DEVELOPING COMPUTATIONAL METHODS FOR STUDYING NONMODEL ORGANISM GENETICS AND HUMAN DISEASE WITH NEXT-GENERATION SEQUENCING DATA

by

Hao Hu

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

December 2012

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Hao Hu**

has been approved by the following supervisory committee members:

Mark Yandell , Chair **7/11/2012**
Date Approved

Alun Thomas , Member **7/11/2012**
Date Approved

Gillian Stanfield , Member **7/11/2012**
Date Approved

Karen Eilbeck , Member **7/11/2012**
Date Approved

Robert Weiss , Member **7/11/2012**
Date Approved

and by **Lynn B. Jorde** , Chair of

the Department of **Human Genetics**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

The rapidly decreasing of costs of sequencing is revolutionizing genetics. Two applications of next-generation sequencing data are of particular importance in this regard. First, high-throughput sequencing now offers a fast and inexpensive means to investigate the genomes and genetics of nonmodel organisms. Second, human personal-genomics data offer a unique opportunity for discovering the genetic basis of human traits and diseases.

My PhD research has focused on developing computational methods to study genetics using next-generation sequencing data. In the first chapter of my thesis, I present a series of genome-based studies of the venomous cone snail *Conus bullatus*, a source of pharmaceutically important small cysteine-rich peptides called conopeptides or conotoxins. Using high-coverage transcriptome sequence from its venom duct together with low-coverage genomic reads, I have developed new methods to characterize key genomic traits in the absence of a complete reference genome, including genome size, sequence diversity, repeat content and mobile element densities. I have also developed an *in silico* transcriptomics pipeline for conotoxin discovery, and have used it to identify novel conotoxins as well as candidate enzymes that are likely to be involved in the post-translational processing of conotoxins.

In the second and the third chapters of my thesis, I describe a probabilistic disease-gene search algorithm VAAST (the Variant Annotation, Analysis and Search

Tool) for finding damaged genes and their disease-causing variants; I also describe a powerful new extension to the original code-base called VAAST 2.0. In these chapters, I demonstrate that VAAST is both an accurate rare Mendelian disease-gene finder and a powerful means for identifying genes and alleles underlying common diseases. I have also carried systematic population-genetic simulations in order to benchmark the performance of VAAST and VAAST 2.0 under different genetic scenarios, and these demonstrate that VAAST 2.0 is the most robust and broadly applicable method available today for identification of genes involved in common genetic diseases such as breast cancer, hypertriglyceridemia and Crohn disease.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Page

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Mark Yandell, for all his guidance during my PhD studies. When I first entered the lab as a graduate student from a pure biology background, I wasn’t fully prepared for the intensity of the computational biology work taking place in the lab. Mark spent many hours each week helping me to develop my programming skills and helping me to learn to think as a computational biologist, for this help I cannot express enough gratitude. I probably benefited even more during daily conversations with Mark in which he shared his valuable insights in biology and informatics. I also deeply appreciate all the interesting projects that Mark sent my way, and the fact that he encouraged me to develop my own interests as well. In summary, I could not have developed the two key qualities every scientist needs—curiosity and logical thinking, without Mark’s help.

I am also deeply grateful for members of my committee, Alun Thomas, Gillian Stanfield, Karen Eilbeck and Robert Weiss for their supervision and advice during my PhD studies. I also would like to thank my former committee member Gerald Spangrude for his kindness of supervising my preliminary exam.

I want to thank current and previous members of Yandell Lab. Especially, Barry Moore and Carson Holt for never saying “no” when I needed any help with my experiments; Chad Huff for always sharing with me his brilliant statistical advice and insight. Also, many thanks to Marc Singleton, Zev Kronenberg, Mike Campbell and

Daniel Ence for all our interesting discussions; these made my graduate student life a much more enjoyable experience.

CHAPTER 1

CHARACTERIZATION OF THE CONUS BULLATUS GENOME AND ITS VENOM-DUCT TRANSCRIPTOME

The following chapter is a reprint of an article coauthored by myself, Pradip K Bandyopadhyay, Baldomero M Olivera and Mark Yandell. This article is originally published in *BMC Genomics* 2011, 12:60.

Hu *et al. BMC Genomics* 2011, **12**:60
http://www.biomedcentral.com/1471-2164/12/60

BMC Genomics

RESEARCH ARTICLE Open Access

Characterization of the *Conus bullatus* genome and its venom-duct transcriptome

Hao Hu[1], Pradip K Bandyopadhyay[2], Baldomero M Olivera[2], Mark Yandell[1*]

Abstract

Background: The venomous marine gastropods, cone snails (genus *Conus*), inject prey with a lethal cocktail of conopeptides, small cysteine-rich peptides, each with a high affinity for its molecular target, generally an ion channel, receptor or transporter. Over the last decade, conopeptides have proven indispensable reagents for the study of vertebrate neurotransmission. *Conus bullatus* belongs to a clade of *Conus* species called *Textilia*, whose pharmacology is still poorly characterized. Thus the genomics analyses presented here provide the first step toward a better understanding the enigmatic *Textilia* clade.

Results: We have carried out a sequencing survey of the *Conus bullatus* genome and venom-duct transcriptome. We find that conopeptides are highly expressed within the venom-duct, and describe an *in silico* pipeline for their discovery and characterization using RNA-seq data. We have also carried out low-coverage shotgun sequencing of the genome, and have used these data to determine its size, genome-wide base composition, simple repeat, and mobile element densities.

Conclusions: Our results provide the first global view of venom-duct transcription in any cone snail. A notable feature of *Conus bullatus* venoms is the breadth of A-superfamily peptides expressed in the venom duct, which are unprecedented in their structural diversity. We also find SNP rates within conopeptides are higher compared to the remainder of *C. bullatus* transcriptome, consistent with the hypothesis that conopeptides are under diversifying selection.

Background

Next-generation sequencing techniques have opened up new opportunities for genomics studies of new model organisms [1]. Many of these organisms are not amenable to classical genetic techniques; thus their sequenced and annotated genomes are the central resource for experimental studies. The popularity of the Planarian *Schmidtea mediterranea*, which can regenerate complete animals from fragments of its body, with stem-cell researchers is one example [2]. The Cone snail is another.

The cone snails (genus *Conus)* belong to the superfamily Conoidea which probably includes over 10,000 venomous gastropods [3]. The venom from each of the species of cone snails includes a mixture of small cysteine-rich peptides, which are used to immobilize their prey. These small peptides (~15 to 40 amino acids in length) have exquisite specificity for different isoforms of ion channels, receptors and transporters [4]. Their disulfide scaffold restricts the conformational space available to a peptide. However, the combination of variable intervening amino acids and their posttranslational modifications enable a spectrum of specific interactions with their target molecules. A typical conopeptide precursor is comprised of three regions: an N-terminal signal peptide, a pro-region, and a mature peptide region. The N-terminal sequence is usually much more conserved than the mature peptide, possibly due to the diversifying selection on the latter [5]. Conopeptides are classified into super-families, mainly based on the conserved signal peptide and different cysteine patterns observed within the mature peptide.

Conopeptides serve as specific neurobiological tools for addressing specific receptors and channels, and are also valuable lead compounds for therapeutic evaluation. A conopeptide, ω-MVIIA (commercially known as

* Correspondence: myandell@genetics.utah.edu
[1]Eccles institute of Human Genetics, University of Utah, and School of Medicine, Salt Lake City, UT 84112, USA
Full list of author information is available at the end of the article

Prialt, ziconotide) isolated from *Conus magus*, has been approved by FDA for the treatment of chronic pain [6,7]. In addition, other conopeptides are also being evaluated for the treatment of pain and epilepsy [8-11]. It is estimated that the venom of a single species of *Conus* may contain as many as 200 different venom peptides [4,12]. This raises the possibility that the 500-700 species of cone snails may provide upwards of 100,000 compounds of potential pharmacological interest, perhaps more when all the members of superfamily *Conoidea* are considered.

We have carried out a sequencing survey of the *Conus bullatus* genome and venom-duct transcriptome. *Conus bullatus* is a fish-hunting cone snail that together with *C. cervus* and *C. dusaveli* are members of the subgenus *Textilia* (Swainson, 1840). This is probably the least understood group of fish-hunting *Conus*. All are from the Indo-Pacific region (Pacific and Indian oceans from Hawaii through South Africa). *Conus bullatus* is the only accessible member of this clade of species; all others are rare and from deep water. *C. bullatus* is found from the intertidal zone to about 240 m, most commonly from slightly subtidal to 50 m, *C. cervus* between 180-400 m and *C. dusaveli* 50-288 m [13].

The pharmacology of the *Textilia* is thus still poorly characterized, and the genomics analyses presented here provide the first step toward a better understanding the enigmatic *Textilia* clade. The biology of the *Conus* species that belong to the Textilia clade is mostly unknown, but we recently documented the prey capture behavior of *Conus bullatus* (Figure 1). The general strategy appears to be analogous to that first established for *Conus purpurascens* [14], with one group of venom peptides causing a rapid tetanic immobilization, and a second set eliciting a block of neuromuscular transmission. Multiple venom peptides that act coordinately to achieve a particular physiological endpoint are referred to as "conopeptides cabals" [15]. The fish-hunting cone snails generally have both a "lightning-strike cabal" and a "motor cabal" leading to the tetanic immobilization and neuromuscular block, respectively. A video of *Conus bullatus* has documented the most rapid tetanic immobilization of prey observed for any fish-hunting cone snail. (http://www.hhmi.org/biointeractive/biodiversity/2009_conus_bullatus.html).

Venom studies in *Conus bullatus* have already yielded results of exceptional pharmacological interest. The best characterized *bullatus* venom component, alpha-conotoxin BuIA is a small peptide antagonist of nicotinic receptors that has become the standard pharmacological tool for differentiating between nicotinic receptors that carry two closely related subunits, β2 and β4. These receptors are of considerable interest in Parkinson's disease [16]. More recently, the μ-conotoxins, peptides with 3 disulfide bonds that are antagonists of voltage-gated Na channels have also been characterized from *Conus bullatus* [17]. These peptides appear to have novel subtype selectivity for the different molecular isoforms of voltage-gated Na channels [17]. Thus, they provide a promising neuropharmacological lead to developing an entirely new pathway to differentiate between different voltage-gated Na channel subtypes. Clearly, better cone snail genomics resources would aid these studies; however, few such resources exist as yet for *Conus* studies, and none for *C. bullatus*.

The cone snails are being extensively investigated as a source of peptidic pharmacological agents (ligands) with exquisite specificity for different subtypes of receptors in the central nervous system. In keeping with this main goal it is not surprising that most of the available nucleic acid sequences from *Conus* are a catalogue of

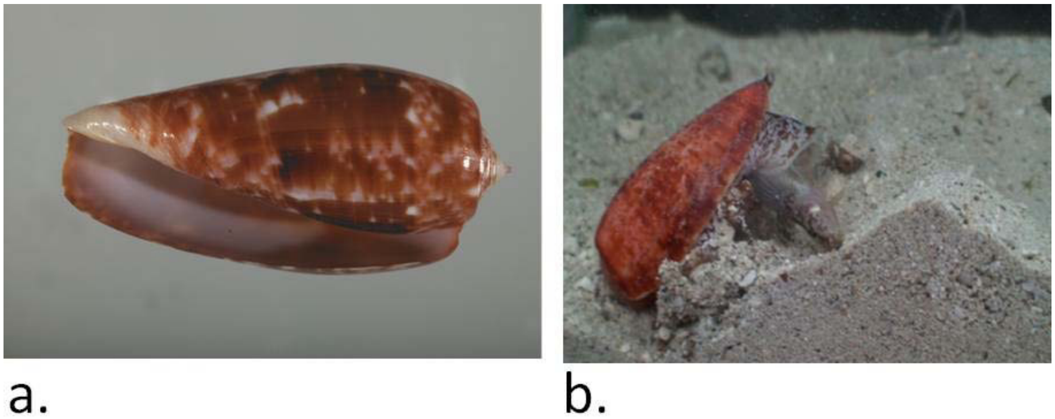


Figure 1 Conus bullatus and its feeding preference. a. Shell of *Conus bullatus*; b. Prey capture by *Conus bullatus*.

these compounds present in the venom. In addition, partial sequences of a few mitochondrial (ribosomal RNA and COI) and nuclear genes [18-23] have also been determined to ascertain the phylogenetic relationship among cone snails.

Previous work has used traditional molecular biology approaches to clone genes encoding members of specific conopeptide super-families [20,24-26], and EST sequencing in another *Conus* snail has identified conopeptides [27,28]. However, to date, no high-throughput sequencing approach on the whole mRNA reservoir of a *Conus* venom-duct has been attempted.

We have used RNA-seq [29] to identify and profile the expression of conopeptides and post-translational modification enzymes implicated in venom production. Our results provide the first global view of venom-duct transcription. Our shotgun genomic survey complements our RNA-seq data, and is also the first reported for a cone snail. Knowledge of several marine gastropod genomes will provide a first step toward the molecular understanding of numerous traits unique to these species. Accordingly, we have used these data to determine the suitability of the genome for sequencing and assembly with 2nd generation technologies, determining genome-wide base composition, sequence heterozygosity, simple repeat, and mobile element densities within the *C. bullatus* genome.

As we show, our RNA-seq and genomic datasets can be combined to enable analyses not possible with either dataset alone. For example, the transcriptome assembly has allowed us to explicitly test the hypothesis that conopeptides are under diversifying selection [5]. We have also developed a novel method for estimating genome size using RNA-seq and genomic shotgun sequences, which we present here. The approach is accurate, and should prove useful for any researcher seeking to determine the size of an emerging model organism [1] genome using 2nd generation sequencing data.

Results

Sequence datasets

We generated 96,379,716 Illumina paired 59-mers and 55,699,572 paired 60-mers for the genome. The average insert size of the paired-end library is 200nt. We also isolated venom-duct poly-A mRNA and sequenced it using both Illumina and Roche technologies. On the Illumina platform, we generated 102,278,116 paired 79-mers with a median insert size of 340bp. The Roche 454 platform generated 848,394 reads with average read length of 248bp. Many cDNA reads from the Illumina platform have low-quality 3' ends, which could be due to either to the small amounts of mRNA used in our experiments, or instrument error during sequencing or processing. We removed 3'end sequences from the reads with phred quality values of 2.

Genome-wide GC content

We randomly selected 30 million genomic reads using the process described in the Methods section (see section Simulated Read Sets) and determined their GC content. This procedure gives an estimated GC content for the *C. bullatus genome* of 42.88%. To validate this method, we also simulated 1 million randomly sampled 60-mers from the *D. melanogaster* genome and performed the same experiment, which gives 41.87%, an estimate in good agreement with the actual GC content (41.74%.) of the *D. melanogaster* genome.

Genome-wide Repeat Content

We took three approaches to characterize the repeat content of the *C. bullatus* genome. First, we ran RepeatMasker on 1 million randomly selected *C. bullatus* genomic reads, comparing the results to a matched human, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Aplysia californica* (a mollusk) datasets of simulated reads, as well as real human genome reads [30] (see Methods for details); these datasets match the *Conus* data precisely as regards number of reads, distance between pairs, read lengths, and (among the simulated sets) base quality (see Methods for details). Comparisons of the simulated human reads to real human reads (purple and grey columns in Figure 2), indicates that the simulated human reads closely match the real reads as regards repeat content for all repeat classes except simple repeats. We speculate that this is because many simple repeats (e.g., those near telomeres and centromeres) are designated as "N" in the reference human genome; hence, a random sampling of segments of the human reference assembly under represents its simple repeat content.

RepeatMasker [31] and RepBase [32] lack extensive libraries of repeats for mollusks, which will compromise the ability of RepeatMasker to identify interspersed repeats in the two mollusk datasets. Although, this fact does not complicate direct comparison of *C. bullatus* and *Aplysia californica*, with regards to the relative numbers of conserved interspersed repeats, it does complicate absolute measurements and comparisons to the other genomes (Figure 2). The ability of RepeatMasker to identify simple repeats, however, is less impacted by the lack of well-characterized repeat libraries for mollusks. This fact together with comparison with J. Flatley's genomic reads [30] (Figure 2) suggests that *C. bullatus* is significantly enriched for simple repeats relative to the other invertebrates, and slightly so (1.44 fold) compared to human.

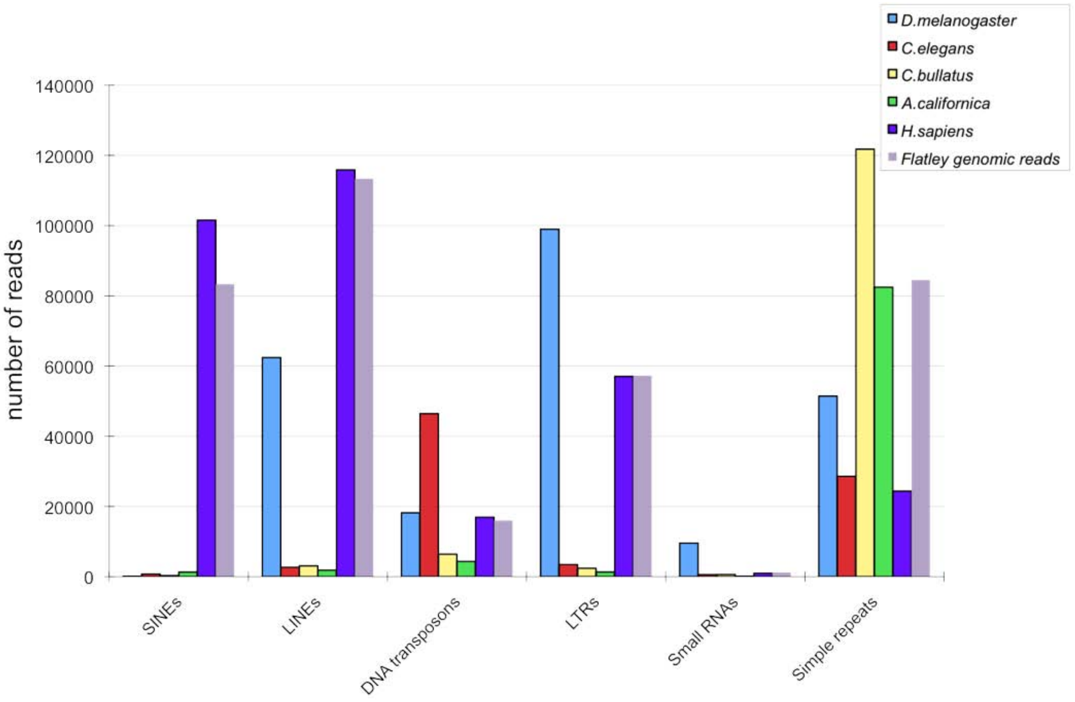


Figure 2 Comparison of Repetitive element counts in 1 million reads drawn from five different genomes. Repeats in 1 million randomly sampled C. bullatus Illumina 80-bp reads were characterized using RepeatMasker and compared to matched datasets manufactured from simulated reads from three other sequenced genomes and real reads from Flatley genome. X-axis: repeat-class. Y-axis: counts.

We also used RECON [33] to identify novel, high-copy genomic sequences that may be interspersed repeats in the *C. bullatus* genome. For this analysis we used our *C. bullatus de novo* genomic assembly (see Methods). In total, we found 115 genomic contigs present in 10 or more copies, with an average length of 544bp. Among these genomic sequences, 5 are homologous with known LINE members that were not detected by RepeatMasker in first repeat analysis. Of the remaining contigs, 9 have significant homology with G-protein receptors; 2 have significant homology with lipoprotein receptors; 1 has a leucine-rich repeat structure. These are probably high-copy number genomic regions but are not interspersed repeats. The remaining contigs have little homology with known interspersed repeats, however, a significant fraction of them have either strong homology to nuclease proteins or weak homology with rRNA and tRNA genes-both common motifs in LINE elements. Running RepClass [33] over these 115 genomic contigs confirmed that 20 contigs have LINE-like structures or are significantly homologous to known LINEs. Including this set would increase the percentage of the *C. bullatus* genome with LINE homology from 0.24% to 0.56%.

Because novel forms of retro-transposons might not have been identified in our RepeatMasker experiment, or some unknown bias in the ABySS [34] assembler might have caused us to underestimate the numbers of novel repeats identified with RECON, we devised a third experiment, that controls for both of these possibilities. In this experiment, we took the same read-datasets used in our RepeatMasker analysis (Figure 2), and performed an all-against-all BLAST [35] search of the *C. bullatus* reads against themselves, and repeated the same experiment for a matched set of simulated reads from *H. sapiens* (see Methods for details). For reasons of computational complexity we choose to limit this analysis to only one target genome: *H. sapiens*, because it is the most repeat rich of any in our dataset and its genome is nearly the same size as the *C. bullatus* genome. We then tallied the percentage of reads having one BLAST hit, two hits and so on. For each read, its number of hits can be used to obtain an estimate of the copy-number of its sequence within the genome (see

Methods). This allows us to estimate the proportion of high-copy number genomic sequences within the *Conus* genome and to make comparisons to the human genome (Figure 3). This experiment presumes no prior knowledge of the repeat content of the genome. We also used the 'SEG' option with WU-BLAST [36] to exclude hits between reads consisting only of low complexity and/or simple sequence repeats. By using BLAST with the SEG option any reads consisting entirely of low complexity or simple sequence repeats will have no hits.

This analysis reveals much about the repeat content of *Conus* compared to that of the Human genome. First, the *Conus* genome has a larger proportion of high copy-number sequences (presumably interspersed repeats) compared to human. This is shown by the fact that 23% of *Conus* reads (compared to 16% in human) have numbers greater than 50. By looking into this group of human reads, we confirmed that 91% of these are homologous to known interspersed repeats. Second, the human dataset (and hence the human genome as compared to the *Conus* genome) has 3× as many genomic sequences with a copy number above 10,000 compared to *Conus* (6.9% versus 2.4%). These sequences are mostly non-LTR elements that exists in extremely high copy number; running Repeatmasker over these human genome reads showed that 75% of these genomic regions are SINEs and another 20% are LINEs, supporting this hypothesis. Taken together, our results show that although the *Conus* genome is enriched for interspersed repeats compared to human, it has far fewer non-LTR repetitive elements.

A partial genome assembly

A previous estimate based upon cytology, placed the *Conus bullatus* genome at around 3 billion base pairs [37]. If true, our 60 bp paired-end Illumina dataset would provide 3× coverage. Although this is insufficient to produce anything near a complete genome assembly,

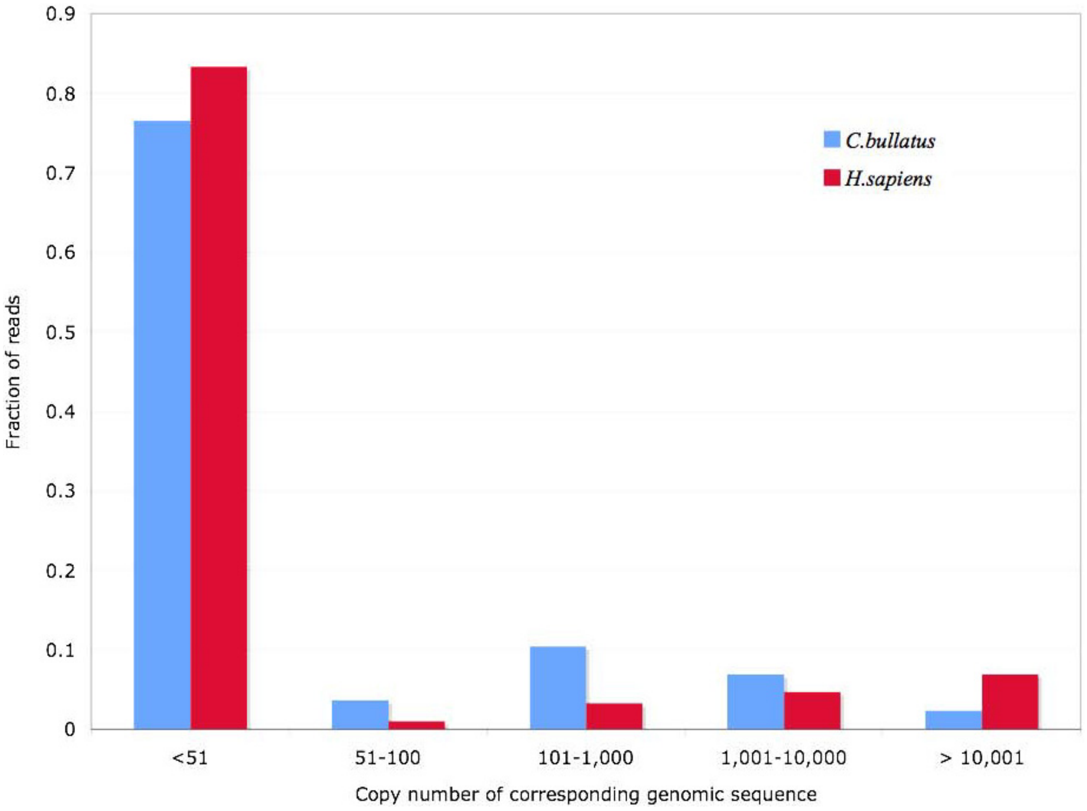


Figure 3 Profile of proportion of the genomic sequences with each copy number. Generated from all-by-all blast analysis of one million *C. bullatus* and *H.sapiens* reads each against themselves. The number of read partners is converted to copy-number of corresponding genomic sequence. X-axis: each bin's label gives the minimum and maximum copy numbers in the genome. Y-axis: fraction of reads falling into that bin.

a partial genome assembly is still desirable for some analyses. We used ABySS to produce a partial assembly 201 million base pairs in length with an N50 value of 182 bp (See Methods for details). This accounts for ~7% of the total length of the *C. bullatus* genome. To estimate the quality of our genome assembly, we simulated 8.7 million 60bp-long Illumina reads from the *D. melanogaster* genome (3× coverage), with the same base-calling accuracy distribution as in our *Conus* genomic reads. To do so we used the procedure described in the Methods section. This process gives 3× coverage over the *Drosophila* genome with the same error rates as our *C. bullatus* reads. Assembling these reads with ABySS with the same parameters produced a *Drosophila* assembly with an N50 of 143 bp and total sequence length of 16 MB, which accounts for roughly 10% of the fly genome. Thus the two assemblies are of comparable quality.

Assembly of the venom-duct transcriptome

We assembled our Illumina RNA-seq reads from the *C. bullatus* venom-duct with ABYSS (see methods for details). This produced 525,537 contigs of 60bp or greater in length and having a total length of 57 MB. We chose 60bp as minimum contig size because conopeptides can be as short as 20 amino acids. The 454 reads were generated and assembled by Roche.

Annotation of transcriptome

To determine the percentage of the total *C. bullatus* proteome sampled one or more times in our Illumina and Roche transcriptome datasets, we took the core eukaryotic protein set from CEGMA [38], which is comprised of 248 core proteins that generally lack paralogs in the eukaryotes [38,39], and asked what percentage of these proteins are found in the combined Illumina or Roche assemblies. Using BLASTX, 211 out of 248 proteins (85%) are found (E < = 1e^{-7}).

To annotate the transcriptome assembly we ran WU-BLASTX on the ABySS Illumina assembly against UniProtKB database [40]. 7,691 unique UniProtKB proteins have significant homology with one or more transcriptome contigs. We also mapped those contigs no shorter than 200bp to GO [41] terms for biological process, molecular function and cellular component. As a control, we applied the same approach to the annotated *C. elegans, D. melanogaster* and *H. sapiens* transcriptomes and compared the proportion of genes assigned to each GO term in these organisms to our transcriptome assembly results (Additional File 1). Note that is not a comparison of expression levels, but rather a comparison using GO of which genes were represented in our transcriptome assembly. In other words, the *relative* proportions of all GO gene categories associated with our *C. bullatus* contigs was found to be similar to the relative proportions of genes assigned to the same GO categories for *C. elegans, D. melanogaster* and *H. sapiens* transcriptomes. We found that the resulting GO profiles are highly similar for all four organisms. This finding, together with our observation that 85% of CEGMA proteins are represented in the assembly, suggests that we have sampled a wide swath of the *C. bullatus* transcriptome.

Identification of Conopeptides in RNA-seq data

We searched our combined Illumina and Roche transcriptome assemblies for significant homology to a set of known conopeptides collected from ConoServer [42], using the procedure described in the Methods section. We find that, as might be expected, conopeptides are transcribed at high levels in the venom duct; the depth of coverage of the putative conopeptides is 102× versus 33× for the remainder of the transcriptome.

Whenever possible, we assigned each of our putative conopeptide contigs to a conopeptide superfamily, by significant homology to signal sequences that are characteristic of each superfamily (see Methods for details). In total, we were able to assign 543 contigs a unique conopeptide super-family. We find that, as in most *Conus* species examined so far, the O1, M, A and T superfamilies were represented by the greatest number of distinct contigs. We also observed that mRNA abundance levels followed this same general pattern with respect to superfamilies (Table 1). Besides these well represented superfamilies, we also found small number of conopeptides belonging to the rarer in I2 and J conopeptide super-families in *Conus bullatus*, which account for ~0.4% of total putative-conopeptide transcripts.

In total, we identified 2,410 putative conopeptide contigs. Most of these contigs are short (with the N50 of 69bp), and do not contain the full-length sequence of the conopeptide precursor. Nevertheless, we were able to identify a few complete conopeptides (mainly from the Roche data), and a selection of 30 putative complete and partial conopeptide sequences are presented in

Table 1 Superfamilies of *C. bullatus* conopeptides identified by RNA-seq

Conopeptide Super Family	*C. bullatus* RNA-seq data	Conoserver reference sequences
T	15%	13%
A	17%	19%
M	20%	9%
O2	4%	5%
O1	44%	40%
Other	< 1%	14%

Percentages for *C. bullatus* refer to percentage of venom-duct RNA-seq reads belonging to a given superfamily. Globally the distribution parallels that for reference conopeptide sequences by class available on Conserver, although rare classes are under-sampled.

(Table 2). The conopeptides listed belong to the O, M, A, J, contryphan and conkunitzin super-families with O- being the most abundant. While conopeptides belonging to the I2, T, con-ikot-ikot, and conantokin super-families could be identified in the Blast analysis; the contig lengths and frameshifts associated these hits precluded the generation of a high confidence protein sequence.

A notable feature of the *Conus bullatus* transcriptome analysis is the breadth of A-superfamily peptides expressed in the venom duct, which are unprecedented in their structural diversity (Table 3). In most Conus species, the predominant structural classes of A-peptides is the α4/7 subfamily; in fish-hunting cone snails, additional subclasses are the α3/5 subfamily and κA conotoxins (in species of the *Pionoconus* clade) and the αA conotoxins (in species of the *Chelyconus* clade). The *Conus bullatus* transcriptome includes an mRNA encoding a κA conotoxin (Bu27), which is unambiguous in its identity. There is also a single member of the α4/7 subfamily (Bu19) of unknown function, which is strikingly different in sequence from all other Conus venom peptides in this group. Although no member of the αA family or the α3/5 subfamilies were found, 8 other A superfamily peptides were identified. Together these comprise a greater range of structural diversity in the

Table 2 Translated transcripts containing putative toxin sequences

O-superfamily: C-C-CC-C-C	
1.	MKLTCVAIVAVLLLTACQLITAEDSRGTQLHRALRKTTKLSVSTRCKGPGAKCLKTMYDCCKYSCSRGRC
2.	MKLTCVLIIAVLFLTAITADDSRDKQVYRAVGLIDKMRRIRASEGCRKKGDRCGTHLCCPGLRCGSGRAGGACRPPYN
3.	MKLMCVLIVSVLVLTACQLSTADDTRDKQKDRLVRLFRKKRDSSDSGLLPRTCVMFGSMCDKEEHSICCYECDYKKGICV
4.	MKLTCVVIVAVLLLTACQLIIAEDSRGTQLHRALRKATKLSVSTRTCVMFGSMCDKEEHSICCYECDYKKGICV
5.	MKLTCVLIVAVLFLTACQLATAENSREEQGYSAVRSSDQIQDSDLKLTKSCTDDFEPCEAGFENCCSKSCFEFEDVYVC*GVSIDYYDSR
6.	MKLICVFIVAVLLLTACQLNAADDSRDTQKHRALRSTTKLSMSKKDSCVPDGDSCLFSRIPCCGTCSSRSKSCV*G
7.	MKLTCMMIVTVLFLTAWTFVTADDSTYGLKNLLPKARHEMMNPEAPKLNKKDECSAPGAFCLIRPGLCCSEFCFFACF [67]
8.	AEDSRGTQLHRALRKATKLSESTRCKRKGSSCRRTSYDCCTGSCRNGKC*G
9.	AVLLLTACQLITAEDSRDTQKHRALRSDTKLSMLTLRCATYGKPCGIQNDCCNICDPARRTCT
10.	DSRGTQLHRALRKATILSVSARCKLSGYRCKRPKQCCNLSCGNYMC*G
11.	ACQLITAEDSRGTQLHRALRSTSKVSKSTSCVEAGSYCRPNVKLCCGFCSPYSKICMNFPKN
12.	TAEDSRGTQLHRALRKATKLPVSTRCITPGTRCKVPSQCCRGPCKNGRCTPSPSEW
13.	AEDSRGTQLHRALRKTTKLSLSIRCKGPGASCIRIAYNCCKYSCRNGKCS
14.	AACQLGTAASFARDKQDYPAVRSDGRQDSKDSTLDRIAKRCSEGGDFCSKNSECCDKKCQDEGEGRGVCLIVPQNVILLH
M-superfamily: CC-C-C-CC	
15.	MLKMGVLLFTFLVLFPLATLQLDADQPVERYADNKQDLNPDERMIFLFGGCCRMSSCQPPPVCNCCAKQDLNPDER
16.	DQPADRPAERMQDDISSEQNPLLEKRVGERCCKNGKRGCGRWCRDHSRCC*GRR [17]
17.	GLYCCQPKPNGQMMCNRWCEINSRCC*GRR
A-superfamily: CC-C-C; CC-C-C-C-C	
18.	MGMRMMFTVFLLIVLATTVVSFSTDDESDGSNEEPSADQTARSSMNRAPGCCNNPACVKHRC*G [68]
19.	MGMRMVFTVFLLVVLATTVVSFTSDRASDGRNAAANDKASDLAALAVRGCCHDIFCKHNNPDIC*G
20.	MGMRMRMMFTVFLLVVLANTVVSFPSDRDSDGADAEASDEPVEFERDENGCCWNPSCPRPRCT*GRR [68]
21.	DGANAEATDNKPGVFERDEKKCCWNRACTRLVPCSK
22.	SDRASDGRNAAANDRASDLVALTVRGCCTYPPCAVLSPLCD
23.	MGMRMMVTVFLLGVLATTVVSLRSNRASDGRRGIVNKLNDLVPQYWTECCGRIGPHCSRCICPEVVCPKN*G
24.	MGMRMMVTVFLLVVLATTVVSLRSNRASDGRRGIVNKLNDLVPKYWTECCGRIGPHCSRCICPEVACPKN*G
25.	MGMRMMVTVFPLVVLATTVVSLRSNRASDGRRGIVNKLNDLVPKYWTECCGRIGPHCSRCICPGVVCPKR*G
26.	LVVLATTVVSFRSNRASDGRKIAVNKRRRELVVPPGKLRECCGRVGPMCPKCMCPPRRC
27.	ASDGRNAVVHERAPELVVTATTTCCGYDPMTICPPCMCTHSCPPKRKP*GRRND
J-superfamily	
28.	MTSVQSATCCCLLWLVLCVQLVTPDSPATAQLSRHLTARVPVGPALAYACSVMCAKGYDTVVCTCTRRRG*VVSSSI
Contryphan	
29.	MGKLTILVLVAAVLLSTQVMGQGDRDQPAARNAVPRDDNPGGASAKLMNLLHRSKCPWSPWC*G
Conkunitzin	
30.	MEGRRFAAVLILPICMLAPGAVASKRWTRPSVCNLPAESGTGTQSLKRFYYNSDKMQCRTFIYKGNGGNDNNFPRTYDCQKKCLYRP*G

Cysteine motifs are shown next to the superfamilies. The underlined residues indicate presumed propeptide cleavage site ascertained by analogy to previously isolated toxins; * indicate probable amidation at the C-terminal residue after cleavage of the following G residue. In the case of 23,24,25,26 where the propeptide cleavage site is uncertain, we have indicated the cleavage site at the basic residues (K) proximal to the presumed toxin sequence. The peptides Bu 7, 16, 18 and 20 have been previously characterized.

Table 3 Sequence diversity and classification of A-superfamily conopeptides from *Conus bullatus*

Subclasses of A-superfamily peptides (Mature toxin sequences)		
α4/4		
	Bu18	APGCCNNPACVKHRC*
	Bu20	DENGCCWNPSCPRPRCT*
α4/5		
	Bu21	CCWNRACTRLVPCSK
α4/6		
	Bu22	GCCTYPPCAVLSPLCD
α4/7		
	Bu19	GCCHDIFCKHNNPDIC*
κA		
	Bu27	APELVVTATTTCCGYDPMTICPPCMCTHSCPPKRKP*
κA-like		
	Bu23	LNDLVPQYWTECCGRIGPHCSRCICPEVVCPKN*
	Bu24	YWTECCGRIGPHCSRCICPEVACPKN*
	Bu25	YWTECCGRIGPHCSRCICPGVVCPKR*
	Bu26	LRECCGRVGPMCPKCMCPPRRC

*C-terminal is amidated. We have assumed that the proteolytic cleavage site is at the basic residue proximal to the presumed toxin sequence.

A-superfamily than has been found in any other venom. Three subclasses of α-conotoxins represented two different α4/4 peptides (Bu18 and 20), one α4/5 peptide (Bu21) and one α4/6 peptide (Bu22). Unique to *C. bullatus* are the four A peptides with 3 disulfide bonds (Bu 23, 24, 25 and 26) which are divergent from both κA and αA families. It is notable that although these comprise a significant fraction of the total complement of A-superfamily peptides in *C. bullatus*, similar peptides have not been reported from any other species thus far. Thus, it appears that *Conus bullatus*, and potentially the *Textilia* clade of Conus species, has explored novel evolutionary pathways in generating their complement of A-gene superfamily peptides.

SNP rates in conopeptides

We also compared the single nucleotide heterozygosity level within the transcripts encoding conopeptides to the rest of the transcriptome. To reduce false negative rates, we restricted our analysis to transcriptome contigs having coverage depths of 10× or more. Our rationale being that SNPs within low-coverage contigs might be missed, leading us to underestimate the actual SNP rate. For the transcriptome as a whole, the SNP rate is 0.0035 (102,955 SNPs in 29.5 MB of high-coverage contigs). By contrast, the single nucleotide polymorphism rate within conotoxin contigs is 0.011 (1146 SNPs in 105,259bp of high-coverage conotoxin contigs; this is 64% of all conotoxin contigs by length). The 3.1-fold higher SNP rate within conopeptides contigs is consistent with the hypothesis that conopeptides are under diversifying selection.

Candidate post-translational processing enzymes

Conopeptides contain post-translationally modified amino acids. These modifications play an important role in conferring target specificity. The most ubiquitous modification is the formation of disulfides leading to proper conotoxin folding; this mediated by disulfide isomerases, chaperones and enzymes involved in redox biochemistry. From an examination of transcriptome sequences we have identified partial and complete sequences of several chaperones and thiol-disulfide oxidoreductases that are likely to be involved in the redox biochemistry of conotoxin folding (Additional File 2).

We identified some of the enzymes that are presumed to catalyze correct disulfide connectivity within conopeptides [43-46]. These include members of the QSOX family of sulfhydryl oxidases, Ero oxidases and protein disulfide isomerases (PDIs). PDIs also have chaperone-like activity and prevent protein aggregation. We have identified three isoforms of protein disulfide isomerase (PDI) and four members belonging to different subfamilies of PDIs. Two of these are members of the P5 subfamily. We also identified a transcript related to human PDIRs, which carry out oxidation-isomerization functions similar to PDI, but are less active. We also identified a transcript encoding a second redox inactive TRX domain b' belong to Ep72 and Ep57 subfamily. In addition, transcriptome contigs with homology to several Chaperones, including 78kDa glucose regulated protein, Hsp70, Hsp60, Hsp90, glucose regulated protein 94, different subunits of the T-complex protein 1, DNA J (Hsp40), calnexin, calreticulin, chaperonin 10kDa subunit, prefoldin superfamily and activator of Hsp90 ATPase I were also identified.

The other enzymes we have identified include a proline hydroxylase related to the enzyme involved in collagen biosynthesis. (Unrelated to the posttranslational modification of peptides, we have also identified the *egl nine* homolog-also a prolyl hydroxylase). We have identified both FK506 binding protein type peptidyl prolyl cis-trans isomerase and the cyclophilin peptidyl prolyl cis-trans isomerase. The latter type has been shown to enhance the rate of correct folding of conopeptides containing proline residues [47]. Other enzymes identified include lysyl hydroxylase, vitamin K dependent γ-glutamyl carboxylase [48,49], vitamin K epoxide reductase and peptidyl glycine alpha amidating monooxygenase.

A large number of hormones and neuro-active peptides require C-terminal amidation for full activity [50-52]; conopeptides are no exception. C-terminal amidation is a two-step process. Peptidylglycine α-hydroxylating monooxygenase (PHM) catalyzes the hydroxylation of the α-carbon of glycine and a second enzyme, peptidyl-α-hydroxy glycine α-amidating lyase (PAL) catalyzes the formation of the amidated product and glyoxylate. In *Drosophila* these two activities are carried out by separate polypeptides, whereas in other organisms (*C.elegans*, *Xenopus laevis*, human and rat) a single polypeptide carries out both activities. We discovered a single transcriptome contig encoding both PHM and PAL domains, thus C-terminal amidation of conopeptides is likely carried out by a single enzyme in *C. bullatus*.

A unique posttranslational modification first identified in *Conus* was the presence of 6-Br tryptophan in conopeptides, e.g. bromocontryphan [53], bromosleeper [54] and light sleeper [55]. Subsequently the modification was also characterized in a peptide isolated from mammalian brain [55-57]. The enzyme responsible for this modification has not been characterized. However, four different classes of haloperoxidases are known [58], which are enzymes that use heme iron/H_2O_2, vanadium/H_2O_2, $FADH_2/O_2$, and non-heme iron/O_2/α-ketoglutarate. In the present analyses we have not identified any of the above classes of enzymes.

Another posttranslational modification is the isomerization of L-amino acids in peptides to the D-conformation [59]. The enzyme has been isolated from the funnel web spider venom [60]. At present we have not identified any transcript possibly encoding the isomerase.

A novel method for estimating genome size

We have developed a novel method for determining genome size, using 2nd generation genomic and RNA-Seq reads (see Methods). For proof of principle, we first estimated the genome size of *D. melanogaster*. To do so, we simulated 4,342,253 59bp genomic reads for the fly-genome, and blasted the annotated fly transcriptome against the simulated reads (red line in Figure 4). The depth of coverage peak is at 1.50 (Figure 4). Thus, the estimated genome size for *D. melanogaster* is 4,342,253*59/1.50 = 170.8 MB. Compared to the current size of fly genome (166.6 MB), the error is 2.5%. We also estimated the genome size of *C. elegans*. This time we randomly sheared the annotated transcriptome of *C. elegans* into short contigs with the same N50 as our *C. bullatus* transcriptome assembly, and randomly selected a 57mb subset of these contigs. We did this to simulate the fragmented nature of our *de novo* transcriptome assembly. We also simulated 2,630,408 genomic *C. elegans* reads, and blasted them to the subset of simulated *C. elegans* transcriptome. As shown in Figure 4 (green line), the peak depth of coverage for the transcriptome is 1.45×. We repeated this experiment three times; there was no variance in this value. This gives us an estimate of genome size of 107.0MB, which is 6.7% higher than estimated genome size (100.3MB), again a good fit to the published genome size. For *Conus bullatus*, the estimated coverage depth is 1.70× from 4.36GB of sequence reads, thus the best estimate for the size of the *Conus bullatus* genome is 2.56 GB.

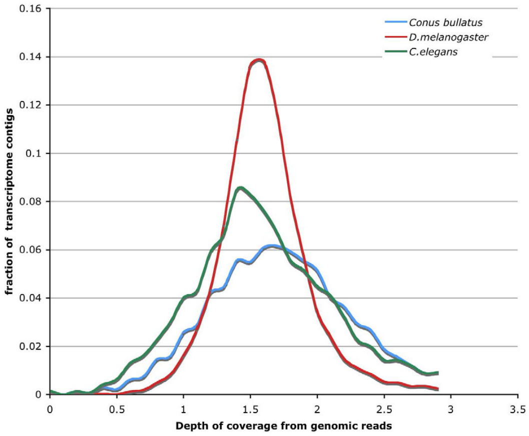


Figure 4 ***C. bullatus*** **genome size estimated using Illumina reads**. Blue-line: *C. bullatus*; Red-line: *D. melanogaster*. Green line: *C. elegans*. x-axis: depth of coverage of transciptome contigs by aligned genomic reads. y-axis: frequency. In all cases the best estimate for genome size is the product of the total length of genomic reads and the mode of the frequency distribution.

Discussion

2nd generation sequencing technologies now make it possible to probe new and emerging model organism genomes in a cost effective manner. This means that genomes and transcriptomes can be rapidly trawled for specific contents, and at the same time the organism can be evaluated for suitability of whole-genome *de novo* assembly. We have tried to accomplish both these tasks in the work reported here.

Our transcriptome analyses provide the first global view of gene expression within a *Conus* venom-duct. Several lines of evidence suggest that our dataset provides a relatively comprehensive view of this pharmacologically important tissue. First, the relative proportion of *C. bullatus* genes (as discovered by annotating out transcriptome data) assigned to different GO terms resemble those of other well annotated transcriptomes. Second, 85% of CEGMA's universally conserved eukaryotic genes are represented by one or more contigs, providing an independent estimate of the degree of completeness of the assembly. One caveat to this conclusion is that highly expressed basic house keeping genes are over represented in the CEGMA set; thus a more precise statement is that 85% of highly expressed genes are present in the RNA-seq data.

Our RNA-seq data are highly enriched for reads with conopeptide homology. The average read depth of contigs homologous to conopeptides is 102× as opposed to 33X for the remaining contigs. Interestingly, their superfamily frequency spectrum roughly approximates that of the Conoserver reference collection in general [42], although some rare classes are missing.

Overall, the distribution and frequencies of GO functions, processes and locations of annotated transcriptome data closely parallel those of various carefully annotated model organism transcriptomes (Additional File 1); this fact suggests that overall, the venom-duct transcriptome is diverse, despite the highly specialized nature of this tissue. Although, as our recovery of numerous conopeptides and post-translational modification (PTM)-enzymes makes clear, its transcription is also clearly geared toward venom production. Our success at characterizing the conopeptide and candidate PTM-enzymes demonstrates the power of the RNA-Seq approach for conopeptide discovery. The conopeptide and PTM-enzymes we have discovered present new avenues for future research, as it is now possible to express these proteins in heterologous cells in order to explore interactions PTM-enzymes and their conopeptide targets [47,48,61,62].

Our genomic shotgun survey data have allowed us to characterize the *C. bullatus* genome. Our analyses indicate that it is enriched for simple repeats relative to the human genome. Characterization of its interspersed repeat populations is complicated by the lack of an adequate repeat library for RepeatMasker. To circumvent this obstacle, we developed a novel analysis method, comparing the inter-read similarity frequency spectrum of our *C. bullatus* genome reads to the inter-read similarity frequency spectrum of matched human dataset. Based upon this analysis we conclude that *C. bullatus* has higher repeat content, yet contains fewer extremely high-copy repeat species. Because this method requires no assembly or prior knowledge of a genome's repeat content, it should prove useful to others seeking to characterize the repeat contents of new and emerging model genomes.

Conclusions

We have carried out the first transcriptome and genomic survey of a *Textilian*, *Conus bullatus*. Our RNA-seq analyses provide the first global view of transcription within a *Conus* venom duct, and demonstrate the feasibility of trawling these data for rapid discovery of new conopeptides and PTM-enzymes. We find that numerous A-superfamily peptides are expressed in the venom duct. These conopeptides are unprecedented in their structural diversity, suggesting that *Conus bullatus*, and potentially the *Textilia* clade in general, has explored novel evolutionary pathways in generating its complement of A-gene super-family peptides. Our data also provide support for the long-standing hypothesis that conopeptides are under diversifying selection. Our genomic analyses have revealed that the *C. bullatus* genome has higher content of interspersed repeats, yet fewer extremely high-copy-number repeats compared to human.

Methods

Preparation of RNA samples

Specimens of *Conus bullatus* were collected in the Philippines. Each specimen was dissected to isolate the venom duct and the duct was immediately suspended in 1.0 mL RNAlater solution (Ambion, Austin, TX) at ambient temperatures, and then stored at -20 degrees Centigrade until used. Total RNA was isolated using *mir*Vana® miRNA isolation kit (Ambion, Applied Biosystems CA USA) according to the manufacturer's recommendation Tissue homogenization was carried out using a tissue tearor (Model 985370, Dremel, WI, USA).

Simulated read sets

To produce the matching sets of reads from other genomes with which to compare our *C. bullatus reads*, we randomly sampled some number of read pairs from our *Conus* dataset. Next we randomly selected substrings from an assembled target genome (e.g. human, *D. melanogaster*, etc.) having the same length and pair distances as our *Conus* reads. This matched dataset mimics the *Conus* data precisely as regards number of reads, distance between pairs, read lengths, and importantly base quality. This last feature is accomplished by mutating the simulated reads from the target genome using the base quality values of the selected *Conus* reads. These matched datasets enable many useful analyses. For example, a set of 1,000,000 randomly selected *Conus* genomic reads can be passed through RepeatMasker and the results directly compared to that produced from its matched human counterpart.

Partial genome assembly

We generated a total of 152 million Illumina genomic reads, with read lengths of either 59bp or 60bp depnding upon run. The reads are paired-end, and have a average insertion size of 200bp. We used the 'quality-Trimmer' algorithm in the EULER-SR software package [63] to remove bad reads and trim low-quality region from reads. We then used ABySS 1.0.15 [34] for assembly, with the following parameters: c = 0, e = 2, n = 2. The k-mer size is an important factor for the quality of assembly, and in order to make an informed decision about the k-mer size, we assembled the *C. bullatus* genome with k = 25, 30, 35, 40, 45 and 50. The k-mer size of 25 generate an assembly with the best total length (201MB) and N50 (182bp). The assembly was filtered so that contigs/scaffolds with lengths less than 100 bp were removed. When aligning the genomic reads back to the *de novo* assembly, 3.6 million reads aligned.

Assembly of the transcriptome

102 million paired-ended RNA-seq reads were generated using the Illumina sequencing platform. The read lengths for these runs were 79bp, with an average insertion size of 340bp. These reads were first filtered with EULER-SR's 'qualityTrimmer' algorithm as above, then assembled by ABySS 1.0.15 using the following parameters: c = 0, e = 2, E = 0. k-mer size of 25, 30, 35, 40, 45, 50 were tested, and the assembly at k = 35 were chosen in consideration for the total assembly size as well as N50. The assembly was filtered so that contigs/scaffolds with lengths less than 60 bp were removed.

To assess the quality of the transcriptome assembly, we aligned the RNA-seq reads back to the assembly with Bowtie. Out of 102 million reads, 31million aligned to the transcriptome under single-end alignment mode. A much smaller portion (3.2 million) of reads were aligned under paired-end mode. This is expected because our library should be enriched for short conopeptide sequences, thus many fragments should be shorter than 340bp, which will produce overlapping paired-reads that won't align under paired-end mode of Bowtie.

Characterization of repeat content in the genomic assembly

We randomly selected 1 million Illumina reads for the genome of *Conus bullatus*. As a control, we used the reference genomes of *Aplysia californica, Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens* from NCBI database. For each of the control genomes, 1 million Illumina reads with the same length and base-calling accuracy distribution were simulated. We also used a second control consisting of 100,000 real Illumina genomic reads randomly sampled from the Flatley genome [30]. We ran RepeatMasker with the '-species all' option in order to characterize all known families of interspersed repeats. These data are shown in Figure 2.

Novel repeat families with *Conus bullatus* genome were identified by running RECON over the longest genomic contigs with a total length of 30MB (masked by RepeatMasker beforehand). We then perfromed an all-by-all BLASTN of the contigs against themselves, using an E-value threshold of $1e^{-8}$. The blastn reports were converted into MSP files and fed to RECON to identiy genomic sequences present in no less than 10 copies in the 30MB sample sequence. 115 high-copy-number sequences were identified, and any of them that have significant homology ($1e^{-5}$) with a UniprotKB or Repbase entry were removed from the novel interspersed repeats collection.

Estimation of the proportion of repetitve regions

1 million genomic reads from the conus genome were randomly selected; 1 million human genomic reads were then simulated with the same length and base-calling accuracy. We aligned each set of reads to themselves with BLASTN to look for significant similarity (M = 1 N = -3 Q = 3 R = 3 W = 15 WINK = 5 filter = seg lcmask V = 1000000 B = 1000000 E = 1e-5 Z = 3000000000). The percentage of reads having each number of BLAST hit were then tallied.

To convert the number of BLAST hits to the copy-number of their corresponding genomic sequence, we simulated a genome with the same size as the human genome and the following features: 38% of this genome are comprised of unique sequence; 20% are sequences with 2 copies; 10% of the genome have 5 copies, 10 copies, 100 copies and 1000 copies each; 1% of the genome have 10,000 and 100,000 copies each. Then we simulated 1 million reads from this genome with the same length and base-calling accuracy as the Conus genomic reads and performed an all-to-all blast approach as described above. For each read generated, we tracked the copy number of the genomic region that it is extracted from. Then we calculated the average number of read partners for reads from different copy-number region. As Additional File 3 shows, the average number of read partners is correlated extremely well with the copy-number of the genomic region the read was drawn from. The equation in Additional File 3 allows us to profile the proportion of genomic regions with different copy-numbers, as shown in Figure 3.

SNP rates

To estimate SNP rates within our transcriptome assembly, Illumina reads were aligned to contigs no shorter than 60bp in the transcriptome assembly, using Bowtie [64] with default parameters. With the samtools package, the resulting Bowtie report was converted into SAM files [65], then used to estimate the SNP ratio with samtools. We used stringent criteria to call SNPs, requiring that: 1) the SNP phred score was higher than 20; and 2) that each SNP variant was supported by at least two reads. The SNP rates within conopeptides were estimated using a same approach. We also calculated the proportion of triallelic SNPs, which is 15%, indicative of the upper bound of the false-positive rate due to mis-alignment.

BLAST searches for conopeptides

We ran BLASTX on our transcriptomal assembly against the combined database of UniProtKB [40] and conotoxins from ConoServer [42], using the following parameters: W = 4 T = 20 filter = seg lcfilter. Contigs that hit a conopeptide as its best hit were collected as the low-stringency conopeptide set, and subsequently translated into peptides according to the reading frame identified by BLASTX. We then ran BLASTP on the low-stringency conopeptides against the combined database, using the following parameters: hitdist = 40 wordmask = seg postsw matrix = BLOSUM80. The results are filtered with E < = $3e^{-5}$.

Assignment of putative conotoxins to superfamilies

We first translated each putative conotoxin conteg into peptide sequence, using the reading-frame predicted from BLASTing the RNA-seq assembly to ConoServer's collection of conopeptides. Each translated putative-conopeptide was then aligned with BLASTP to conotoxin signal peptides sequences, downloaded from ConoServer. We required all aligments to have Expect < = 1e-4, and to have at least 7 identical amino acids aligned. The best hit for each putative conopeptide is used to predict its superfamily. Overall, we were able to assign 543 putative conopeptides to a superfamily. As a control, we downloaded previously reported conopeptides from ConoServer, and randomly sheared these sequences into short oligos with the same N50 as our putative conopeptide contigs. We applied the same approach to assign these to superfamiles. Out of 3274 oligos, we were able to assign 449 to a superfamily, of which 443 (98.7%) were correct. Thus, we believe our assignment method is reasonably accurate.

Genome size estimation

We ran WU-BLASTN over all transcriptomal contigs longer than 300bp against 73,898,732 59-mer genomic reads, with the following parameters: M = 1 N = -3 Q = 3 R = 1 wordmask seg lcmask. The coverage depth for each transcript was calculated from dividing total length of reads mapped to this transcript by its transcript length. Then the frequency distribution is shown in Figure 4. The estimated coverage depth for the genome is determined as the coverage depth with the highest frequency, which is 1.70×. The estimated genome size for *Conus bullatus* is thus 73,898,732*59/1.70 = 2.56×10^9 bp.

Significance of conopeptide BLAST hits

The short reads and base quality issues combine with the short lengths of conopeptides to make identification of conopeptides in RNA-seq data difficult. Because many conopeptide transcript species are represented by only one or a few reads, the base-quality of the resulting contig is often low, especially as regards indels. All of these facts combine to make the detection of even highly conserved conopeptides problematic, because BLASTX is unable to take into account indel induced frameshifts in the contigs when calculating the significance of a hit [35], thus many real hits are not detected. Also problematic is the cysteine-rich nature of conopeptides, leading to spuriously significant hits against other non-homologous but cysteine-rich proteins, and protein domains. To control for these issues we performed a simulation to help us determine the best E-value threshold for a conopeptide hits in RNA-seq data. We first ran WU-BLASTP [36] on our transcriptome assembly against the combined database of UniProtKB [40] and conopeptides from ConoServer [42]. In total, 6,677 peptides were found to have a known conopeptide as its best hit. We then plotted the E-value distribution of the BLAST results for the best HSPs (Additional File 4). Next, we randomly permuted the sequences of each of our 6,677 *C. bullatus* contigs with conopeptide hits using a Fisher-Yates shuffle [66]. We then ran BLASTP using the permuted peptides against the combined UniProtKB and conotoxins database, and plotted the E-value distribution for all hits. Presumably, the latter plot should represent the background distribution of insignificant BLAST hits. We found that only 5% of the hits in the permuted peptide set have an E-value of lower than 3e-5, while in the putative conopeptide set, the percentage is 48%. Thus we used E < = $3e^{-5}$ as the E-value threshold for our BLASTP searches for conopeptides.

Data and software availability

The read-simulation tool and data (transcriptome assembly, genomic assembly, putative conotoxin sequences and post-translational modification enzymes) can be downloaded at http://derringer.genetics.utah.edu/conus/. The software is open source.

Additional material

Additional File 1: GO analyses. GO term abundance for molecular function. In each organism (colored as in the legend), each transcript was assigned applicable high-level generic GO slim terms. The occurrence of each GO term was counted and converted into frequency among all GO terms. Similar congruency between transcriptomes was seen for GO process and location terms.

Additional File 2: Proteins involved in post-translational modification. Annotated list of proteins that are presumed to participate in conotoxin synthesis and posttranslational modification. Deduced from conceptual translation of transcripts (ESTs) present in the venom duct.

Additional File 3: Correlation between Average read partner number (from all-by-all BLAST) and actual copy number of corresponding genomic sequence. A human-size genome is simulated so that certain fractions of the sequence are present in 1 copy, 2 copies, 5 copies, 10 copies, 100 copies, 1000 copies, 10,000 copies and 100,000 copies. The average read partner count for reads simulated from each group is calculated and used for the plot.

Additional File 4: Determining the appropriate BLAST E-value for identification of conotopeptides. Red-line: E-value frequencies for all contigs with conopeptide homology. Blue-line:E-value frequencies for the same set of contigs after permutation. X-axis: frequency; y-axis E-value. 5% of the permuted contigs have an E-value of less than 3e-5, compared to 45% of the native set. Thus, we choose 3e-5 as our cutoff threshold for a 0.05 confidence level.

Acknowledgements
We thank Roche 454 sequencing for helping to make this work possible by contributing transcriptome data to the project; especially C. Kodira for help and support. We also acknowledge Cofactor Genomics for their help and support in obtaining the Illumina RNA-seq and Genomic data. This work was supported in part by NHGRI NIH grant 1R01HG004694 to MY, GM48677 (BMO, PB) and a University of Utah seed grant to MY and PB.

Author details
[1]Eccles institute of Human Genetics, University of Utah, and School of Medicine, Salt Lake City, UT 84112, USA. [2]Department of Biology, University of Utah, Salt Lake City, UT 84112, USA.

Authors' contributions
HH, PB and MY wrote the paper. HH wrote software and carried out experiments. PB annotated and analyzed results. MY, PB and BO conceived of the project and oversaw the experiments. All Authors read and approved the final manuscript.

Received: 23 August 2010 Accepted: 25 January 2011
Published: 25 January 2011

References

1. **Tools for genetic and genomic studies in emerging model organisms.** [http://grants.nih.gov/grants/guide/pa-files/PA-04-135.html].
2. Alvarado AS, Newmark PA, Robb SMC, Juste R: **The Schmidtea mediterranea database as a molecular resource for studying platyhelminthes, stem cells and regeneration.** *Development* 2002, **129(24)**:5659-5665.
3. Bouchet P, Rocroi JP: **Malacologia: International Journal of Malacology, Classification and Nomenclator of Gastropod Families.** Conch Books; 2005**47.**
4. Terlau H, Olivera BM: ***Conus* venoms: a rich source of novel ion channel-targeted peptides.** *Physiological Reviews* 2004, **84**:41-68.
5. Conticello SG, Gilad Y, Avidan N, Ben-Asher E, Levy Z, Fainzilber M: **Mechanisms for evolving hypervariability: the case of conopeptides.** *Mol Biol Evol* 2001, **18(2)**:120-131.
6. Lynch SS, Cheng CM, Yee JL: **Intrathecal ziconotide for regractory chronic pain.** *Ann Pharmacother* 2006, **40**:1293-1300.
7. Miljanich GP: **Ziconotide: neuronal calcium channel blocker for treating severe chronic pain.** *Current Medicinal Chemistry* 2004, **11**:3029-3040.
8. Han TS, Teichert RW, Olivera BM, Bulaj G: **Conus venoms - a rich source of peptide-based therapeutics.** *Curr Pharm Des* 2008, **14(24)**:2462-2479.
9. Lewis RJ, Garcia ML: **Therapeutic Potential of Venom Peptides.** *Nat Rev Drug Discov* 2003, **2(10)**:790-802.
10. Olivera BM, Teichert RW: **Diversity of the neurotoxic Conus peptides: a model for concerted pharmacological discovery.** *Molecular Interventions* 2007, **7(5)**:251-260.
11. Wang CZ, Chi CW: ***Conus* Peptides - A rich Pharmaceutical Treasure.** *Acta Biochimica et Biophysica Sinica* 2004, **36(11)**:713-723.
12. Olivera BM: **Conus peptides: biodiversity-based discovery and exogenomics.** *Journal of Biological Chemistry* 2006, **281(42)**:31173-31177.
13. Röckel D, Korn W, Kohn AJ: **Manual of the living Conidae.** Hackenheim, Germany: Verlag Christa Hemmen; 1995.
14. Terlau H, Shon KJ, Grilley M, Stocker M, Stuhmer W, Olivera BM: **Strategy for rapid immobilization of prey by a fish-hunting marine snail.** *Nature* 1996, **381(6578)**:148-151.
15. Olivera BM: ***Conus* venom peptides, receptor and ion channel targets and drug design: 50 million years of neuropharmacology (E.E. Just Lecture, 1996).** *Mol Biol Cell* 1997, **8**:2101-2109.
16. Azam L, Dowell C, Watkins M, Stitzel JA, Olivera BM, McIntosh JM: **Alpha-conotoxin BuIA, a novel peptide from Conus bullatus, distinguishes among neuronal nicotinic acetylcholine receptors.** *J Biol Chem* 2005, **280(1)**:80-87.
17. Holford M, Zhang MM, Gowd KH, Azam L, Green BR, Watkins M, Ownby JP, Yoshikami D, Bulaj G, Olivera BM: **Pruning nature: Biodiversity-derived discovery of novel sodium channel blocking conotoxins from Conus bullatus.** *Toxicon* 2009, **53(1)**:90-98.
18. Bandyopadhyay P, Stevenson BJ, Ownby JP, Cady MT, Watkins M, Olivera BM: **The Mitochondrial Genome of Conus textile, coxI-coxII Intergenic Sequences and Coinoidean Evolution.** *Molecular Phylogenetics and Evolution* 2007.
19. Bandyopadhyay PK, Stevenson BJ, Cady MT, Olivera BM, Wolstenholme DR: **Complete mitochondrial DNA sequence of a Conoidean gastropod, *Lophiotoma (Xenuroturris) cerithiformis*: Gene order and gastropod phylogeny.** *Toxicon* 2006, **48**:29-43.
20. Biggs JS, Olivera BM, Kantor YI: **Alpha-conopeptides specifically expressed in the salivary gland of Conus pulicarius.** *Toxicon* 2008, **52(1)**:101-105.
21. Duda TF Jr, Palumbi SR: **Evolutionary diversification of multigene families: allelic selection of toxins in predatory cone snails.** *Mol Biol Evol* 2000, **17**:1286-1293.
22. Duda TF Jr, Kohn AJ: **Species-level phylogeography and evolutionary history of the hyperdiverse marine gastropod genus Conus.** *Mol Phylogenet Evol* 2005, **34(2)**:257-272.
23. Espiritu DJD, Watkins M, Dia-Monje V, Cartier GE, Cruz LJ, Olivera BM: **Venomous cone snails: molecular phylogeny and the generation of toxin diversity.** *Toxicon* 2001, **39**:1899-1916.
24. Santos AD, McIntosh JM, Hillyard DR, Cruz LJ, Olivera BM: **The A-superfamily of conotoxins: structural and functional divergence.** *Journal of Biological Chemistry* 2004, **279**:17596-17606.
25. Twede VD, Teichert RW, Walker CS, Gruszczynski P, Kazmierkiewicz R, Bulaj G, Olivera BM: **Conantokin-Br from Conus brettinghami and selectivity determinants for the NR2D subunit of the NMDA receptor.** *Biochemistry* 2009, **48(19)**:4063-4073.
26. Walker C, Steel D, Jacobsen RB, Lirazan MB, Cruz LJ, Hooper D, Shetty R, DelaCruz RC, Nielsen JS, Zhou L, *et al*: **The T-superfamily of conotoxins.** *J Biol Chem* 1999, **274**:30664-30671.
27. Pi C, Liu J, Peng C, Liu Y, Jiang X, Zhao Y, Tang S, Wang L, Dong M, Chen S, *et al*: **Diversity and evolution of conotoxins based on gene expression profiling of *Conus litteratus*.** *Genomics* 2006, **88(6)**:809-819.
28. Pi C, Liu Y, Peng C, Jiang X, Liu J, Xu B, Yu X, Yu Y, Jiang X, Wang L, *et al*: **Analysis of expressed sequence tags from the venom ducts of *Conus striatus*: focusing on the expression profile of conotoxins.** *Biochimie* 2005, **88(2)**:131-140.
29. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M: **Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing.** *Biotechniques* 2008, **45(1)**:81-94.
30. [http://developer.amazonwebservices.com/connect/entry.jspa?externalID=3357].
31. Smit A, Hubley R, Green P: **RepeatMasker Open-3.0.**1996-2004.
32. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110(1-4)**:462-467.
33. Bao Z, Eddy SR: **Automated de novo identification of repeat sequence families in sequenced genomes.** *Genome Res* 2002, **12(8)**:1269-1276.
34. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19(6)**:1117-1123.
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215(3)**:403-410.
36. [http://blast.wustl.edu/].
37. Hinegardner R: **Cellular DNA content of the Mollusca.** *Comp Biochem Physiol A Comp Physiol* 1974, **47(2)**:447-460.
38. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23(9)**:1061-1067.
39. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
40. Apweiler R, Bairoch A, Wu CH: **Protein sequence databases.** *Curr Opin Chem Biol* 2004, **8(1)**:76-80.

41. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1)**:25-29.
42. Kaas Q, Westermann JC, Halai R, Wang CK, Craik DJ: **ConoServer, a database for conopeptide sequences and structures.** *Bioinformatics* 2008, **24(3)**:445-446.
43. Appenzeller-Herzog C, Ellgaard L: **The human PDI family: Versatility packed into a single fold.** *Biochimica et Biophysica Acta* 2007, **1783**:535-548.
44. Hatahet F, Ruddock LW: **Protein Disulfide Isomerase: A Critical Evaluation of Its Function in Disulfide Bond Formation.** *Antioxidants & Redox Signaling* 2009, **11(11)**:2807-2839.
45. Thorpe C, Coppock DL: **Generating Disulfides in Multicellular Organisms: Emerging Roles for a New Flavoprotein Family.** *Journal of Biological Chemistry* 2007, **282(19)**:13929-13933.
46. van Anken E, Braakman I: **Versatility of the Endoplasmic Reticulum Protein Folding Factory.** *Critical Reviews in Biochemistry and Molecular Biology* 2005, **40**:191-228.
47. Safavi-Hemami HBG, Olivera BM, Williamson NA, Purcell AW: **Identification of Conus peptidyl Prolyl cis-trans isomerases (PPIases) and assessment of their role in the oxidative folding of conotoxins.** *J Biol Chem* 2010.
48. Bandyopadhyay PK, Garrett JE, Shetty RP, Keate T, Walker CS, Olivera BM: **γ-Glutamyl carboxylation: an extracellular post-translational modification that antedates the divergence of molluscs, arthropods and chordates.** *Proc Natl Acad Sci USA* 2002, **99**:1264-1269.
49. Czerwiec E, Begley GS, Bronstein M, Stenflo J, Taylor KL, Furie BC, Furie B: **Expression and characterization of recombinant vitamin K-dependent γ-glutamyl carboxylase from an intvertebrate, *Conus textile*.** *Eur J Biochem* 2002, **269**:6162-6172.
50. De M, Ciccotosto GD, Mains RE, Eipper BA: **Trafficking of a Secretory Granule Membrane Protein Is Sensitive to Copper.** *Journal of Biological Chemistry* 2007, **282(32)**:23362-23371.
51. Eipper BA, Milgram SL, Husten EJ, Yun HY, Mains RE: **Peptidylglycine α-amidating monooxygenase: a multifunctional protein with catalytic, processing, and routing domains.** *Protein Sci* 1993, **2**:489-497.
52. Prigge ST, Mains RE, Eipper BA, Amzel LM: **New insights into copper monooxygenases an peptide amidation: structure, mechanism and function.** *Cellular and Molecular Life Sciences* 2000, **57**:1236-1259.
53. Jimenez EC, Craig AG, Watkins M, Hillyard DR, Gray WR, Gulyas J, Rivier JE, Cruz LJ, Olivera BM: **Bromocontryphan: post-translational bromination of tryptophan.** *Biochemistry* 1997, **36**:989-994.
54. Craig AG, Jimenez EC, Dykert J, Nielsen DB, Gulyas J, Abogadie FC, Porter J, Rivier JE, Cruz LJ, Olivera BM, *et al*: **A novel post-translational modification involving bromination of tryptophan: identification of the residue, L-6-bromotryptophan, in peptides from *Conus imperialis* and *Conus radiatus* venom.** *J Biol Chem* 1997, **272**:4689-4698.
55. Jimenez EC, Watkins M, Olivera BM: **Multiple 6-bromotryptophan residues in a sleep-inducing peptide.** *Biochemistry* 2004, **43**:12343-12348.
56. Fujii R, Yoshida H, Fukusumi S, Habata Y, Hosoya M, Kawamata Y, Yano T, Hinuma S, Kitada C, Asami T, *et al*: **Identification of a neuropeptide modified with bromine as an endogenous ligand for GPR7.** *J Biol Chem* 2002, **277**:34010-34016.
57. Tanaka H, Yoshida T, Miyamoto N, Motoike T, Kurosu H, Shibata K, Yamanaka A, Williams SC, Richardson JA, Tsujino N, *et al*: **Characterization of a family of endogenous neuropeptide ligands for the G protein-coupled receptors GPR7 and GPR8.** *Proc Natl Acad Sci USA* 2003, **100**:6251-6256.
58. Vaillancourt FH, Yeh E, Vosburg DA, Garneau-Tsodikova S, Walsh CT: **Nature's Inventory of Halogenation Catalysts: Oxidative Strategies Predominate.** *Chem Rev* 2006, **106**:3364-3378.
59. Jimenez EC, Olivera BM, Gray WR, Cruz LJ: **Contryphan is a D-tryptophan-containing *Conus* peptide.** *J Biol Chem* 1996, **281**:28002-28005.
60. Shikata Y, Watanabe T, Teramoto T, Inoue A, Kawakami Y, Nishizawa Y, Katayama K, Kuwada M: **Isolation and characterization of a peptide isomerase from funnel web spider venom.** *J Biol Chem* 1995, **270**:16719-16723.
61. Bulaj G, Buczek O, Goodsell I, Jimenez EC, Kranski J, Nielsen JS, Garrett JE, Olivera BM: **Efficient oxidative folding of conotoxins and the radiation of venomous cone snails.** *Proc Natl Acad Sci USA* 2003, **100(Supp 2)**:14562-14568.
62. Zhi-Qiang W, Yu-Hong H, Xiao-Xia S, Cheng-Wu C, Zhan-Yu G: **Molecular cloning, expression and characterization of protein disulfide isomerase from *Conus marmoreus*.** *FEBS J* 2007, **274**:4778-4787.
63. Chaisson MJ, Brinza D, Pevzner PA: **De novo fragment assembly with short mate-paired reads: Does the read length matter?** *Genome Res* 2009, **19(2)**:336-346.
64. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10(3)**:R25.
65. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25(16)**:2078-2079.
66. Fisher RA, Yates F: **Statistical Tables.** London; 1938.
67. Bulaj G, DeLa Cruz R, Azimi-Zonooz A, West P, Watkins M, Yoshikami D, Olivera BM: **δ-Conotoxin structure/function through a cladistic analysis.** *Biochemistry* 2001, **40**:13201-13208.
68. Puillandre N, Watkins M, Olivera BM: **Evolution of *Conus* Peptide Genes: Duplication and Positive Selection in the A-Superfamily.** *J Mol Evol* 2010, **70**:190-202.

doi:10.1186/1471-2164-12-60
Cite this article as: Hu *et al.*: Characterization of the *Conus bullatus* genome and its venom-duct transcriptome. *BMC Genomics* 2011 **12**:60.

CHAPTER 2

A PROBABILISTIC DISEASE-GENE FINDER FOR PERSONAL GENOMES

The following chapter is a reprint of an article coauthored by Mark Yandell, Chad Huff, myself, Marc Singleton, Barry Moore, Jinchuan Xing, Lynn B. Jorde and Martin G. Reese. This article is originally published in *Genome Research* 2011, 21:1529-1542.

Resource

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).

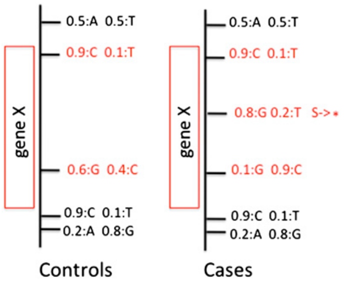


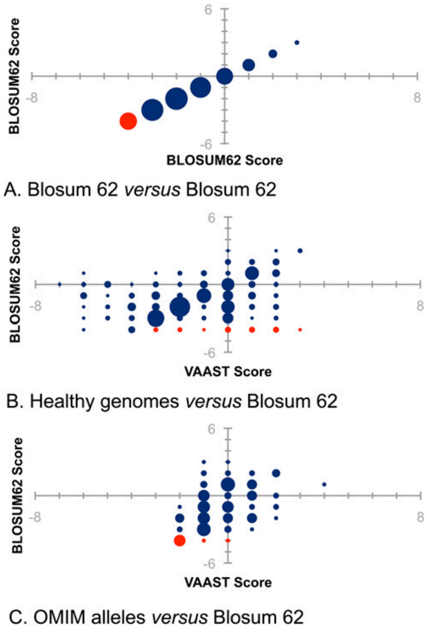
Figure 1. VAAST uses a feature-based approach to prioritization. Variants along with frequency information, e.g., 0.5:A 0.5:T, are grouped into user-defined features (red boxes). These features can be genes, sliding windows, conserved sequence regions, etc. Variants within the bounds of a given feature (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 = 2.4×10^{-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 = 2.4×10^{-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).

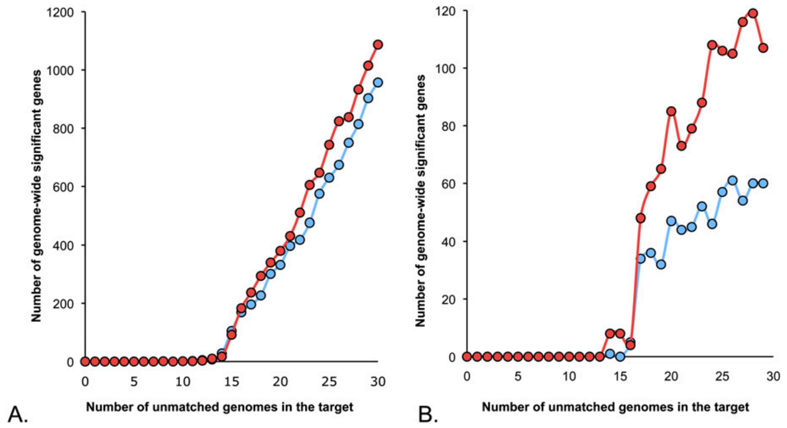


Figure 3. Impact of population stratification and platform bias. Numbers of false positives with and without masking. (*A*) Effect of population stratification. (*B*) Effect of heterogeneous platform and variant calling procedures. (Red line) Number of false positives without masking; (blue line) after 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 × 10^{-6}), compared to the 1087 observed in our population stratification experiments (alpha = 2.4 × 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	7.92×10^{-7}	32	1.98×10^{-6}
MSK	17	14	9.93×10^{-7}	19	5.79×10^{-5}
Mixed ethnicities (189 genomes)					
UMSK	16	9	6.78×10^{-9}	5	2.00×10^{-9}
MSK	9	4	7.60×10^{-9}	5	1.18×10^{-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 $2.4{\times}10^{-6}$. 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.

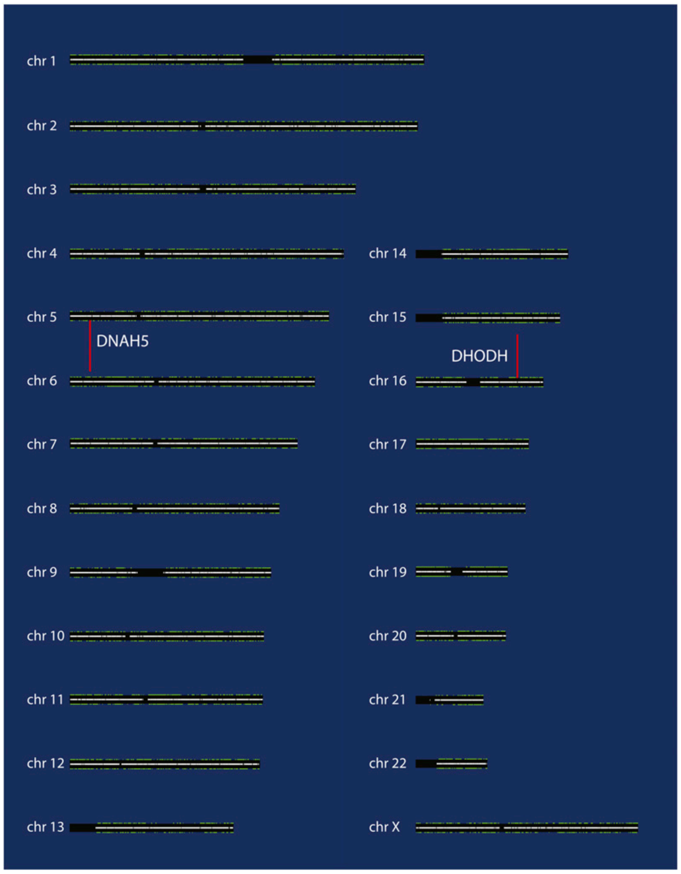


Figure 4. Genome-wide VAAST analysis of Utah Miller Syndrome Quartet. VAAST was run in its quartet mode, using the genomes of the two parents to improve specificity when scoring the two affected siblings. Gray bars along the center of each chromosome show the proportion of unique sequence along the chromosome 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	2.36×10^{-4}	5.26×10^{-3}
2 Compound heterozygotes	4	3	0	2	2.81×10^{-8}	5.51×10^{-5}
3 Compound heterozygotes	2	2	1	1	2.61×10^{-11}	8.61×10^{-7}
4 Compound heterozygotes	1	1	1	1	1.99×10^{-15}	1.78×10^{-8}
5 Compound heterozygotes	1	1	1	1	6.95×10^{-15}	4.60×10^{-10}
6 Compound heterozygotes	1	1	1	1	5.79×10^{-17}	1.42×10^{-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.

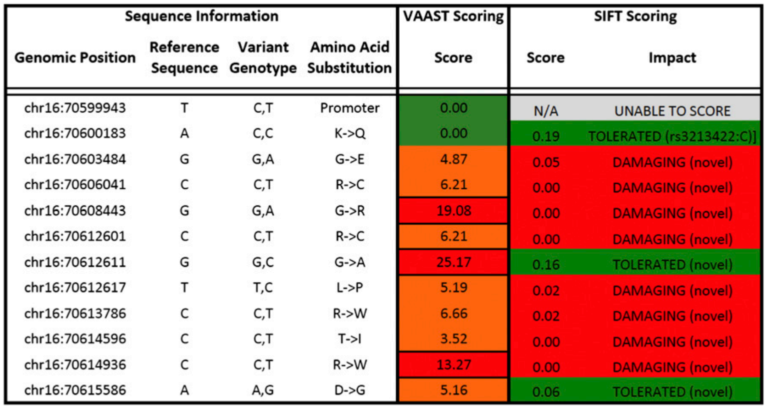
Figure 5A, running all three tools on a case cohort of six and three individuals for the dominant and recessive comparisons, respectively (for details, see Methods). In these analyses, all members of the case cohort contain disease-causing alleles. For ANNOVAR, we set the expected combined disease-allele 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	1.22×10^{-5}	5.26×10^{-3}
Union 2 homozygotes	7	7	0	3	4.74×10^{-10}	5.51×10^{-5}
Intersection 2 homozygotes	3	3	0	1	7.47×10^{-10}	5.51×10^{-5}
Union 3 homozygotes	2	2	2	1	2.83×10^{-13}	8.61×10^{-7}
Intersection 3 homozygotes	1	1	1	1	1.29×10^{-13}	8.61×10^{-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 = 2.4×10^{-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 = 2.4×10^{-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

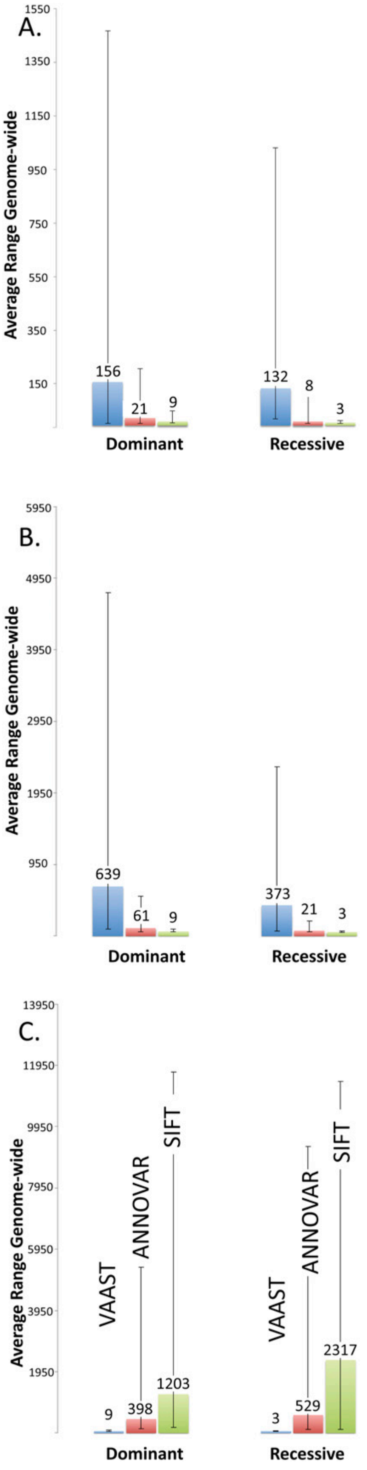
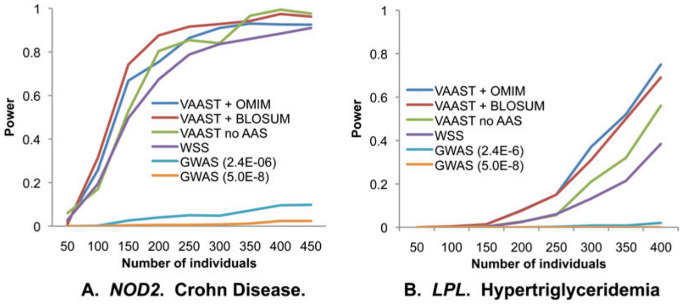


Figure 5. Benchmark analyses using 100 different known disease genes. In each panel the *y*-axis denotes the average rank of the disease gene among 100 searches for 100 different disease genes. Heights of boxes are proportional to the mean rank, with the number above each box denoting the mean rank of 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

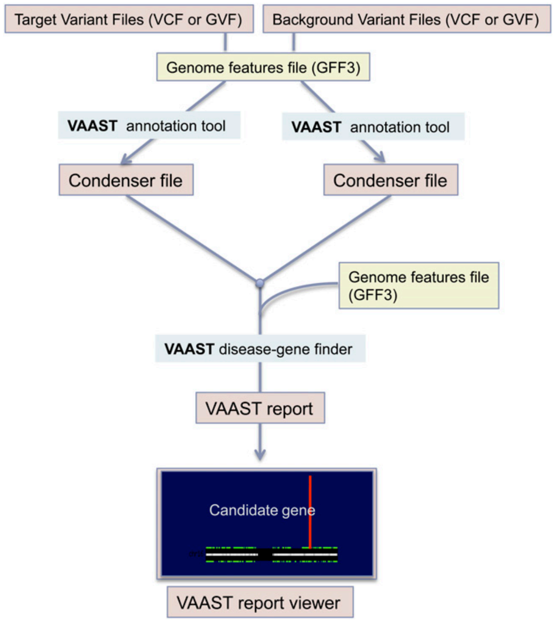

Target Variant Files (VCF or GVF)
Background Variant Files (VCF or GVF)
Genome features file (GFF3)
VAAST annotation tool
VAAST annotation tool
Condenser file
Condenser file
Genome features file (GFF3)
VAAST disease-gene finder
VAAST report
Candidate gene
VAAST report viewer

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:

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

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, -2λ approximately follows a χ^2 distribution with $k + m$ degrees of freedom. We report the non-LD-corrected χ^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 χ^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 χ^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_in_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:

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

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.

CHAPTER 3

VAAST 2.0: IMPROVED VARIANT CLASSIFICATION AND DISEASE-GENE IDENTIFICATION IN PERSONAL GENOMES DATA USING A CONSERVATION-CONTROLLED AMINO ACID SUBSTITUTION MATRIX

The following chapter is a manuscript coauthored by myself, Chad D. Huff, Barry Moore, Martin G. Reese and Mark Yandell. This article is submitted for publication.

Abstract

The need for improved algorithmic support for variant prioritization and disease-gene identification in personal genomes data is widely acknowledged. The Variant Annotation, Analysis, and Search tool (VAAST) employs an aggregative variant association test that combines amino acid substitution (AAS) and allele frequencies. Here we describe and benchmark VAAST 2.0, which uses a novel conservation-controlled amino acid substitution 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*). Moreover, VAAST 2.0 outperforms these other methods 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 also perform remarkably well for rare Mendelian disease gene identification. In addition to CASM, VAAST 2.0 has other new functionalities as well, including native support for additional aggregative association-test methods, support for indels, and a new 'single-case' mode, designed for maximal performance when only a single affected genome is available. VAAST 2.0 thus provides a highly accurate, comprehensive and unified framework for identifying disease-causing variants in personal genomes.

Introduction

Traditionally, genome wide association studies (GWAS) have been used to identify disease-associated variants using sets of 'tagging' SNPs distributed across the genome. GWAS approaches, however, are underpowered to detect the effects of rare causal variants because they are usually in poor linkage disequilibrium (LD) with the tagging SNPs [1]. New sequencing technologies have significantly reduced the price of human genome re-sequencing, 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 [2] and AlignGV-GD [3], 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 [4] and MutationTaster [5], 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 [2].

A major weakness of many variant prioritization tools is that they can only prioritize variants within phylogenetically conserved 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 [4]. 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 (e.g., [6]) . Thus, the growing size of publicly-available human-genome databases (e.g., HapMap [7], the 1000 genomes project [8] and dbSNP [9]) all provide valuable frequency information that could, in principle, be used for variant prioritization. VAAST [10] is a step forward in both regards in that it uses an approach to variant classification that combines both AAS information with variant frequency information, allowing it to score all variants no matter where they lie in the genome and with greater accuracy [10].

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 amino acid substitution matrix (CASM) to overcome this shortcoming. The CASM approach allows VAAST 2.0 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 [11], which allows users to impose a threshold on variant frequencies as observed in dbSNP or in the 1000 Genomes Project [8,9] dataset, 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 this approach renders the tool ineffective for searching datasets containing disease-causing alleles distributed across a range of population frequencies[10]. 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[12], CMC[13], WSS[14], KBAC[15], VT[16], SKAT [17] and VAAST [10]. Although, the different algorithms approach the problem differently, all of these approaches either explicitly or implicitly use the MAF information to weight variants. In addition, a few methods, including VT and VAAST 2.0, can also use functional predictions from 3rd party variant prioritization tools such as PolyPhen and PhastCons [18] to weight variants [16]. 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, but recent work has demonstrated VAAST's applicability to rare-disease gene searches as well [19]. 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 Percent Attributable Risk (PAR) and allelic and locus heterogeneity on their ability to identify genes and alleles responsible both rare and common disease[14].

Here we describe 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 [10] and to other published association tests, including WSS [14], KBAC[15], SKAT [17] and VT[16]. Our results demonstrate the improvements to VAAST made possible by the CASM approach; they also provide a general framework in 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.

Methods

The CASM approach

VAAST uses an extended composite likelihood ratio (CLR) test method to determine a severity score for genomic variants [10]. The null model of the CLR states that the frequency of a variant or variant group are the same between 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 is further updated by the amino acid substitution severity parameter (a_i/ h_i), where h_i is the likelihood that an amino acid substitution 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 [18] 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 variants at any conservation level within a disease and a neutral 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 amino acid substitution *i* (*i* = 1, 2… 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 [18], the likelihood that a variant is disease causing can be estimated by:

$$a_{i1} = (\sum_{j=1}^{n_j} P_{ij}) / n_j, \tag{1}$$

for variants with a PhastCons score of 1, and

$$a_{i0} = (\sum_{j=1}^{n_j} (1 - P_{ij})) / n_j, \tag{2}$$

for variants with a PhastCons score of 0. Similarly using a database of n_k neutral variants, the likelihood that a variant is not disease causing can be estimated by:

$$h_{i1} = (\sum_{j=1}^{n_k} P_{ij}) / n_k, \tag{3}$$

for variants with a PhastCons score of 1, and

$$h_{i0} = (\sum_{j=1}^{n_k} (1 - P_{ij})) / n_k, \tag{4}$$

for variants with a PhastCons score of 0.

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}) terms, namely,

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

where a_{ix}/h_{ix} are the terms in the Conservation-controlled Amino Acid Substitution Matrix, or CASM. This provides an estimate of likelihood ratio of a given amino acid change being disease-causal versus being neutral, controlled for the phylogenetic conservation level in the gene context.

Unless otherwise noted, we calculated the severity parameter using HGMD [20] variants as the disease allele database, and used variants from 1000 genomes project phase I data [8] with MAFs >=0.05 as the neutral allele database. Empirically, we found the PhastCons scores generated from the UCSC vertebrate genome alignment [21] performed best (data not shown). Thus we used these PhastCons scores throughout this paper.

Indel support in VAAST 2.0

VAAST 2.0 also has support for small insertion and deletion (indel) mutations; this is invoked by using the –indel option. The Variant Annotation Tool (VAT) component of the VAAST 2.0 package [10,22] now annotates the functional impact of indels on protein-coding genes in GVF format [22]. 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 amino acid substitutions, additions and deletions. VAAST 2.0 then scores indels with the same CLR test as SNVs, i.e., it calculates the likelihood ratio (LR) of null model versus 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 following. 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 reading frame of protein translation. 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 neutral 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 [10]. This is especially important for indels, since the boundary calling of indel variants is often imprecise. Collapsing variants thus allows VAAST to assess the impact of multiple overlapping indels in the cases.

Comparing VAAST 2.0 to other variant prioritization tools

In order to benchmark VAAST 2.0 as a variant prioritization tool, we used HGMD disease variants [20] and 800 genomes from [8]. We first randomly selected the 400 of these genomes for training (training-set 1). Because VAAST 2.0 uses a control genome set (which we refer to as the background) as an input to improve its accuracy, the remaining 400 genomes not included in training-set 1 were split into two sets comprised of 350 genomes (testing-set 1) and 50 genomes (testing-set 2). We randomly selected 10,000 common SNVs (MAF >=0.05) from training-set 1 as the neutral variant training set (described in the section above), and another 2,000 randomly selected SNVs (common and rare) from testing-set 2 for testing. We chose only common SNVs for training, because MAF is generally negatively correlated with purifying selection strength; thus common SNVs are more likely to be neutral. However, because VAAST 2.0 uses allele frequency to as part of its variant prioritization process, if we included only common variants in testing-set 2 set, the comparison could be biased toward VAAST 2.0; thus we also included rare variants in the testing-set 2. The testing and training sets did not have any variants in common, and we removed any variants present in OMIM or HGMD database from the neutral training set to minimize the chances of including deleterious variants.

Similarly, disease-causal SNVs from the HGMD database were split into two sets with their size ratio being approximately 9:1. The first set (about 44,000 variants) was used for training and the second set (about 5,100 variants) was used for testing. These two sets also do not have any overlap.

We ran VAAST 2.0 over each of the variants in the test set, with "–g 0" and otherwise default parameters to calculate its score. "-g 0" disables the variant grouping functionality so that the score is an accurate measurement of each individual variant. To benchmark of other three algorithms (SIFT, Polyphen-2 and Mutation-taster), we used pre-computed scores downloaded from: http://www.openbioinformatics.org/annovar/annovar_download.html.

For the evaluation of variants in the *BRCA1* and *BRCA2* genes, we used set of 1,433 genetic variants collected by Easton et al[23]. Easton et al. calculated odds ratios for breast cancer causality based upon 1) co-occurrence *in trans* with known deleterious mutations; 2) personal/family history of cancer; and 3) co-segregation of disease in pedigrees. In this study, 133 variants were found to have odds of at least 100:1 in favor of neutrality and another set of 43 have odds of at least 20:1 in favor of causality (Tables 3 and 4 in [23]). We used the 143 missense mutations from these two sets for our benchmark analysis.

Comparing the power of VAAST 2.0 to other aggregative variant association tests

To benchmark VAAST 2.0, we compared it with four other recently published aggregative variant association test algorithms (WSS, VT, KBAC and SKAT). WSS has been shown to have superior power compared to CMC [13], and CAST [12], so we did not include these two tools in our benchmark analyses. We used PolyPhen-2 scores for VT throughput these analyses, since this improved performance [16]. The VAAST 2.0 package provides native support for all of these association tests. Thus VAAST 2.0 users can directly employ WSS, VT, KBAC and SKAT, supplementing them with VAAST 2.0's many other features to improve performance.

Our benchmark used a previously published simulation framework described in [14]. Briefly, we simulated several scenarios, each controlling for 1) genetic model (dominant or recessive); 2) number of causal variants; 3) number of cases and controls and 4) total population attributable risk (PAR) [14] of the causal variants. All parameters used to generate these datasets are described in [14]. For each scenario, we performed 100 simulations and measured the power of each method according to the proportion of trials reaching a significance level of $0.05/21000=2.4\text{x}10^{-6}$ (assuming approximately 21,000 genes in the human genome).

For our investigations of the impact of PAR on each test's performance, we assume that each causal variant has the same individual PAR; hence, each deleterious variant's PAR is the total PAR divided by number of causal variants in the dataset. Importantly this is not true of real datasets we benchmark here, and likely is responsible for some of the performance differences between the simulated and these real datasets.

For each causal variant, its PAR value can be converted to odds ratio (*r*) with the following formula [14]:

$$r = \frac{\alpha}{(1-\alpha)q_u} + 1 \tag{6}$$

where as α is PAR for individual variant and q_U is the genotype frequency in the unaffected population. With this equation, rare variants tend to have higher odds ratios than more common variants at the same PAR. As in [14], we investigated different levels of total PAR and numbers of causal variants.

For each experiment, we also added an equal number of simulated neutral variants to the case datasets, as justified by [14]. The allele frequencies of simulated variants are sampled from the probability density function given by Wright's formula [24] using parameters for mildly deleterious mutations [14]. In control genomes, the genotypes of simulated variants conform to Hardy-Weinberg Equilibrium. In case genomes, the phenotypes of neutral variants have the same probability density distribution as in the control genomes, but the causal variants occur more frequently, according their respective genetic model and risk ratio (calculated from corresponding PAR value; see [14]).

Under the dominant model, both heterozygous and homozygous causal alleles have the same elevated risk level. For recessive cases, we extended the original simulation pipeline in [14], so that our recessive model comprises both simple recessive cases and recessive set cases, i.e., both homozygous and compound heterozygous phenotypes. We thus did not constrain p_M values (the probability that a haplotype contains at least one disease-risk mutation in unaffecteds [14]). Note that the simulation procedure assumes no linkage disequilibrium for simulated variants [14]; our benchmarks on real data assess the impact of this factor on performance.

To simulate the PhastCons scores and amino acid changes, which are inputs to VAAST 2.0, we randomly bootstrapped variants from HGMD database (for causal variants) and from 1000 genome database (for neutral variants), and used their PhastCons scores and amino acid changes for our simulated variants. We removed any variants that were included in the training-sets for VAAST 2.0. The Variable Threshold (VT) method can also use external AAS scores (Polyphen-2 scores) to boost its power [16], accordingly we also bootstrapped the PolyPhen-2 scores from the HGMD and 1000 genome variants that we sampled above and used this information for our benchmarks of the Variable Threshold method.

Benchmark comparisons were preformed using the weighted sum statistics (WSS) and Variable Threshold (VT) methods as implemented in VAAST 2.0 package according to original publications. The performance of VT is also compared to the implementations in plink-seq package [25] and no discrepancies are observed. SKAT and KBAC were benchmarked as implemented by the original authors in the R environment: (http://code.google.com/p/kbac-statistic-implementation/; http://www.hsph.harvard.edu/research/skat/download/).

They were run with using a wrapper script available in the VAAST 2.0 package. For SKAT, “linear.weighted” kernel is used as we simulated no variant-epistasis effects.

Although VAAST 2.0 can employ user-specified genetic inheritance models to increase accuracy, most of the other methods have no such functionalities. Thus in this simulation study we did not provide genetic model information, even though doing so would likely further improve the performance of VAAST 2.0.

We also made sure that our simulation pipeline was behaving correctly. We did so by checking that the distribution of p-values conformed to a uniform distribution supported on [0,1] when the null hypothesis is true [26]. That is, when there is no association between disease phenotype and the genotype. We validated this by setting PAR value to 0 and calculated p-values from VAAST 2.0 in 10,000 simulations, each simulating 1000 cases and 1000 controls and assuming 100 mutation sites exist in the simulated gene. Indeed, the distribution of p-values agrees very well with its theoretical distribution.

Benchmark VAAST 2.0 as a rare Mendelian disease gene finder

For these analyses, we first randomly selected a known disease gene from OMIM, together with its published disease-causing alleles. We then inserted these alleles at their reported positions into different whole genome sequences drawn from the Complete Genomics Diversity Panel [27]. The control (background) genomes dataset consisted of a total of 443 genomes, drawn from multiple sources, consisting of (1) low-coverage exome sequencing data from the 1000 genome project pilot phase [8]; (2) low coverage Danish exome data [28]; (3) 10 genomes sequenced with various platforms

[29]; and (4) Complete genomics diversity panel genomes [27]. This control dataset thus contains a variety of sequencing platforms and ethnicities, and as such presents a realistic snap shot of publically-available genomes. We ran VAAST 2.0 and the other tools and recorded the rank of the disease gene genomewide, repeating the analyses for 100 different known disease genes. This process is described in detail in [10].

The command line used for VAAST1.0 was:

```
VAAST -k -d 2e6 -o <output ID> -m lrt -iht <dominant/recessive> <feature definition file> <control cdr file> <case cdr file>.
```

For VAAST2.0, we used the following command line:

```
VAAST -l <PhastCons score file> -k -d 2e6 -o <output ID> -m lrt -iht <dominant/recessive> <feature definition file> <control cdr file> <case cdr file>.
```

We also used VAAST 2.0's optional single-case mode (sc-mode) to enforce a stringent filtering step in some of our analyses, as this improve its power for disease-gene hunting using one case genome. The VAAST single-case mode assumes complete penetrance of causal variants and no locus heterogeneity. This is achieved by adding "-lh no –pnt c" options to the VAAST2.0 command:

```
VAAST -lh no –pnt c -l <PhastCons score file> -k -d 2e6 -o <output ID> -m lrt -iht <dominant/recessive> <feature definition file> <control cdr file> <case cdr file>.
```

Results

Variant prioritization

We compared the performance of VAAST 2.0 to other variant classifiers. Whereas tools such as SIFT, PolyPhen, and Align-GD [2,3,4] cannot score regions lacking 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 a basis for variant prioritization; in regions where conservation information is available, VAAST 2.0 supplements this information with PhastCons scores [18], 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 VAAST2.0).

To evaluate the prioritization performance of each tool, we plotted the Receiver Operator Curve (ROC) for each algorithm using a set of neutral variants (drawn randomly from 1000 genome pilot phase [8]) and a set of disease-causal variants (from HGMD database). **Figure 3.1A** demonstrates that the accuracy of VAAST 2.0 and 1.0 is considerably better than 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 only 5%. The third best tool is Mutation-taster, 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 **Figure 3.1**) is the 4th best performing approach, followed by PolyPhen-2 and SIFT. We also calculated the Area Under the Curve (AUC) value and the accuracy at FPR = 0.05 for each algorithm, which demonstrates the same trends (**Table 3.1**).

For a second variant prioritization benchmark, we compared the performance 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 [23]. This variant set differs from HGMD/1000 genome variants used to produce **Figure 3.1A** in that the

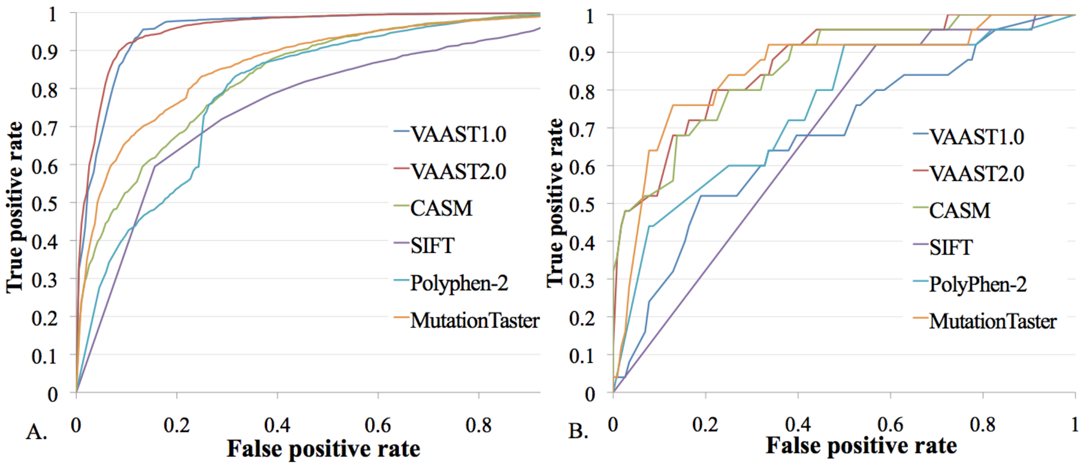


Figure 3.1. Receiver Operator Curves (ROC) for the Variant Prioritization tools. Shown are ROCs for VAAST1.0, VAAST2.0, CASM, SIFT, PolyPhen-2 and MutationTaster, using two benchmark datasets: A) common and rare variants from HGMD and 1000 genomes project; B) BRCA1 and BRCA2 rare variant set. X-axis: false positive rate; y-axis: true positive rate.

Table 3.1. Variant prioritization performance benchmarks. Top half of the table reports Area Under the ROC shown in figure 3.1 (AUC); bottom half the Accuracy of each tool at a false positive rate (FPR) of 0.05. Benchmarks are reported for both HGMD and 1000 genomes data, and for rare *BRCA* 1 and 2 variants.

AUC						
	VAAST1.0	VAAST2.0	CASM	SIFT	PolyPhen-2	Mutation Taster
Data Set1 (HGMD+1kg)	0.95	0.96	0.83	0.76	0.8	0.87
Data Set2 (rare *BRCA* variants)	0.68	0.87	0.86	0.73	0.76	0.85
Accuracy at FPR of 0.05						
	VAAST1.0	VAAST2.0	CASM	SIFT	PolyPhen-2	Mutation Taster
Data Set1 (HGMD+1kg)	0.81	0.86	0.68	0.57	0.62	0.74
Data Set2 (rare *BRCA* variants)	0.53	0.72	0.72	0.52	0.62	0.68

data set used to produce Figure 3.1A contains both common and rare variants for neutral and deleterious alleles, whereas this set only contains very rare variants (MAF<<1%). The results of this benchmark analysis are shown in **Figure 3.1B** and **Table 3.1**. Since majority of the variants in this set are observed only once, VAAST 2.0 cannot use allele frequency information to leverage 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, 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 HGMD/1000 genomes dataset (**Figure 3.1A** and **Table 3.1**) 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. This is illustrated by the *BRCA* variants benchmark dataset in **Figure 3.1B** and **Table 3.1**.

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 [30]; *LPL*, implicated in hypertriglyceridemia [31]; and *CHEK2* a gene involved in breast cancer [32]. 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. Summary statistics for each of the three datasets are presented in **Table 3.2**. We calculated power using a bootstrap approach for varying numbers of cases, with a genomewide significance level of $2.4x10^{-6}$ for *NOD2* and *LPL*. For *CHEK2*, we set the significance level to 0.0005 for *CHEK2* in concordance with the original study [32].

In all three datasets VAAST 2.0 is consistently the most powerful association test (**Figure 3.2**). 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, seems to perform very 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 (*NOD2*). WSS, on the other hand, performs well under both scenarios, and outperforms KBAC, SKAT, and VT when common variants are observed (e.g., the *NOD2* data).

We also benchmarked VAAST 2.0 on the Dallas Heart Study dataset [33], 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 [33]. 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

Table 3.2. Characteristics of the *NOD2*, *LPL* and *CHEK2* datasets. The number of unique multisite genotypes is the number of chromosomes with distinct combinations of variants. The Population Attributable Risk (PAR) is calculated as the sum of PAR values of all susceptibility variants.

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

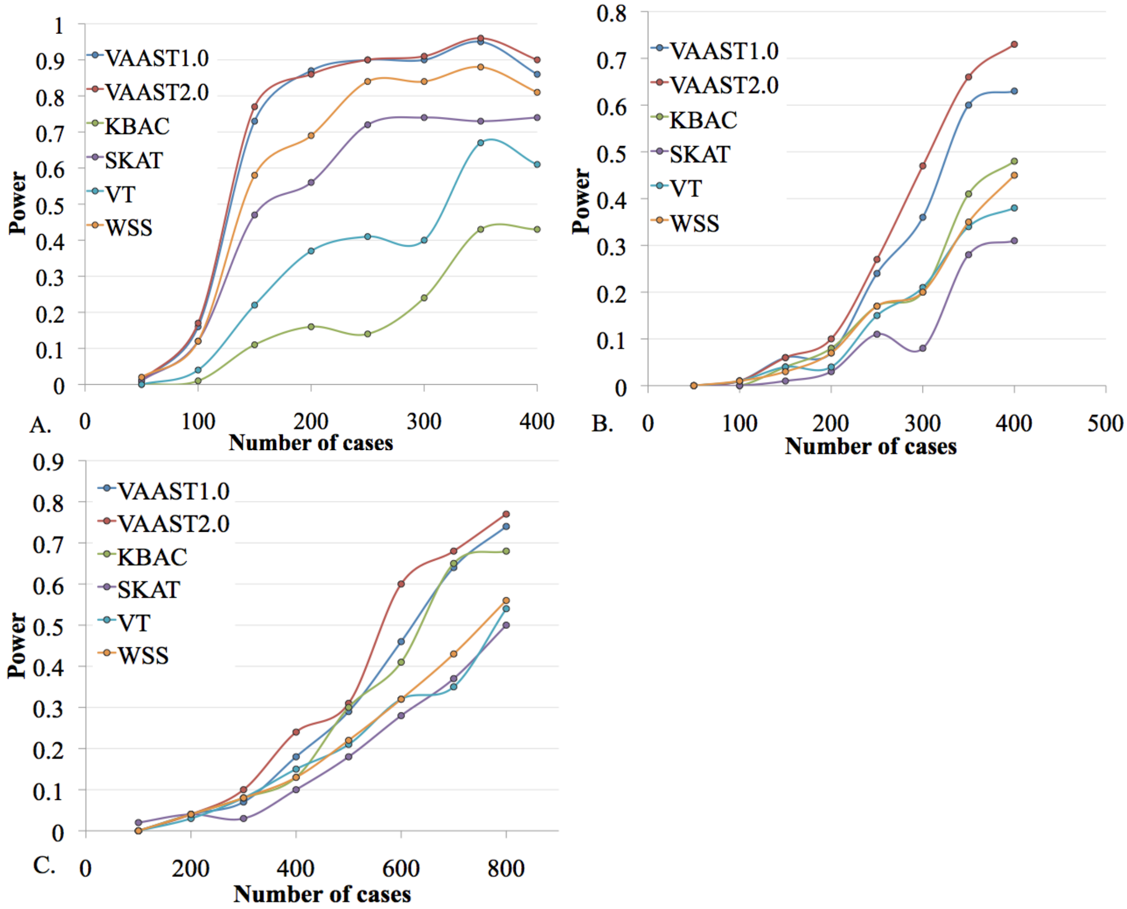


Figure 3.2. Power comparisons over three published common disease datasets. A) NOD2, B) LPL, C) CHEK2. The x-axis shows the number of case genomes and y-axis shows the statistical power. The power is calculated based on 100 bootstraps.

experimental sampling design using bootstraps. The uncorrected significance values for each test are reported in **Table 3.3**. 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 [14] to compare the power of six aggregative variant association tests (see Methods section for additional details).

We first benchmarked the power of these tests under different aggregated Population Attributable Risk (PAR) [14] values, which reflects the aggregated disease risk of all simulated mutations. These results are shown in **Figure 3.3**. 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% to 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 (**Figure 3.3B)**. Note that in this experiment we assumed equal number of causal and noncausal mutation sites, but we also explored other proportions (**Figure 3.4)**.

Both VAAST 2.0 and WSS can use user-specified inheritance models (e.g., dominant or recessive) to boost power. However, for the analyses presented in **Figure**

Table 3.3. Significance of associations between low-triglyceride-levels and rare variants in the *ANGPTL4* gene. Shown are p-values from the dichotomous tests conducted by each method. Note that VT is run with PolyPhen-2 scores.

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

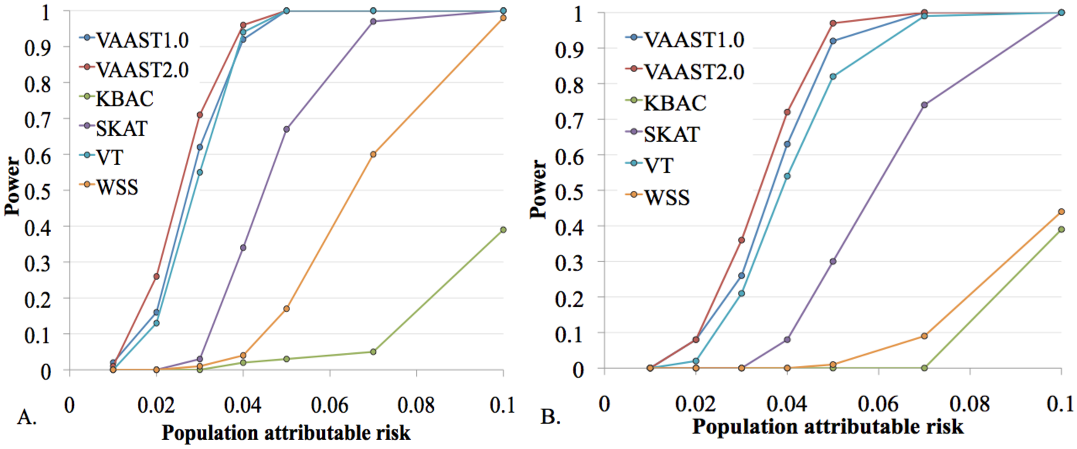


Figure 3.3. Impact of PAR. Shown is the power of six association tests under different total Population Attributable Risk (PAR) levels. x-axis shows the total PAR values from all contributing variants; y-axis shows the statistical power based on 100 bootstraps. A) Dominant model, B) recessive model. The number of cases and control are set at 1000, with the number of disease-causal alleles and noncausal alleles both fixed at 50.

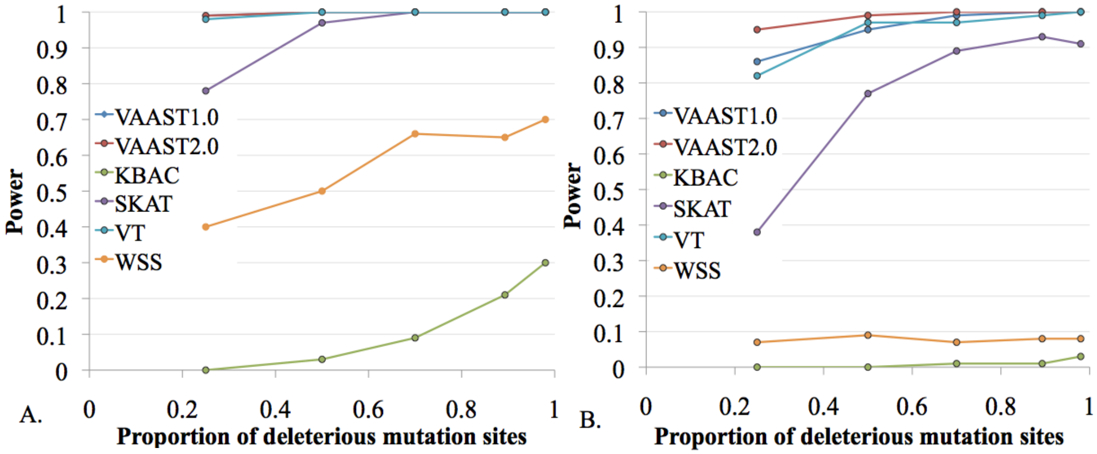


Figure 3.4. Impact of different proportions of deleterious mutation sites contributing to the disease risk. x-axis is the proportion of deleterious mutation sites among all simulated sites; y-axis statistical power. A) Dominant model; B) recessive model. Total PAR is fixed at 10%; the numbers of case/controls are set at 500; the number of casual variants is 50 with varying number of noncasual variants.

3.3, 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 [14] 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.

We then 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 approaches. These results are shown in **Figure 3.5**. 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 other methods. We summarize the number of cases/controls required for each algorithm to achieve 80% power in **Table 3.4** 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 [15]. 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 causal alleles or common causal alleles are present, the number of multisite genotypes grows very rapidly, with a concomitant loss in power. This behavior can be seen quite clearly in **Figure 3.5**.

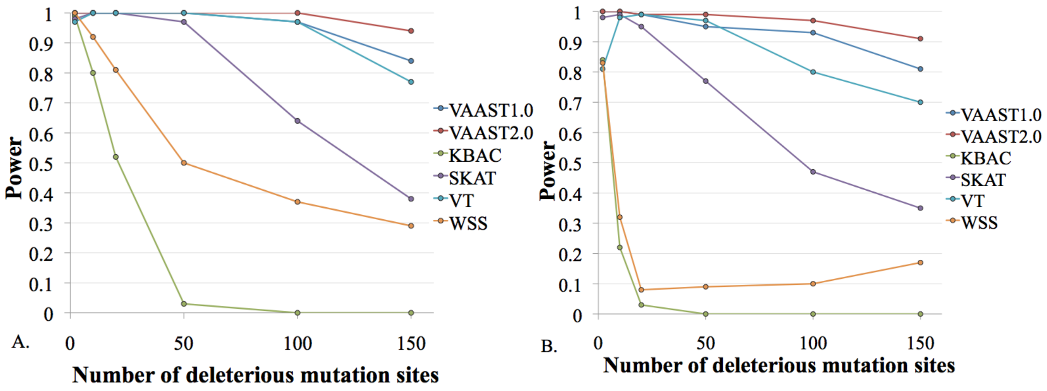


Figure 3.5. Impact of differing numbers of deleterious mutation sites. X-axis is the number of deleterious mutation sites (ND); y-axis shows the statistical power based on 100 bootstraps. A) Dominant model, B) recessive model. The number of cases and control are set at 500, and the total PAR value is set at 10%.

Table 3.4. Numbers of cases and controls required for 80% power. Total PAR is set at 10%.

	Dominant		Recessive	
	ND=5	ND=50	ND=5	ND=50
VAAST1.0	150	300	300	500
VAAST2.0	150	300	300	400
KBAC	300	>1000	800	>1000
SKAT	200	400	300	600
VT	200	300	400	500
WSS	300	700	800	>1000

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 3.2**). We tested this hypothesis by comparing the power of WSS and KBAC under different numbers of deleterious alleles. 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 loses 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 3.2** and **3.5**.

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 [10,19]. 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 [10]. 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 with Complete Genomics platform [27] in order to simulate diseased individuals. 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 3.6**. Figure 3.6A and 3.6B reports the proportion of the 100 OMIM 'target' genes falling into 4 bins based upon rank; these are bin A: 1 to 10, bin B: 11 to 100, bin C: 101 to 1000, and bin D: greater than 1000 among all protein coding genes.

For the dominant disease scenario, with only one case genome, VAAST 2.0 ranked 40% of disease-genes among the top 100 candidates genomewide. Performance improved dramatically as the number of case genomes increases. With only two case genomes, the mean ranking for the disease-gene is 55, and 67% of disease-genes are ranked within top 10, genomewide; with three case genomes, the mean ranking is 10 and 92% 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 83% of the disease-genes are ranked among top 100, and with two cases, the mean ranking was 9, with 95% of the disease genes ranked among top 10. We note that in this benchmark analysis, the performance of VAAST 2.0 is only slightly better than 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 interesting aspects of this analysis is the general finding that all of the association tests do relatively well on these datasets. For example, using top 10

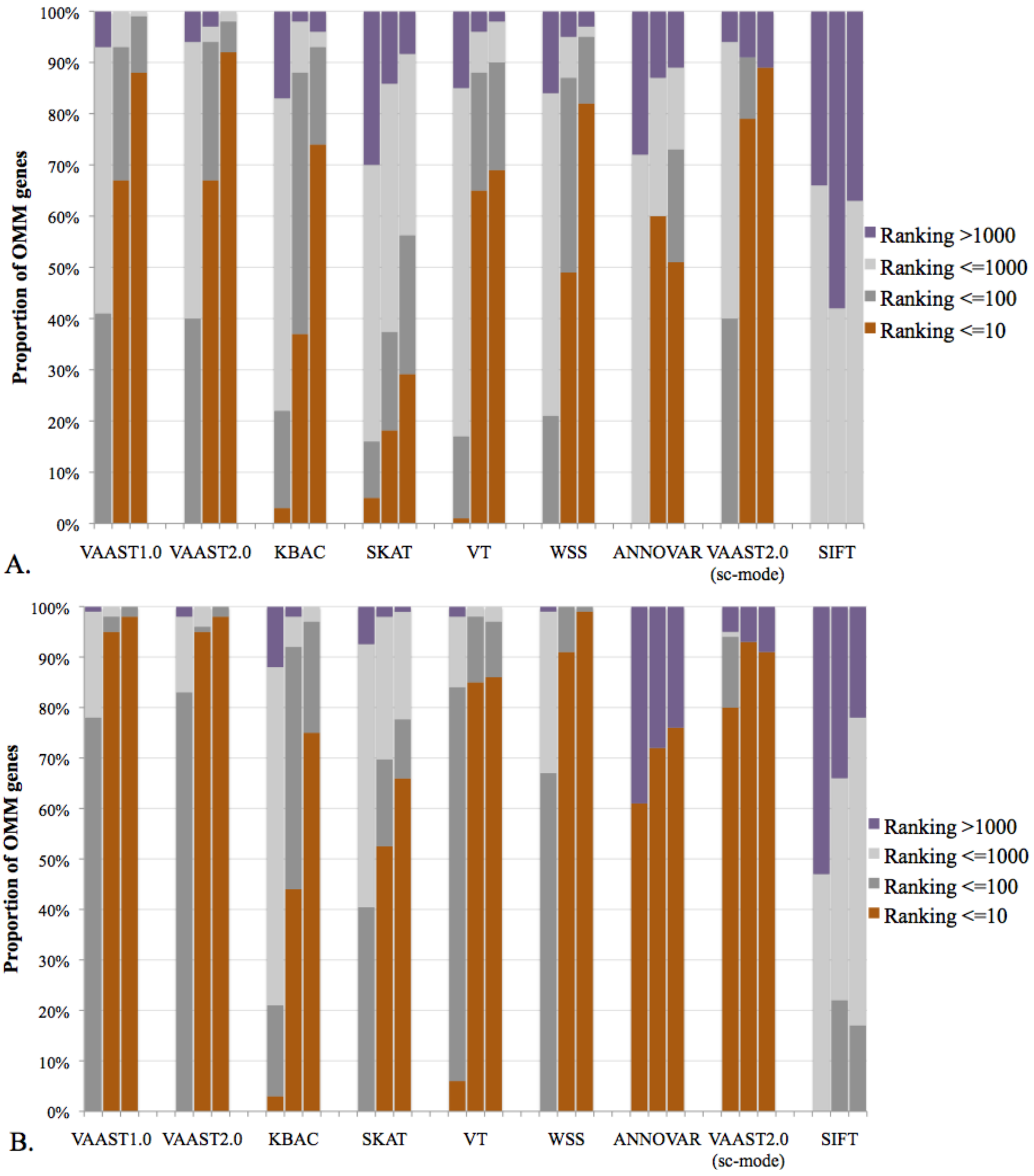


Figure 3.6. Rankings for 100 different genomewide searches for known rare disease-genes. Panel A) and B) show the rankings under dominant and recessive models, respectively. The different colors denote the proportion of the 100 OMIM 'target' genes falling into four bins based upon genomewide rank (see insert legend), with Orange, denoting the percentage cases for which the disease-gene was ranked among the top ten candidates genomewide. 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 to sample sizes of 1, 2 and 3. Panel C) and D) show the mean rankings for the same datasets, under dominant and recessive models respectively. Error bars represent the standard deviation of the rankings.

ranking as an empirical significance level with a sample size of 3, under the dominant model, VAAST 2.0 achieves 92% power, WSS 82% and KBAC 74% power. Under recessive model, VAAST 2.0, achieves 98% power, WSS 99% and KBAC 75% 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 [2,11]. As would be expected, SIFT, does poorly compared to the other tests. Notably, ANNOVAR, a filtering based approach, does very well with a sample size of 1 under recessive model, with only VAAST 2.0 run in the 'single-case' mode outperforming it (see Methods section for the commands). These results also illustrate another important difference between aggregative tests and filter-based approaches: While ANNOVAR's performance is generally very good as regards the proportion of OMIM genes ranked as top 10, when it fails, it usually fails completely. For examplein cases where ANNOVAR fails to rank the gene in the top ten, the disease gene in never found among the top 1000 genes (data not shown).

Discussion

Phylogenetic conservation is a valuable source of information for distinguishing between benign and disease-causing variation. However, how best to make use of this information—for purposes of variant prioritization and for association testing—is still an open question. Variant prioritization tools, such as SIFT[2], 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 [4]. VAAST takes a different approach. Rather than looking at individual columns of multiple alignments in order to judge the impact of a coding variant, VAAST uses the global, genomewide frequency of observing an amino acid substitution (AAS) in any gene, anywhere in the genome. This means that VAAST can score every coding change, regardless of whether or not a particular gene, or that particular region of its protein is conserved. Although it 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 develop a method that can make use of the detailed information provided by multiple alignments, and at the same time still score every coding variant. As **Figure 3.1** 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 genomewide search tool. As such, VAAST is one of several aggregative variant association tests published in the last few years [12,13,14,15,16,17]. Although several benchmarks have been published [14,15,16,17,34], ours is the first to systematically compare of the power 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 3.2** 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-causal alleles; (2) their frequencies; and (3) their collective attributable risk (PAR).

The performance curves of 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 (**Figure 3.5**). This is likely explained by the increasing number of multisite genotypes associated with number of causal sites. Since KBAC estimates the sample risk for each unique multisite genotype, when the space of multisite genotypes is large and each genotype has relatively low risk, the power of KBAC is compromised. This is consistent with its poor performance on the *NOD2* dataset, compared to its much better power on the *CHEK2* and *LPL* datasets, as 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 3.2**), and all of their deleterious variants are rare.

Although SKAT performed well in our simulation studies, it did much less well on the three real datasets. Its performance the *LPL* and *CHEK2* datasets, for example, suggest that SKAT is not well suited for analyses of datasets having modest numbers of causal 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 dramatically improved SKAT's statistical power, from 31% to 45% at maximal sample size. Moreover, SKAT is a supervised method, requiring users to choose kernels and weights, as the default parameters can be suboptimal in certain cases. This also presents challenges. For example, on the *NOD2* dataset, the 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 both low and high ratio of disease-causing and neutral alleles in the case dataset under both dominant and recessive modes of inheritance (Figures 3.3, 3.4 and 3.5). These strengths likely result from two features shared by the two approaches. 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 the power [16].

Despite their similar performance characteristics on simulated data, VAAST and VT behave very differently from one another on real datasets. One possible explanation 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. In contrast, VT assumes that for any given disease dataset, a single optimal MAF threshold exists, and less frequent variants are more likely to be deleterious. It thus explores all possible thresholds to find the MAF that maximizes the contrast between cases and controls [16]. This assumption is generally true for our simulated datasets, and is probably true in expectation for most disease-causing loci. However, because that genetic drift is a stochastic process, the distribution of disease-causing variation at any given locus can deviate from its theoretical expectation. In addition, our theoretical understanding of the expected distribution of disease-causing variation is also far from complete, given the complexities of demographic history, natural selection, and a complex, changing environment. 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.

With the exception of VAAST, 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 genomewide 99%, 98%, and 75% of the time, respectively, using only three case genomes under a recessive model. ANNOVAR—a filtering approach—performed very well, relative to other methods, when only one case genome was used. However, this performance advantage fell off quickly as additional case genomes are added, demonstrating that filtering-based approaches scale poorly with increasing sample size, whereas the opposite is true for the association tests. VAAST 2.0 has a unique flexibility in this regard. It can be run in ‘single-case’ mode when only a single affected individual is available. When run in this mode it not only outperformed the other association-test methods, but ANNOVAR as well.

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 scenarios from both simulated and observed case/control datasets.

References

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

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

3. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, et al. (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43: 295-305.

4. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.

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

6. 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: 727-739.

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

8. Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

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

10. Yandell M, Huff C, Hu H, Singleton M, Moore B, et al. (2011) A probabilistic disease-gene finder for personal genomes. Genome Res 21: 1529-1542.

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

12. 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.

13. 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.

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

15. 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.

16. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832-838.

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

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

19. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, et al. (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: 28-43.

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

21. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32: D493-496.

22. Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, et al. (2010) A standard variation file format for human genome sequences. Genome Biol 11: R88.

23. Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, et al. (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: 873-883.

24. Wright S (1990) Evolution in Mendelian populations. 1931. Bull Math Biol 52: 241-295; discussion 201-247.

25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
26. Murdoch DJ, Tsai Y-L, Adcock J (2008) P-values are random variables. The American Statistician 62: 242-245.

27. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327: 78-81.

28. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, et al. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet 42: 969-972.

29. Moore B, Hu H, Singleton M, De La Vega FM, Reese MG, et al. (2011) Global analysis of disease-related DNA sequence variation in 10 healthy individuals: implications for whole genome-based clinical diagnostics. Genet Med 13: 210-217.

30. Lesage S, Zouali H, Cezard JP, Colombel JF, Belaiche J, 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.

31. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, et al. (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 42: 684-687.

32. Le Calvez-Kelm F, Lesueur F, Damiola F, Vallee M, Voegele C, et al. (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: R6.

33. Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, et al. (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119: 70-79.

34. 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: e1002496.

CHAPTER 4

CONCLUSIONS AND PERSPECTIVES

Next-generation sequencing techniques are revolutionizing genetic studies

The next-generation sequencing (NGS) techniques have altered genetics studies in a few important ways. First of all, it brought in a series of techniques that have become essential in current biological studies, such as RNA-seq[1], Chip-Seq[2], bisulfite pyrosequencing[3], Hi-C[4], etc. Compared to previous approaches, they are usually far more cost-effective, and achieve a higher accuracy and a finer scale of resolution. Take RNA-seq for example; it enabled simultaneous expression profiling of whole transcriptome and the characterization of novel transcripts. Compared to gene expression microarrays, it measures the transcript levels more accurately, and in addition allows the profiling of alternative promoter usages, uncharacterized transcripts and various splice forms[5]; when compared to previous EST sequencing approaches, it generates a much higher volume of sequences, making it possible to detect transcripts with modest or low expression levels. Because of these obvious advantages, RNA-seq has gradually become the prevailing expression profiling method. In Chapter 1, I presented a series of applications of RNA-seq in the genomic study of *Conus bullatus*.

The high-throughput nature of NGS has also created a vast amount of personal genomics data, which are likely to dramatically improve our understanding of evolutionary process. Most population genetics studies prior to the era of NGS either contain only a handful of sequenced loci, or use fully sequenced genomes for very limited amount of individuals. Since each individual typically possesses only a small proportion of the gene pool of the species, our understanding of the evolution gleaned from small-scale studies could be limited. It is not until the advent of NGS techniques that researchers are able to sequence a population of individuals and examine the signatures of selections at whole-genome scale. This phenomenon has at least two important implications. First, it allows us to gain a more quantitative understanding of molecular events that shape our genome landscape, such as meiosis recombinations and *de novo* mutations (e.g., see [6]). The determination of these basic statistics in molecular genetics will in turn allows us to create better models for human evolution that accounts for individual and population variations. Second, the steadily increasing sample size of human genomes will eventually provide sufficient power for exploring the genetic bases of currently undercharacterized complex traits, including many common genetic diseases. Especially, NGS is capable of genotyping rare and private genomic variants, which are likely to account for a significant portion of disease risk but are usually unseen by DNA microarray approaches[7].

The massive amount of data generated by NGS presents challenges to the field of bioinformatics, and necessitates the development of computational tools for each application. My main interest in bioinformatics lies in two aspects: 1) developing methods for studying the genomics of nonmodel organisms, using NGS data; and 2) creating a unified statistical test for the identification of causal genes for both rare Mendelian diseases and complex diseases, using personal genomics data. These efforts are presented in the previous chapters; here I summarize the results and provide perspectives for each section, as follows.

Application of next-generation sequencing techniques for cone snail studies

Interest in cone snail venoms has grown steadily over the last a few decades [8,9,10,11,12], largely as a result of their pharmaceutical importance due to their specificity for particular ion channels, receptors and transporters [13]. Traditional approaches to characterize conopeptides include protein mass spectrometry and Sanger sequencing of the cDNA libraries [14]; however, these techniques suffer from relatively low throughputs and are suboptimal for identification of less abundantly expressed toxins. Chapter 1 of my thesis presented the first application of next-generation sequencing technique to cone snail research [15]. There I presented several methodological advances. First, using only very low-coverage (3x) genomic reads, I developed techniques that allowed me to characterize several key features of the *Conus bullatus* genome, including its genomewide repeat content, polymorphism rate and genome size. I also used RNA-seq data to estimate the relative abundance for the major conopeptide superfamilies. Because these estimates are based on high-throughput sequencing data, our confidence is much higher compared to previous estimates. Second, using a novel *in silico* pipeline that I developed, I identified 2,410 putative conopeptides, 30 of which are represented by complete or near-complete sequences. Collectively, these results demonstrate the power of next-generation sequencing techniques for investigation of nonmodel organisms for which there is little prior knowledge of its genome contents. We are currently performing whole-genome sequencing of *Conus bullatus* and will begin annotating this genome once the sequencing and assembly are finished. The availability of the reference genomic sequence for a cone snail species in the very near future will further advance our understanding of conopeptide biology.

Using VAAST to identify disease genes

The VAAST algorithm is a search tool for genomewide disease-gene finding. It is not merely a test that calculates that significance of a disease-to-gene association, but an *ab initio* search tool that can also incorporate diverse information (e.g., inheritance model, penetrance, prevalence, locus heterogeneity etc.) to maximize its accuracy. VAAST is also aware of problems in experimental design such as population stratification and sequencing-platform biases. Collectively, these features of VAAST allow it to achieve the highest accuracy of any currently available association test, even in cases where data-set sizes limit statistical power. Since its initial publication, close to 200 groups worldwide have licensed VAAST, and VAAST has already been used to identify a new genetic disease [16].

Prior to the creation of VAAST, most researchers have relied on filtering-based approaches (e.g., ANNOVAR [17]) to identify genes causal for Mendelian disorders. As my work has demonstrated, there are two inherent disadvantages in such an approach. First, the false negative rate (FNR) is not strictly controlled; this is a real problem because with increasing numbers of filtering steps, FNR will increase exponentially. Second, despite the constantly increasing size of public genome databases, the accuracies of filtering methods do not necessarily improve because of the FNR problem. Third, filtering methods cannot be used to identify genes and alleles involved with common diseases. In contrast, VAAST offers a statistically robust approach that is applicable to both common and rare genetic diseases, one that can leverage publicly available database such 1000 genome project [18] for disease gene finding.

Using VAAST to identify the genetic basis for Mendelian traits

In addition to rare Mendelian diseases, VAAST can also be used to identify genes and Quantitative Trait Loci (QTL) associated Mendelian traits in the same manner as Genome Wide Association Studies (GWAS), e.g., [19]. This application is very similar to Mendelian disease gene finding in the sense that in both applications the goal is to identify significantly different distributions in cases vs. controls of variants in pre-defined genomic features (e.g., nucleotide-sites, genes, coding regions, etc.). Compared to GWAS and other association tests, my work has demonstrated that VAAST has several inherent advantages. First, we have shown that VAAST has better power under a variety of genetic parameter spaces. Second, VAAST has a wider scope of applicability, having been designed as a generic disease-gene finder suitable for scenarios in which both rare and common variants contribute to disease. Third, VAAST is able to consider external information (e.g., disease prevalence, inheritance mode, shared genomic segment regions, etc.) to improve its accuracy; such information is readily available for many genetic studies, but VAAST is the first tool capable of using it in a non *ad hoc* fashion.

Future challenges for disease-gene finding in human

One of the most intriguing aspects of personal genomics study is that, the ever-increasing size of publically available genomes could directly improve the chance of detecting disease-causal mutations, since the power of association tests increases with the sample sizes of both controls and cases. The possibility of using public genomes as controls will also likely simplify future study designs and reduce the overall experimental cost. This methodology has been adopted for discovering the genetics basis for rare Mendelian diseases (e.g., [20]); theoretically it also applies to common genetic conditions, when the size of public genome database is large enough to provide sufficient matched controls. In Chapter 2 and 3, we have demonstrated the potential of using public genomes for performing association tests for both rare and common genetic diseases.

However, in order to incorporate public genomes into disease-gene finding, a strategy to control for population stratification becomes a necessity. The concern is that many genetic diseases have different prevalence across populations; by randomly collecting samples or using all samples from public genomes, we will likely get imbalanced ratios of ethnicities between cases and controls, which could lead to the identification of variants that have high F_{st} values [21] between populations rather than real disease-causal variants. This has indeed been a common concern about case-control study designs in 1990s before better strategies for controlling population stratification comes into play [22]. While researchers can choose to stratify the genomes and uses only ethnically matched cases and controls, this could reduce the power of the study. In GWAS, several methods have been proposed to reduce the false-positive rate due to population stratification. For example, Yu et al. [23] use a mixed-model approach to control for population structures; in contrast, Coop et al. [24] use a Bayesian framework to incorporate empirical MAF differences between populations into association tests. Most current aggregative association tests, however, are developed for matched case-control studies and few have implemented methods to alleviate the population stratification issue. In near future, being able to effectively control for ethnicity will likely become an essential feature for disease-gene finding algorithms. This is a relatively easy task for VAAST; because it calculates the likelihood ratio first separately for each variant, it is straightforward to apply established GWAS false-positive-control methods to revise the individual-variant scores before they are combined. For example, one option is to incorporate the population identity as a Bayes factor in the likelihood calculation. Alternatively, we can directly model the genetic drift of variant allele frequencies between populations with Brownian motion in a similar fashion as in [25]; the likelihood ratio will then be calculated based on a robust model that has accounted for population stratification.

An equally serious challenge in disease-gene discovery in the future is controlling for platform differences. Currently, the most widely used sequencing platforms include Illumina Hi-Seq, Roche 454, SOLiD and Complete Genomics; for each platform there against typically exists several computational pipelines for variant calling. In our experience, different sequencing and computational pipelines often produce a substantial set of discrepant variant calls for the same genome (unpublished data), which may cause a serious false-positive issue for association tests when cases and controls are called differently. This becomes a concern for researchers using public genomes as controls, which are usually genotyped with heterogeneous pipelines. See Figure 2.3 for an example. While the constantly improving quality of sequencing techniques and variant-calling algorithms may ultimately solve this issue, in near future controlling for platform difference at run time will still be a desirable feature for disease-gene finders. However, none of the previous aggregative association tests has such functionality. On the other hand, we have observed that these discrepant variant calls occur repeatedly at certain genomic locations and also tend to cluster at given genomic regions, which justifies empirical approaches such as masking error-prone variant sites (adopted in both VAAST1.0 and 1000 genome project), or adding a post-association-test step to adjust the significance levels of genes according to observed local false-positive rate.

With more and more high-risk disease variants being identified with the NGS data, the need for methodologies that accurately find modest-risk variants keeps growing. This is partially addressed by the CASM algorithm presented in Chapter 3; however, human geneticists are still struggling to answer a few fundamental questions in modest-risk variant classifications. First, to what extent do intergenic regions participate in the pathogenesis of complex diseases, and what is the best way to classify the noncoding variants? This problem is much harder than the classification of missense mutations because 1) the Dirichlet distributions [26] of noncoding DNA are less well-defined than amino acids, and 2) in sharp contrast to protein-coding genes, the functional elements in intergenic regions are a mosaic of evolutionary units, such as miRNAs, lincRNAs and enhancers, most of which are still poorly annotated and understood. This situation will likely improve as the ENCODE project [27] proceeds, but a statistical strategy that explicitly models regulatory variants within association tests is still very desirable. Second, are all variants equally pathogenic, if given that they cause the same type of amino acid substitution and have the same conservation level? And to what extent are the gene/pathway contexts and the variants' locations within the CDS relevant to the pathogenesis? These questions have immediate implications in the design of association tests, as the answers will determine how variants are clustered and handled in order to achieve the maximal power.

Besides case-control study designs, another popular strategy for human disease gene finding is linkage analysis using pedigree genotyping data. This method measures the strength of association between disease status and genotypes in pedigrees, and usually involves the calculation of a LOD score that compares the likelihood of "linkage" model and "no-linkage" model [28]. It is especially helpful for genetic diseases with very high locus-heterogeneity, where case-control study design may have low power even with relatively large sample size. The next-generation sequencing techniques present both challenges and opportunities to linkage analysis methods. First, with the possibility of directly genotyping causal variants, a statistically robust approach of incorporating functional predictions of variants into linkage score will potentially improve the performance. Second, ideally, new linkage algorithms should be able to leverage their power from public genomes, as an external control set. Third, the ability to combine case-control association tests and linkage studies into one statistical framework will also be important, since for many common genetic diseases we will likely have data from both designs. We are currently developing an extension to the VAAST framework that should possess these features.

References

1. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, et al. (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45: 81-94.

2. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.

3. Colella S, Shen L, Baggerly KA, Issa JP, Krahe R (2003) Sensitive and quantitative universal Pyrosequencing methylation analysis of CpG sites. Biotechniques 35: 146-150.

4. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289-293.

5. Liu S, Lin L, Jiang P, Wang D, Xing Y (2011) A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. Nucleic Acids Res 39: 578-588.

6. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-639.

7. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18-21.

8. Olivera BM (1997) Conus venom peptides, receptor and ion channel targets and drug design: 50 million years of neuropharmacology (E.E. Just Lecture, 1996). Mol Biol Cell 8: 2101-2109.

9. Jimenez EC, Olivera BM, Gray WR, Cruz LJ (1996) Contryphan is a D-tryptophan-containing Conus peptide. J Biol Chem 281: 28002-28005.

10. Endean R, Duchemin C (1967) The venom apparatus of Conus magus. Toxicon 4: 275-284.

11. Conticello SG, Gilad Y, Avidan N, Ben-Asher E, Levy Z, et al. (2001) Mechanisms for evolving hypervariability: the case of conopeptides. Mol Biol Evol 18: 120-131.

12. Bandyopadhyay PK, Colledge CJ, Walker CS, Zhou LM, Hillyard DR, et al. (1998) Conantokin-G precursor and its role in gamma-carboxylation by a vitamin K-dependent carboxylase from a Conus snail. J Biol Chem 273: 5447-5450.

13. Terlau H, Olivera BM (2004) Conus venoms: a rich source of novel ion channel-targeted peptides. Physiological Reviews 84: 41-68.

14. Pi C, Liu Y, Peng C, Jiang X, Liu J, et al. (2005) Analysis of expressed sequence tags from the venom ducts of Conus striatus: focusing on the expression profile of conotoxins. Biochimie 88: 131-140.

15. Hu H, Bandyopadhyay PK, Olivera BM, Yandell M (2011) Characterization of the Conus bullatus genome and its venom-duct transcriptome. BMC Genomics.

16. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, et al. (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: 28-43.

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

18. Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

19. Huang X, Wei X, Sang T, Zhao Q, Feng Q, et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42: 961-967.

20. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30-35.

21. Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet 10: 639-650.

22. Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65: 220-228.

23. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208.

24. Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. Genetics 185: 1411-1423.

25. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20: 393-402.

26. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, et al. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12: 327-345.

27. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, et al. (2011) ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res 39: D871-875.

28. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7: 277-318.