIM "target" genes falling into four bins based upon rank; these are bin A: 1–10, bin B: 11–100, bin C: 101–1,000, and bin D: greater than 1,000 among all protein coding genes. Supplementary Figure S3 reports the mean values for these same analyses.

For the dominant disease scenario, with only one case genome (one individual), VAAST 2.0 ranked 19% of disease genes among the top 10 candidates genome-wide and 55% as top 100 candidates. Performance improved dramatically as the number of case genomes increases. With only two case genomes, 79% of disease genes are ranked within top 10,

Table 4. Numbers cases and controls required for 80% power in simulations[a]

	Dominant		Recessive	
	ND = 5	ND = 50	ND = 5	ND = 50
VAAST1.0	150	300	300	500
VAAST2.0	150	300	300	400
KBAC	300	>1,000	800	>1,000
SKAT	200	400	300	600
VT	200	300	400	500
WSS	300	700	800	>1,000

[a] Total PAR is set at 10%.

genome-wide; with three case genomes, 90% of disease genes are among top 10. VAAST 2.0's performance is even better under recessive model. For example, with only one case genome 73% of the disease genes are ranked among top 10, and with two cases, 97% of the disease genes ranked among top 10. We note that in this benchmark analysis, the performance of VAAST 2.0 is very similar to VAAST 1.0 in most cases, suggesting that the CASM approach improves performance primarily on datasets containing common causal variants or complex disease cases.

One of the most intriguing aspects of this analysis is the general finding that most association tests do well on these datasets, despite the fact we included common nonsynonymous SNVs in this analysis, which is expected to violate the underlying hypothesis of many burden tests. For example, using top 10 ranking as an empirical significance level with a sample size of 3, under the dominant model, VAAST 2.0 achieves 90% power, WSS 82%, and KBAC 73% power. Under recessive model, VAAST 2.0, achieves 100% power, WSS 99%, and KBAC 74% power. These analyses thus make it clear that some aggregative variant association tests are excellent rare Mendelian disease-gene finders (e.g., WSS, VT, and KBAC), despite having been developed for common, multigenetic diseases. For purposes of comparison, we also assayed the performance of SIFT and ANNOVAR for rare disease disease-gene identification [Ng and Henikoff, 2006; Wang et al., 2010]. As would be expected, SIFT does poorly compared to the other tests. ANNOVAR, as a state-of-art approach for the identification of rare Mendelian disease genes, compares favorably to SIFT but in general does not perform well relative to association tests (see supplementary Fig. S3). For example, the power with only one recessive case using VAAST 2.0 and ANNOVAR are 73% and 14%, respectively. This indicates filtering based approaches are suboptimal for the identification of rare Mendelian disorders, when compared to more sophisticated rare-variant association tests. This is true even when the case sample size is as small as one genome.

Inflation of FPR in Stratified Populations

A potential problem with disease-gene discovery using rare-variant association tests is that if cases and controls are sampled from a stratified population, an inflation of the FPR can result if the proportion of samples from each subpopulation differs between cases and controls [Nelson et al., 2012]. This situation can occur for a variety of reasons, including biased sampling procedures and differences in disease incidence rates among subpopulations. To explore the behavior of rare-variant association tests when cases and controls are not equally stratified, we simulated case-control data for 1,000 cases and 1,000 controls from the 202 drug target genes presented in Nelson et al. [2012]. We evaluated three scenarios: cases and controls from a balanced mixture of two subpopulations (western and northwestern Europeans); modest levels of imbalance between cases and controls (ratio of western to northwestern European genomes of 3:7 in controls and 7:3 in cases), and high levels of imbalance (ratio of 1:9 in controls and 9:1 in cases). As expected, we did not observe an inflation of the FPR in the balanced scenario. Most tests behaved reasonably well when the imbalance between controls and cases was modest, but all tests suffered from substantial FPR inflation when the imbalance was high. SKAT can incorporate demographic information as a covariate matrix, which eliminated the population stratification problem (supplementary Fig. S4). In theory, KBAC can incorporate a covariate matrix as well, although this functionality is not yet available in the KBAC public release. VAAST does not currently provide a solution for this issue. These results indicate that a careful examination of potential population stratification issues is warranted in large studies.

Discussion

Phylogenetic conservation is a valuable source of information for distinguishing between benign and disease-causing variation. Determining the best way to make use of this information—for purposes of variant prioritization and for association testing—is however, still an open question. Variant prioritization tools, such as SIFT [Ng and Henikoff, 2006], use multiple alignments of homologous proteins and judge a human variant damaging if it alters a highly conserved amino acid. PolyPhen-2 goes one step further, making use of protein structural information where available [Adzhubei et al., 2010]. Because SIFT and PolyPhen-2 rely on conservation information, they can only score variants at conserved positions. VAAST 1.0 implemented a different approach. Rather than evaluating individual columns of multiple alignments in order to judge the impact of a coding variant, VAAST 1.0 used the global, genome-wide frequency of observing a given AAS in any gene, anywhere in the genome. In this regard, the AAS matrix in VAAST 1.0 shared some similarity to Grantham's matrix [Grantham, 1974], which quantifies the property differences between amino acid changes, except that VAAST 1.0 empirically estimated the likelihood of being deleterious or neutral for every AAS using known disease-causing alleles. This means that VAAST could score every coding change, regardless of whether or not a particular gene, or that a particular region of its protein is conserved. Although this approach casts a wider net, VAAST 1.0 was unable to take advantage of position-specific conservation information. Thus, the basic motivation of this work has been to preserve the

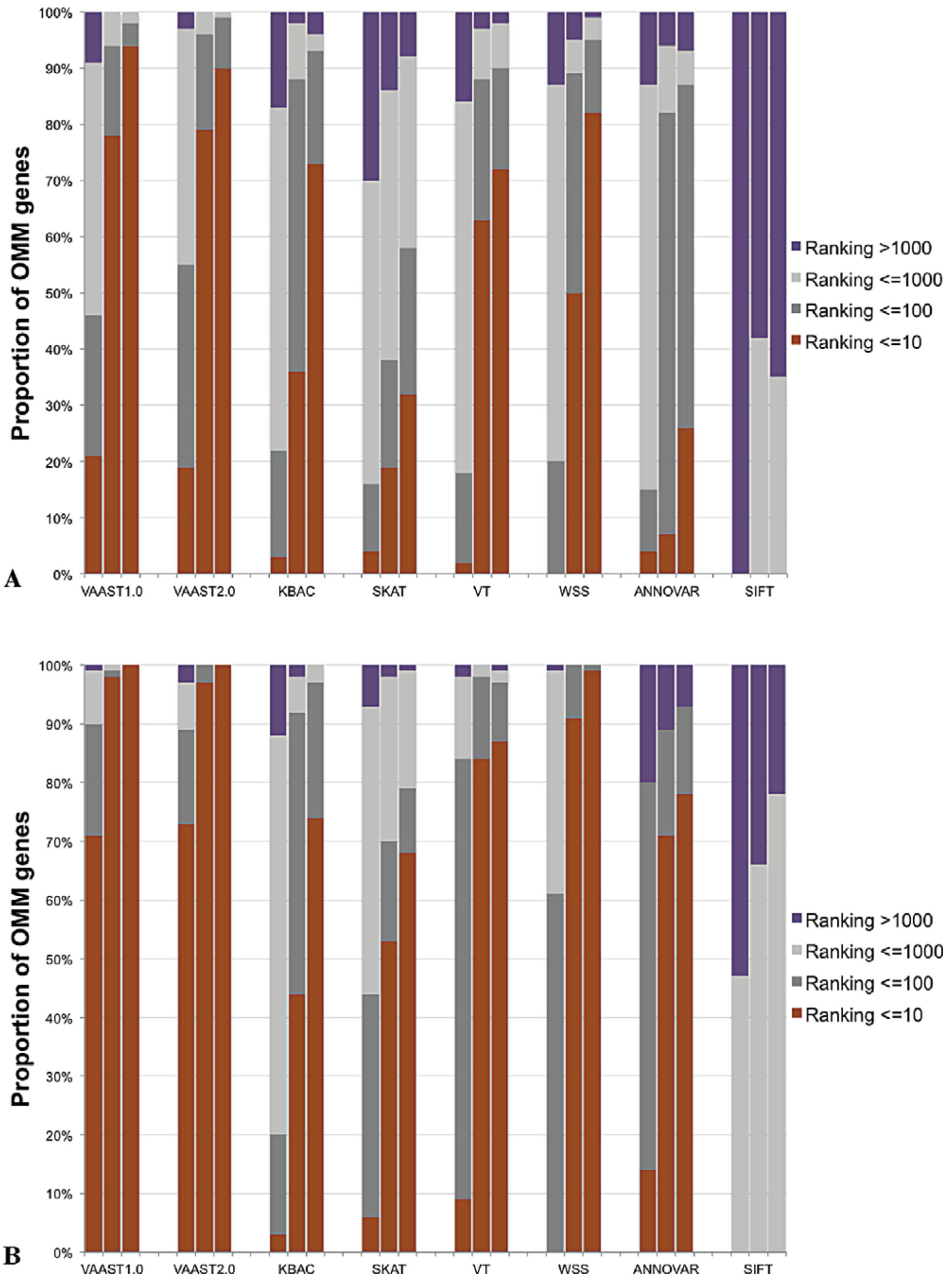

Figure 6. Rankings for 100 different genome-wide searches for known rare disease genes. Panels (A) and (B) shows dominant and recessive models, respectively. The different colors denote the proportion of the 100 OMIM "target" genes falling into four bins based upon genome-wide rank (see insert legend), with orange, denoting the percentage cases for which the disease gene was ranked among the top 10 candidates genome-wide. Dominant and recessive disease scenarios are investigated separately. To model the dominant diseases, one causal variant was inserted into the gene of interest, and in the recessive cases two different alleles are inserted (per case genome). For each algorithm, three columns are shown, corresponding one individual, two individuals, and three individuals as cases.

ability to provide informative scores for every coding variant in the genome while taking advantage of the detailed information provided by multiple sequence alignments wherever they are available. CASM evaluates the global frequency of AAS in the whole proteome, and thus is still capable of making an inference even if there are few or no local alignments. Although the PhastCons scores at these positions are less informative, CASM can still infer severity from type of the amino acid change. As Figure 1 demonstrates, the CASM approach provides an effective solution to this problem, granting VAAST 2.0 a significant advantage in variant prioritization compared to other state-of-the-art tools.

VAAST 2.0, however, is more than a tool for variant prioritization; it is also a tool for genome-wide searches. As such, VAAST is one of several aggregative variant association tests published in the last few years [Li and Leal, 2008; Liu and Leal, 2010; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Price et al., 2010; Wu et al., 2011]. Although several benchmarks have been published [Ladouceur et al., 2012; Liu and Leal, 2010; Madsen and Browning, 2009; Price et al., 2010; Wu et al., 2011], ours is the first to systematically compare the performance of these methods across heterogeneous disease datasets—both real and simulated, and for both common and rare diseases. VAAST 2.0 consistently outperforms VAAST 1.0, WSS, VT, KBAC, and SKAT in these analyses, but performance advantages vary across the datasets. Indeed, an important conclusion of our benchmarking analyses is that no single dataset—real or simulated—is sufficient for benchmarking aggregative variant association tests because of the complex behaviors exhibited by these tools. Figure 2 provides an excellent case in point. Collectively, our analyses show how three basic characteristics of case-control datasets impact the performance of the different tools. These are (1) the number of disease-casual alleles; (2) their allele frequencies; and (3) their collective PAR.

The performance curves for KBAC and SKAT serve to highlight the general sensitivity of all the association tests to these three factors. KBAC, for instance, is clearly very sensitive to numbers of deleterious alleles at a given PAR (Fig. 5). Because KBAC estimates the sample risk for each unique multisite genotype, when the number of multisite genotypes is large and each genotype has a relatively low disease risk, the power of KBAC is compromised. KBAC's poor performance on the *NOD2* dataset, compared to its much better power on the *CHEK2* and *LPL* datasets, further illustrates this behavior. The *NOD2* dataset contains 566 unique multisite genotypes, including a single common variant (MAF 27.7%) that explains 47% of the total PAR of this dataset. In contrast, the *LPL* and *CHEK2* case datasets contain only 14 and 30 distinct genotypes, respectively (Table 2) and all of their deleterious variants are rare.

Although SKAT performed well in our simulation studies, it did less well on the three real datasets. Its performance on the *LPL* and *CHEK2* datasets, for example, suggest that SKAT is not well suited for analyses of datasets having modest numbers of casual variants that contribute to a relatively small total PAR (8.4% for *LPL* and 3.81% for *CHEK2*). To test whether SKATs poor performance on these datasets might be due to the fact that it does not group low-risk rare variants, we used VAAST to group variants in *LPL* and passed this information to SKAT at run-time. This approach improved SKAT's statistical power, from 31% to 45% at maximal sample size (see supplementary Table S3). The fact that SKAT is a supervised method, requiring users to choose kernels and weights, also presents challenges, as the default parameters can be suboptimal for certain cases. This additional complication is demonstrated, for example, on the *NOD2* dataset. SKAT's default weight resulted in low power (<40% at sample size of 450) because it severely down-weighted common variants, which contribute to a large proportion of disease risk in this dataset. For this reason we used a beta weight value of (1,1) for SKAT for the *NOD2* data, which greatly improved its performance.

In contrast to the other tools, VT and VAAST, when run on simulated data, exhibited very robust and similar performance across a wide range of PARs and allelic heterogeneities, at varying ratios of disease-causing and neutral alleles, and under both dominant and recessive modes of inheritance (Figs. 3–5). These strengths likely result from two features both tools share in common. First, they directly compare the variant MAF between cases and controls at each site to weight variants. Second, they make use of external predictors of variant function to improve power [Price et al., 2010].

Despite their similar performance characteristics on simulated data, VAAST outperformed VT on the real datasets. One possible explanation for this fact is that VAAST 2.0 employs a more flexible variant-weighting method, one that does not rely on a priori assumptions about variant severity and MAF. VT, in contrast, assumes that less frequent variants are more likely to be deleterious, and that a single optimal MAF threshold exists. It thus explores all possible thresholds to find the MAF that maximizes the contrast between cases and controls [Price et al., 2010]. This assumption is valid for our simulated datasets, but not necessarily true for the real disease dataset. Consistent with these observations, VAAST is the best overall performing tool on every dataset—simulated and real—demonstrating that VAAST 2.0 can cope effectively with the diverse parameter spaces that characterize real case-control datasets.

One limitation of VAAST is that it is currently only supports the analysis of dichotomous traits in a case-control study design, while WSS, VT, KBAC, and SKAT support quantitative traits. For continuous phenotypes, it is possible to categorize samples into two groups (perhaps using only phenotypic extremes), although doing so will often compromise statistical power.

With the exception of VAAST, which has been shown effective for rare-disease gene searches as well [Rope et al., 2011], the aggregative variant association tests benchmarked here were developed to identify genes involved in common disease. Our analyses demonstrate that these tests are also applicable to the identification of rare Mendelian disease genes. For example, WSS, VAAST 2.0, and KBAC ranked the disease gene in the top 10 genes genome-wide 99%, 100%, and 74% of the time, respectively, using only three case genomes under a recessive model. The performance compares very favorably to traditional filtering-based approaches, suggesting the necessity of employing association tests in these scenarios.

Conclusions

Collectively, our analyses illustrate the unexpectedly complex performance characteristics of aggregative variant association tests. They also demonstrate that VAAST 2.0 is a powerful disease-gene finder that performs robustly across a wide variety of simulated and real-world case-control datasets.

Software and Data Access

VAAST 2.0 is available for download at http://www.yandell-lab.org/software/vaast.html with an academic user license.

Acknowledgments

This work was supported by NIH SBIR grant 1R4HG003667 to M.G.R. and M.Y., NIH ARRA GO grant 1RC2HG005619 to M.Y. and M.G.R., all administered by the National Human Genome Research Institute (NHGRI). C.D.H. was supported by NIH grants R01 DK091374, R01 CA164138, and 1T32HL105321-01. H.H. was supported by the MD Anderson Cancer Center Odyssey Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank the Center for High Performance Computing at the University of Utah for an allocation of computer time. The authors have declared that no competing interests exist.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* 7(4):248–249.

Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Collins FS, Vega FMDl, Donnelly P, Egholm M and others. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061–1073.

Cooper DN, Ball EV, Krawczak M. 1998. The human gene mutation database. *Nucleic Acids Res* 26(1):285–287.

Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G and others. 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327(5961):78–81.

Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, Tavtigian SV, Monteiro AN, Iversen ES, Couch FJ and others. 2007. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. *Am J Hum Genet* 81(5):873–883.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185(4154):862–864.

Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, Martins RA, Kennedy BA, Hassell RG, Visser ME and others. 2010. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat Genet* 42(8):684–687.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32(Database issue):D493–D496.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 80(4):727–739.

Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB. 2012. The empirical power of rare variant association methods: results from Sanger sequencing in 1,998 individuals. *PLoS Genet* 8(2):e1002496.

Le Calvez-Kelm F, Lesueur F, Damiola F, Vallee M, Voegele C, Babikyan D, Durand G, Forey N, McKay-Chopin S, Robinot N and others. 2011. Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study. *Breast Cancer Res* 13(1):R6.

Lesage S, Zouali H, Cezard JP, Colombel JF, Belaiche J, Almer S, Tysk C, O'Morain C, Gassull M, Binder V and others. 2002. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* 70(4):845–857.

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

Liu DJ, Leal SM. 2010. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet* 6(10):e1001156.

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

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 615(1–2):28–56.

Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D and others. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337(6090):100–104.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 7:61–80.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86(6):832–838.

Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K. 2010. A standard variation file format for human genome sequences. *Genome Biol* 11(8):R88.

Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen JC. 2009. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *J Clin Invest* 119(1):70–79.

Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, Johnson WE, Moore B, Huff CD, Bird LM and others. 2011. Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am J Hum Genet* 89(1):28–43.

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 7(8):575–576.

Smigielski EM, Sirotkin K, Ward M, Sherry ST. 2000. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 28(1):352–355.

Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. 2006. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* 43(4):295–305.

Thorisson GA, Smith AV, Krishnan L, Stein LD. 2005. The International HapMap Project Web site. *Genome Res* 15(11):1592–1593.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* 90(1):7–24.

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

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82–93.

Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, Jorde LB, Reese MG. 2011. A probabilistic disease-gene finder for personal genomes. *Genome Res* 21(9):1529–1542.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139(2):993–1005.