nature biotechnology

A unified test of linkage analysis and rare-variant association for analysis of pedigree sequence data

Hao Hu[1], Jared C Roach[2], Hilary Coon[3], Stephen L Guthery[4], Karl V Voelkerding[5,6], Rebecca L Margraf[6], Jacob D Durtschi[6], Sean V Tavtigian[7], Shankaracharya[1], Wilfred Wu[8], Paul Scheet[1], Shuoguo Wang[9], Jinchuan Xing[9], Gustavo Glusman[2], Robert Hubley[2], Hong Li[2], Vidu Garg[10,11], Barry Moore[8], Leroy Hood[2], David J Galas[12,13], Deepak Srivastava[14], Martin G Reese[15], Lynn B Jorde[8], Mark Yandell[8] & Chad D Huff[1]

High-throughput sequencing of related individuals has become an important tool for studying human disease. However, owing to technical complexity and lack of available tools, most pedigree-based sequencing studies rely on an ad hoc combination of suboptimal analyses. Here we present pedigree-VAAST (pVAAST), a disease-gene identification tool designed for high-throughput sequence data in pedigrees. pVAAST uses a sequence-based model to perform variant and gene-based linkage analysis. Linkage information is then combined with functional prediction and rare variant case-control association information in a unified statistical framework. pVAAST outperformed linkage and rare-variant association tests in simulations and identified disease-causing genes from whole-genome sequence data in three human pedigrees with dominant, recessive and *de novo* inheritance patterns. The approach is robust to incomplete penetrance and locus heterogeneity and is applicable to a wide variety of genetic traits. pVAAST maintains high power across studies of monogenic, high-penetrance phenotypes in a single pedigree to highly polygenic, common phenotypes involving hundreds of pedigrees.

Linkage analysis evaluates recombination events between genetic markers and potential causal alleles in families to map phenotypic loci[1]. In comparison, genetic association tests detect genetic markers that are correlated with phenotypes among unrelated individuals. Traditionally, both types of analyses use genetic markers such as microsatellites or single nucleotide polymorphisms (SNPs). Thus, the corresponding statistical methods usually test against the null hypothesis that the focal variants are in linkage or linkage disequilibrium with causal variants and do not assume that causal variants are directly observable. High-throughput sequencing techniques now allow comprehensive detection of rare and private variants throughout the exome or whole genome. To take advantage of the increased availability of sequencing data, rare-variant association tests (RVATs) have been developed to aggregate rare variants in each gene, which reduces multiple comparison problems and increases the statistical power for discovering disease-associated genes[2–4]. Once disease loci have been identified through association or linkage studies, variant classifiers such as SIFT[5] and PolyPhen-2 (ref. 6) are often used to prioritize rare mutations that are likely to be damaging.

Association tests and linkage analysis use two different types of information to perform disease locus mapping. Both methods take advantage of genetic recombination information; however, association signals derive mostly from the historical recombination events in the population, whereas linkage analysis makes use only of recombination events that occurred in the pedigree under investigation. In a biological sense, these two types of data are related; yet, from a statistical point of view, they provide orthogonal and thus complementary information about the disease locus. Currently, comprehensive analysis of pedigree sequencing data is a labor-intensive process that requires an array of bioinformatics tools (linkage analysis, association tests and variant classifiers). Given these challenges, most pedigree sequencing studies apply a simplified and suboptimal approach involving a series of ad hoc filtering criteria[7]. A few existing tests use family data in rare-variant association tests (for example, refs. 8 and 9). By accounting for pedigree relationships using an appropriate covariance matrix, these tests use information from related pedigree members without inflating type I error with large sample sizes. However, these methods capture only association signals and do not incorporate linkage or variant-classification information.

[1]Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [2]Institute for Systems Biology, Seattle, Washington, USA. [3]Department of Psychiatry, University of Utah, Salt Lake City, Utah, USA. [4]Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA. [5]Department of Pathology, University of Utah School of Medicine, Salt Lake City, Utah, USA. [6]ARUP Institute for Clinical and Experimental Pathology, ARUP Laboratories, Salt Lake City, Utah, USA. [7]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah, USA. [8]Department of Human Genetics and USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, Utah, USA. [9]Department of Genetics, Rutgers, the State University of New Jersey, Piscataway, New Jersey, USA. [10]Department of Pediatrics, The Ohio State University, Columbus, Ohio, USA. [11]Center for Cardiovascular and Pulmonary Research, Research Institute at Nationwide Children's Hospital, Columbus, Ohio, USA. [12]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. [13]Pacific Northwest Diabetes Research Institute, Seattle, Washington, USA. [14]Gladstone Institute of Cardiovascular Disease and University of California, San Francisco, San Francisco, California, USA. [15]Omicia, Inc., Oakland, California, USA. Correspondence should be addressed to M.Y. (myandell@genetics.utah.edu) or C.D.H. (chad@hufflab.org).

Received 17 October 2013; accepted 4 April 2014; published online 18 May 2014; doi:10.1038/nbt.2895

One particular challenge in pedigree analysis lies in mapping *de novo* causal mutations, i.e., private mutations that occurred in the germline of affected individuals. *De novo* mutations can cause rare Mendelian diseases[10] as well as common complex diseases such as autism[11]. However, the analyses of *de novo* mutations face a few nontrivial challenges: (i) *De novo* mutations are not in linkage with any other genetic markers; as a result, traditional linkage methods cannot analyze them; (ii) sequencing technologies will generate a number of erroneous variant calls that resemble *de novo* mutations, and failing to properly account for the platform-specific genotyping errors may introduce either type I or type II errors; (iii) in large-scale pedigree studies of complex genetic diseases, both *de novo* and inherited mutations can contribute to the disease prevalence; separately analyzing the risk of these two types of disease mutations will result in a loss of power.

Previously, we developed the Variant Annotation, Analysis and Search Tool (VAAST)[12,13]. VAAST implements an RVAT that uses a composite likelihood ratio test ($CLRT_v$) to incorporate two types of genetic information: allele frequency differences between cases and controls and variant classification information from phylogenetic conservation and predicted biochemical function. VAAST performs variant classification in conjunction with the association test. Variants with a high likelihood under the disease model (for example, variants with large differences in case and control frequencies and producing non-conservative amino acid changes) receive high CLRT scores, whereas variants predicted as neutral by VAAST receive a score of 0. For this reason, VAAST is robust to inclusion of common variants. More recently, we demonstrated that VAAST is applicable to a wide array of disease scenarios using both simulations and empirical data sets[13].

Here we present pVAAST, a tool that combines linkage analysis, case-control association and functional variant prediction in a unified statistical framework that offers much higher power relative to each of the individual methods. We demonstrate the utility of pVAAST in a variety of simulated and real data sets involving dominant, recessive and *de novo* patterns of inheritance across a broad range of family-based study designs.

RESULTS

pVAAST

pVAAST searches through the personal genomic data from disease pedigrees, sporadic cases and unaffected controls to identify genes associated with disease. To do so, it combines logarithm of odds (lod) scores with association signals to generate a unified test statistic that offers a higher power compared to either method alone. Unlike lod scores in traditional parametric linkage analysis, the lod score in pVAAST is designed for sequence data. Specifically, the statistical model assumes that the dysfunctional variants influencing disease-susceptibility can be directly detected. As a result, the pVAAST lod score is in general more powerful than traditional linkage analysis with sequencing data, as we show below. Moreover, this assumption allows us to calculate lod scores for *de novo* mutations, which is not possible with traditional linkage analysis, given that *de novo* mutations are not in linkage with other markers. pVAAST is built upon the CLRT used in VAAST, but in addition integrates the linkage information (quantified by a lod score) as a separate log likelihood ratio in the pVAAST CLRT ($CLRT_p$) (**Fig. 1**). pVAAST evaluates the significance of the $CLRT_p$ score using a combination of a randomization test and a gene-drop simulation[14] (Online Methods).

Simulated family data

We first evaluated the performance of pVAAST to identify variants causing rare Mendelian diseases using simulated family data and unaffected control genomes (we recorded the parameterization of all pVAAST experiments in this manuscript in **Supplementary Note 1**). We investigated three disease models using both association- and pedigree-based approaches:

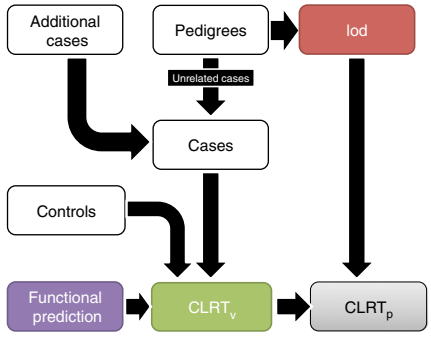


Figure 1 A schematic illustration of pVAAST. The three components of the pVAAST $CLRT_p$ are binomial likelihood test based on alleles counts in cases and controls ($CLRT_v$), functional prediction likelihood ratio and lod score. These are summed to generate the central test statistic of pVAAST.

dominant, recessive and dominant resulting from *de novo* mutations. In all models, we compared pVAAST with two rare-variant association tests, VAAST[12] and SKAT-O[3] (version 0.91; using the 'linear.weighted' kernel and 'optimal.adj' method). For comparison, we also included a nonparametric linkage method based on an idealized scenario with perfect knowledge of identity-by-descent (IBD) states in all families and a two-point parametric linkage analysis using Superlink[15] for dominant and recessive models (**Supplementary Note 2**). For the *de novo* model, we included a Poisson-based test, which detects excess inheritance error in cases (**Supplementary Note 2**). pVAAST correctly controls for type I error in all three scenarios (**Supplementary Fig. 1**).

We used each method to analyze the required sample size at four different levels of population-attributable risk (PAR)[16] (**Fig. 2a–c**). Under all disease models, pVAAST was consistently the most powerful approach. The required sample size of pVAAST was usually an order of magnitude lower than for nonparametric linkage analysis, demonstrating the value of case-control sequencing data in the identification of genes associated with rare Mendelian diseases. Under dominant and *de novo* models, pVAAST typically required half the sample size of VAAST, and one-fifth the sample size of SKAT-O. Under the *de novo* model, the Poisson-based test was more powerful than rare-variant association tests alone (VAAST and SKAT-O), but substantially less powerful than pVAAST. In general, parametric linkage analysis performed worse than nonparametric (**Fig. 2a–b** and **Supplementary Table 1**), which is expected given that our nonparametric test was based on perfect knowledge of IBD states.

We also benchmarked the performance of pVAAST in common, complex diseases by simulating four-generation families (**Fig. 2d**). We compared the relative performance of four different choices of sequenced pedigree members: affected parent-offspring pairs, affected first-cousin pairs, affected second-cousin pairs and the entire pedigree. We simulated mildly deleterious risk alleles with a selection coefficient of 0.001, which resulted in an average MAF of 1.9×10^{-3} (**Fig. 2e**). With all pedigree members shown in **Figure 2e**, pVAAST required only 66% of the sample size of VAAST, and with affected first- or second-cousin pairs, pVAAST required 79% the sample size of VAAST (**Fig. 2e**). We observed no performance improvement with affected parent-offspring pairs in pVAAST compared to VAAST. With a selection coefficient of 0.01 (average MAF = 2.2×10^{-4}) (**Fig. 2f**), we observed a similar trend but with slightly better pVAAST performance in all scenarios. pVAAST correctly controlled for type I error in all scenarios (**Supplementary Fig. 2**). For both the rare and common disease simulations, we also compared the performance of pVAAST to ASKAT[9] (version 1.2d, build 2013-09-05), an extension of SKAT

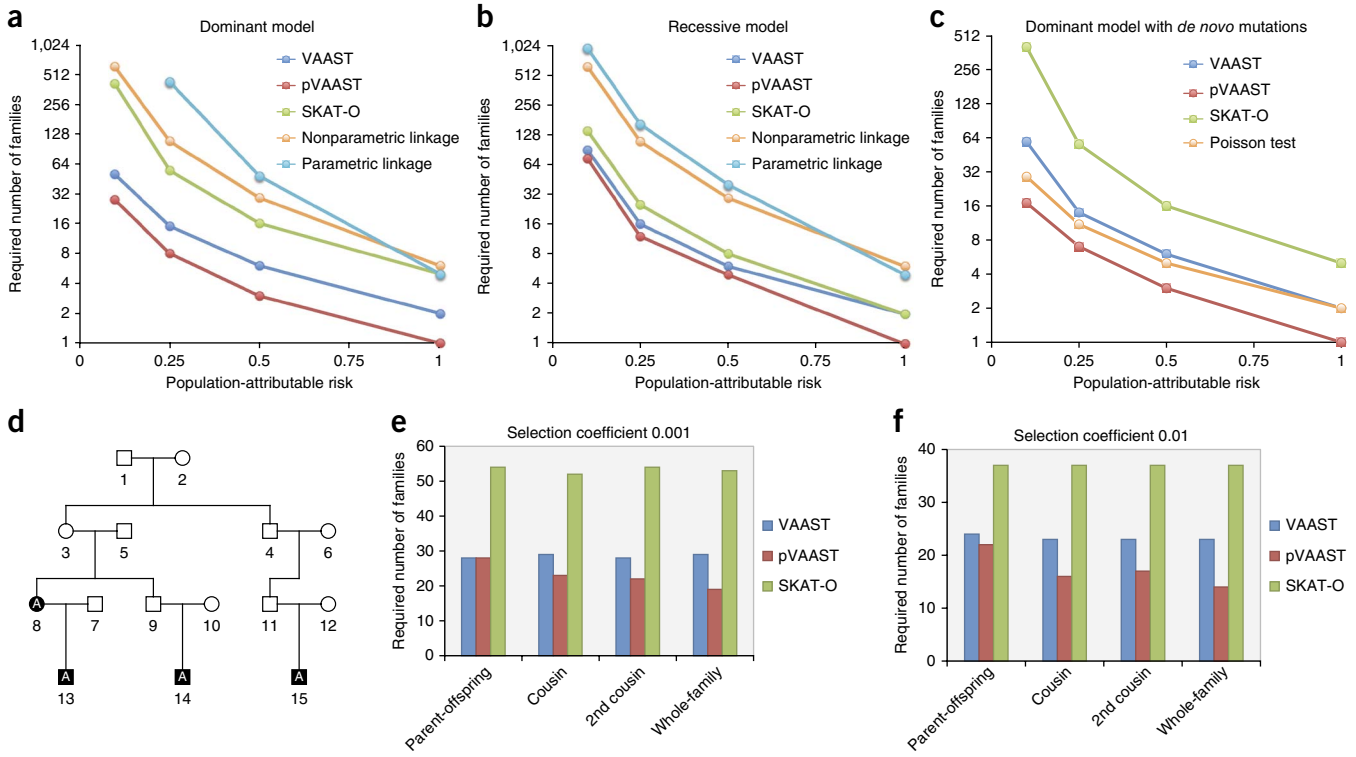


Figure 2 Rare Mendelian and common complex disease simulations. (**a**–**c**) Sample sizes required to achieve 80% power by VAAST, pVAAST, SKAT-O, parametric linkage, nonparametric linkage and a Poisson-based test, in rare Mendelian disease simulations. (**a**) A dominant model simulation, assuming two affected cousins from each pedigree are sequenced. (**b**) A recessive model simulation, assuming two affected siblings from each pedigree are sequenced. (**c**) A *de novo* mutation model simulation, assuming the whole trio is sequenced and genotyping error rate is 1×10^{-5}. At PAR = 0.1 in **a**, the required sample size to achieve 80% by the parametric linkage test is greater than the maximal sample size that we evaluated (1,000); thus we did not show this data point. (**d**–**f**) Benchmark experiments on simulated common complex disease pedigrees. (**d**) Simulated pedigree structure. Individuals labeled 'A' were always affected; other individuals were allowed to be either affected or unaffected in the rejection sampling. (**e**) Required sample size to achieve 80% power when selection coefficient is 0.001. (**f**) Required sample size to achieve 80% power when selection coefficient is 0.01. In **e** and **f**, PAR was 0.05. Sample size is defined as the number of pedigrees used for the analysis. Type I error was set to 5×10^{-4}. In all experiments 1,000 control genomes were used.

that accommodates family-based studies. However, ASKAT controls for familial relationships through asymptotic assumptions, and for the relatively small sample sizes that we evaluated, the type I error of ASKAT was inflated (**Supplementary Fig. 3a–f**).

De novo inheritance in an enteropathy pedigree

We performed whole-genome sequencing on a family quartet and used pVAAST to identify the potential causal mutation for a child with undiagnosed enteropathy (**Fig. 3a**). The proband was a 12-year-old male with severe diarrhea, total villous atrophy and hypothyroidism. Both parents and the sibling of the proband were unaffected. The phenotype was most consistent with the IPEX syndrome (OMIM 304790), but clinical sequencing of the *FOXP3* and *IL2RA* genes revealed no pathogenic mutations.

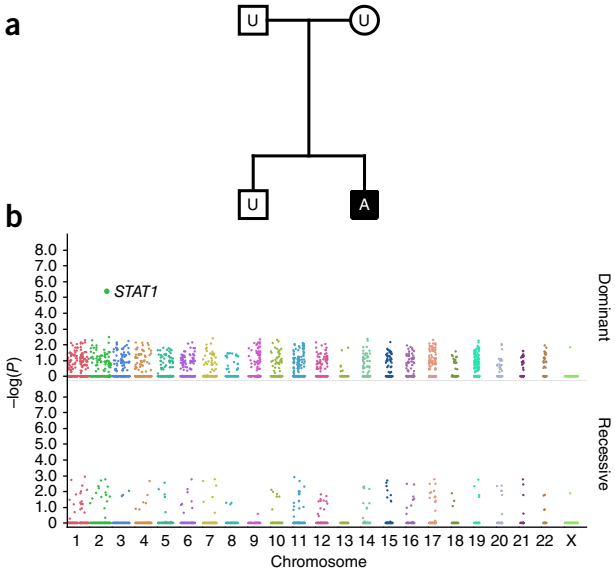
We analyzed this pedigree using both the dominant and recessive models in pVAAST. Under the dominant model, the highest-ranking gene, *STAT1*, had a *P* value of 3.97×10^{-6}. The only variant in this gene is a *de novo* mutation in the affected child, with a lod score of 0.70 and a $CLRT_p$ score of 11.724. The second ranking gene was *PAX3* ($P = 3.33 \times 10^{-3}$; lod score = 0 and $CLRT_p$ score = 11.047). *STAT1* was the only gene in the genome with a lod score >0.1; genes with lod scores between 0.1 and 0 fit an inheritance pattern of dominance with incomplete penetrance. Under the recessive model, no gene has a *P* value $<1.18 \times 10^{-3}$ (**Fig. 3b**). We validated the *de novo* inheritance pattern by genotyping the offspring and parental genotypes with Sanger sequencing. Other than this mutation, we did not identify any exonic variation in *STAT1* in the family. This heterozygous mutation is observed only in the proband but not in the parents or unaffected sibling.

The *de novo* mutation found in the affected child is a single-nucleotide guanine-to-adenine mutation, causing the amino acid change T385M in the DNA-binding motif of *STAT1*; the reference allele–encoded threonine is conserved among almost all sequenced vertebrate genomes[17]. *STAT1* encodes a transcription factor belonging to the signal transducers and activator of transcription family; both gain- and loss-of-function mutations in *STAT1* cause human disease[18]. Gain-of-function mutations in *STAT1* cause autosomal dominant chronic mucocutaneous candidiasis (CMC)[19–21] and an IPEX-like phenotype[22]. The T385M mutation was reported as a cause of CMC in a Japanese patient[23] and a Ukrainian patient[24]. These data support T385M as the causative mutation for this patient's phenotype, and demonstrate pVAAST's ability to identify a causal *de novo* mutation from a family quartet with a single affected proband.

Dominant inheritance in a cardiac septal defect pedigree

We analyzed whole-genome sequencing data from a previous study[25] on a single pedigree affected with cardiac septal defects and having an autosomal dominant pattern of inheritance (**Fig. 4a**). Previously[25], the G296S mutation in GATA-binding protein 4 (encoded by *GATA4*) was identified as the cause of cardiac septal defects in this pedigree using genome-wide linkage mapping followed by sequencing of the *GATA4* coding region and functional studies. pVAAST successfully identified *GATA4* with genome-wide significance ($P = 2.0 \times 10^{-9}$; **Fig. 4b**). The mutation encoding G296S had a $CLRT_p$ score of 38.4

Figure 3 pVAAST results on the enteropathy pedigree. (**a**) The pedigree structure. A, affected; U, unaffected. (**b**) The genome-wide gene *P* values reported by pVAAST under dominant and recessive models. The *x* axis shows the genomic locations arranged by chromosome.

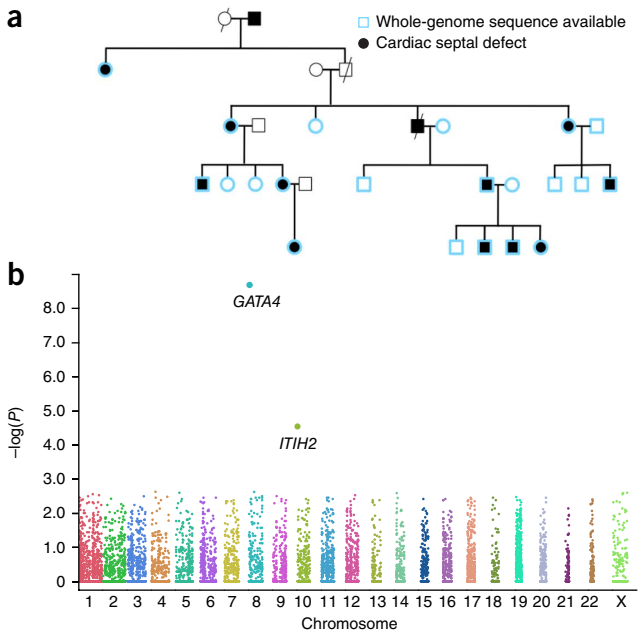


($CLRT_v$ score = 13.2; lod score = 5.47), and no other variants received a positive $CLRT_v$ or lod score in *GATA4*. The second-ranking gene was *ITIH2*, with a *P* value of 2.3×10^{-5} and a lod score of 1.51. Because the prevalence parameter (disease prevalence in general population) was set to 0.01 to match that of cardiac septal defects, no other gene received a positive lod score in the pedigree. ASKAT was not applicable to this example owing to the small sample size (**Supplementary Fig. 3g**). When VAAST analyzed the genomic sequence of a single affected individual in the cardiac septal defect pedigree (the affected individual in the second generation), *GATA4* was ranked forty-first genome-wide, with a *P* value of 2.0×10^{-3} (**Supplementary Fig. 4**).

We also analyzed the cardiac septal defect pedigree using a two-point parametric linkage test implemented in Superlink[15]. The mutation encoding G296S in *GATA4* has a lod score of 5.13 and was the highest-scoring variant genome-wide. Assuming $2\ln(10^{lod})$ is χ^2 distributed with two degrees of freedom (penetrance and recombination frequency), the *P* value of the mutation encoding G296S from two-point linkage analysis was 7.32×10^{-6}.

Recessive inheritance in a Miller's syndrome pedigree

We investigated the performance of pVAAST on a recessive disease, Miller's syndrome, using previously generated[7] whole-genome sequencing data from a two-generation pedigree (**Fig. 5a**). The two offspring are affected with Miller's syndrome and primary ciliary dyskinesia, both of which are rare recessive Mendelian diseases. The two diseases are caused by compound heterozygous mutations in the *DHODH* and *DNAH5* genes, respectively[7]. All four individuals in the family quartet were sequenced. pVAAST identified only five genes with positive lod scores, and the two disease-causal genes (*DHODH* and *DNAH5*) were ranked first and second genome-wide (**Fig. 5b**), with *P* values of 3.3×10^{-5} and 1.3×10^{-4}, and $CLRT_v$ scores of 27.9 and 30.8, respectively. The lod scores were 1.204 in both cases. In both genes, only the two causal mutations received positive scores; all other variants had scores of 0.



We also explored the performance of pVAAST after removing one affected child (B01) from the pedigree. That is, we converted the original Miller's syndrome pedigree to a trio family with two unaffected parents and one affected child. In this scenario, *DHODH* and *DNAH5* were ranked first and thirteenth genome-wide, respectively (**Supplementary Fig. 5a**), both with lod scores of 0.602. We also ran VAAST over the genome-sequencing data of only one affected child (i.e., not using the data from the parents and the affected sibling). *DHODH* and *DNAH5* were ranked tenth and twenty-seventh, respectively (**Supplementary Fig. 5b**). In our previous work, by enforcing a strict filtering method based on inheritance patterns and minor allele frequencies, VAAST was also able to identify the correct causal genes in this pedigree but was unable to produce an accurate *P* value that accounted for the familial relationships[12].

Challenging situations in pedigree studies

In linkage analysis, factors such as incomplete penetrance, locus heterogeneity and missing phenotypes negatively affect linkage signals and thus reduce disease-gene identification power. The cardiac septal defect pedigree data presented above (**Fig. 4**) is a large pedigree with no locus heterogeneity and very high penetrance (93.3%) for the mutation encoding G296S. We modified the genotype and phenotype data from this pedigree (**Supplementary Note 3**) to benchmark pVAAST in four scenarios: (i) missing phenotypes, (ii) reduced penetrance, (iii) locus heterogeneity and (iv) reduced number of informative meioses in the family. For each test case, we evaluated the lod score and the genome-wide ranking of *GATA4* (ranked by *P* values). The lod score reported by pVAAST was approximately a monotonic function of each of the four parameters and was highly correlated with the classic two-point parametric lod score (**Fig. 6**). pVAAST was robust to pedigrees with missing phenotype data. For example, when 82% of pedigree members had unknown phenotypes, the lod score of *GATA4* was 1.5 and genome-wide ranking was first (**Fig. 6a**,**b**). Reduced penetrance generally decreased the lod score without significantly compromising

Figure 4 pVAAST identifies the dominant causal gene *GATA4* in cardiac septal defect pedigree. (**a**) Illustration of the cardiac septal defect pedigree. (**b**) Manhattan plot of the *P* values of all protein-encoding genes from the pVAAST run; each dot in the plot represents one gene. The *x* axis shows the genomic locations arranged by chromosome.

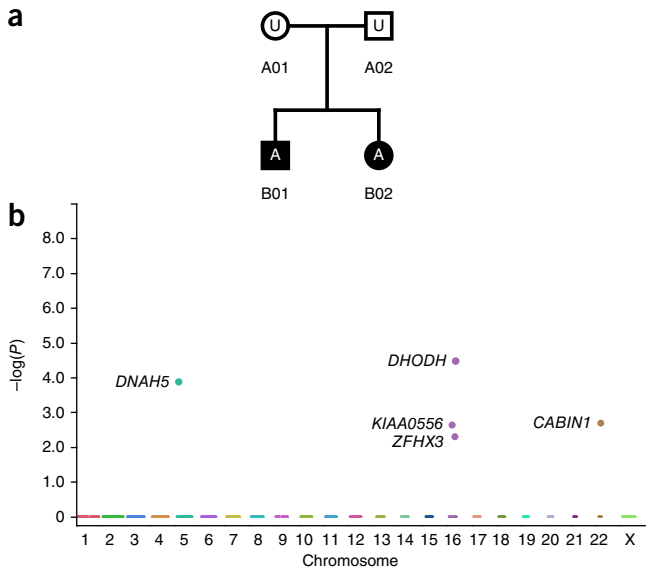


Figure 5 pVAAST identifies the recessive causal genes for Miller's syndrome (*DHODH*) and primary ciliary dyskinesia (*DNAH5*) with a two-generation pedigree. (**a**) Pedigree structure. 'A' denotes affected individuals; 'U' denotes unaffected individuals. (**b**) Manhattan plot of the *P* values of all protein-encoding genes in the whole-genome run of pVAAST. Each dot represents one gene. The *x* axis shows the genomic locations arranged by chromosome. All four individuals in the family quartet were sequenced.

the genome-wide ranking (**Fig. 6c**,**d**). Specifically, the genome-wide ranking of *GATA4* was consistently first until the penetrance dropped below 40%; even with penetrance of 20%, *GATA4* was ranked eighth genome-wide. In comparison, locus heterogeneity had a greater impact on power (**Fig. 6e**,**f**). When locus heterogeneity was modest, *GATA4* always ranked first or second. However, when the proportion of affected individuals carrying G296S fell to 50%, the lod score dropped below 0.2, and the genome-wide ranking was beyond fiftieth. The original family has 20 informative meiosis events, and our results show pVAAST ranked *GATA4* first genome-wide even when there are only 11 informative meioses in the family (**Fig. 6g**). Furthermore, with only six meioses, pVAAST still ranked *GATA4* second genome-wide. This suggests that for a rare Mendelian disease with high penetrance and low locus heterogeneity within the family, the risk gene can often be identified among the top hits genome-wide using a typical three-generation pedigree.

For comparison, we evaluated the genome-wide ranking of *GATA4* with three alternative approaches. In the first approach, we calculated a two-point parametric lod score at each polymorphism site with Superlink[15] and designated the lod score from the best-scoring site overlapping a protein-encoding gene as the gene lod score. We then ranked all genes by the gene lod scores. We also attempted to perform multipoint linkage analysis with Merlin[26], but this proved computationally infeasible. In the second approach, we applied the same procedure to the pVAAST lod score to calculate the ranking (**Fig. 6**). Finally, we evaluated a hard-filtering approach that only considered variants that perfectly fit the expected inheritance pattern with minor allele frequencies below 0.5% (**Supplementary Note 3**).

We found that the pVAAST lod score was consistently more robust than the classic two-point parametric lod score in challenging scenarios such as low penetrance, high locus heterogeneity, small sample size and large fraction of unknown phenotypes. The ranking of *GATA4* with pVAAST lod scores was usually one order of magnitude higher than with Superlink. This performance difference is perhaps not surprising given that traditional linkage analysis tests the hypothesis of disease linkage rather than disease causation and was developed for sparse marker data rather than complete sequence data. Ranking using pVAAST *P* values instead of lod scores further improved the accuracy of disease-gene identification, and the improvement was pronounced when the penetrance was low or the phenotypes were missing for a large fraction of the pedigree. Hard-filtering makes strict assumptions about the expected inheritance pattern and minor allele frequency of the causal mutation. When these assumptions hold, hard-filtering has comparable performance to traditional two-point linkage analysis but is less robust compared to pVAAST and pVAAST lod scores. However, hard filtering performed very poorly when any of these assumptions were violated (**Fig. 6e**).

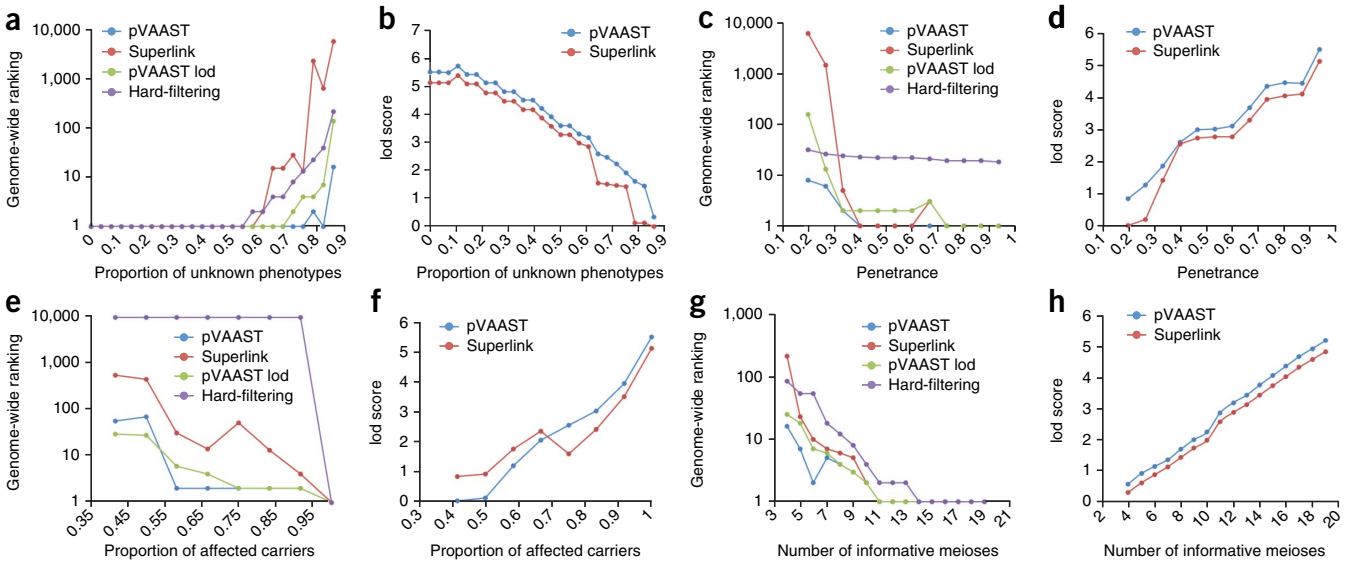


Figure 6 The genome-wide ranking and lod score of *GATA4* in challenging situations of pedigree studies. (**a**–**h**) lod scores and genome-wide rankings corresponding to differing levels of unknown phenotypes (**a**,**b**), degrees of penetrance (**c**,**d**), proportion of affected individuals being G296S mutation carriers (**e**,**f**) and number of informative meioses (**g**,**h**). For genome-wide rankings, *y* axis is shown in log scale, and four methods were compared (pVAAST, Superlink, pVAAST lod and a hard-filtering approach).

We also investigated the impact of incomplete penetrance, locus heterogeneity and unknown phenotypes in conjunction with smaller family sizes. To do so, we used only a subset of the individuals in the original cardiac septal defect pedigree to reduce the number of informative meiosis. We evaluated the genome-wide ranking of *GATA4* using pVAAST, pVAAST lod scores, two-point linkage analysis in Superlink, multipoint linkage analysis in Merlin[26] and hard filtering (**Supplementary Figs. 6–8**). The ranking of *GATA4* was highest when using pVAAST in almost all scenarios, which is consistent with the results involving the entire family (**Fig. 6**).

DISCUSSION

Because pVAAST employs the same CLRT framework as its predecessor, VAAST, a comparison of these two algorithms demonstrates the power gained by using inheritance information from pedigrees. In dominant rare Mendelian diseases, the improvement is remarkable: when an additional affected cousin was sequenced, pVAAST required only half the number of families as VAAST (**Fig. 2a**), regardless of the level of locus heterogeneity. These results demonstrate that although linkage analysis is usually substantially less powerful than a rare-variant association test (RVAT) alone, in these scenarios, linkage provides orthogonal information for disease-gene identification, and this information can greatly improve the power of association tests. Although RVATs were initially developed for common genetic disorders, we previously demonstrated that they are more powerful than standard hard-filtering approaches often used to analyze rare Mendelian diseases[12,13]. The current study extends this work and provides a unified test that computes a single *P* value over the combined linkage and association evidence.

Classic linkage methods were designed for sparse genetic-marker data and model the recombination frequencies between genetic markers and disease to identify large genomic regions in the family that may harbor a causal mutation. In contrast, pVAAST is designed for sequence-based studies and assumes that the causal mutations can be directly assayed. Our model also incorporates an additional unobserved risk locus (latent locus) to capture an additional layer of genetic architecture of the disease, enabling pVAAST to accurately model complex diseases in families with phenocopies or locus heterogeneity. For these reasons, the pVAAST lod score typically outperformed both the classic two-point (**Fig. 6**) and multipoint (**Supplementary Figs. 6–8**) parametric lod scores in the scenarios we evaluated, particularly in challenging scenarios relevant to common, complex disease involving reduced penetrance, locus heterogeneity, small sample size or missing phenotypes.

Our results from the enteropathy, cardiac septal defect and Miller's syndrome pedigrees demonstrate that pVAAST can successfully identify rare, Mendelian disease-causing variants from genome-wide searches involving only a single pedigree. In particular, the identification of *STAT1* as the likely cause of enteropathy in a small pedigree establishes that excellent statistical resolution can be achieved in a small family with a disease-causing *de novo* mutation (**Fig. 3**). It should be noted, however, that in *de novo* disease models the genotyping error rate has a large impact on power (**Supplementary Fig. 9**), and with higher genotyping error rates that can result from earlier sequencing or variant-calling technologies, a potential *de novo* mutation is more likely to be a sequencing error and less likely to be a true *de novo* event. The results shown in **Figures 3–5** also show that pVAAST is robust to technical complications that are present in real genomic data but not represented in simulations, such as genotyping errors, missing genotype calls and differences in sequencing platforms between cases and publicly available controls.

An important practical consideration is which family members to sequence to achieve optimal power. For rare Mendelian diseases with high penetrance, the choice is straightforward given that the inheritance path of the causal mutation can be inferred. However, for common genetic disorders, determining the optimal choice of family members is more complex. Sequencing more distantly related individuals increases the number of informative meioses in the pedigree but also increases the probability of phenocopies. Here we show that in a common complex disease with a modest level of locus heterogeneity (PAR = 0.05 and only 40% of affected individuals carrying mutations with odds ratio >1.1 in the gene of interest; see also **Supplementary Note 2**), sequencing affected first- or second-cousin pairs yields substantially better results than sequencing affected parent-offspring pairs in the same family (**Fig. 2e–f**). Sequencing the entire extended family offers a modest improvement over cousin pairs, consistent with previous findings[27].

If sample size is not a limiting factor, another consideration is the cost effectiveness of sequencing pedigrees versus unrelated cases. For example, as shown in the simulations of dominant inheritance, pVAAST requires half the number of pedigrees as VAAST but requires two individuals per pedigree to be sequenced (**Fig. 2a**). Thus, with affected cousin pairs, the two approaches are equally cost effective. However, in rare Mendelian diseases with high penetrance, because the *P* value decreases exponentially with the number of informative meiosis (**Supplementary Fig. 10**), sequencing affected pairs more distant than the first cousin is more cost-effective than sequencing only unrelated index cases from each pedigree. A two-stage design can also be cost effective. Specifically, in the first stage, only unrelated cases are sequenced, and VAAST prioritizes genes according to their significance levels. In the second stage, candidate risk variant in the relatives of affected carriers are genotyped, and pVAAST analyzes the original sequence data with the additional genotype information. This approach can be economical given the relative costs genotyping and whole-exome sequencing. Although pVAAST is primarily designed for sequence data, it is also applicable to exome chip genotyping data. pVAAST was recently used to identify candidate genes associated with an increased risk of suicide from exome chip data in extended high-risk pedigrees[28].

Because pVAAST combines linkage analysis and case-control association, all the caveats from these methods are applicable. In particular, loci not causally related to disease may be in linkage disequilibrium with a causal locus in association studies. Therefore, as with traditional linkage analysis and association tests, rejection of the null hypothesis in pVAAST can establish disease-gene association but cannot rule out the possibility that the association results from a linked locus that is causal. As with other case-control association tests, uncontrolled confounding covariates can potentially inflate type I error rates in pVAAST. To control for covariates, pVAAST can interface with the BiasedUrn package[29] (http://cran.r-project.org/web/packages/BiasedUrn/index.html) to conduct a covariate adjusted randomization test (**Supplementary Note 4**).

Existing family-based sequence-analysis approaches are typically applicable to only a narrow range of studies. Hard filtering approaches that enforce strict inheritance patterns are appropriate for studies involving small families with rare Mendelian diseases but do not provide robust statistical interpretations and do not scale to large families or common, complex diseases[7,30]. Sequence analysis in large families typically involve multistep ad hoc procedures in which linkage analysis or IBD mapping is used to identify large genomic regions followed by the application of a series of hard filters based on inheritance patterns, variant annotations and population allele frequencies[31]. In addition, approaches that rely primarily on hard filters do not scale well to multifamily studies[12,13]. ASKAT is a family-based rare-variant association test that is designed for large, multifamily studies but is not presently applicable to studies involving relatively small sample sizes. Methods used to identify disease-causing *de novo* mutations can efficiently combine statistical evidence from multiple families but require parent-offspring trios and cannot incorporate evidence from families with inherited disease[32].

pVAAST is also applicable to non-disease trait mapping in nonhuman species. In typical genetic screens in model organisms, researchers cross-breed individuals with different phenotypes for generations and then map the locations of possible causal variants using linkage analysis. When sequencing data are available, pVAAST could be an attractive alternative to traditional mutation mapping in these studies, as it incorporates additional information from association signals and functional predictions of the mutations. This is especially true for species with high levels of genetic diversity such as rice[33] and maize[34], where a large proportion of near-neutral variants may complicate the identification of mutations responsible for the phenotype. The integrated variant classification functionality in VAAST and pVAAST may mitigate these challenges[35,36].

In contrast to existing methods, pVAAST performs well across a wide range of study designs, from a single small family with a rare, Mendelian disease to hundreds of families with common, complex genetic diseases and arbitrary pedigree structures. pVAAST is a flexible, general-purpose tool for identifying disease-associated genes that combines variant classification, rare-variant association testing and linkage analysis in a unified statistical framework to increase the power and reduce the technical complexity of family-based sequencing studies.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The human genome sequencing data for the enteropathy and cardiac septal defects pedigrees have been submitted to the database of Genotypes and Phenotypes (dbGaP), and accession codes will be provided as soon as they are available. Meanwhile, inquires about the data should be directed to M.Y. or C.D.H.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

An allocation of computer time on the University of Texas MD Anderson Research Computing High Performance Computing (HPC) facility is gratefully acknowledged. This work was supported by US National Institutes of Health grants R01 GM104390 (M.Y., L.B.J., C.D.H. and H.H.), R01 DK091374 (S.L.G., C.D.H. and L.B.J.), R01 CA164138 (S.V.T. and C.D.H.), R44HG006579 (M.G.R. and M.Y.) and R01 GM59290 (L.B.J.) as well as the University of Luxembourg—Institute for Systems Biology Program. D.S. was supported by grants from the NHLBI (UO1 HL100406 and U01 HL098179) related to this project. H.C. was supported by NIH grants R01 MH094400 and R01 MH099134. H.H. was supported by the MD Anderson Cancer Center Odyssey Program. J.X. was supported by NIH grant R00HG005846.

AUTHOR CONTRIBUTIONS

C.D.H. conceived of the project. C.D.H. oversaw and coordinated the research. C.D.H. and H.H. designed the algorithms. H.H. and B.M. wrote the software. C.D.H., H.H. and P.S. contributed to the statistical development. C.D.H., H.H., J.C.R., M.Y., S.V.T., D.S., K.V.V., L.H., L.B.J., M.G.R. and S.L.G. designed the experiments. H.H., H.C., W.W., R.L.M., J.D.D., S.W., H.L., J.X., Shankaracharya, R.H., B.M., J.C. and G.G. performed the experiments. H.H., C.D.H., M.Y., S.V.T., S.L.G. and L.B.J. analyzed and interpreted the data. H.H. generated the figures. H.H., C.D.H., L.B.J., M.Y., S.L.G., P.S., and S.V.T. wrote the paper. S.L.G., D.S., V.G., D.J.G., L.H., H.L., R.H., K.V.V., R.L.M., J.D.D., G.G. participated in pedigree identification, recruitment and validation.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Borecki, I.B. & Province, M.A. Linkage and association: basic concepts. *Adv. Genet.* **60**, 51–74 (2008).
2. Muller, H.J. Our load of mutations. *Am. J. Hum. Genet.* **2**, 111–176 (1950).
3. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
4. Neale, B.M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7**, e1001322 (2011).
5. Ng, P.C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80 (2006).
6. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
7. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
8. Schaid, D.J., McDonnell, S.K., Sinnwell, J.P. & Thibodeau, S.N. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet. Epidemiol.* **37**, 409–418 (2013).
9. Oualkacha, K. *et al.* Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet. Epidemiol.* **37**, 366–376 (2013).
10. Hoischen, A. *et al. De novo* mutations of *SETBP1* cause Schinzel-Giedion syndrome. *Nat. Genet.* **42**, 483–485 (2010).
11. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
12. Yandell, M. *et al.* A probabilistic disease-gene finder for personal genomes. *Genome Res.* **21**, 1529–1542 (2011).
13. Hu, H. *et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet. Epidemiol.* **37**, 622–634 (2013).
14. Jung, J., Weeks, D.E. & Feingold, E. Gene-dropping vs. empirical variance estimation for allele-sharing linkage statistics. *Genet. Epidemiol.* **30**, 652–665 (2006).
15. Fishelson, M. & Geiger, D. Exact genetic linkage computations for general pedigrees. *Bioinformatics* **18** (suppl. 1), S189–S198 (2002).
16. Rosner, B. Fundamentals of biostatistics, edn. 7 (Cengage Learning, Boston, 2011).
17. Dreszer, T.R. *et al.* The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* **40**, D918–D923 (2012).
18. Boisson-Dupuis, S. *et al.* Inborn errors of human *STAT1*: allelic heterogeneity governs the diversity of immunological and infectious phenotypes. *Curr. Opin. Immunol.* **24**, 364–378 (2012).
19. Hori, T. *et al.* Autosomal-dominant chronic mucocutaneous candidiasis with *STAT1*-mutation can be complicated with chronic active hepatitis and hypothyroidism. *J. Clin. Immunol.* **32**, 1213–1220 (2012).
20. Liu, L. *et al.* Gain-of-function human *STAT1* mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. *J. Exp. Med.* **208**, 1635–1648 (2011).
21. van de Veerdonk, F.L. *et al. STAT1* mutations in autosomal dominant chronic mucocutaneous candidiasis. *N. Engl. J. Med.* **365**, 54–61 (2011).
22. Uzel, G. *et al.* Dominant gain-of-function *STAT1* mutations in *FOXP3* wild-type immune dysregulation-polyendocrinopathy-enteropathy-X-linked-like syndrome. *J. Allergy Clin. Immunol.* **131**, 1611–1623 (2013).
23. Takezaki, S. *et al.* Chronic mucocutaneous candidiasis caused by a gain-of-function mutation in the *STAT1* DNA-binding domain. *J. Immunol.* **189**, 1521–1526 (2012).
24. Soltész, B. *et al.* New and recurrent gain-of-function *STAT1* mutations in patients with chronic mucocutaneous candidiasis from Eastern and Central Europe. *J. Med. Genet.* **50**, 567–578 (2013).
25. Garg, V. *et al.* GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* **424**, 443–447 (2003).
26. Abecasis, G.R., Cherny, S.S., Cookson, W.O. & Cardon, L.R. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101 (2002).
27. Feng, B.J., Tavtigian, S.V., Southey, M.C. & Goldgar, D.E. Design considerations for massively parallel sequencing studies of complex human disease. *PLoS ONE* **6**, e23221 (2011).
28. Coon, H. *et al.* Genetic risk factors in two Utah pedigrees at high risk for suicide. *Transl. Psychiatr.* **3**, e325 (2013).
29. Epstein, M.P. *et al.* A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *Am. J. Hum. Genet.* **91**, 215–223 (2012).
30. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
31. Marchani, E.E. *et al.* Identification of rare variants from exome sequence in a large pedigree with autism. *Hum. Hered.* **74**, 153–164 (2012).
32. Heinzen, E.L. *et al. De novo* mutations in *ATP1A3* cause alternating hemiplegia of childhood. *Nat. Genet.* **44**, 1030–1034 (2012).
33. Zhao, K. *et al.* Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.* **2**, 467 (2011).
34. Vigouroux, Y. *et al.* Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. *Am. J. Bot.* **95**, 1240–1253 (2008).
35. Shapiro, M.D. *et al.* Genomic diversity and evolution of the head crest in the rock pigeon. *Science* **339**, 1063–1067 (2013).
36. Domyan, E.T. *et al.* Epistatic and combinatorial effects of pigmentary gene mutations in the domestic pigeon. *Curr. Biol.* **24**, 459–464 (2014).

ONLINE METHODS

Basic lod score calculation in pVAAST. In classic two-point parametric linkage analysis, the marker under investigation is usually assumed not to be causal but rather linked with the actual causal variant with a certain recombination probability (r). Under the null hypothesis, $r = 0.5$, which indicates that the marker and causal mutation are unlinked. Under the alternative hypothesis, r is a free parameter. Given the disease prevalence, allele frequency of the marker and causal allele and the penetrance of the causal allele, the likelihood of alternative and null model can be calculated for given values of r using the Elston-Stewart algorithm[37]. The $\log_{10}$ ratio of the maximum likelihood of the alternative and null model is the lod score.

For simplicity, we use the term causal to refer to any variant that directly increases disease risk, regardless of penetrance. Our model assumes that the disease is caused by either the locus under investigation (current locus) or some other unlinked locus in the genome (latent locus). In both models, the current and latent loci are unlinked, and there is no epistatic interaction between the alleles. The null model states that variant(s) in the latent locus cause the disease with some probability, and the current locus is not causal. The alternative model states that variants in both the current and latent loci can independently cause the disease, with different probabilities. In other words, the null model attributes the disease phenotype solely to the latent locus, and the alternative model allows variants in both the current and latent loci to be independently causal. We then maximize the likelihoods of the alternative and null models over ρ_c (genotype disease probability vector for the current variant), ρ_l (genotype disease probability vector for the latent locus) and f_l (minor allele frequency of the latent locus) and calculate the log10 likelihood ratio as the lod score. Formally,

$$\text{lod} = \log_{10} \max L(alt) - \log_{10} \max L(null)$$

and the likelihood for both null model and alternative model has the form

$$L = P(g_c, g_l, p \mid \rho_c, \rho_l, f_c, f_l)$$

Here g_c and g_l are the genotype vectors (with values of 0, 1 and 2 corresponding to homozygous-reference, heterozygous and homozygous-nonreference genotypes) of the current and latent variant sites; p is the phenotype vector of the pedigree; f_c and f_l are the allele frequencies of the current and latent alleles. Under the null model, the expression can be further decomposed into

$$P_{null}(g_c, g_l, p \mid \rho_c, \rho_l, f_c, f_l) = P(g_c \mid f_c) P(g_l, p \mid \rho_l, f_l)$$

because only the latent allele is causal for the disease under the null model, and g_c is thus independent from p, g_l and ρ_l.

Given ρ_c, ρ_l, f_c and f_l, all of the aforementioned probabilities can be calculated with the Elston-Stewart algorithm[15] in linear computational time relative to the family size. We estimate f_c from the allele frequency in a control population and perform a grid search over ρ_c, ρ_l and f_l in the specified order to maximize the likelihood. By default, we explore ρ_c and ρ_l values ranging from 0 to 1 with increment of 0.1, and in addition the following values: 0.001, 0.01 and 0.999. We explored the following f_l values: 5×10^{-7}, 5×10^{-4}, 5×10^{-3}, 0.01, 0.02, 0.05, 0.5 and 0.999. The resolution of the grids is tunable. In all our experiments, the aforementioned parameters offer a good balance between algorithm efficiency and statistical power, although using a finer grid may increase the power of pVAAST at the cost of longer computation time. For the dominant model, a heterozygous genotype is sufficient to be considered as a risk genotype; for a simple recessive model, a homozygous genotype is required, with the exception of sex chromosomes. Compound heterozygous scenarios are discussed below.

If more than one family is present, for each variant, we maximize the likelihood under the assumptions that ρ_c is consistent across families but ρ_l and f_l varies between families using a nested grid search. Then, within each family, the lod of one variant is chosen to be the gene lod score of this family. By default, the variant with the highest CLRT_v score is chosen, but the user can opt to use CLRT_p score or lod score alone as well. In practice, we found that in large pedigrees, using the CLRT_p score as a selection criterion may yield more favorable results. Finally, we sum the gene lod score from multiple families to generate the overall pVAAST lod score.

Extending the dominant model to *de novo* mutations. We accommodate *de novo* mutations in our model by allowing Mendelian inheritance errors to occur in the pedigree likelihood calculation. Specifically, in the Elston-Stewart algorithm[37], if the offspring carries a mutation absent from both parents, then this transmission has a probability of m (mutation rate per site per generation in human genome; default 1.2×10^{-8} (ref. 7)). Accordingly, we also randomly introduce Mendelian inheritance error in our gene-drop simulations[14] with probability equal to the genotyping error rate.

Extending the recessive model to compound heterozygotes. Compound heterozygotes require special attention because the genotype vectors (g_l and g_c) now involve more than one variant site. Under the recessive model we are specifically interested in the situation where two deleterious mutations occur at two different chromosomes of the same gene, so that both copies are defective. To illustrate, consider a gene with three polymorphism sites, i, j and k. A straightforward approach to calculate the gene lod score would be to calculate the lod for all pairwise combination of heterozygous variant sites within the gene (i.e., i + j; i + k; and j + k) separately and then select the highest lod score. This requires the evaluation of $n(n-1)/2$ combinations, where n is the number of variant sites in the gene. However, this approach is flawed because it assumes the genotype disease probabilities for all pairs of sites are independent, which is incorrect. Instead, we assume that any variant in the gene is either causal (D-variants) or neutral (N-variants)[38] with the same relative risk. For example, if a gene has four heterozygous sites i, j, k and l, within which i, j, and k are causal, then an individual with at least two mutations at i, j and k sites on two different chromosomes would be at risk; otherwise she will not be at risk.

Under this model, we can construct a Boolean risk vector for a gene to denote whether each variant within the gene is a D-variant or N-variant. If we know the underlying risk vector for some gene, then we can easily determine the genotype of an individual by evaluating whether he or she carries at least one D-variant on each chromosome. Then the calculation of lod score is reduced to the simple recessive case described above. However, finding the optimal risk vector is not trivial, as a brute-force approach to find the risk vector maximizing the lod score has a complexity of $O(2^n)$, where n is the number of sites in the gene. To make this more efficient, we use an MCMC method[39] to approximate the optimal risk vector. Briefly, given a particular risk vector and the phenotype probability for each genotype, the joint likelihood for all sequenced and phenotyped individuals can be calculated as

$$L = \rho_r^{n_a} (1-\rho_r)^{n_b} \rho_n^{n_c} (1-\rho_n)^{n_d}$$

where ρ_r is the probability that an individual with a risky genotype is affected; ρ_n is the probability that an individual with a neutral genotype is affected; n_a and n_b are the total numbers of affected and unaffected individuals with a risky genotype, respectively; n_c and n_d are the total number of affected/unaffected individuals with neutral genotypes, respectively. Both ρ_r and ρ_n are configurable parameters, although we found that the performance of our MCMC method was usually insensitive to these parameters.

We start with a random risk vector, and randomly select a variant site to switch to the opposite value (neutral to risky and risky to neutral). The likelihoods for both risk vectors are calculated, and we selectively accept the new risk vector according to the Metropolis-Hastings method[39]. This process is repeated until convergence or the maximal number of iterations is achieved. Lastly, we select the most likely risk vector from the Markov chain and calculate the lod score as described in the previous section.

Optionally, the joint likelihood can incorporate an empirical functional score. Let $I_D(k)$ be an indicator function for whether the kth site is a D-variant. The empirical functional score (F score) is a function of VAAST CLRT_v scores across all sites in the current gene

$$F = \frac{\sum_k CLRT_{v(k)} \times I_D(k)}{2 \sum_k I_D(k)}$$

and the updated likelihood is calculated as $L^* = L \times e^F$.

doi:10.1038/nbt.2895

The calculation of $CLRT_v$ score is detailed in (ref. 12). Briefly, it is twice the log-scale composite likelihood ratio of disease model versus null model, incorporating the mutation frequency in the control genome and the functional impact of the mutation on the protein sequence. This option (mcmc_use_functional_score) can be switched on or off. We used the updated likelihood function throughout the present study, although in our recessive model simulations, these two likelihood functions generated similar results.

Integrating lod scores into the CLR test. pVAAST is built on the framework of VAAST[12], which uses an extended CLRT to determine a severity score for genomic variants. The null model of the CLRT states that the frequency of a variant or variant group is the same in the control population (background genomes) and the case population (target genomes), whereas the alternative model allows these two frequencies to differ. Under a binomial distribution, the likelihood for both models can be calculated on the basis of observed allele frequencies in the control and case data sets. This likelihood is further updated by calibrated amino acid substitution and insertion and deletion (indel) severity weights.

To integrate genetic linkage information into the CLRT, we select only one sequenced and affected individual from each pedigree (pedigree representative) to establish a group of cases. The identifiers of the selected individuals can be provided, but if such information is absent, pVAAST will randomly choose one individual carrying the highest-scoring variant in the current gene. Additional affected individuals not related to any other individuals in the study can also be included among the cases. λ represents the natural log of the composite likelihood ratio calculated as previously described[12]. We calculate the pVAAST CLRT ($CLRT_p$) score as

$$CLRT_p = c\sum_{i=1}^{n} LOD_i - 2\lambda$$

where LOD_i is the lod score for the ith family and

$$c = 2 * ln(10)$$

To avoid confusion, we denote the original CLRT score in VAAST (without the linkage component) as $CLRT_v$ in this manuscript. **Figure 1** provides a schematic diagram for the calculation of the $CLRT_p$ scores in pVAAST.

Evaluating the significance of the test statistic. c represents the two parental haplotypes at the current gene locus in a particular individual. Let subscript p, pf, b and sc represent a vector of cs among all pedigree members, pedigree founders, background (control) individuals or sporadic cases, respectively, and a superscript r or s represent real data and simulated data, respectively. For example, c^r_{pf} represents the vector of haplotypes in all pedigree founders in the real data. T represents the unordered set of chromosomes among pedigree founders, background genomes, and sporadic cases in the real situation. Our null hypothesis is that pedigree founders, controls and sporadic cases are derived from the same population and that haplotypes in pedigree offspring randomly segregate according to Mendel's law. When the two haplotypes in each sequenced individual are known and all pedigree founders are sequenced, a combination of a randomization test and gene-drop simulation can be used to evaluate any statistic that is a function of the genotype and phenotype data in the pedigree and controls.

We first sample (without replacement) N_{pf} (the cardinality of the set c^r_{pf}, i.e., $|c^r_{pf}|$) individuals from T as the pedigree founder (denoted by c_{pf}); N_{ct} ($<= |c^r_b|$) individuals as the control set for $CLRT_v$ calculation (denoted by c_{ct}); and N_{sc} ($|c^r_{sc}|$, which can be 0) individuals as the sporadic cases (denoted by c_{sc}). We then generate the c_p from c_{pf} via gene-drop simulation[14]. Briefly, we simulate the two haplotypes of each offspring by randomly sampling one of each parent's two haplotypes with equal probability. The gene dropping starts from the first generation of the pedigree and is repeated until all pedigree members are simulated. $g(c_p, c_{sc}, c_{ct})$ represents the desired test statistic. In pVAAST, this test statistic is $CLRT_p$. The real data in this procedure are represented as c^r_p, c^r_{ct} and c^r_{sc}, where c^r_{ct} is a random subset of c^r_b with size N_{ct}. If we calculate

$$P = P(\{c_p, c_{sc}, c_{ct} : CLRT_p(c_p, c_{sc}, c_{ct}) \geq CLRT^r_p\})$$

within the described sampling space, we will have a valid P value with specified type I error under the null model. This holds because the real data are one realization of the described sampling scheme with probability equal to any other realization under the null hypothesis.

In reality, because enumerating all values of c_p, c_{sc} and c_{ct} is computationally intractable, we use a Monte Carlo method to sample n realizations of the described procedure and calculate

$$P = \frac{1 + \sum_{i=1}^{n} I(CLRT^s_{p,i} \geq CLRT^r_p)}{n+1}$$

(I is an indicator function) and report this as the gene-level P value. Alternatively, P value can be calculated using the lod score instead of $CLRT_p$ score as the test statistic.

We emphasize two points: (i) the number of sporadic cases can be 0 and (ii) the choice of N_{ct} is free and does not affect the validity of the P value.

To sample from T, the above procedure requires that the haplotypes of all pedigree founders are known. In reality, c^r_{pf} can be unknown or partially known because pedigree founders may not have been sequenced, thus we may not be able to directly sample from T. To accommodate this situation, we define a new set T^* to be the unordered set of haplotypes among pedigree representatives (one affected sequencing individual in each pedigree, as denoted in the pVAAST parameter file), background genomes and sporadic cases in the real situation. Obviously we have

$$T^* \subset T$$

We propose sampling our test-statistics $CLRT_p$ from the cumulative distribution function

$$F_{CLRT_p}(CLRT_p(c_p, c_{sc}, c_{ct}) | c_{pf}, c_{ct}, c_{sc} \subset T^*)$$

during the simulation to approximate the distribution

$$F_{CLRT_p}(CLRT_p(c_p, c_{sc}, c_{ct}) | c_{pf}, c_{ct}, c_{sc} \subset T)$$

The approximation becomes more accurate when the $|T - T^*| << |T|$, or in other words, when the number of unsequenced founders is small compared to the total number of sequenced background individuals, pedigree representatives and sporadic cases. Our implementation also approximates the idealized procedure owing to haplotype phase uncertainty. Despite these approximations, we observed no inflation in type I error rate in any of the experiments we evaluated (**Supplementary Figs. 1** and **2**).

We also documented the implementation of our simulation procedure in pVAAST in **Supplementary Note 4**.

Genomic data. For the enteropathy pedigree, whole-genome sequencing was performed on all four pedigree members using the Illumina HiSeq platform. We followed the Genome Analysis Toolkit (GATK) best practice to perform variant-calling steps[40]. Briefly, we used Burrows-Wheeler Aligner to align reads[41], GATK[40] to remove PCR duplicates and perform indel realignment and UnifiedGenotyper in GATK[40] to jointly call the genotypes in the sequenced pedigree members and 136 exomes from the 1000 Genomes Project[42]. The 136 exomes used as controls include individuals with western European ancestry (CEU) and British in England and Scotland (GBR). Potential disease-causing mutations were validated with Sanger sequencing at the University of Utah sequencing core. For the cardiac septal defect pedigree, Complete Genomics performed whole-genome sequencing and variant calling on selected pedigree members.

For the results presented the sections on cardiac septal defects, Miller Syndrome and challenging situations in pedigree studies, we used the control genome set consisting of 1,057 exomes from 1000 Genomes Project phase I data[43], 54 genomes from the Complete Genomics Diversity Panel[44], 184 Danish exomes[45] and nine nonduplicative genomes from the 10Gen data set[46], representing a wide variety of ethnicities and sequencing platforms. To include a wider set of variants that are unlikely to be causal for rare Mendelian

diseases, we further collected high-quality variants (defined as polymorphism sites with sample sizes no smaller than 100 chromosomes) from dbSNP build 137 (http://www.ncbi.nlm.nih.gov/SNP/) and NHLBI exome sequencing data (http://evs.gs.washington.edu/EVS). We then randomly inserted these variants into the control genome set, setting the minor allele frequency equal to the reported value.

Secondary analysis studies were approved by the Western Institutional Review Board for the cardiac septal defects and Miller's syndrome (dbGAP phs000244.v1.p1) pedigrees after initial studies were approved by local institutional review boards at sites interacting with participants. Procedures followed were in accordance with institutional and national ethical standards of human experimentation. Proper informed consent was obtained.

pVAAST runtime. pVAAST supports multithreading parallelization. The computational time is proportional to the size of pedigree and to the rounds of randomization tests being performed. On our Linux server with Intel Xeon 2.00 GHz CPUs, the enteropathy pedigree took 0.6 h (clock time) using 40 threads (1×10^8 maximum randomizations). The cardiac septal defect pedigree took 181 h (clock time) using 40 threads (maximum randomizations: 1×10^9). The Miller's syndrome pedigree took 0.3 h (clock time) using 70 threads (maximum randomizations: 1×10^6).

Software access. pVAAST is available for download at http://www.yandell-lab.org/software/vaast.html with an academic user license. The source code for pVAAST is included as **Supplementary Software**.

37. Elston, R.C. & Stewart, J. A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542 (1971).
38. Madsen, B.E. & Browning, S.R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
39. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
40. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
42. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
43. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
44. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
45. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
46. Reese, M.G. *et al.* A standard variation file format for human genome sequences. *Genome Biol.* **11**, R88 (2010).

doi:10.1038/nbt.2895