Published in final edited form as:

Nat Genet. 2016 December ; 48(12): 1455–1461. doi:10.1038/ng.3697.

Genomewide significance testing of variation from single case exomes

Amy B. Wilfert¹, Katherine R. Chao², Madhurima Kaushal³, Sanjay Jain^{3,5}, Sebastian Zöllner⁴, David R. Adams², and Donald F. Conrad^{1,5}

¹Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA

²NIH Undiagnosed Diseases Program, NIH, Bethesda, Maryland, USA

³Department of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri, USA

⁴Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA

⁵Department of Pathology & Immunology, Washington University School of Medicine, St. Louis, Missouri, USA

Abstract

Standard techniques from genetic epidemiology are ill-suited to formally assess the significance of variants identified from a single case. We developed a statistical inference framework for identifying unusual functional variation from a single genome, what we refer to as the “n-of-one” problem. Using this approach we assess our ability to identify the causal genotypes in over 5 million simulated cases of Mendelian disease, identifying 39% of disease genotypes as the most damaging unit in a typical exome background. We apply our approach to 129 n-of-one families from the Undiagnosed Diseases Program, nominating 60% of 30 disease genes determined to be diagnostic by a standard clinical workup. Our method can currently produce well calibrated p-values when applied to single genomes, can facilitate integration of multiple data types for n-of-one analyses, and, with further work, could become a widely used epidemiological method like linkage analysis or GWAS.

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

Corresponding Author: Dr. Donald F. Conrad, Department of Genetics, Washington University School of Medicine, Campus Box 8232, St. Louis MO, 63110, USA, dconrad@genetics.wustl.edu.

URLs.

Software and data for implementing PSAP are available from <http://github.com/awilfert/PSAP-pipeline>; ExAC, <http://exac.broadinstitute.org>; CADD, <http://cadd.gs.washington.edu>; Human Gene Mutation Database (HGMD), <http://www.hgmd.cf.ac.uk>; ANNOVAR, <http://annovar.openbioinformatics.org/en/latest/>.

Author Contributions

D.F.C. designed the study. S.Z. provided helpful conceptual guidance on modeling and interpretation. D.F.C. wrote the simulation code. A. B. W. developed the PSAP pipeline, performed spike-in analyses, and evaluated the impact of population structure on PSAP values. D.F.C., A.B.W., D.R.A., and K.C. performed the UDP analyses. M.K. and S.J. contributed CAKUT samples and data. A.B.W. and D.F.C. wrote the manuscript with input from all authors.

Competing Financial Interests

D.F.C. is funded by a research contract with PierianDx to develop novel methods for clinical exome analysis.

Introduction

The utility of exome sequencing as a diagnostic tool relies heavily on our ability to distinguish rare benign variants from disease causing mutations in small sample sizes. Recently, filtering based analysis frameworks have been developed to generate short lists of candidate disease variants based on *ad hoc* criteria such as inheritance patterns, predicted pathogenicity, and population allele frequencies¹. This approach has proven to be a reliable alternative when studies are underpowered for more statistically rigorous association or linkage analyses but is marred with uncertainty as there is no formal way to assess the relative interest of candidates. Identification of the “best” candidate is subjective and biased toward variants and genes that are well characterized, resulting in a 20–30% diagnostic rate from exome sequencing, despite the abundance of databases available to interpret sequencing data^{2,3}.

Although these databases have provided a more comprehensive picture of observed genetic variation, it has become more difficult to differentiate newly arisen neutral alleles from disease alleles under purifying selection. For instance, the Exome Aggregation Consortium (ExAC) has recently released the largest publicly available set of allele frequency data derived from human exome sequencing. This database is comprised of over 61,000 individuals, trumping other published resources in size by ten-fold, and represents six demographic backgrounds from around the globe. Despite its size, very conservative calculations based on standard coalescent theory predict that 19–21% of individuals have benign exome variants that have yet to be reported in the ExAC database, suggesting that the filtering approach will still be limited by imprecision and uncertainty⁴ (Supplementary Figure 1). In response, efforts have been made to establish more formal guidelines for implicating sequencing variants in disease when assessing only a small number of patients⁵. Much of this work has highlighted the need for statistical methods that would allow for an unbiased and rigorous assessment of the variant in question that does not rely on large sample sizes and can be applied in a high-throughput manner⁶.

Here we describe a model-based framework to assess the significance of genotypes ascertained from a nuclear family or even a single case of Mendelian disease, often referred to as the “n-of-one” problem. In order to make inferences on never-before-seen genotypes, we use a single-sided testing framework that determines the probability of sampling a genotype or set of genotypes from the unaffected population. To do this we have parameterized gene-specific null models based on the pathogenicity scores and frequencies of variants observed in the unaffected population. Critically, the release of the ExAC dataset provides the first opportunity to accurately parameterize gene-specific population genetic models, as prior databases were too sparse to estimate the rare end of the site frequency spectrum with precision. Conceptually our approach is similar to recently described approaches for studying *de novo* mutations, which utilize large databases of *de novo* mutations to assess the significance of observed *de novo* mutations in cancer and constitutional diseases.^{7,8} Our method is a generalization that makes no assumptions about the age of the variants under study, explicitly accommodates pathogenicity scores and can be applied to arbitrary disease models, including the *de novo* case.

We use this framework to formally study the n-of-one problem across a variety of known Mendelian diseases, and ask, how well can we identify the causal variant from a single patient genome? What are reasonable thresholds for genomewide significance for an n-of-one study? We address possible limitations to our approach and investigate extensions to multiple types of genetic variants and other data types. Finally, we apply our approach to analyze real n-of-one cases, including 129 nuclear families from the NIH's Undiagnosed Diseases Program (UDP).² The UDP is essentially a “mystery disease” clinic run at the NIH's Clinical Center, whose mission is to provide a genetic diagnosis for individuals with severe diseases that have remained undiagnosed despite a prior exhaustive clinical workup. This cohort of samples represents a sizeable number of n-of-one cases of ultra-rare disease.

Results

Intuition for the approach

Given a sequencing dataset from a single case of rare disease, we want to perform genome-wide tests of the null hypothesis that the variation in the patient was drawn from a “normal” population (Fig. 1A–C). Furthermore, we require the following features in our testing framework: (a) the predicted functional consequences of each variant are taken into account; (b) the method is *gene-based*, capable of jointly assessing all the variation in a gene at once, and (c) the method is *genetic*, meaning that it would consider the ploidy of the variants in a gene.

To apply our method to a patient genome sequence, we construct three test statistics for each gene in the genome, denoted X_1 , X_2 and X_3 . We omit a gene-specific subscript for clarity. These tests correspond to three primary disease models for Mendelian disorders, respectively: autosomal dominant, single variant autosomal recessive (for simplicity, we refer to this as the “autosomal recessive” model henceforth), and compound heterozygote autosomal recessive (CHET). If c_i is the pathogenicity score associated with variant i , then we calculate X_1 as the maximum c_i over all sites in the gene that are heterozygous in the patient; X_2 as the maximum c_i over all homozygous (non-reference) genotypes; and X_3 as the second largest c_i over all heterozygous sites. In order to evaluate the significance of the test statistics, we compare each one to a suitable null distribution that is parameterized from a large database of exome data from “normal” individuals (**Methods**, Fig. 1C). Specifically, we assess the probability of drawing a test statistic as large or larger than what we observe in the case. We refer to this probability as the Population Sampling Probability (PSAP), which represents a p-value based on a null distribution built from healthy individuals. These p-values can be used to immediately prioritize genes or compare the support for different disease models within the context of a genetic analysis, and they can be combined with other forms of gene-based data to provide a more comprehensive workup of a patient or family. Below, we explore the idea of genomewide significance when evaluating PSAP p-values from the genome data of a single patient.

In the following analyses, we use the Combined Annotation Dependent Depletion (CADD) score as our metric for functional consequence (c_i)⁹, although comparisons of results based on different functional prediction methods and suggests that other schemes can produce similar results (Supplementary Note, Supplementary Figure 2).

PSAP calibration and impact of population structure

To ensure our gene-specific null models were well calibrated, we compared our fitted models to empirical data from healthy exomes from three control cohorts. We found that our null models produce reasonably well calibrated p-values when used to analyze these population controls, especially in the extreme right tail of the distribution (Fig. 2, Supplementary Figs. 3–5). In each combination of disease model and cohort, over 97% of PSAP p-values fall within a factor of 2 of the expected 95% confidence intervals for a uniform [0,1] distribution. The most poorly calibrated distribution was the autosomal recessive model for Swedish population controls – here the maximum deviation from the expected p-values reached one order of magnitude in the range of 10^{-4} . This could be due to the quality of the sequence data, as the sequencing and variant calling were from a very early exome sequencing study, or, perhaps, population structure.

Previous studies have reported differences in the frequency of deleterious variation among human populations¹⁰. This raises the concern that accurate calibration of PSAP p-values for an individual may be contingent on that individual's genetic ancestry, as our null models are largely derived from individuals of Caucasian ancestry. We constructed qq-plots from PSAP analysis of 189 African American exomes and observed an inflation of the resulting p-values across multiple disease models (Fig. 2A). This inflation affects many sites within each AA exome, suggesting that PSAP p-values may be amenable to the same type of ancestry correction as p-values derived from a standard GWAS. Spike-in analyses show that real disease mutations will still rank highly within AA genomes using PSAP, but the overall trend of inflation seen in the control analyses means that caution must be used in the context of population structure, especially when interpreting the absolute magnitude of PSAP values. (Supplementary Figure 6).

Encouraged that we were able to parameterize realistic, gene-specific null distributions for CADD, we selected as a test case the well-studied gene *PTEN*, which encodes a tumor suppressor implicated in numerous somatic and constitutional diseases. We found that when using a genome-wide empirical null distribution for CADD, we assessed the significance of a typical heterozygous *PTEN* mutation at $p < 0.01$; however, when using a gene-specific null model, we assessed the majority of *PTEN* mutations as $p < 10^{-6}$, which is the maximum precision of our simulations (Fig. 3A). These results highlight one important advantage of the PSAP approach-PSAP provides a standardized way of assessing the significance of pathogenicity scores when there is heterogeneity in calibration of these scores among genes (Supplementary Figs. 7–8).

Benchmarking our ability to address the n-of-1 problem

To benchmark the performance of our population-based approach across all genes, we simulated over 30 million cases of the n-of-one problem using nearly 92,000 known Mendelian disease mutations in the Human Gene Mutation Database (HGMD)¹¹ and real exome data from 189 putatively healthy individuals (**Methods**). Similar experiments using disease-causing variants present in ClinVar¹², a clinically curated disease mutation database, indicate that the source of disease mutation information is not a cause of bias for this analysis (Supplementary Note, Supplementary Figure 9). First, we evaluated the utility of

the PSAP as a prioritization tool by rank ordering all genotypes in each simulated case exome and recording the rank of the known disease genotype(s) in each simulated case. A useful prioritization tool should reliably assign a disease genotype a high rank (e.g. <10) within the background of benign variation. As a well-studied point of reference, we also assessed the utility of CADD alone in this analysis. In order to present a fair comparison, we perform the same gene-based calculations for CADD as for PSAP, using the test statistics X1, X2 and X3 as the priority score, instead of the PSAP p-values corresponding to these test statistics.

PSAP performs as well as or better than CADD across all three genetic models that we considered, although there are marked differences in our ability to rank known disease genes (Fig. 4A, Supplementary Figure 10, Table 1). In many clinical settings, reliably placing the causal gene into the top 10 most damaged genes within the genome is very useful. PSAP is highly successful at the top-10 criteria: PSAP ranked 35% of 28,498 autosomal dominant genotypes, 89% of 25,567 autosomal recessive genotypes and 86% of 25,567 compound heterozygote genotypes (Table 1). Very interestingly, the autosomal dominant model performed markedly worse than the autosomal recessive and CHET models, despite roughly equivalent calibration. This could be a sign of sequencing and/or informatics errors creating a background of low p-value heterozygous genotypes in our test data.

When evaluating the more rigorous task of correctly identifying the disease gene as the most likely candidate in the entire genome, PSAP performs remarkably well for the recessive models-ranking 78% of autosomal recessive genotypes and 71% of compound heterozygous diplotypes as the most unusual in the genome. The autosomal dominant model remains a challenge with 12% of genotypes ranking as the most unusual in the genome on average. When filtering on variants with an allele frequency less than 1% in ExAC, PSAP performs better than CADD and comparably to PSAP without allele frequency filter; however the allele frequency filter helps improve the rank of disease SNVs that are difficult to predict (Fig. 4A, Supplementary Figure 10).

In order to better understand the features of PSAP that influence sensitivity, we studied the relationship between PSAP and features of the ExAC database. We found that the autosomal dominant model was especially sensitive to the gene-specific singleton rate, suggesting that sensitivity and specificity of PSAP with the autosomal dominant model will be influenced by sequencing error and mis-estimation of the singleton rate (Fig. 3B). Although the autosomal recessive and CHET models perform better overall when compared to autosomal dominant, they too are underpowered to detect mutations in a small number of genes. Closer inspection of ExAC variants in these poor performing genes revealed that many had high frequency, high CADD variants. This highlights an important limitation of the PSAP approach-the smallest possible PSAP for a gene is constrained by the CADD scores in the database used to parameterize the PSAP null; thus it will be difficult to detect true disease mutations in a genic context with other apparently damaging mutations at high frequency (Fig. 3C).

Over 15% of disease variants in HGMD are indels. We developed a way to extend CADD, and by result, PSAP, to annotate frameshift indels and large deletions by mapping the

functional consequences of deletions to the same space as SNVs (**Methods**, Supplementary Figure 11). We used PSAP values to prioritize 6,326 frameshift indels from HGMD amongst the background of all SNVs and short indels identified in a typical exome. As a class, these HGMD indels were prioritized higher than HGMD SNVs, with 1,574/2,900 (54%) of heterozygous genotypes in autosomal dominant genes recognized as the highest priority (rank = 1), 99.7% of homozygous genotypes in autosomal recessive genes, and 100% of compound heterozygous diplotypes in autosomal recessive genes (Supplementary Figure 12).

Thresholds for genomewide significance

A routine benchmark of pathogenicity prediction algorithms is to assess the ability of the scoring algorithm to discriminate between disease variants and non-disease variants. A more ambitious analysis is to discriminate healthy individuals from cases of Mendelian disease based strictly on whole-genome or whole-exome genetic data. Here we set out to determine if thresholds of genomewide significance for PSAP could be developed to this end. We created a simple classifier, PSAP-MIN, which summarizes the complete set of genotypes in a genome as the smallest PSAP value in that genome. We simulated 1,500 Mendelian disease exomes, as discussed above, and by comparison to 1,500 exomes of controls, constructed ROC curves for PSAP-MIN for each of our three disease models (Fig. 4B, Table 2, Supplementary Table 1). To our surprise, PSAP-MIN performed reasonably well at the task of classifying individuals for both the single variant and compound-het recessive disease models (AUCs of 0.81 and 0.76, respectively). On the other hand, identifying individuals under the dominant disease model was far more challenging with an AUC of 0.11. This approach is truly hypothesis free; we reasoned that classification could be more accurate if we restricted calculation of PSAP-MIN to subsets of the genome that are *a priori* more likely to harbor disease causing mutations. We found this to be the case, with greatly improved results when limiting analysis to known Mendelian disease genes (AUC=0.36–0.98) or genes with strong selective constraint (RVIS<0, AUC=0.16–0.90). For an approach based on PSAP-MIN to be truly useful for unsupervised diagnosis of Mendelian disease in a clinical setting, the false positive rate of the classifier will likely need to be less than 1%. In Table 2 and Supplementary Table 2, we report the frequency of individuals detected with various PSAP-MIN thresholds while analyzing control populations – these frequencies can be interpreted as conservative estimates of false positive rates. When used in combination with the results from the HGMD spike-in experiments (Supplementary Table 1), it is possible to establish thresholds for declaring genomewide significance using PSAP that corresponds to a specified sensitivity and specificity.

The motivating n=1 case

Next, we evaluated the utility of PSAP for identifying diagnostic variants in real n=1 cases. In a previous study of men with idiopathic non-obstructive azoospermia¹³, we identified a case with an extremely rare condition known as uniparental isodisomy of chromosome 2 (Fig. 5A). Fewer than 10 cases of this cytogenetic anomaly had ever been reported, all of which were ascertained through studies of Mendelian disease¹⁴. We performed exome sequencing on this case and applied the PSAP approach to the resulting genotype calls (Fig. 5B). Strikingly, the 6 smallest PSAP p-values corresponded to genes on chromosome 2, and

the smallest PSAP was the result of a homozygous M360T substitution in the gene Inhibin β B (*INHBB*), a biomarker of Sertoli cell function sometimes used in the clinical management of male infertility¹⁵.

UDP analyses

To explore the performance of our approach on a broader cohort of clinical exomes we applied PSAP to 129 families from the NIH Undiagnosed Diseases Program (**Methods**). Each UDP case represents a real-life n-of-one problem, ascertained as having a suspected genetic disorder with no previous diagnosis despite extensive clinical workup. Thirty of the families had received a genetic diagnosis from the UDP (Supplementary Table 3). For each disease model, we identified PSAP p-value thresholds corresponding to a false positive rate of 5% when applied to population controls (1×10^{-5} for autosomal dominant and autosomal recessive models, 2.25×10^{-5} for CHET model). We observed an excess of cases with at least one PSAP p-value less than the 5% false positive thresholds for each disease model (Fig. 5C). Moreover, for the autosomal recessive disease models we found an excess of HGMD genes with PSAP p-values $< 1 \times 10^{-5}$ in unaffected relatives, which could be a sign of incomplete penetrance of disease-causing mutations in some unaffecteds ($p < 0.05$ Fisher exact test, Fig. 5C).

The distribution of extreme PSAP p-values across UDP individuals could not be explained by geographic ancestry (Supplementary Figure 13). We first evaluated our ability to identify the diagnostic genotypes in the hardest situation: analyzing the case genomes in the absence of family data. Of our cases with a genetic diagnosis from UDP, 8 presented the diagnostic gene as the smallest PSAP in the genome, and 19 in the top 10 (median rank 6, max 45). By comparison, only 1 diagnostic gene was top-ranked by CADD alone and 12 in the top 10 (median rank 18, max 46).

Next, we took advantage of the family data to perform more aggressive filtering, removing genotypes inconsistent with each disease model under evaluation. As a result, 18/30 (60%) presented the diagnostic gene as the smallest PSAP in the genome, and 29/30 (97%) were in the top 10. We manually reviewed all genes with PSAP values $< 10^{-3}$ in the 99 undiagnosed families. Sixty-two of the families had at least one gene to review, with an average of 1.98 genes to review per family. We identified variants from 14 families that, although not considered diagnostic at the moment, were deemed strong candidates for being causal mutations based on the gold-standard clinical curation of the same exomes by a team of experts at the UDP. The identification of additional families harboring mutations in these genes and/or extensive experimentation, are required to elevate these candidate mutations to the level of diagnostic status.

The Undiagnosed Diseases Program prioritizes DNA sequence variants using a pipeline with several components (**Methods**), typically yielding 0 – 6 variants of high interest per family. PSAP p-values were used to assess 130 genotypes or genotype pairs that had been prioritized at some point by the UDP. PSAP p-values were generated for the autosomal dominant, autosomal recessive, and CHET models. PSAP p-values of 0.01 or greater were marked as “uninteresting” (low pathogenicity potential). Seventeen of the 130 genes met this criterion. Of the 17 genes, 11 had eventually been deprioritized by the UDP based on new

information not available at the time of initial prioritization, e.g. the ExAC database. The remaining 6 genes remained in a prioritized status. The prioritized variants in two of these genes, *HAS1* and *CCAR1* had been the subjects of extensive research investment, including model organism generation for the latter. No clear association with human disease was established for either gene. Utilization of the PSAP metric would have deprioritized these variants during initial assessment.

CAKUT Analyses

As further validation of the approach, we applied our method to a set of 122 exomes generated from a cohort of children born with severe congenital anomalies of the kidney and urinary tract (CAKUT). As a control cohort, we processed and analyzed 463 Caucasian control exomes from the Genotype Tissue Expression (GTEx) Consortium.

It is estimated that 30% of cases of severe CAKUT have a recognized monogenic genetic basis. As in the UDP cohort, we found an excess of affected individuals with PSAP p-values when using significance threshold corresponding to a 5% false positive rate for each disease model (Supplementary Figure 14). The autosomal dominant model showed an 8% excess over the null expectation (i.e. 13% versus 5%, $p < 0.05$), the autosomal recessive model a 4.5% excess (p-value n.s.), and the CHET model a 5.5% excess (p-value n.s.). If one interprets this excess of low PSAP p-values as the fraction of cases with at least one CAKUT risk mutation, a total of 18% of the affected individuals may have genetic causes identifiable by PSAP p-value analysis.

Extension to integrative analysis

Using the gene-based analysis framework described here naturally lends itself to integrating PSAP results with multiple data types such as RNAseq, protein-protein interaction data, model organism and cell-based assays. We used GTEx gene expression data to summarize the likelihood that each gene in the genome plays an essential role in the tissue(s) affected by disease in each case (**Methods**). Intuitively, a gene with no observable expression in a tissue is less likely to have a role in the pathobiology of that tissue. Encouragingly, by combining PSAP with GTEx expression summaries we were able to further improve the identification of diagnostic genes, moving the mean rank from 2.4 to 1.9, and the maximum rank from 14 to 5. (Fig. 6A, Supplementary Table 4).

Discussion

A critical challenge for medical genetics is the identification of actionable genetic variants from personal genome data. We have described a framework that allows calculation of well-calibrated population sampling probabilities for genotypes identified in a single genome. The most practical result of our approach is to provide a simple and interpretable way to describe how unusual a genotype is in light of a very large population database such as ExAC. While our original focus was rare monogenic disease, our approach also shows promise for identifying monogenic or oligogenic causes of more complex diseases in the context of cohort studies (Figs. 5A, 5B). Given further study and characterization, PSAP p-

values could become accepted epidemiological evidence for the identification of causal rare disease mutations.

The statistical analysis of rare disease genomes is one of the most active areas of human genetics today. Many groups are developing gene-based constraint scores, which quantify the ability of a gene to “tolerate” deleterious variation^{7,16,17}. A primary application of these scores is to evaluate the significance of genotypes observed in studies of disease, although to our knowledge no formal framework for doing so has yet to be described for inherited variation. The p-values that we obtain using our approach do appear moderately correlated with several popular constraint metrics (Pearson correlations ranging from -0.3 to 0.3 , Supplementary Figure 15, Supplementary Tables 5–7), and PSAPs themselves could be used as a sort of constraint metric (Supplementary Figure 11). Other groups have described probabilistic methods for identifying disease mutations that, like our approach, allow incorporation of functional annotation as weights or priors for each test^{18,19}. One representative and commonly used approach, VAAST, uses a very general likelihood ratio framework to test for association between variation and a phenotype. The likelihood ratio requires an explicit alternative hypothesis, which is problematic when dealing with single cases, as our head-to-head comparisons with VAAST indicate (Supplementary Note).

During the course of this project, we identified specific limitations of the PSAP approach that should be kept in mind during a rare disease analysis, as well workarounds for these problems (Supplementary Note). These include: sensitivity to population structure, sensitivity to sequencing error, sensitivity to idiosyncrasies in the annotation method used (e.g. CADD), and lack of phase information in CHET calculations. Beyond addressing these limitations, a number of other areas beckon for future work. The autosomal dominant model is one of the most important areas of future development for the PSAP approach, and integrating recent progress in mutation rate modeling to provide more detailed and dynamic predictions about position-specific singleton rate may improve the goodness-of-fit (Supplementary Figure 16, Supplementary Note)^{7,20}. PSAP calculations for non-coding regions will soon be feasible based on the rapid increase in published whole genome sequences. Using the gene-based analysis framework described here naturally lends itself to integrating multiple data types during an n-of-one analysis, such as RNAseq, pathway analysis protein-protein interaction data, model organism and cell-based assays (Fig. 6B).

Each diagnosis made by the UDP represents the culmination of a week of clinical workup and many hours of data analysis and discussion by an interdisciplinary team of specialists. After careful validation and in-depth characterization of the PSAP approach on thousands of cases of Mendelian disease with known diagnoses, as well as thousands of controls, it may be possible to establish thresholds of genomewide significance that will allow at least some of the burden of genetic disease diagnosis to be shifted to purely computational assessment of patient phenotypes and genetic variants.

Online Methods

Datasets used for simulation and analysis

A dataset with all potential alternative alleles within the gene coding regions of hg19 Gencode version 19 (GencodeV19) was generated to represent the total search space for all variation within gene coding regions.

The core model used for PSAP calculations was parameterized using ExAC release v0.2 from the Exome Aggregation Consortium (Cambridge, MA). We extracted all sites from the ExAC database that both passed the GATK tranche filter and were represented by at least one individual with a read depth greater than 20 ($AC_Adj > 0$). Sites with more than one alternative allele were parsed onto separate lines, and the remaining data was annotated with ANNOVAR. The average coverage of each gene included in the ExAC data was calculated by averaging the mean number of reads at each base pair in the gene. Genes with an average coverage that was less than 10 were added to a low coverage gene list.

We used the Human Gene Mutation Database (HGMD; release: April 2013) as our source of disease mutations and extracted all variants that were tagged as “DM” for the analyses described here. We annotated each gene in HGMD with a putative disease inheritance model (e.g. “Autosomal dominant”, “Autosomal recessive”, etc.) using a manually curated version of OMIM²¹. Some genes were associated with multiple disease inheritance models, and those genes without a known inheritance model were excluded from further analysis.

Exomes and demographic data from the Women’s Health Initiative Sequencing Project (WHISP) were downloaded from the NCBI Database of Genotypes and Phenotypes (dbGaP accession phs000200.v10.p3.c1 and phs000200.v10.p3.c2)²². Genotypes were recalled, jointly, from 450 WHI BAM files using Haplotype Caller and cleaned according to GATK best practices using GATK-3.2.2. Principal component analysis was performed on the recalled exomes to validate the self-reported demography of the samples. Control individuals from the European American and African American populations in this dataset were used in analyses assessing model performance in different demographic backgrounds.

Families collected through the Undiagnosed Diseases Program (UDP) were exome sequenced and called for variants by the NIH. Variants that did not pass the tranche filter were removed and highly variable genes shared across the families were compiled into a list, referred to as the blacklist. PCA was performed to evaluate the demographic background of each of the families.

We acquired specimens from 122 unrelated patients seen at the adult, pediatric nephrology, pathology and urology clinics at the Washington University School of Medicine between 2007 and 2013. All patients consented to the study approved by the Institutional Review Board. The CAKUT phenotypes in the cohort reported here included renal agenesis, duplication, hypoplasia or dysplasia. Over 95% of patient families self-reported as Caucasian. As a control population, we used genotype calls from the exome sequences of 463 Caucasian individuals generated by the GTEx consortium (dbGAP project

phs000424). The genotype calls are available through dbGAP as GTEx data release version 6, 2015-01-12.

Data annotations

The publicly available program ANNOVAR (Version date: July 14, 2014) was used to annotate all datasets with gencodeV19 genes and functional consequences, allele frequencies from NHLBI Exome Sequencing Project 6500 (ESP6500), the 1000 Genomes Project phase 3 (1000GP), and the Exome Aggregation Consortium release 0.2 (ExAC) were annotated, and SNV pathogenicity was estimated by CADD scores (both raw and phred-scaled) and predictions available through dbNSFPv2.6^{9,23–25}. The ExAC allele frequency annotation table used by ANNOVAR was generated in-house, as described in the data information section above. All other tables were previously generated by ANNOVAR and downloaded for use.

PSAP calculations

The population sampling probability (PSAP) is a general way of describing the probability of sampling a set of variants from a given region of a test genome, conditional on their pathogenicity scores. More formally, we construct a test statistic from a set of pathogenicity scores, and then evaluate the probability of sampling that test statistic under the null distribution that the subject being assessed came from a “healthy” population. In the present work we define three test statistics, X1, X2 and X3, which are meant to evaluate three different popular disease models. X1, X2 and X3 are calculated for each gene in the genome but, for clarity, we omit a gene specific subscript in the following description. Define c_i as the pathogenicity score associated with variant i in the gene of interest. X1 is the test statistic for the dominant disease model and is defined as the maximum c_i at heterozygous sites; X2, for the monoallelic recessive model, the maximum c_i at homozygous sites; and X3, for the compound heterozygous recessive model, the second largest c_i at heterozygous sites. Note that this modeling framework is generic and could be used to evaluate an arbitrary test statistic other than X1, X2 or X3.

We derive our null distributions for X1, X2 and X3 using simulation. In order to perform these simulations, we first parameterize a simple population genetic model using a large reference database. Simulations are done separately for each gene in the genome. For each gene, we define a set of k nucleotides that span the coding region and splice sites of that gene, and we define a “variant” as any of the possible $3k$ departures from the reference genome sequence in our gene. We define p_i as the expected frequency of variant i in the healthy population, and c_i as the score associated with variant i . If variant i has been previously observed in the reference database (in our case, ExAC) then p_i is simply the observed frequency. If variant i has not been observed then we estimate p_i as the gene-specific per-nucleotide rate of singletons observed in the reference database. Explicitly, let n_l be the number of singleton loci in the gene, and N the total number of individuals in the database, then we set $p_i = n_l/3Nk$. These singleton rates are estimated separately for homozygous and heterozygous genotypes. If no singletons are observed for a gene then we estimate the singleton rate as the average singleton rate for the entire genome. We assume

that the probability of drawing x copies of variant i is binomially distributed, that is,

$$\Pr(X=x) = \binom{2}{x} p_i^x (1-p_i)^{2-x}.$$

Using this simple model, we simulated 1 million sets of genotypes for the k bases in each gene, representing the unphased sequences of 1 million simulated diploid individuals. For each simulated individual, we record 3 different gene-specific summaries of the functional impact of their variation, corresponding to 3 different disease models. These can be considered the test statistics that are being evaluated under each null model. For the dominant disease models we record the maximum c_i at heterozygous sites; for the monoallelic recessive model, the maximum c_i at homozygous sites; and for the compound heterozygous recessive model, second largest c_i at heterozygous sites. The result of this procedure is a set of 3 vectors of 1 million CADD scores for a given gene, each of which defines a null model for that gene.

Inspection of empirical p-values from the CHET test statistic calculated from the GTEx donors showed a modest but consistent inflation of low p-values compared to the null expectation. This is due to the fact that our CHET p-values are calculated under the assumption of complete independence between all pairs of heterozygous sites in a gene; thus, any linkage disequilibrium between sites will create an excess of CHET pairs above the null expectation. To address this, we simply recalibrated the CHET test statistic empirically. We fit a linear model, regressing the observed p-values onto the expected, and used the resulting model to provide recalibrated p-values (Supplementary Figure 5).

Extension to indels and CNVs

We developed a way to extend CADD, and by result, PSAP, to annotate frameshift indels and large deletions by mapping the functional consequences of deletions to the same space as SNVs. We reasoned that protein-truncating indels and CNVs should have the same functional consequences as protein-truncating SNVs, and created a gene-specific lookup table of average CADD scores for all predicted protein-truncating variants. While we did not perform a formal analysis of PSAP as a metric for prioritizing CNVs, we did observe a moderate correlation between PSAP values for heterozygous protein truncating genotypes and a popular metric for CNV pathogenicity assessment known as P(HI)¹⁷ (Spearman correlation=0.37, Supplementary Figure 11).

Model quality control

Unaffected and unrelated individuals from the WHI were used to determine the observed distribution of CADD scores across a population of 434 healthy individuals. Specifically, each gene represented in gencodeV19 was assessed for variation in each individual. If the gene was variable in the individual, summary scores were calculated, as described in the population model, for all appropriate disease models (dominant, homozygous recessive, and compound heterozygous recessive) based on the genotypes of the variation. This observed data was compared to the data simulated from the population model according to the gene and disease model to evaluate the fit of the population model, and the accuracy and reliability of the simulated null distributions.

In the analysis of real data, we perform several QC steps prior to calculating a final set of population sampling probabilities. We filter all variants from genes with average per-base coverage of 10 reads per sample in ExAC, and we also filter all variants corresponding to sites with an average per-sample read depth less than 20 in ExAC. The purpose of these steps is to align the “callable” regions of both the ExAC database and the test dataset. Finally, based on empirical analysis of each cohort to be analyzed with PSAP, we identified genes with poorly calibrated null models that produce an excess of small PSAPs. We refer to these gene sets as “blacklists”. We defined a blacklist for each cohort, consisting of 100–300 genes each. It is important to emphasize that definition of a blacklist may be highly contingent on the specifics of a production pipeline and thus, a blacklist should not be assumed to be portable among projects (Supplementary Figure 17).

Simulating disease exomes (spike-ins)

In order to evaluate the sensitivity and specificity of PSAP p-values as a metric for identifying disease mutations from real exome data, we simulated over 1 million exomes with Mendelian disease mutations. Each simulated exome was generated by placing one or two Mendelian disease genotypes into an exome genotype dataset (the “background”) obtained from a healthy control. Known disease variants were obtained from HGMD (n=30,222 for autosomal dominant variants, n=27,110 for both autosomal recessive and CHET variants), while the exome sequencing data from 189 healthy individuals from the were used to obtain the exome genotype “background” for each simulated case. Variants from HGMD and WHI that did not reside in a splice site or exon of the gene, or had a missing allele frequency in the ExAC data but were reported as having common population frequencies, above 1%, in the ESP6500 or 1000GP populations were removed. Variants in blacklisted genes were also removed.

Analysis of UDP exomes

The Undiagnosed Diseases Program prioritizes DNA sequence variants using a pipeline with several components. Variants are initially filtered in an unattended manner using familial variant segregation and matched population frequency criteria. Remaining variants are subjected to direct inspection of BAM-file alignments. Finally, variants are prioritized based on predicted pathogenicity, phenotype and pathway associations (Exomiser) and known biological characteristics of the associated gene. The final assessment is carried out by graduate and medically-trained staff. Details of the diagnostic genes are disclosed in Supplementary Table 3.

Families from the UDP were reanalyzed using the PSAP pipeline to identify candidate causal disease mutations in affected individuals. Poorly calibrated genes were identified for each disease model and compiled into a blacklist (n=102 for autosomal dominant, n=167 for compound het, and n=50 for autosomal recessive). Variants in each family that were in genes on the blacklist, genes on the low coverage list, did not reside in a splice site or exon of the gene, had a missing allele frequency in the ExAC data but were reported as having common population frequencies, above 1% in the ESP6500 or 1000GP populations, or that failed Mendelian consistency checks were removed. Population sampling probabilities were calculated for each family member. According to the disease model, the variant with the

smallest population sampling probability from each family was intersected with a list of known dominant or recessive HGMD genes, and the candidate from each family was reviewed by an expert.

Integration of gene expression data with PSAP

One of the strengths of the PSAP statistic, is that, as a gene-based p-value, it can be easily combined with other gene-based measurements that may be informative about the functional relevance of a gene. We wished to demonstrate this in the current analysis by combining PSAPs with a summary statistic derived from gene expression data. We obtained version 6 of the GTEx data, which consist of RNAseq measurements on 53 tissues taken from a total of 451 donors (although note, not all tissues have been assessed in all donors). Expression levels for 20,657 genes were summarized as Transcripts per million reads mapped (TPM) using RSEM for each RNAseq library, counting only uniquely mapping reads. We only use data from the tissues with at least 25 RNAseq libraries available, and we also excluded the two cell lines being sequenced by GTEx. This left 46 tissues which we use for the following analyses.

While conceptually straightforward, there has been little work done on the optimal way of summarizing gene expression data from a control cohort, for the purposes of Mendelian disease gene identification. We explored a small number of approaches and present results from one approach in the main text. We reasoned that three summaries of expression may be informative about the importance of gene i to the normal functioning of tissue j :

The average expression level of that gene in tissue j ,

$$x_{ij} = \frac{\sum_{k=1}^n x_{ijk}}{n}$$

The coefficient of variance of the that gene in tissue j ,

$$cv_{ij} = \frac{x_{ij}}{\sigma_{ij}^2}$$

and the specificity of that gene's expression to tissue j ,

$$si_{ij} = \frac{x_{ij}}{\sum_{j=1}^t x_{ij}}$$

In order to combine these statistics with the PSAP probabilities, we need a way to obtain a p-value associated with each one. We use rank-based p-values for this initial analysis, that is to say, if the rank of x_{ij} is r_{ij} out of k genes measured, then the p-value for that observation is r_{ij}/k (for the mean and specificity index) or $1 - r_{ij}/k$ (for the coefficient of variance). Further, we used two methods for calculating ranks: version 1, we rank all r_{ij} from all tissues in a

single “cross-tissue” vector; and version 2, we only calculate ranks by comparing r_{ij} from within the same tissue. The potential benefit of version 1 is that it allows estimation of smaller p-values, with the smallest p-value being on the order of 1×10^{-6} (12000 genes $\times$ 46 tissues = 550k and $1/550k = 1.8 \times 10^{-6}$) and thus potential could improve the power of this approach for identifying genes with truly exceptional properties. The benefit of version 2 is that, by “normalizing” within tissues, this estimate of significance is more robust to large variation among tissues (for instance, if some tissue is particularly odd, and has a very high preponderance of highly expression genes compared to all other tissues in the body, calculating a global rank across all tissues could unduly mask an otherwise interesting tissue-specific observation).

We use these empirical p-values as “priors” on the extent to which each gene is essential for the function of each tissue in the GTEx dataset. While this may seem like an *ad hoc* modeling choice, there are numerous reports that indicate that Mendelian disease genes often have higher mean expression levels in or specificity to affected tissues.^{26,27}

As discussed in the main text, we use these expression priors to re-evaluate the 30 diagnosed cases from the UDP, in hopes that integration of expression information would upweight tissue-relevant genes and downweight genes with low or no expression relevant to the tissue of interest. We manually curated each disease diagnosis for the 30 diagnosed UDP cases, assigning a GTEx tissue label for the “primary” affected tissue in each case. For each case, we then combine the gene expression p-values from the relevant tissue with the PSAP p-value for each gene in the genome, using Fisher’s method.

In Supplementary Table 4, we compare the overall ranking of the diagnostic variant using PSAP alone, and then the ranking that resulted from the various methods that we devised for summarizing and combining expression p-values with PSAP.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Daniel Wilson, Dr. Matthew Stephens, and three anonymous reviewers for helpful comments. We thank Dr. Ni Huang for useful discussions and for providing updated versions of some annotations used in this work, and Dr. Katinka Vigh-Conrad for assistance in preparing the figures. We thank Dr. Daniel MacArthur, Dr. Monkol Lek and the members of the ExAC Consortium for generous prepublication sharing of their data. We thank Mary Hoffmann and WU Kidney Translational Research Core (KTRC) for patient enrolment and Genome Technology Access Center (GTAC) for exome sequencing of CAKUT patients. Our work was supported by US NIH grant R01MH101810 (to D.F.C.), March of Dimes Foundation grant #6-FY14-430 (to S.J.).

References

1. Bamshad MJ, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet.* 2011; 12:745–755. [PubMed: 21946919]
2. Gahl WA, et al. The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet Med.* 2012; 14:51–59. [PubMed: 22237431]
3. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet.* 2011; 12:628–640. [PubMed: 21850043]

4. Mitchell AA, Chakravarti A, Cutler DJ. On the probability that a novel variant is a disease-causing mutation. *Genome Research*. 2005; 15:960–966. [PubMed: 15965029]
5. Richards S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015; 17:405–424. [PubMed: 25741868]
6. MacArthur DG, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014; 508:469–476. [PubMed: 24759409]
7. Samocha KE, et al. A framework for the interpretation of de novo mutation in human disease. *Nat Genet*. 2014; 46:944–950. [PubMed: 25086666]
8. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013; 499:214–218. [PubMed: 23770567]
9. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014; 46:310–315. [PubMed: 24487276]
10. Lohmueller KE. The distribution of deleterious genetic variation in human populations. *Current Opinion in Genetics & Development*. 2014; 29:139–146. [PubMed: 25461617]
11. Stenson PD, et al. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current Protocols in Bioinformatics*. 2012; 39:1.13.1–1.13.20.
12. Landrum MJ, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*. 2014; 42
13. Lopes AM, et al. Human spermatogenic failure purges deleterious mutation load from the autosomes and both sex chromosomes, including the gene DMRT1. *PLoS Genetics*. 2013; 9:e1003349. [PubMed: 23555275]
14. Kantarci S, et al. Donnai-Barrow syndrome (DBS/FOAR) in a child with a homozygous LRP2 mutation due to complete chromosome 2 paternal isodisomy. *Am J Med Genet A*. 2008; 146A: 1842–1847. [PubMed: 18553518]
15. Rey RA, et al. Male hypogonadism: an extended classification based on a developmental, endocrine physiology-based approach. *Andrology*. 2013; 1:3–16. [PubMed: 23258624]
16. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genetics*. 2013; 9
17. Huang N, Lee I, Marcotte EM, Hurles ME. Characterising and Predicting Haploinsufficiency in the Human Genome. *PLoS Genetics*. 2010; 6:e1001154. [PubMed: 20976243]
18. Yandell M, et al. A probabilistic disease-gene finder for personal genomes. *Genome Research*. 2011; 21:1529–1542. [PubMed: 21700766]
19. Lee S, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*. 2012; 91:224–237. [PubMed: 22863193]
20. Michaelson JJ, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*. 2012; 151:1431–1442. [PubMed: 23260136]
21. Blekhman R, et al. Natural selection on genes that underlie human disease susceptibility. *Current Biology: CB*. 2008; 18:883–889. DOI: 10.1016/j.cub.2008.04.074 [PubMed: 18571414]
22. Tryka KA, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic acids research*. 2014; 42:D975–979. DOI: 10.1093/nar/gkt1211 [PubMed: 24297256]
23. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*. 2010; 38:e164. [PubMed: 20601685]
24. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Human mutation*. 2013; 34:E2393–2402. DOI: 10.1002/humu.22376 [PubMed: 23843252]
25. Genomes Project C et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491:56–65. DOI: 10.1038/nature11632 [PubMed: 23128226]
26. Lage K, et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111:12345–12350. [PubMed: 24812345]

- States of America. 2008; 105:20870–20875. DOI: 10.1073/pnas.0810772105 [PubMed: 19104045]
27. Koboldt DC, et al. Exome-based mapping and variant prioritization for inherited Mendelian disorders. American journal of human genetics. 2014; 94:373–384. DOI: 10.1016/j.ajhg.2014.01.016 [PubMed: 24560519]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. Approach to the n-of-one problem

We break the n-of-one problem down into two parts: **a**, the variant annotation problem, which is to replace a generic “reference” and “alternate” allele labeling with a score that reflects the potential of each variant for damaging molecular function, and **b**, significance testing—evaluating the likelihood that a pair of haplotypes observed in the n-of-one case contains the causal genotype(s), conditional on their variant annotations. Standard approaches in genetic epidemiology, such as GWAS and linkage analysis, typically evaluate the relative likelihood of the data under a null and alternative model. In most rare disease situations, it is infeasible to define a likelihood function that reflects the likelihood of the data when sampled from the disease population (ie the alternative model). Thus, our approach is to simplify significance testing to evaluating the likelihood of the patient’s data under the null. **c**, To implement our approach, three test statistics are constructed for an annotated haplotype pair corresponding to each of three disease models: autosomal dominant, single variant recessive and compound heterozygous disease models. The test statistic for each disease model is then evaluated using gene-specific null models parameterized by population genetic data from over 61,000 exomes. Note that the linear depiction of alleles is not meant to imply that phase is known or required in the implementation of the method as described here. Integration of phase information should be helpful but would require a phased population reference to implement.

Figure 2. PSAP calibration

We calculated PSAP p-values for each disease model using exome data from three independent control cohorts (A, WHI; B, GTEx; C, Swedish population controls). For each combination of cohort and disease model, we have pooled all PSAP p-values from all individuals (e.g. n individuals $\times k$ genes = nk total PSAP p-values) and plotted these against quantiles of the uniform distribution. Row (A): using data from the WHI, we compared p-value calibration from 189 European American (EA) individuals to that of 189 African American (AA) individuals for each of the three disease models evaluated in the main text. QQ-plots show that p-values from our null model are well calibrated when calculated using

control exomes of European descent, but are sensitive to population structure. **(B)** QQ-plots for 418 European Americans sequenced by the GTEx project. **(C)** QQ-plots for 2,200 Swedish population controls. In each panel, the grey shaded area represents a conservative 95% CI for the expected distribution of p-values. These are computed from the beta distribution, assuming independent tests, and are dependent on the number of tests being plotted. For each combination of disease model and cohort, over 97% of PSAP p-values fall within a factor of 2 of these confidence intervals. The most poorly calibrated distribution was the AR model for Swedish population controls –the maximum deviation from the expected p-values reached one order of magnitude in the range of 10^{-4} . AD=autosomal dominant disease model; AR=autosomal recessive disease model; CHET= compound heterozygote disease model.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 3. Factors influencing PSAP

Three primary sources of information contribute to the performance of PSAP values- the use of gene specific-models, modeling gene-specific singleton rates and integration of frequency information from ExAC. **A** We illustrate the advantage of gene-specific models using the well-studied disease gene PTEN. Using 1 million simulated individuals, we constructed null models for the worst CADD score across the exome (black line) and within the gene PTEN (blue line). CADD scores for 223 known disease variants within the PTEN gene are indicated with black hashmarks along the x-axis. The significance of most known PTEN disease mutations are on the order of 0.01 when evaluated with the genomewide null, but many have significance of 10^{-6} when evaluated with a PTEN-specific null. **B**. The probability of observing a heterozygous genotype with a CADD score greater than 5 or 35 is plotted against the gene-specific singleton rate for each gene in the genome. **C**. The minimum observed PSAP value for a gene is constrained by the frequency of other apparently deleterious variants in that gene. For each of 396 autosomal recessive genes, we calculated the PSAP values for all homozygous mutations reported in HGMD. We then compared the minimum HGMD PSAP for each gene to the maximum ExAC frequency of an apparently deleterious variant in the same gene. Here “apparently deleterious” is defined as a variant with an equal or greater CADD score than the worst HGMD disease mutation for each gene.

Figure 4. Benchmarking our ability to identify the causal gene in simulated n-of-one cases
 (A) We simulated 4,832,163 cases of recessive disease by adding pairs of heterozygous disease genotypes to the real exome genotype calls from healthy individuals (**Methods**). For each simulated disease exome, genes were rank ordered according to (i) PSAP p-value (CHET model), (ii) a gene-based statistic using CADD, or (iii) the combination of CHET PSAP p-value and CADD. We also plot the performance of each statistic when applied to the subset of exome variants with minor allele frequency < 1% in each simulated case (“+ AF”). The rank of the disease gene within the simulated case is shown on the y-axis, while the simulated disease genes are ranked against each other based on the CADD scores of their disease mutations (x-axis). Analogous plots for the autosomal dominant and autosomal recessive disease models are shown in Figure S10. (B) We constructed ROC curves that quantify the ability of PSAP-MIN to discriminate cases from controls using a set of 1,500 simulated cases and 1,500 ethnicity-matched controls. Here we show the ROC curves for PSAP-MIN when evaluating recessive disease mutations under the compound heterozygote disease model-the area under the curve was 0.76 when evaluating all genes in the genome (n=19,068 genes), 0.84 when evaluating only those genes with strong selective constraint (RVIS<0, n=8,146 genes) and 0.95 when restricting analysis to recessive disease genes in HGMD (n=507). Full results from all three disease models are in Table 2.

Figure 5. Application of PSAP to real cases of the n=1 problem

(A). In a previous cohort study examining homozygous-by-descent regions (HBDRs) in 300 cases of male infertility¹³, we identified a case of uniparental isodisomy spanning all of chromosome 2 (histogram). Fewer than 10 cases of this cytogenetic anomaly had ever been reported, all of which were ascertained through studies of Mendelian disease¹⁴. Inset: the B-allele frequency data from a SNP array run on this case. (B) PSAP analysis of exome sequencing data from this individual revealed an excess of low p-values on chromosome 2, and identified the most unusual genotype in this case to be a homozygous change in the gene *INHBB*, a biomarker of Sertoli cell function sometimes used in the clinical management of male infertility¹⁵. Each point in the QQ-plot represents the test of a single gene; genes are colored by genomic location. (C). Enrichment of low PSAP p-values in the UDP families. For each disease model we identified PSAP p-value thresholds corresponding to a 5% false positive rate when applied to population controls ($p=1 \times 10^{-5}$ for autosomal dominant and autosomal recessive models, $p=2.25 \times 10^{-5}$ for CHET model). These barplots indicate the percent of UDP individuals with at least one PSAP p-value less than these thresholds. Results are stratified by disease model and gene set evaluated. Fisher exact tests were used to identify cohorts that were significantly enriched for individuals with low PSAP p-values compared to expectation. * = $p\text{-value} < 0.05$, ** = $p\text{-value} < 0.01$

Figure 6. PSAP facilitates integrative analysis of rare disease patients

(A). One of the strengths of the PSAP statistic, is that, as a gene-based p-value, it can be easily combined with other gene-based measurements that may be informative about the functional relevance of a gene. We manually curated each disease diagnosis for the 30 diagnosed UDP cases, assigning a tissue label for the “primary” affected tissue in each case, and then combined gene expression p-values from the relevant tissue with the PSAP p-value for each gene in the genome, using Fisher’s method for data fusion (Supplementary Note). Here we plot the improvement in rank for each of 30 diagnostic genes when using this new “fused” p-value compared to the use of PSAP alone. (B) It should be feasible to generalize this integrative approach to many sources of molecular information on each gene in the genome. While the causal gene in a rare disease patient may not be definitively identified by a single measurement, joint analysis of a set of measurements may provide better resolution (red dots = measurements on the causal gene).

Table 1

Percentage of causal genes ranked within the 10 top of all genes in simulated cases of Mendelian disease. This table summarizes the results of the disease exome simulations shown in Figure 3; here we tabulate the proportion of HGMD genes identified as one of the top-10 most likely candidates within a real exome background. In order for a gene to be considered in the “top 10”, its mean rank had to be less than 10 across 189 unrelated EA exome backgrounds.

Disease Model	PSAP	CADD	PSAP + MAF < 1%	CADD + MAF < 1%
AD (n = 28,493)	35.1%	15.0%	35.1%	22.6%
AR (n = 25,567)	89.4%	21.0%	99.9%	99.9%
CHET (n = 25,567)	86.3%	61.9%	97.3%	98.3%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Empirical distributions of PSAP-MIN in a control population

Here, we tabulate the proportion of 189 EA individuals from the WHISP exome data with PSAP-MIN < 10⁻⁴, 10⁻⁵, or 10⁻⁶ for three disease models. We further tabulate these rates for different subsets of genes – “All genes”, genes with an RVIS value <0, and genes with at least one disease variant listed in HGMD. Under each column heading we provide the number of genes evaluated for each combination of disease model and gene set (n=). Finally, we list the area under the curve when using PSAP-MIN to discriminate cases from controls for each combination of disease model and gene set using simulated cases as in Fig. 4B. PSAP-MIN distributions for other control cohorts are very similar (Supplementary Table 2). AR=autosomal recessive disease model; CHET= compound heterozygote disease model; AD=autosomal dominant disease model; AUC=area under the curve.

All Genes	AR	CHET	AD
	n=18,973	n=19,068	n=19,167
PSAP-MIN	1.00E 04	0.326	0.085
	1.00E 05	0.065	0.015
	1.00E 06	<0.005	0.005
AUC	0.81	0.76	0.11

Genes with RVIS <0	AR	CHET	AD
	n=8,099	n=8,146	n=8,184
PSAP-MIN	1.00E 04	0.189	0.055
	1.00E 05	0.034	0.010
	1.00E 06	<0.005	0.005
AUC	0.90	0.84	0.16

HGMD genes	AR	CHET	AD
	n=500	n=507	n=344
PSAP-MIN	1.00E 04	0.023	0.010
	1.00E 05	0.005	<0.005
	1.00E 06	<0.005	<0.005
AUC	0.98	0.95	0.35