CSH PRESS

GENOME RESEARCH

# A probabilistic disease-gene finder for personal genomes

Mark Yandell, Chad Huff, Hao Hu, et al.

*Genome Res.* 2011 21: 1529-1542 originally published online June 23, 2011
Access the most recent version at doi:10.1101/gr.123158.111

# A probabilistic disease-gene finder for personal genomes

Mark Yandell,[1,3,4] Chad Huff,[1,3] Hao Hu,[1,3] Marc Singleton,[1] Barry Moore,[1] Jinchuan Xing,[1] Lynn B. Jorde,[1] and Martin G. Reese[2]

*[1]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah and School of Medicine, Salt Lake City, Utah 84112, USA; [2]Omicia, Inc., Emeryville, California 94608, USA*

**VAAST (the Variant Annotation, Analysis & Search Tool) is a probabilistic search tool for identifying damaged genes and their disease-causing variants in personal genome sequences. VAAST builds on existing amino acid substitution (AAS) and aggregative approaches to variant prioritization, combining elements of both into a single unified likelihood framework that allows users to identify damaged genes and deleterious variants with greater accuracy, and in an easy-to-use fashion. VAAST can score both coding and noncoding variants, evaluating the cumulative impact of both types of variants simultaneously. VAAST can identify rare variants causing rare genetic diseases, and it can also use both rare and common variants to identify genes responsible for common diseases. VAAST thus has a much greater scope of use than any existing methodology. Here we demonstrate its ability to identify damaged genes using small cohorts ($n$ = 3) of unrelated individuals, wherein no two share the same deleterious variants, and for common, multigenic diseases using as few as 150 cases.**

**[Supplemental material is available for this article.]**

The past three decades have witnessed major advances in technologies for identifying disease-causing genes. As genome-wide panels of polymorphic marker loci were developed, linkage analysis of human pedigrees identified the locations of many Mendelian disease-causing genes (Altshuler et al. 2008; Lausch et al. 2008). With the advent of SNP microarrays, the principle of linkage disequilibrium was used to identify hundreds of SNPs associated with susceptibility to common diseases (Wellcome Trust Case Control Consortium 2007; Manolio 2009). However, the causes of many genetic disorders remain unidentified because of a lack of multiplex families, and most of the heritability that underlies common, complex diseases remains unexplained (Manolio et al. 2009).

Recent developments in whole-genome sequencing technology should overcome these problems. Whole-genome (or exome) sequence data have indeed yielded some successes (Choi et al. 2009; Lupski et al. 2010; Ng et al. 2010; Roach et al. 2010), but these data present significant new analytic challenges as well. As the volume of genomic data grows, the goals of genome analysis itself are changing. Broadly speaking, discovery of sequence dissimilarity (in the form of sequence variants) rather than similarity has become the goal of most human genome analyses. In addition, the human genome is no longer a frontier; sequence variants must be evaluated in the context of preexisting gene annotations. This is not merely a matter of annotating nonsynonymous variants, nor is it a matter of predicting the severity of individual variants in isolation. Rather, the challenge is to determine their aggregative impact on a gene's function, a challenge unmet by existing tools for genome-wide association studies (GWAS) and linkage analysis.

Much work is currently being done in this area. Recently, several heuristic search tools have been published for personal genome data (Pelak et al. 2010; Wang et al. 2010). Useful as these tools are, the need for users to specify search criteria places hard-to-quantify limitations on their performance. More broadly, applicable probabilistic approaches are thus desirable. Indeed, the development of such methods is currently an active area of research. Several aggregative approaches such as CAST (Morgenthaler and Thilly 2007), CMC (Li and Leal 2008), WSS (Madsen and Browning 2009), and KBAC (Liu and Leal 2010) have recently been published, and all demonstrate greater statistical power than existing GWAS approaches. But as promising as these approaches are, to date they have remained largely theoretical. And understandably so: creating a tool that can use these methods on the very large and complex data sets associated with personal genome data is a separate software engineering challenge. Nevertheless, it is a significant one. To be truly practical, a disease-gene finder must be able to rapidly and simultaneously search hundreds of genomes and their annotations.

Also missing from published aggregative approaches is a general implementation that can make use of Amino Acid Substitution (AAS) data. The utility of AAS approaches for variant prioritization is well established (Ng and Henikoff 2006); combining AAS approaches with aggregative scoring methods thus seems a logical next step. This is the approach we have taken with the Variant Annotation, Analysis & Search Tool (VAAST), combining elements of AAS and aggregative approaches into a single, unified likelihood framework. The result is greater statistical power and accuracy compared to either method alone. It also significantly widens the scope of potential applications. As our results demonstrate, VAAST can assay the impact of rare variants to identify rare diseases, and it can use both common and rare variants to identify genes involved in common diseases. No other published tool or statistical methodology has all of these capabilities.

To be truly effective, a disease-gene finder also needs many other practical features. Since many disease-associated variants are located in noncoding regions (Hindorff et al. 2009), a disease-gene finder must be able to assess the cumulative impact of variants in

**[3]These authors contributed equally to this work.**
**[4]Corresponding author.**
**E-mail myandell@genetics.utah.edu.**

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.123158.111. Freely available online through the *Genome Research* Open Access option.

both coding and noncoding regions of the genome. A disease-gene finder must also be capable of dealing with low-complexity and repetitive genome sequences. These regions complicate searches of personal genomes for damaged genes, as they can result in false-positive predictions. The tool should also be capable of using pedigree and phased genome data, as these provide powerful additional sources of information. Finally, a disease-gene finder should have the same general utility that has made genomic search tools such as BLAST (Altschul et al. 1990; Korf et al. 2003), GENSCAN (Burge and Karlin 1997), and GENIE (Reese et al. 2000) so successful: It must be portable, easily trained, and easy to use; and, ideally, it should be an ab initio tool, requiring only very limited user-specified search criteria. Here we show that VAAST is such a tool.

We demonstrate VAAST's ability to identify both common and rare disease-causing variants using several recently published personal genome data sets, benchmarking its performance on more than 100 Mendelian conditions including congenital chloride diarrhea (Choi et al. 2009) and Miller syndrome (Ng et al. 2010; Roach et al. 2010). We also show that VAAST can identify genes responsible for two common, complex diseases, Crohn disease (Lesage et al. 2002) and hypertriglyceridemia (Johansen et al. 2010).

Collectively, our results demonstrate that VAAST provides a highly accurate, statistically robust means to rapidly search personal genome data for damaged genes and disease-causing variants in an easy-to-use fashion.

# Results

## VAAST scores

VAAST combines variant frequency data with AAS effect information on a feature-by-feature basis (Fig. 1) using the likelihood ratio (λ) shown in Equations 1 and 2 in Methods. Importantly, VAAST can make use of both coding and noncoding variants when doing so (see Methods). The numerator and denominator in Equation 1 give the composite likelihoods of the observed genotypes for each feature under a healthy and disease model, respectively. For the healthy model, variant frequencies are drawn from the combined control (background) and case (target) genomes ($p_i$ in Eq. 1); for the disease model, variant frequencies are taken separately from the control genomes ($p_i^U$ in Eq. 2) and the case genomes file ($p_i^A$ in Eq. 1), respectively. Similarly, genome-wide Amino Acid Substitution (AAS) frequencies are derived using the control (background) genome sets for the healthy model; for the disease model, these are based either on the frequencies of different AAS observed for OMIM (Yandell et al. 2008) alleles or from the BLOSUM (Henikoff and Henikoff 1992) matrix, depending on user preference. Figure 2 shows the degree to which AAS frequencies among known disease-causing alleles in OMIM and AAS frequencies in healthy personal genomes differ from the BLOSUM model of amino acid substitution frequencies. As can be seen, the AAS frequency spectra of these data sets differ markedly from one another. The differences are most notable for stop codons, in part because stop gains and losses are never observed in the multiple protein alignments used by AAS methods and LOD-based scoring schemes such as BLOSUM (Henikoff and Henikoff 1992).

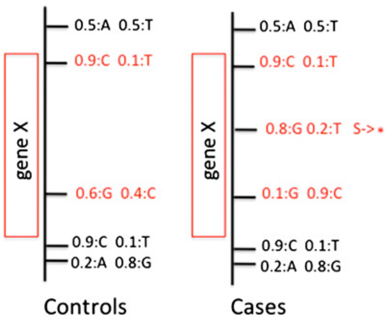


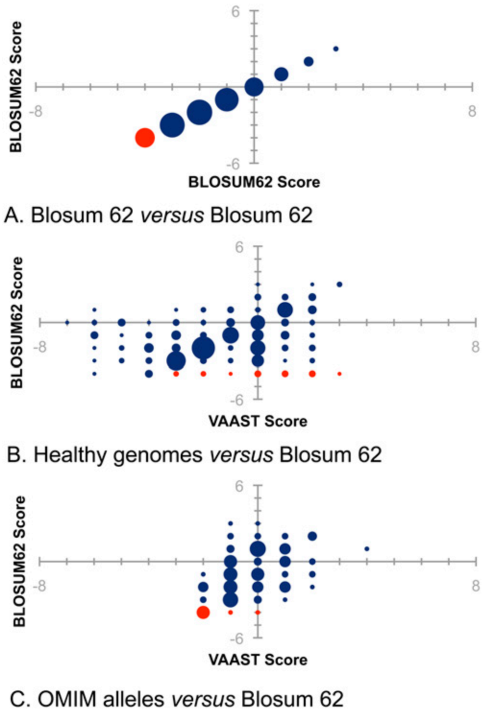
**Figure 1.** VAAST uses a feature-based approach to prioritization. Variants along with frequency information, e.g., 0.5:A 0.5:T, are grouped into user-defined features (red boxes). These features can be genes, sliding windows, conserved sequence regions, etc. Variants within the bounds of a given feature (shown in red) are then scored to give a composite likelihood for the observed genotypes at that feature under a healthy and disease model by comparing variant frequencies in the cases (target) compared to control (background) genomes. Variants producing nonsynonymous amino acid changes are simultaneously scored under a healthy and disease model.

VAAST aggregately scores variants within genomic features. In principle, a feature is simply one or more user-defined regions of the genome. The analyses reported here use protein-coding human gene models as features. Each feature's significance level is the one-tailed probability of observing λ, which is estimated from a randomization test that permutes the case/control status of each individual. For the analyses reported below, the genome-wide statistical significance level (assuming 21,000 protein-coding human genes) is 0.05/21,000 = 2.4 × $10^{-6}$.

## Comparison to AAS approaches

Our approach to determining a variant's impact on gene function allows VAAST to score a wider spectrum of variants than existing AAS methods (Lausch et al. 2008) (for more details, see Eq. 2. in Methods). SIFT (Kumar et al. 2009), for example, examines nonsynonymous changes in human proteins in the context of multiple alignments of homologous proteins from other organisms. Because not every human gene is conserved and because conserved genes often contain unconserved coding regions, an appreciable fraction of nonsynonymous variants cannot be scored by this approach. For example, for the genomes shown in Table 2, ~10% of nonsynonymous variants are not scored by SIFT due to a lack of conservation. VAAST, on the other hand, can score all nonsynonymous variants. VAAST can also score synonymous variants and variants in noncoding regions of genes, which typically account for the great majority of SNVs (single nucleotide variants) genome-wide. Because AAS approaches such as SIFT cannot score these variants, researchers typically either exclude them from the search entirely or else impose a threshold on the variants' frequencies as observed in dbSNP or in the 1000 Genomes Project data set (The 1000 Genomes Project Consortium 2010). VAAST takes a more rigorous, computationally tractable approach: The VAAST score assigned to a noncoding variant is not merely the reciprocal of the variant's frequency; rather, the noncoding variant's score is a log-likelihood ratio that incorporates an estimate of the severity of the substitution as well as the allele frequencies in the control and case genomes (for details, see Scoring Noncoding Variants section in Methods).

To illustrate the consequences of VAAST's novel approach to nonsynonymous variant scoring, we compared it to two widely used tools for variant prioritization, SIFT (Kumar et al. 2009) and ANNOVAR (Wang et al. 2010). Using a previously published data set of 1454 high-confidence known disease-causing and predisposing coding variants from OMIM (Yandell et al. 2008), we asked what fraction were identified as deleterious by each tool. SIFT correctly identified 69% of the disease-causing variants ($P < 0.05$),



A. Blosum 62 *versus* Blosum 62

B. Healthy genomes *versus* Blosum 62

C. OMIM alleles *versus* Blosum 62

**Figure 2.** Observed amino acid substitution frequencies compared to BLOSUM62. Amino acid substitution frequencies observed in healthy and reported for OMIM disease alleles were converted to LOD-based scores for purposes of comparison to BLOSUM62. The BLOSUM62 scores are plotted on the *y*-axis throughout. (Red circles) stops; (blue circles) all other amino acid changes. The diameter of the circles is proportional to the number of changes with that score in BLOSUM62. (*A*) BLOSUM62 scoring compared to itself. Perfect correspondence would produce the diagonally arranged circles shown. (*B*) Frequencies of amino acid substitutions in 10 healthy genomes compared to BLOSUM62. (*C*) OMIM nonsynonymous variant frequencies compared to BLOSUM62.

ANNOVAR (Wang et al. 2010) identified 71%, and VAAST identified 98.0% (Table 1). We then carried out the same analysis using 1454 nonsynonymous variants, randomly drawn from five different European-American (CEU) genome sequences by the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010). These variants are unlikely to be disease-causing given that the individuals are healthy adults. SIFT incorrectly identified 18% of the "healthy" variants as deleterious ($P < 0.05$), ANNOVAR (Wang et al. 2010) identified 1%, and VAAST identified 8%. Under the assumption that there are 1454 true positives and an equal number of true negatives, these two analyses indicate that overall the accuracy [(Sensitivity + Specificity/2)] of SIFT was 75%, ANNOVAR 85%, and VAAST 95% (Table 1). Figure 5C below provides a comparison of the same three tools in the context of genome-wide disease-gene hunts.

We also used these data to investigate the relative contribution of AAS and variant frequency information to VAAST's allele prioritization accuracy. Running VAAST without using any AAS information, its accuracy decreased from 95% to 80%, demonstrating that the AAS information contributes significantly to VAAST's accuracy in identifying deleterious alleles.

## Population stratification

The impact of population stratification on VAAST's false-positive rate is shown in Figure 3A (red line). In this test we used 30 European-American genomes as a background file and various mixtures of 30 European-American and Yoruban (African) genomes as targets. We then ran VAAST on these mixed data sets and observed the number of genes with VAAST scores that reached genome-wide significance, repeating the process after replacing one of the target or background genomes with a Yoruban genome from the 1000 Genomes data set (The 1000 Genomes Project Consortium 2010), until the target contained 30 Yoruban genomes and the background set contained 30 European-American genomes. The resulting curve shown in red in Figure 3A thus reports the impact of differences in population stratification in cases and controls on VAAST's false-positive prediction rate. With complete stratification (e.g., all genomes in the target are Yoruban and all background genomes are CEU), 1087 genes have LD-corrected genome-wide statistically significant scores (alpha = $2.4 \times 10^{-6}$).

## Platform errors

We also investigated the impact of bias in sequencing platform and variant-calling procedures on false-positive rates, using a similar approach to the one we used to investigate population stratification effects. Here we varied the number of case genomes drawn from different sequencing platforms and alignment/variant-calling pipelines. We began with 30 background genomes drawn from the CEU subset of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010) initial release. All of the selected genomes were sequenced to ~6× and called using the 1000 Genomes Project variant-calling pipeline. The target file in this case consisted of 30 similar 1000 Genomes Project CEU genomes that were not included in the background file. This was the starting point for these analyses. We then ran VAAST and recorded the number of genes with LD-corrected genome-wide statistically significant scores (alpha = $2.4 \times 10^{-6}$), repeating the process after substituting one of the target genomes with a non–1000 Genomes Project European-American (CEU) genome (Reese et al. 2000; Li et al. 2010). We repeated this process 30 times. These results are shown in Figure 3B (red line). Taken together, these results (Fig. 3) quantify the impact of population stratification and the cumulative effects of platform differences, coverage, and variant-calling procedures on false-positive rates and allow comparisons of the relative magnitude of platform-related biases to population stratification effects. With all background genomes from the subset of the 1000 Genomes Project data (The 1000 Genomes Project Consortium

**Table 1.** Variant prioritization accuracy comparisons

| | Percent judged deleterious | | |
|---|---|---|---|
| | SIFT | ANNOVAR | VAAST |
| Diseased | 69% | 71% | 98% |
| Healthy | 18% | 1% | 8% |
| Accuracy | 75% | 85% | 95% |

SIFT, ANNOVAR, and VAAST were run on a collection of 1454 known disease-causing variants (Diseased) and 1454 presumably healthy variants randomly chosen from five different CEU genomes (Healthy). The top portion of the table reports the percentage of variants in both sets judged deleterious by the three tools. The bottom row reports the accuracy of each tool. The filtering criteria used in ANNOVAR excluded all variants present in the 1000 Genomes Project data and dbSNP130 as well as any variant residing in a segmentally duplicated region of the genome. For the "Diseased" category, the VAAST control data set contained 196 personal genomes drawn from the 1000 Genomes Project and 10Gen data sets and dbSNP130. For the "Healthy" category, the VAAST control data set contained 55 other European-American genomes drawn from the 1000 Genomes Project data set (to match the ethnicity of the 1454 CEU alleles).

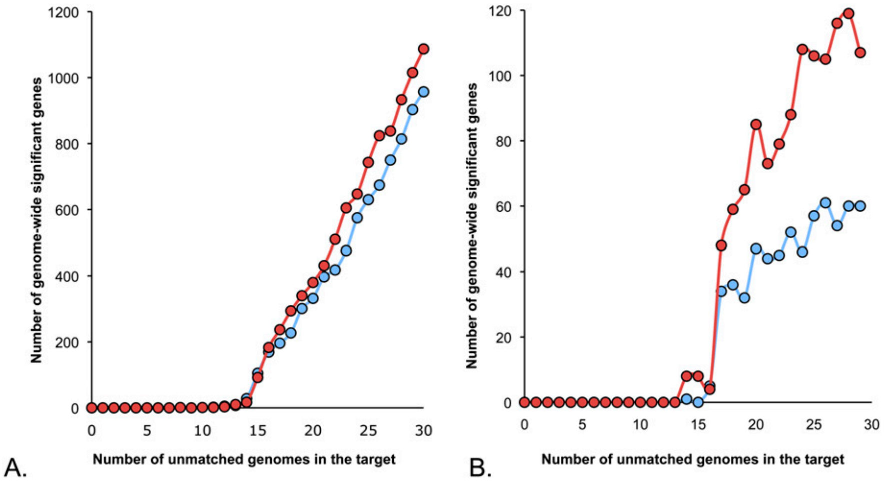


**Figure 3.** Impact of population stratification and platform bias. Numbers of false positives with and without masking. (*A*) Effect of population stratification. (*B*) Effect of heterogeneous platform and variant calling procedures. (Red line) Number of false positives without masking; (blue line) after masking. Note that although masking has little effect on population stratification, it has a much larger impact on platform bias. This is an important behavior: Population stratification introduces real, but confounding signals into disease gene searches; these signals are unaffected by masking (*A*); in contrast, VAAST's masking option removes false positives due to noise introduced by systematic errors in platform and variant calling procedures (*B*).

2010) described above and all target genomes from data sets other than the 1000 Genomes data set (Reese et al. 2000; Li et al. 2010), 107 genes have genome-wide LD-corrected statistically significant scores (alpha = $2.4 \times 10^{-6}$), compared to the 1087 observed in our population stratification experiments (alpha = $2.4 \times 10^{-6}$).

## Variant masking

The limited number of personal genomes available today necessitates comparisons of genomes sequenced on different platforms, to different depths of coverage, and subjected to different variant-calling procedures. As shown in Figure 3B, these factors can be a major source of false positives in disease-gene searches. Based on an analysis of these data, we found variant-calling errors to be overrepresented in low-complexity and repetitive regions of the genome, which is not unexpected. We therefore developed a VAAST runtime option for masking variants within these regions of the genome. VAAST users specify a read length and paired or unpaired reads. VAAST then identifies all variants in non-unique regions of the genome meeting these criteria and excludes them from its calculations. The blue lines in Figure 3 plot the number of genes attaining LD-corrected genome-wide significance after masking. As can be seen, whereas masking has negligible impact on false positives due to population stratification, it has a much larger impact on sequencing platform and variant-calling bias. This is a desirable behavior. Population stratification introduces real, but confounding, signals into disease-gene searches, and these real signals are unaffected by masking (Fig. 3A). In contrast, masking eliminates many false positives due to noise introduced by systematic errors in sequencing platform and variant-calling procedures (Fig. 3B).

## Identification of genes and variants that cause rare diseases

### *Miller syndrome*

Our targets in these analyses were the exome sequences of two siblings affected with Miller syndrome (Ng et al. 2010; Roach et al. 2010). Previous work (Ng et al. 2010; Roach et al. 2010) has shown that the phenotypes of these individuals result from variants in two different genes. The affected siblings' craniofacial and limb malformations arise from compromised copies of *DHODH*, a gene involved in pyrimidine metabolism. Both affected siblings also suffer from primary ciliary dyskinesia as a result of mutations in another gene, *DNAH5*, that encodes a ciliary dynein motor protein. Both affected individuals are compound heterozygotes at both of these loci. Thus, this data set allows us to test VAAST's ability to identify disease-causing loci when more than one locus is involved and the mutations at each locus are not identical by position or descent.

### *Accuracy on the Miller syndrome data*

We carried out a genome-wide search of 21,000 protein-coding genes using the two affected Miller syndrome exomes as targets and using two different healthy-genome background files. The first background file consists of 65 European-American (CEU) genomes selected from the 1000 Genomes Project data (The 1000 Genomes Project Consortium 2010) and the 10Gen data set (Reese et al. 2010). The second, larger background file consists of 189 genomes selected from the same data sources, but, in distinction to the first, is ethnically heterogeneous and contains a mixture of sequencing platforms, allowing us to assay the impact of these factors on VAAST's performance in disease-gene searches. In these experiments, we ran VAAST using its recessive disease model option (for a description of VAAST disease models, see Methods), and with and without its variant-masking option. Depending on whether or not its variant-masking option was used, VAAST identified a maximum of 32, and a minimum of nine, candidate genes. Variant masking, on average, halved the number of candidates (Table 2). The best accuracy was obtained using the larger background file together with the masking option. *DHODH* ranked fourth and *DNAH5* fifth among the 21,000 human genes searched. This result demonstrates that VAAST can identity both disease genes with great specificity using a cohort of only two related individuals, both compound heterozygotes for a rare recessive disease. Overall, accuracy was better using the second, larger background file, demonstrating that, for rare diseases, larger background data sets constructed from a diverse set of populations and sequencing platforms improve VAAST's accuracy, despite the stratification issues these data sets introduce.

We also took advantage of family quartet information (Ng et al. 2010; Roach et al. 2010) to demonstrate the utility of pedigree information for VAAST searches. When run with its pedigree and variant-masking options, only two genes are identified as candidates: *DNAH5* is ranked first, and *DHODH* is ranked second, demonstrating that VAAST can achieve perfect accuracy using only a single family quartet of exomes (Fig. 4). Our previously published analysis (Roach et al. 2010) identified four candidate genes, and further, expert post hoc analyses were required to identify the two actual disease-causing genes. The results shown in Figure 4 thus demonstrate that VAAST can use pedigree data to improve its accuracy, even in the face of confounding signals due to relatedness

**Table 2.** Effect of background file size and stratification on accuracy

| | Genome-wide significant genes | *DHODH* | | *DNAH5* | |
|---|---|---|---|---|---|
| | | Rank | *P*-value | Rank | *P*-value |
| Caucasian only (65 genomes) | | | | | |
| UMSK | 32 | 25 | $7.92 \times 10^{-7}$ | 32 | $1.98 \times 10^{-6}$ |
| MSK | 17 | 14 | $9.93 \times 10^{-7}$ | 19 | $5.79 \times 10^{-5}$ |
| Mixed ethnicities (189 genomes) | | | | | |
| UMSK | 16 | 9 | $6.78 \times 10^{-9}$ | 5 | $2.00 \times 10^{-9}$ |
| MSK | 9 | 4 | $7.60 \times 10^{-9}$ | 5 | $1.18 \times 10^{-8}$ |

Results of searching the intersection of two Utah Miller Syndrome affected genomes against two different background files, with and without masking. (Caucasians only) 65 Caucasian genomes drawn from six different sequencing/alignment/variant calling platforms; (mixed ethnicities) 189 genomes (62 YRI, 65 CAUC, 62 ASIAN), from the 1000 Genomes Project and 10Gen data set; (UMSK) unmasked; (MSK): masked; (genome-wide significant genes) number of genes genome-wide attaining a significant non-LD corrected *P*-value; (rank) gene rank of *DHODH* and *DNAH5* among all scored genes; (*P*-value) non-LD corrected *P*-value; genome-wide significant alpha is $2.4\times10^{-6}$. Data were generated using a fully penetrant, monogenic recessive model. The causative allele incidence was set to 0.00035.

of target exomes, significant population stratification, and platform-specific noise.

### Impact of noncoding SNVs

We used these same data sets to investigate the impact of using both coding and noncoding variants in our searches. To do so, we extended our search to include all SNVs at synonymous codon positions and in conserved DNase hypersensitive sites and transcription factor-binding sites (for details, see Methods). Doing so added an additional 36,883 synonymous and regulatory variants to the 19,249 nonsynonymous changes we screened in the analyses reported above. Using only the two Utah siblings, 189 candidate genes are identified. *DHODH* is ranked 15th and *DNAH5* is sixth among them. Repeating the analysis using family quartet information, 23 candidate genes are identified; *DHODH* is ranked fourth and *DNAH5* is ranked first. Thus, increasing the search space to include almost 37,000 additional noncoding variants had little negative impact on accuracy.

### Impact of cohort size

We also used the Miller syndrome data to assess the ability of VAAST to identify disease-causing genes in very small case cohorts wherein no two individuals in the target data set are related or share the same disease-causing variants. We also wished to determine the extent to which the relatedness of the two siblings introduced spurious signals into the analyses reported in Table 2. We used information from additional Miller syndrome kindreds (Ng et al. 2010; Roach et al. 2010) to test this scenario. To do so, we used a publicly available set of Danish exome sequences (Li et al. 2010). We added two different disease-causing variants in *DHODH* reported in individuals with Miller syndrome (Ng et al. 2010; Roach et al. 2010) to six different Danish exomes to produce six unrelated Danish exomes, each carrying two different Miller syndrome causative alleles. The background file consisted of the same 189 genome equivalents of mixed ethnicities and sequencing platforms used in Table 2. We then used VAAST to carry out a genome-wide screen using these six exomes as targets. We first used one exome as a target, then the union of two exomes as a target, and so on, in order to investigate VAAST's performance in a series of case cohorts containing pools of one to six exomes. The results are shown in Table 3.

*DHODH* is the highest ranked of two candidates for a cohort of three unrelated individuals and the only candidate to achieve LD-corrected genome-wide statistical significance (Table 3). In this data set no two individuals share the same variants, nor are any homozygous for a variant. This data set thus demonstrates VAAST's ability to identify a disease-causing gene in situations in which the gene is enriched for rare variants, but no two individuals in the case data set share the same variant, and the cohort size is as small as three unrelated individuals. VAAST's probabilistic framework also makes it possible to assess the relative contribution of each variant to the overall VAAST score for that gene, allowing users to identify and prioritize for follow-up studies those variants predicted to have the greatest functional impact on a gene. Table 4 shows these scored variants for the Miller syndrome alleles of all six affected individuals.

### Congenital Chloride Diarrhea (CCD) data set

We tested VAAST's ability to identify the genetic variant responsible for a rare recessive disease using the whole-exome sequence of a patient diagnosed with congenital chloride diarrhea (CCD) due to a homozygous D652N mutation in the *SLC26A3* gene (Choi et al. 2009). In this analysis the background data set consisted of 189 European-American genomes (Table 5). Using the single affected exome as a target, *SLC26A3* is ranked 21st genome-wide. We also evaluated the impact of bias in platform and variant-calling procedures on this result. To do so, we added the CCD causative allele as a homozygote to an ethnically matched genome drawn from the 1000 Genomes data set (Table 5; The 1000 Genomes Project Consortium 2010), in the same manner that was used to generate the data in Table 3. Under the assumption that this rare recessive disease is due to variants at the same location in each affected genome (intersection by position), only a single pair of unrelated exomes is required to identify CCD with perfect specificity. Adding a third affected exome is sufficient to obtain LD-corrected genome-wide statistical significance, even when the selection criteria are relaxed to include the possibility of different disease-causing alleles at different positions in different individuals (union of variants by position).

### Impact of recessive modeling on accuracy

We also investigated the impact of VAAST's recessive inheritance model on our rare disease analyses (Supplemental Tables 2, 3). In general, running VAAST with this option yielded improved specificity but had little impact on gene ranks. For a cohort of three unrelated Miller syndrome individuals, the recessive inheritance model had no impact on rank or specificity (Supplemental Table 2). For CCD, using a cohort of three unrelated individuals, *SLC26A3* was ranked first in both cases, but the recessive model decreased the number of candidate genes from seven to two (Supplemental Table 3). These results demonstrate VAAST's ab initio capabilities: It is capable of identifying disease-causing alleles with great accuracy, even without making assumptions regarding mode of inheritance. Our large-scale performance analyses, described below, support and clarify these conclusions.

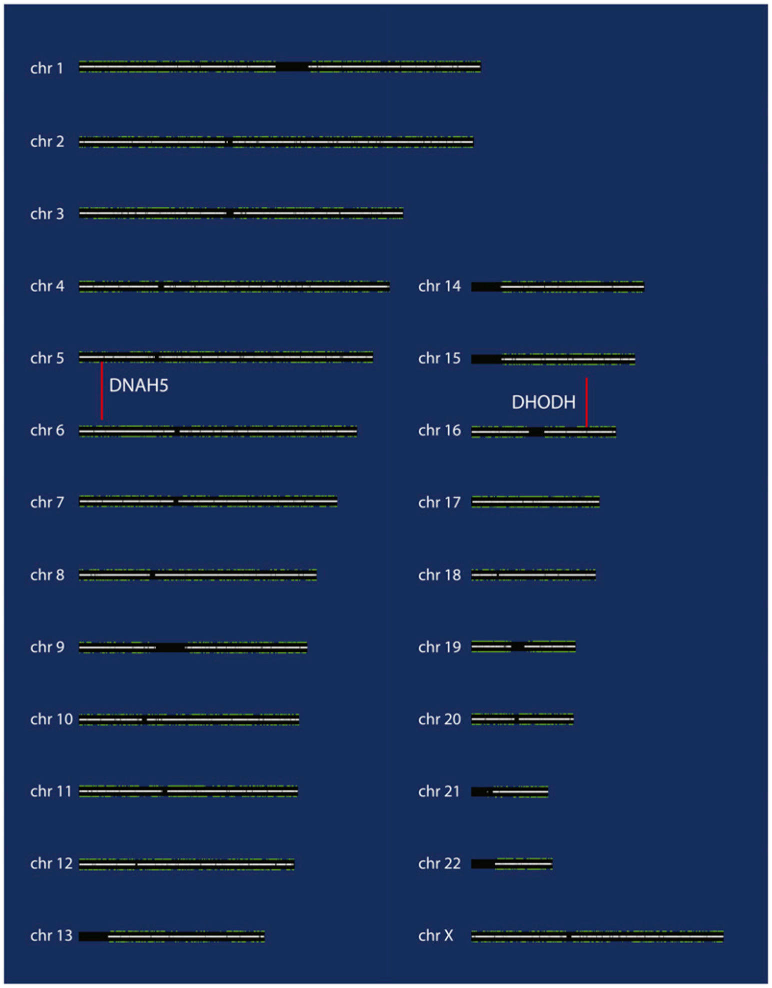


**Figure 4.** Genome-wide VAAST analysis of Utah Miller Syndrome Quartet. VAAST was run in its quartet mode, using the genomes of the two parents to improve specificity when scoring the two affected siblings. Gray bars along the center of each chromosome show the proportion of unique sequence along the chromosome arms, with white denoting completely unique sequence; black regions thus outline centromeric regions. Colored bars above and below the chromosomes (mostly green) represent each annotated gene; plus strand genes are shown above and minus strand genes below; their width is proportional to their length; height of bar is proportional to their VAAST score. Genes colored red are candidates identified by VAAST. Only two genes are identified in this case: *DNAH5* and *DHODH*. Causative allele incidence was set to 0.00035, and amino acid substitution frequency was used along with variant-masking. This view was generated using the VAAST report viewer. This software tool allows the visualization of a genome-wide search in easily interpretable form, graphically displaying chromosomes, genes, and their VAAST scores. For comparison, the corresponding figure, without pedigree information, is provided as Supplemental Figure 1.

### *Benchmark on 100 different known disease genes*

To gain a better understanding of VAAST's performance characteristics, we also evaluated its ability to identify 100 different known disease-causing genes in genome-wide searches. For these analyses, we first randomly selected (without replacement) a known disease-causing gene from OMIM for which there existed at least six different published nonsynonymous disease-causing alleles. See Supplemental File 2 for a complete listing of diseases, genes, and alleles. Next we randomly selected known disease-causing alleles at the selected locus (without replacement) and inserted them at their reported positions within the gene into different whole-genome sequences drawn for the Complete Genomics Diversity Panel (http://www.completegenomics.com/sequence-data/download-data/). We then ran VAAST under a variety of scenarios (e.g., dominant, recessive, and various case cohort sizes) and recorded the rank of the disease gene, repeating the analyses for 100 different known disease genes. We also compared the performance of VAAST to SIFT and ANNOVAR using these same data sets. (Details of the experimental design can be found in the Methods section.) The results of these analyses are shown in Figure 5. In this figure the height of each box is proportional to the mean rank of the disease-causing gene for the 100 trials, and the number shown above each box is the mean rank from among 17,293 RefSeq genes. The error bars delimit the spread of the ranks, with 95% of the runs encompassed within the bars.

Figure 5A summarizes VAAST's performance on this data set under both dominant and recessive disease scenarios. For these experiments, we assayed the average rank for three different cohort sizes: two, four, and six individuals for the dominant scenario, and one, two, and three individuals for the recessive analyses. For both scenarios, the mean and variance rapidly decrease as the cohort size increases. For the dominant scenario, using a case cohort of six unrelated individuals, each carrying a different disease-causing allele, VAAST ranked the disease-causing gene on average ninth out of 17,293 candidates with 95% of the runs having ranks between 5 and 40 in 100 different genome-wide searches. For the recessive scenario, using a case cohort of three unrelated individuals each carrying two different disease-causing variants at different positions (all compound heterozygotes), VAAST ranked the disease-causing gene on average third out of 17,293 candidates, with 95% of the runs having ranks between 2 and 10. None of the individuals had any disease-causing alleles in common.

Figure 5B summarizes VAAST's performance when only a subset of the case cohort contains a disease-causing allele, which could result from (1) no calls at the disease-causing allele during variant calling; (2) the presence of phenocopies in the case cohort; and (3) locus heterogeneity. As can be seen in Figure 5B, averages and variances decrease monotonically as increasing numbers of individuals in the case cohort bear disease-causing alleles in the gene of interest. Moreover, for dominant diseases, even when one-third of the cases lack disease-causing alleles in the selected OMIM disease gene, VAAST achieves an average rank of 61 with 95% of the runs having ranks between 5 and 446. For recessive diseases the average was 21, with 95% of the disease genes ranking between 7 and 136 out of 17,293 genes, genome-wide.

Figure 5C compares VAAST's accuracy to that of ANNOVAR and SIFT. For these analyses, we used the same data used to produce

**Table 3.** **Impact of cohort size on VAAST's ability to identify a rare disease caused by compound heterozygous alleles**

| | Genome-wide | | | *DHODH* rank | | |
|---|---|---|---|---|---|---|
| | | Significant genes | | | *P*-value | |
| Target genome(s) | Genes scored | Non-LD-corrected | LD-corrected | Rank | Non-LD-corrected | LD-corrected |
| 1 Compound heterozygote | 92 | 67 | 0 | 86 | $2.36 \times 10^{-4}$ | $5.26 \times 10^{-3}$ |
| 2 Compound heterozygotes | 4 | 3 | 0 | 2 | $2.81 \times 10^{-8}$ | $5.51 \times 10^{-5}$ |
| 3 Compound heterozygotes | 2 | 2 | 1 | 1 | $2.61 \times 10^{-11}$ | $8.61 \times 10^{-7}$ |
| 4 Compound heterozygotes | 1 | 1 | 1 | 1 | $1.99 \times 10^{-15}$ | $1.78 \times 10^{-8}$ |
| 5 Compound heterozygotes | 1 | 1 | 1 | 1 | $6.95 \times 10^{-15}$ | $4.60 \times 10^{-10}$ |
| 6 Compound heterozygotes | 1 | 1 | 1 | 1 | $5.79 \times 10^{-17}$ | $1.42 \times 10^{-11}$ |

The background file consisted of 189 genomes of mixed ethnicity from the 1000 Genomes Project combined with nine additional genomes of mixed ethnicity and sequencing platforms drawn from the 10Gen genome set (Reese et al. 2010). Causative alleles reported in the six individuals described in Ng et al. (2010) were added to unrelated exomes from re-sequenced individuals from Denmark reported in Li et al. (2010). Data were generated using a fully penetrant monogenic recessive model (see Supplemental Table 2). Causative allele incidence was set to 0.00035 (for details, see Supplemental Table 2), and amino acid substitution frequency was used along with masking of repeats. (Genes scored) Number of genes in the genome with variant distributions consistent with VAAST's fully penetrant monogenic recessive model and causative allele incidence threshold. Scoring was evaluated by permutation by gene and permutation by genome.

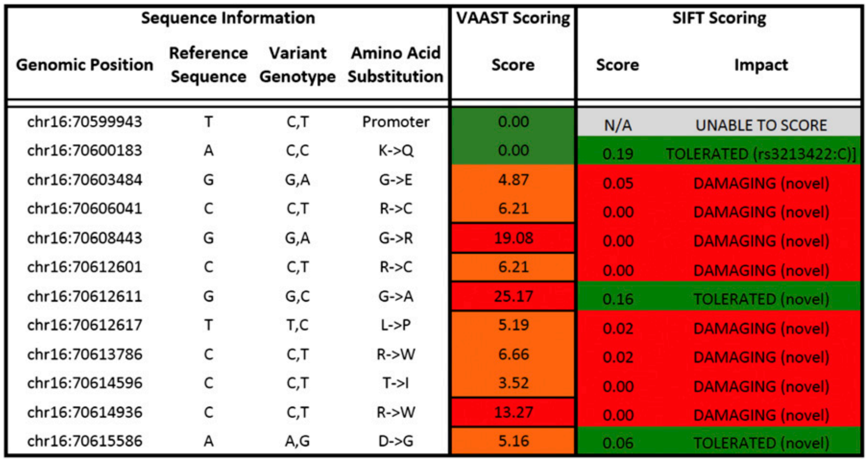
Figure 5A, running all three tools on a case cohort of six and three individuals for the dominant and recessive comparisons, respectively (for details, see Methods). In these analyses, all members of the case cohort contain disease-causing alleles. For ANNOVAR, we set the expected combined disease-allele frequency at <5% (see Methods) as this improved ANNOVAR's performance (data not shown), but for VAAST no prior assumptions were made regarding the disease-causing alleles' frequencies in the population. VAAST outperforms both SIFT and ANNOVAR—both as regards to mean ranks and variances. VAAST, for example, achieves a mean rank of 3 for recessive diseases using three compound heterozygous individuals as a case cohort. SIFT achieves an average rank of 2317, and ANNOVAR an average rank of 529. There is also much less variance in the VAAST ranks than in those of the other tools. For example, in the recessive scenario, using three compound heterozygous individuals as a case cohort, in 95% of the VAAST runs the rank of the disease-causing gene was between ranks 2 and 10. By comparison, ANNOVAR's ranks varied between 67 and 8762 on the same data sets, and SIFT's varied between 66 and 9107. See Supplemental Figures 2 and 3 for the complete distributions. We also investigated the possibility that taking the intersection of ANNOVAR and SIFT calls might improve accuracy compared to either of these tools alone. It did not; see Supplemental Figure 4. Closer inspection of these data revealed the reasons for the high variances characteristic of SIFT and ANNOVAR. In SIFT's case, the variance is due to failure to identify one or more of the disease-causing alleles as deleterious, a finding consistent with our accuracy analysis presented in Table 1. This, coupled with its inability to make use of variant frequencies, means that SIFT also identifies many very frequent alleles genome-wide as deleterious, increasing the rank of the actual disease-causing gene. ANNOVAR's performance, because it can filter candidate variants based on their allele frequencies, is thus better than SIFT's (average rank of 529 vs. 2317). However, its variance from search to search remains high compared to VAAST, as the OMIM alleles in the analysis are distributed across a range of frequencies, and unlike VAAST, ANNOVAR is unable to leverage this information for greater accuracy.

## Identification of genes and variants causing common multigenic diseases

### *Power analyses*

Our goal in these analyses was twofold: first, to benchmark the statistical power of VAAST compared to the standard single nucleotide variation (SNV) GWAS approach; and second, to determine the relative contributions of variant frequencies and amino acid substitution frequencies to VAAST's statistical power. We also compared the statistical power of VAAST's default scoring algorithm to that of WSS (Madsen and Browning 2009), one of the most accurate aggregative methods to date for identifying common disease genes using rare variants. Figure 6A shows the results for the *NOD2* gene, implicated in Crohn's disease (CD) (Lesage et al. 2002). This data set contains both rare (minor allele frequency [MAF] <5%) and common variants. Figure 6B shows the same power analysis

**Table 4.** **Relative impacts of observed variants in *DHODH***



| Sequence Information | | | | VAAST Scoring | SIFT Scoring | |
|---|---|---|---|---|---|---|
| Genomic Position | Reference Sequence | Variant Genotype | Amino Acid Substitution | Score | Score | Impact |
| chr16:70599943 | T | C,T | Promoter | 0.00 | N/A | UNABLE TO SCORE |
| chr16:70600183 | A | C,C | K->Q | 0.00 | 0.19 | TOLERATED (rs3213422:C)] |
| chr16:70603484 | G | G,A | G->E | 4.87 | 0.05 | DAMAGING (novel) |
| chr16:70606041 | C | C,T | R->C | 6.21 | 0.00 | DAMAGING (novel) |
| chr16:70608443 | G | G,A | G->R | 19.08 | 0.00 | DAMAGING (novel) |
| chr16:70612601 | C | C,T | R->C | 6.21 | 0.00 | DAMAGING (novel) |
| chr16:70612611 | G | G,C | G->A | 25.17 | 0.16 | TOLERATED (novel) |
| chr16:70612617 | T | T,C | L->P | 5.19 | 0.02 | DAMAGING (novel) |
| chr16:70613786 | C | C,T | R->W | 6.66 | 0.02 | DAMAGING (novel) |
| chr16:70614596 | C | C,T | T->I | 3.52 | 0.00 | DAMAGING (novel) |
| chr16:70614936 | C | C,T | R->W | 13.27 | 0.00 | DAMAGING (novel) |
| chr16:70615586 | A | A,G | D->G | 5.16 | 0.06 | TOLERATED (novel) |

The "score contribution" column shows the magnitude of impact of each observed variant in *DHODH* to its final score. (Red) Most severe; (green) least severe. For comparison, SIFT values are also shown. Note that SIFT judges two of the known disease-causing alleles as tolerated and is unable to score the noncoding SNV. The target file contains six unrelated individuals with the compound heterozygous variants described in Table 3. The background file consisted of 189 genomes of mixed ethnicity from the 1000 Genomes Project combined with nine additional genomes of mixed ethnicity and sequencing platforms drawn from the 10Gen set (Reese et al. 2010). Data were generated using VAAST's fully penetrant monogenic recessive model and masking. Causative allele incidence was set to 0.00035.

**Table 5.** Impact of cohort size on VAAST's ability to identify a rare recessive disease

| | Genome-wide | | | *SLC26A3* | | |
|---|---|---|---|---|---|---|
| | | Significant genes | | | *P*-value | |
| Target genome(s) | Genes scored | Non-LD-corrected | LD-corrected | Rank | Non-LD-corrected | LD-corrected |
| 1 Homozygote | 127 | 69 | 0 | 21 | $1.22 \times 10^{-5}$ | $5.26 \times 10^{-3}$ |
| Union 2 homozygotes | 7 | 7 | 0 | 3 | $4.74 \times 10^{-10}$ | $5.51 \times 10^{-5}$ |
| Intersection 2 homozygotes | 3 | 3 | 0 | 1 | $7.47 \times 10^{-10}$ | $5.51 \times 10^{-5}$ |
| Union 3 homozygotes | 2 | 2 | 2 | 1 | $2.83 \times 10^{-13}$ | $8.61 \times 10^{-7}$ |
| Intersection 3 homozygotes | 1 | 1 | 1 | 1 | $1.29 \times 10^{-13}$ | $8.61 \times 10^{-7}$ |

The background file consists of 189 genomes of mixed ethnicity from the 1000 Genomes Project combined with nine additional genomes of mixed ethnicity and sequencing platforms drawn from the 10Gen set (Reese et al. 2010). (Targets) The first homozygote affected is the single CCD affected exome reported in Choi et al. (2009); (remaining target genomes) unrelated exomes from re-sequenced individuals from Denmark reported in Li et al. (2010) with the causative allele added. Data were generated on either the union or intersection of affecteds using VAAST's fully penetrant monogenic recessive model. Causative allele incidence was set to 0.013; masking was also used. Scoring was evaluated by non-LD and LD-corrected permutation. (Genes scored) The number of genes in the genome receiving a score >0.

using *LPL*, a gene implicated in hypertriglyceridemia (HTG) (Johansen et al. 2010). This analysis uses a data set of 438 re-sequenced subjects (Johansen et al. 2010). For the *LPL* gene, only rare variants (MAF < 5%) were available; therefore, this analysis tests VAAST's ability to detect disease genes for common diseases in which only rare variants contribute to disease risk. To control for Type I error in this analysis, we applied a Bonferroni correction, with the number of tests approximately equal to the number of genes that would be included in a genome-wide analysis (alpha = $0.05/21{,}000 = 2.4 \times 10^{-6}$).

VAAST rapidly obtains good statistical power even with modest sample sizes; its estimated power is 89% for *NOD2* using as few as 150 individuals (alpha = $2.4 \times 10^{-6}$). By comparison, the power of GWAS is <4% at the same sample size. Notably, for *NOD2*, nearly 100% power is obtained with VAAST when a GWAS would still have <10% power. Also shown is VAAST's power as a function of sample size without the use of amino acid substitution data. The red and blue lines in Figure 6A show the power curves for VAAST using OMIM and BLOSUM, respectively, for its AAS disease models. As can be seen, power is improved when AAS information is used.

In general, the *LPL* results mirror those of *NOD2*. Although VAAST obtained less power using the *LPL* data set compared to *NOD2*, this was true for every approach. Interestingly, for *NOD2*, BLOSUM attains higher power using smaller sample sizes compared to OMIM. The fact that the trend is reversed for *LPL*, however, suggests that the two AAS models are roughly equivalent. We also compared VAAST's performance to that of WSS (Madsen and Browning 2009), another aggregative prioritization method. VAAST achieves greater statistical power than WSS on both data sets, even when VAAST is run without use of AAS information.

## Discussion

VAAST uses a generalized feature-based prioritization approach, aggregating variants to achieve greater statistical search power. VAAST can score both coding and noncoding variants, evaluating the aggregative impact of both types of SNVs simultaneously. In this first study, we have focused on genes, but in principle, the tool can be used to search for disease-causing variants in other classes of features as well; for example, regulatory elements, sets of genes belonging to a particular genetic pathway, or genes belonging to a common functional category, e.g., transcription factors.

In contrast to GWAS approaches, which evaluate the statistical significance of frequency differences for individual variants in cases versus controls, VAAST evaluates the likelihood of observing the aggregate genotype of a feature given a background data set of control genomes. As our results demonstrate, this approach greatly improves statistical power, in part because it bypasses the need for large statistical corrections for multiple tests. In this sense, VAAST resembles several other methods that aggregate variants: CAST (Morgenthaler and Thilly 2007), CMC (Li and Leal 2008), WSS (Madsen and Browning 2009), and KBAC (Liu and Leal 2010). However, in contrast to these methods, VAAST also uses AAS information. Moreover, it uses a new approach to do so, one that allows it to score more SNVs than existing AAS methods such as SIFT (Kumar et al. 2009) and Polyphen (Sunyaev et al. 2001).

Much additional statistical power and accuracy are also gained from other components of the VAAST architecture, such as its ability to use pedigrees, phased data sets, and disease inheritance models. No existing AAS (Ng and Henikoff 2006) or aggregating method (Morgenthaler and Thilly 2007; Li and Leal 2008; Madsen and Browning 2009; Liu and Leal 2010) has these capabilities. The power of VAAST's pedigree approach is made clear in the quartet-based Miller syndrome analysis shown in Figure 4, where genome-wide only the two disease-causing genes are identified in a genome-wide screen of 19,249 nonsynonymous variants. Another important feature of VAAST is its ability to identify and mask variants in repetitive regions of the genome. As our results show, this provides a valuable method for mitigating platform-specific sequencing errors in situations in which it is cost-prohibitive to obtain a sufficiently large control set of genomes matched with regard to sequencing and variant calling pipeline. VAAST also differs in important ways from published heuristic search tools such as ANNOVAR (Wang et al. 2010). Unlike these tools, VAAST is not designed specifically to identify rare variants responsible for rare diseases. Instead, VAAST can search any collection of variants, regardless of their frequency distributions, to identify genes involved in both rare and common diseases.

Collectively, our results make clear the synergy that exists between these various components of the VAAST architecture. For example, they grant VAAST several unique features that distinguish it from commonly used AAS methods such as SIFT. Unlike AAS approaches, VAAST can score all variants, coding and noncoding, and in nonconserved regions of the genome. In addition, VAAST can obtain greater accuracy in judging which variants are deleterious. Comparison of the two Utah Miller syndrome exomes serves to highlight these differences. The two Miller syndrome exomes (Ng et al. 2010; Roach et al. 2010), for example, share 337 SNVs that are judged deleterious by SIFT; these 337 shared SNVs are distributed among 277 different genes. Thus, although AAS tools such as SIFT are useful for prioritizing the variants within a single known disease gene for follow-up studies, they are of limited use when carrying out genome-wide disease-gene searches, especially when the affected individuals are compound heterozygotes, as in the Miller syndrome examples.

In comparison to SIFT, VAAST scores 10% more nonsynonymous SNVs but identifies only nine candidate genes (Table 2), with the two disease-causing genes ranked fourth and fifth. When run in its pedigree mode, only the four disease-causing variants in the two disease genes are judged deleterious by VAAST genome-wide. The original analysis (Roach et al. 2010) of the family of four required 3 mo and identified eight potential disease-causing variants in four genes. An exome analysis required four affected individuals in three families to identify *DHODH* as the sole candidate for Miller syndrome (Ng et al. 2010). In contrast, using only the data from the family of four, VAAST identified the two disease genes in ~11 min using a 24-CPU compute server, and with perfect accuracy. Even when an additional 36,883 synonymous and noncoding regulatory variants are included in this genome-wide screen, only 23 candidate genes are identified, with *DHODH* still ranked fourth and *DNAH5* ranked first.

Our benchmark analyses using 100 different known diseases and 600 different known disease-causing alleles make it clear that our Miller syndrome and CCD analyses are representative results, and that VAAST is both a very accurate and a very reliable tool. VAAST consistently ranked the disease gene in the top three candidates genome-wide for recessive diseases and in the top nine gene candidates for dominant diseases. Equally important is reliability. VAAST has a much lower variance than either SIFT or ANNOVAR. In the recessive scenario, using three compound heterozygous individuals as a case cohort, for 95% of the VAAST runs, the disease-causing gene was ranked between second and 10th genome-wide; in comparison, ANNOVAR's ranks varied between 67 and 8762 on the same data sets, and SIFT's varied between 66 and 9107. Thus, VAAST is not only more accurate, it is also a more reliable tool. These same analyses also demonstrate that VAAST remains a reliable tool even when confronted with missing data due to phenomena such as missed variants, locus heterogeneity, and phenocopies in the case cohorts. Even when one-third of the cohort lacked disease-causing alleles at the locus, the average rank was still 61 for dominant diseases and 21 for recessive diseases (Fig. 5B).

VAAST can also be used to search for genes that cause common diseases and to estimate the impact of common alleles on gene function, something tools like ANNOVAR are not able to do. For example, when run over a published collection of 1454

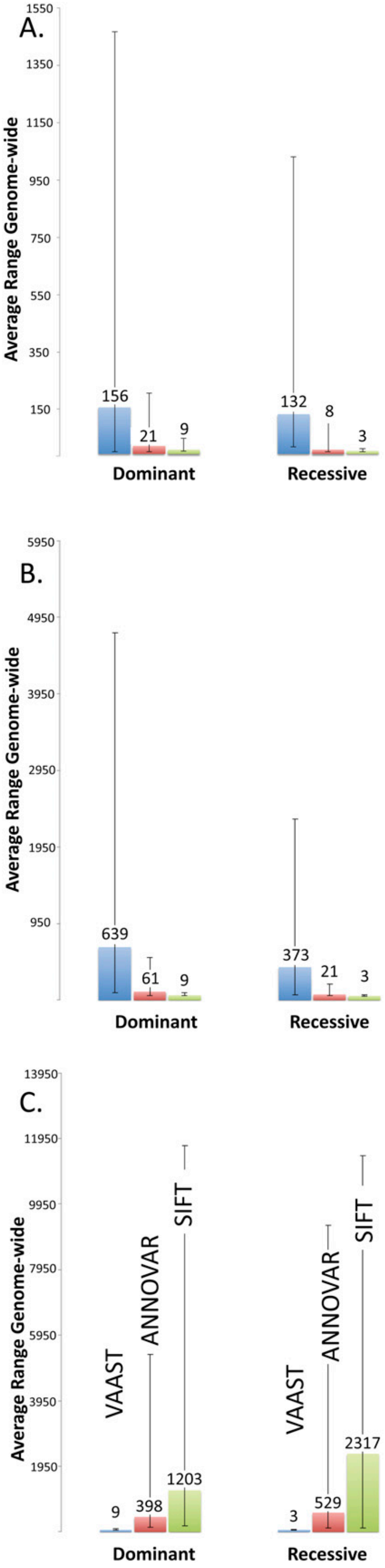
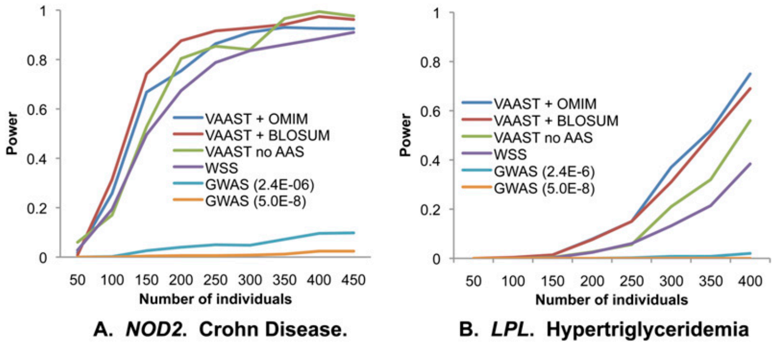


**Figure 5.** Benchmark analyses using 100 different known disease genes. In each panel the *y*-axis denotes the average rank of the disease gene among 100 searches for 100 different disease genes. Heights of boxes are proportional to the mean rank, with the number above each box denoting the mean rank of the disease gene among all RefSeq annotated human genes. Error bars encompass the maximum and minimum observed ranks for 95% of the trials. (*A*) Average ranks for 100 different VAAST searches. (*Left* half of panel) The results for genome-wide searches for 100 different disease genes assuming dominance using a case cohort of two (blue box), four (red box), and six (green box) unrelated individuals. (*Right* half of panel) The results for genome-wide searches for 100 different recessive disease genes using a case cohort of 1 (blue box), 2 (red box), and 3 (green box). (*B*) Impact of missing data on VAAST performance. (*Left* and *right* half of panel) Results for dominant and recessive gene searches as in panel *A*, except in this panel the case cohorts contain differing percentages of individuals with no disease-causing variants in the disease gene. (Blue box) Two-thirds of the individuals lack a disease-causing allele; (red box) one-third lack a disease-causing allele; (green box) all members of the case cohort contain disease-casing alleles. (*C*) Comparison of VAAST performance to that of ANNOVAR and SIFT. (*Left* half of panel) The results for genome-wide searches using VAAST, ANNOVAR, and SIFT to search for 100 different dominant disease genes using a case cohort of six unrelated individuals. (*Right* half of panel) The results for genome-wide searches using VAAST, ANNOVAR, and SIFT to search for 100 different recessive disease genes using a case cohort of three unrelated individuals.



**A.** ***NOD2*****. Crohn Disease.** **B.** ***LPL*****. Hypertriglyceridemia**

**Figure 6.** Statistical power as a function of number of target genomes for two common disease genes. (*A*) *NOD2*, using a data set containing rare and common nonsynonymous variants. (*B*) *LPL*, using a data set containing only rare nonsynonymous variants. For each data point, power is estimated from 500 bootstrapped resamples of the original data sets, with $\alpha = 2.4 \times 10^{-6}$ except where specified. *y*-axis: probability of identifying gene as implicated in disease in a genome-wide search; *x*-axis: number of cases. The number of controls is equal to the number of cases up to a maximum of 327 for *LPL* (original data set) and 163 for *NOD2* (original data set + 60 Europeans from 1000 Genomes). (VAAST + OMIM) VAAST using AAS data from OMIM as its disease model; (VAAST + BLOSUM) VAAST using BLOSUM62 as its disease model; (VAAST no AAS) VAAST running on allele frequencies alone; (WSS) weighted sum score of Madsen and Browning (2009); (GWAS) single variant GWAS analysis. *NOD2* and *LPL* data sets were taken from Lesage et al. (2002) and Johansen et al. (2010), respectively.

high-confidence disease-causing and predisposing SNVs from OMIM (Yandell et al. 2008), VAAST identifies all but 29 (2%) of these SNVs as damaging. ANNOVAR (Wang et al. 2010), in comparison, excludes 427 (29%) of these SNVs from further analysis because they are present in the 1000 Genomes Project data (The 1000 Genomes Project Consortium 2010), dbSNP130, or in segmentally duplicated regions. These results underscore the advantages of VAAST's probabilistic approach. VAAST can assay the impact of rare variants to identify rare diseases and both common and rare variants to identify the alleles responsible for common diseases, and it operates equally well on data sets (e.g., *NOD2*) wherein both rare and common variants are contributing to disease. Our common-disease analyses serve to illustrate these points. These results demonstrate that VAAST can achieve close to 100% statistical power on common-disease data sets, where a traditional GWAS test has almost no power. We also demonstrate that VAAST's own feature-based scoring significantly outperforms WSS (Madsen and Browning 2009), which, like all published aggregative scoring methods, does not use AAS information. These analyses also demonstrate another key feature of VAAST: While the controls in the Crohn's disease data set were fully sequenced at *NOD2*, only a small subset of the cases was sequenced, and the rest were genotyped at sites that were polymorphic in the sample. VAAST does well with this mixed data set. It is likely that VAAST would do even better using a data set of the same size consisting only of sequence data, as such a cohort would likely contain additional rare variants not detectable with chip-based technologies. Consistent with this hypothesis, VAAST also attains high statistical power compared to traditional GWAS methods on the *LPL* data set, which only contains alleles with a frequency of <5%. This demonstrates that VAAST can also identify common-disease genes even when they contain no common variants that contribute to disease risk.

These results suggest that VAAST will prove useful for re-analyses of existing GWAS and linkage studies. Targeted VAAST analyses combined with region-specific resequencing around GWAS hits will allow smaller Bonferroni corrections (Nicodemus et al. 2005) than the genome-wide analyses presented here, resulting in still greater statistical power, especially in light of VAAST's feature-based approach. The same is true for linkage studies. In addition, because much of the power of VAAST is derived from rare variants and amino acid substitutions, the likelihood of false positives due to linkage disequilibrium with causal variants is low. Thus, it is likely that VAAST will allow identification of disease genes and causative variants in GWAS data sets in which the relationships of hits to actual disease genes and the causative variants are unclear, and for linkage studies, where only broad spans of statistically significant linkage peaks have been detected to date.

VAAST is compatible with current genomic data standards. Given the size and complexity of personal genome data, this is not a trivial hurdle for software applications. VAAST uses GFF3 (http://www.sequenceontology.org/resources/gff3.html), and GVF (Reese et al. 2010) and VCF (http://www.1000genomes.org/wiki/Analysis/vcf4.0), standardized file formats for genome annotations and personal genomes data. The size and heterogeneity of the data sets used in our analyses make clear VAAST's ability to mine hundreds of genomes and their annotations at a time. We also point out that VAAST has a modular software architecture that makes it easy to add additional scoring methods. Indeed, we have already done so for WSS (Madsen and Browning 2009). This is an important point, as aggregative scoring methods are a rapidly developing area of personal genomics (Morgenthaler and Thilly 2007; Li and Leal 2008; Madsen and Browning 2009; Liu and Leal 2010). VAAST thus provides an easy means to incorporate and compare new scoring methods, lending them its many other functionalities.

Although there exist other tools with some of its features, to our knowledge, VAAST is the first generalized, probabilistic ab initio tool for identifying both rare and common disease-causing variants using personal exomes and genomes. VAAST is a practical, portable, self-contained piece of software that substantially improves on existing methods with regard to statistical power, flexibility, and scope of use. It is resistant to no calls, automated, and fast; works across all variant frequencies; and deals with platform-specific noise.

## Methods

### Inputs and outputs

The VAAST search procedure is shown in Figure 7. VAAST operates using two input files: a background and a target file. The background and target files contain the variants observed in control and case genomes, respectively. Importantly, the same background file can be used again and again, obviating the need—and expense—of producing a new set of control data for each analysis. Background files prepared from whole-genome data can be used for whole-genome analyses, exome analyses and for individual gene analyses. These files can be in either VCF (http://www.1000genomes.org/wiki/Analysis/vcf4.0) or GVF (Reese et al. 2010) format. VAAST also comes with a series of premade and annotated background condenser files for the 1000 genomes data (The 1000 Genomes Project Consortium 2010) and the 10Gen data set (Reese et al. 2010). Also needed is

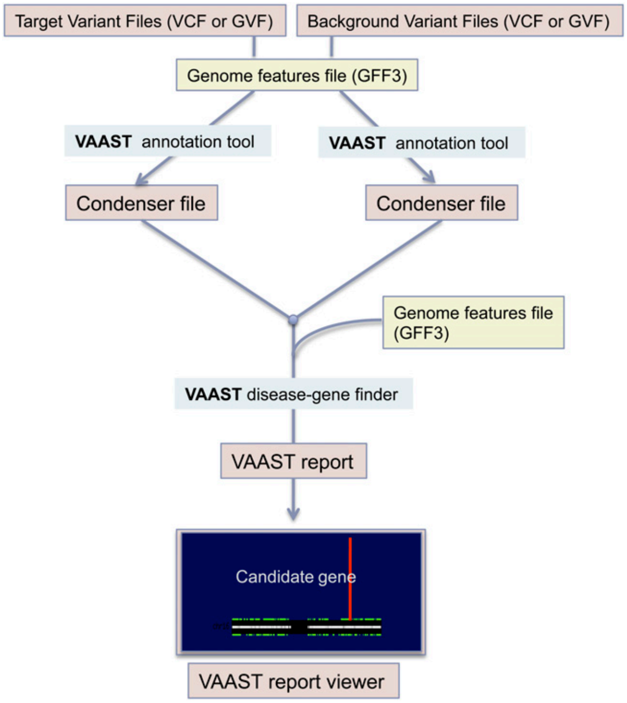

Target Variant Files (VCF or GVF)
Background Variant Files (VCF or GVF)
Genome features file (GFF3)
VAAST annotation tool
VAAST annotation tool
Condenser file
Condenser file
Genome features file (GFF3)
VAAST disease-gene finder
VAAST report
Candidate gene
VAAST report viewer

**Figure 7.** VAAST search procedure. One or more variant files (in VCF or GVF format) are first annotated using the VAAST annotation tool and a GFF3 file of genome annotations. Multiple target and background variant files are then combined by the VAAST annotation tool into a single condenser file; these two files, one for the background and one for the target genomes, together with a GFF3 file containing the genomic features to be searched are then passed to VAAST. VAAST outputs a simple text file, which can also be viewed in the VAAST viewer.

a third file in GFF3 (http://www.sequenceontology.org/resources/gff3.html) containing genome features to be searched.

## Basic CLR method

The composite likelihood ratio (CLR) test is designed to evaluate whether a gene or other genomic feature contributes to disease risk. We first calculate the likelihood of the null and alternative models assuming independence between nucleotide sites and then evaluate the significance of the likelihood ratio by permutation to control for LD. The basic method is a nested CLR test that depends only on differences in allele frequencies between affected and unaffected individuals. In a manner similar to the CMC method (Li and Leal 2008), we collapse sites with rare minor alleles into one or more categories, but we count the total number of minor allele copies among all affected and unaffected individuals rather than just the presence or absence of minor alleles within an individual. For our analyses, we set the collapsing threshold at fewer than five copies of the minor allele among all affected individuals, but this parameter is adjustable. Let $k$ equal the number of uncollapsed variant sites among $n_i^U$ unaffected and $n_i^A$ affected individuals, with $n_i$ equal to $n_i^U + n_i^A$. Let $l_{k+1} \ldots l_{k+m}$ equal the number of collapsed variant sites within $m$ collapsing categories labeled $k + 1$ to $m$, and let $l_1 \ldots l_k$ equal 1. Let $X_i$, $X_i^U$, and $X_i^A$ equal the number of copies of the minor allele(s) at variant site $i$ or collapsing category $i$ among all individuals, unaffected individuals, and affected individuals, respectively. Then the log-likelihood ratio is equal to:

$$\lambda = \ln\left(\frac{L_{Null}}{L_{Alt}}\right) = \sum_{i=1}^{k+m} \ln\left[\frac{(\hat{p}_i)^{X_i}(1-\hat{p}_i)^{2l_i n_i - X_i}}{\left(\hat{p}_i^U\right)^{X_i^U}\left(1-\hat{p}_i^U\right)^{2l_i n_i^U - X_i^U}\left(\hat{p}_i^A\right)^{X_i^A}\left(1-\hat{p}_i^A\right)^{2l_i n_i^A - X_i^A}}\right], \quad (1)$$

where $p_i$, $p_i^U$, and $p_i^A$ equal the maximum-likelihood estimates for the frequency of minor allele(s) at variant site $i$ or collapsing category $i$ among all individuals, unaffected individuals, and affected individuals, respectively. When no constraints are placed on the frequency of disease-causing variants, the maximum-likelihood estimates are equal to the observed frequencies of the minor allele(s). Assuming that variant sites are unlinked, $-2\lambda$ approximately follows a $\chi^2$ distribution with $k + m$ degrees of freedom. We report the non-LD-corrected $\chi^2$ $P$-value as the VAAST score to provide a statistic for rapid prioritization of disease-gene candidates. To evaluate the statistical significance of a genomic feature, we perform a randomization test by permuting the affected/unaffected status of each individual (or each individual chromosome, when phased data are available). Because the degrees of freedom can vary between iterations of the permutation test, we use the $\chi^2$ $P$-value as the test statistic for the randomization test.

## Extensions to the basic CLR method

In the basic CLR method, the null model is fully nested within the alternative model. Extensions to this method result in models that are no longer nested. Because the $\chi^2$ approximation is only appropriate for likelihood ratio tests of nested models, we apply Vuong's closeness test in extended CLR tests using the Akaike Information Criterion correction factor. Thus, the test statistic used in the permutation tests for these methods is $-2\lambda - 2(k + m)$. To efficiently calculate the non-LD-corrected $P$-value for non-nested models, we use an importance sampling technique in a randomization test that assumes independence between sites by permuting the affected/unaffected status of each allele at each site. To evaluate the LD-corrected statistical significance of genomic features for these models, we permute the affected/unaffected status of each individual (or each individual chromosome).

For rare diseases, we constrain the allele frequency of putative disease-causing alleles in the population background such that $p_i^U$ cannot exceed a specified threshold, $t$, based on available information about the penetrance, inheritance mode, and prevalence of the disease. With this constraint, the maximum-likelihood estimate for $p_i^U$ is equal to the minimum of $t$ and $X_i/l_i n_i$.

The framework can incorporate various categories of indels, splice-site variants, synonymous variants, and noncoding variants. Methods incorporating amino acid severity and constraints on allele frequency can result in situations in which the alternative model is less likely than the null model for a given variant. In these situations, we exclude the variant from the likelihood calculation, accounting for the bias introduced from this exclusion in the permutation test. For variants sufficiently rare to meet the collapsing criteria, we exclude the variant from the collapsing category if the alternative model is less likely than the null model prior to variant collapse.

## Severity of amino acid changes

To incorporate information about the potential severity of amino acid changes, we include one additional parameter in the null and alternative models for each variant site or collapsing category. The parameter $h_i$ in the null model is the likelihood that the amino acid change does not contribute to disease risk. We estimate $h_i$ by setting it equal to the proportion of this type of amino acid change in the population background. The parameter $a_i$ in the alternative model is the likelihood that the amino acid change contributes to disease risk. We estimate $a_i$ by setting it equal to the proportion of this type of amino acid change among all disease-causing mutations in OMIM (Yandell et al. 2008). Incorporating information about amino acid severity, λ is equal to:

$$\lambda = \ln\left(\frac{L_{Null}}{L_{Alt}}\right) = \sum_{i=1}^{k+m} \ln\left[\frac{h_i(\hat{p}_i)^{X_i}(1-\hat{p}_i)^{2l_i n_i - X_i}}{a_i\left(\hat{p}_i^U\right)^{X_i^U}\left(1-\hat{p}_i^U\right)^{2l_i n_i^U - X_i^U}\left(\hat{p}_i^A\right)^{X_i^A}\left(1-\hat{p}_i^A\right)^{2l_i n_i^A - X_i^A}}\right]. \quad (2)$$

To include the severity of amino acid changes for collapsed rare variants, we create $m$ collapsing categories that are divided according to the severity of potential amino acid changes. To create the collapsing categories, we first rank all possible amino acid changes according to their severity. We then assign an equal number of potential changes to each category, with the first category receiving the least severe changes and each subsequent category receiving progressively more severe changes. Each rare variant is then included in the category with its corresponding amino acid change (Tavtigian 2009). For each collapsing category $i$, we set the parameters $h_i$ and $a_i$ equal to their average values among all variants present in the category. We first calculate the likelihood of the null and alternative models assuming independence between nucleotide sites and then evaluate the significance of the likelihood ratio by permutation to control for LD.

## Scoring noncoding variants

The VAAST CLR framework can also score noncoding variants and synonymous variants within coding regions. Because ascertainment bias in OMIM can cause a bias against such variants, we took an evolutionary approach to estimate the relative impacts of noncoding and synonymous variants using the vertebrate-to-human genome multiple alignments downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/multiz44way/maf/). For each codon in the human genome, we calculated the frequency in which it aligns to other codons in primate genomes (wherever an open reading frame [ORF] in the corresponding genomes is available). Then for every codon alignment pair involving one or fewer nucleotide changes, we calculated its Normalized Mutational Proportion (NMP), which is defined as the proportion of occurrences of each such codon pair among all codon pairs with the identical human codon and with one or fewer nucleotide changes. For example, suppose the human codon GCC aligned to codons in primate genomes with the following frequencies: GCC → GCC: 1000 times; GCC → GCT: 200 times; GCC → GCG: 250 times; GCC → GGG: 50 times. The NMP value of GCC → GCT would be 0.134 [i.e., 200/(1000 + 200 + 250)]. For every codon pair that involves a nonsynonymous change, we then calculated its severity parameter from the OMIM database and 180 healthy genomes from the 1000 Genomes Project ($a_i/h_i$ in Eq. 2). Linear regression analysis indicates that $\log(a_i/h_i)$ is significantly correlated with log(NMP) ($R^2 = 0.23$, $p < 0.001$). This model allows us to estimate the severity parameter of synonymous variants (again by linear regression), which by this approach is 0.01 (100 times less severe than a typical nonsynonymous variant). We used a similar approach to derive an equivalent value for SNVs in noncoding regions. To do so, we again used the primate alignments from UCSC, but here we restricted our analysis to primate clustered DNase hypersensitive sites and transcription factor binding regions as defined by ENCODE regulation tracks, calculating NMP for every conserved trinucleotide. The resulting severity parameter for these regions of the genome is 0.03.

## Inheritance and penetrance patterns

VAAST includes several options to aid in the identification of disease-causing genes matching specific inheritance and penetrance patterns. These models enforce a particular disease model within a single gene or other genomic feature. Because the disease models introduce interdependence between sites, VAAST does not provide a site-based non-LD-corrected *P*-value for these models.

For recessive diseases, VAAST includes three models: recessive, recessive with complete penetrance, and recessive with no locus heterogeneity. In the basic recessive model, the likelihood calculation is constrained such that no more than two minor alleles in each feature of each affected individual will be scored. The two alleles that receive a score are the alleles that maximize the likelihood of the alternative model. The complete penetrance model assumes that all of the individuals in the control data set are unaffected. As the genotypes of each affected individual are evaluated within a genomic feature, if any individual in the control data set has a genotype exactly matching an affected individual, the affected individual will be excluded from the likelihood calculation for that genomic feature. This process will frequently remove all affected individuals from the calculation, resulting in a genomic feature that receives no score. In the recessive with no locus heterogeneity model, genomic features are only scored if all affected individuals possess two or more minor alleles at sites where the alternative (disease) model is more likely than the null (healthy) model. The two alleles can be present at different nucleotide sites in each affected individual (i.e., allelic heterogeneity is permitted), but locus heterogeneity is excluded. The models can be combined, for example, in the case of a completely penetrant disease with no locus heterogeneity.

The three dominant disease models parallel the recessive models: dominant, dominant with complete penetrance, and dominant with no locus heterogeneity. For the basic dominant model, only one minor allele in each feature of each affected individual will be scored (the allele that maximizes the likelihood of the alternative model). For the complete penetrance dominant model, alleles will only be scored if they are absent among all individuals in the control data set. For the dominant with no locus heterogeneity model, genomic features are only scored if all affected individuals posses at least one minor allele at variant sites where the alternative model is more likely than the null model.

## Protective alleles

For non-nested models, the default behavior is to only score variants in which the minor allele is at higher frequency in cases than in controls, under the assumption that the disease-causing alleles are relatively rare. This assumption is problematic if protective alleles are also contributing to the difference between cases and controls. By enabling the "protective" option, VAAST will also score variants in which the minor allele frequency is higher in controls than in cases. This option also adds one additional collapsing category for rare protective alleles. Because we have no available AAS model for protective alleles, we set $h_i$ and $a_i$ equal to 1 for these variants.

## Variant masking

The variant-masking option allows the user to exclude a list of nucleotide sites from the likelihood calculations based on information obtained prior to the genome analysis. The masking files used in these analyses exclude sites where short reads would map to more than one position in the reference genome. This procedure mitigates the effects introduced by cross-platform biases by excluding sites that are likely to produce spurious variant calls due to improper alignment of short reads to the reference sequence. The three masking schemes we used were (1) 60-bp single-end reads, (2) 35-bp single-end reads, and (3) 35-bp paired-end reads separated by 400 bp. These three masking files are included with the VAAST distribution, although VAAST can mask any user-specified list of sites. Because variant masking depends only on information provided prior to the genome analysis, it is compatible with both nested and non-nested models CLR models.

### Trio option

By providing the genomes of the parents of one or more affected individuals, VAAST can identify and exclude Mendelian inheritance errors for variants that are present in the affected individual but absent in both parents. Although this procedure will exclude both de novo mutations and sequencing errors, for genomes with an error rate of ~1 in 100,000, ~99.9% of all Mendelian inheritance errors are genotyping errors (Roach et al. 2010). This option is compatible with both nested and non-nested models.

### Minor reference alleles

Most publicly available human genome and exome sequences do not distinguish between no calls and reference alleles at any particular nucleotide site. For this reason, VAAST excludes reference alleles with frequencies of <50% from the likelihood calculation by default. This exclusion can be overridden with a command-line parameter.

VAAST options, including command lines used to generate each table and figure, are provided in the Supplemental Material.

### Benchmark analyses

We assayed the ability of VAAST, SIFT, and ANNOVAR to identify mutated genes and their disease-causing variants in genome-wide searches. To do so, we randomly selected a set of 100 genes, each having at least six SNVs that are annotated as deleterious by OMIM. For each run, the OMIM variants from one of the 100 genes were inserted into the genomes of healthy individuals sampled from the Complete Genomics Diversity Panel (http://www.completegenomics.com/sequence-data/download-data/). For the partial representation panel (Fig. 5B), we inserted the OMIM variants into only a partial set of the case genomes. For example, in the panel of 66% partial representation and dominant model, we inserted four OMIM variants into four of the six case genomes for each gene, so that 66% of the case genomes have deleterious variants; for 66% representation under the recessive model, we inserted four OMIM variants into two of the three case genomes.

We ran VAAST using 443 background genomes (including 180 genomes from the 1000 Genomes Project pilot phase, 63 Complete Genomics Diversity panel genomes, nine published genomes, and 191 Danish exomes) and with the inheritance model option (-iht). We ran SIFT using its web service (http://sift.jcvi.org/www/SIFT_chr_coords_submit.html, as of 5/3/2011). For ANNOVAR, we used version: 2011-02-11 00:07:48 with the 1000 Genomes Project 2010 July release as the variant database. We used its automatic annotation pipeline (auto_annovar.pl) and default parameters for annotation, setting its -maf option to the upper 99% confidence interval of the expected minor allele frequency (MAF), such that the combined MAF for inserted alleles did not exceed 5%. The dbSNP database was not used in this analysis because ANNOVAR's dbSNP130 database does not provide MAF information, and a portion of the disease-causing OMIM alleles are collected by dbSNP130. We found that setting -maf and excluding dbSNP130 for this analysis greatly improved the accuracy of ANNOVAR in comparison to its default parameters (data not shown); thus we used these more favorable parameters for our comparisons.

To compare the performance of the three algorithms with a sample size of six under a dominant model, for each of the 100 genes, we inserted the six different OMIM variants located in this gene into six different healthy genomes, making all of them heterozygous for a different disease-causing SNV at that locus. Under the recessive model, with a sample size of two, for example, we inserted four different OMIM variants located in each gene into two healthy genomes, so that each case genome carries two different OMIM variants in this gene, i.e., the individuals are compound heterozygotes.

### Scalability

VAAST computes scale linearly with the number of features (genes) being evaluated and the number of variants in the targets. The maximum number of permutations needed is bounded by $O(n^k)$, where $n$ equals the number of background and target genomes, and $k$ equals the number of target genomes. VAAST is a multithreaded, parallelized application designed to scale to cohorts of thousands of genomes.

## Data access

VAAST is available for download at http://www.yandell-lab.org 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., NIH GM59290 to L.B.J., and NIH K99HG005846 to J.X., all administered by the National Human Genome Research Institute (NHGRI). C.H. was supported by 1RC2HG005619, the University of Luxembourg–Institute for Systems Biology Program, and NIH 1T32HL105321-01. An allocation of computer time from the Center for High Performance Computing at the University of Utah is gratefully acknowledged.

*Authors' contributions:* M.Y. and M.G.R. conceived of the project. M.Y. oversaw and coordinated the research. M.Y. designed the software architecture. M.Y., H.H., C.H., and B.M. wrote the software package. C.H., H.H., and M.Y. contributed to the statistical method development. M.S., H.H., C.H., and J.X. performed experiments. M.Y., C.H., L.B.J., and M.G.R. wrote the paper.

## References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* **322:** 881–888.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78–94.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Özen S, Sanjad S, et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci* **106:** 19096–19101.

Henikoff S, Henikoff J. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci* **89:** 10915–10919.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci* **106:** 9362–9367.

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

Korf I, Bedell J, Yandell M. 2003. *BLAST: An essential guide to the Basic Local Alignment Search Tool*. O'Reilly, Beijing.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4:** 1073–1081.

Lausch E, Hermanns P, Farin HF, Alanay Y, Unger S, Nikkel S, Steinwender C, Scherer G, Spranger J, Zabel B, et al. 2008. TBX15 mutations cause craniofacial dysmorphism, hypoplasia of scapula and pelvis, and short stature in Cousin syndrome. *Am J Hum Genet* **83:** 649–655.

Lesage S, Zouali H, Cézard JP, Colombel JF, Belaiche J, Almer S, Tysk C, O'Morain C, Gassull M, Binder V, et al. 2002. CARD15/NOD2 mutational analysis and genotype–phenotype correlation in 612 patients with inflammatory bowel disease. *Am J Hum Genet* **70:** 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:** 311–321.

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, et al. 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* **42:** 969–972.

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:** e1001156. doi: 10.1371/journal.pgen.1001156.

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, et al. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* **362:** 1181–1191.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* **5:** e1000384. doi: 10.1371/journal.pgen.1000384.

Manolio TA. 2009. Cohort studies and the genetics of complex disease. *Nat Genet* **41:** 5–6.

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

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

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

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al. 2010. Exome sequencing identifies the cause of a Mendelian disorder. *Nat Genet* **42:** 30–35.

Nicodemus KK, Liu W, Chase GA, Tsai YY, Fallin MD. 2005. Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms. *BMC Genet* (Suppl 1) **6:** S78. doi: 10.1186/1471-2156-6-S1-S78.

Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE, et al. 2010. The characterization of twenty sequenced human genomes. *PLoS Genet* **6:** e1001111. doi: 10.1371/journal.pgen.1001111.

Reese MG, Kulp D, Tammana H, Haussler D. 2000. Genie–gene finding in *Drosophila melanogaster. Genome Res* **10:** 529–538.

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:** R88. doi: 10.1186/gb-2010-11-8-r88.

Roach J, Glusman G, Smit A, Huff C, Hubley R, Shannon P, Rowen L, Pant K, Goodman N, Bamshad M, et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328:** 636–639.

Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. *Hum Mol Genet* **10:** 591–597.

Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N, et al. 2009. Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am J Hum Genet* **85:** 427–446.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38:** e164. doi: 10.1093/nar/gkq603.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678.

Yandell M, Moore B, Salas F, Mungall C, MacBride A, White C, Reese MG. 2008. Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. *PLoS Comput Biol* **4:** e1000218. doi: 10.1371/journal.pcbi.1000218.

*Received March 9, 2011; accepted in revised form June 8, 2011.*