Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome

Zemojtel, T., Koehler, S., Mackenroth, L., Jaeger, M., Hecht, J., Krawitz, P., Graul-Neumann, L., Doelken, S., Ehmke, N., Spielmann, M., Oien, N.C., Schweiger, M.R., Krueger, U., Frommer, G., Fischer, B., Kornak, U., Floettmann, R., Ardeshirdavani, A., Moreau, Y., Lewis, S.E., Haendel, M., Smedley, D., Horn, D., Mundlos, S., Robinson, P.N.

Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome

Tomasz Zemojtel,[1,2,3§] Sebastian Köhler,[1,§] Luisa Mackenroth,[1,§] Marten Jäger,[1] Jochen Hecht,[4,5] Peter Krawitz,[1,4] Luitgard Graul-Neumann,[1] Sandra Doelken,[1] Nadja Ehmke,[1] Malte Spielmann,[1,4] Nancy Christine Øien,[1,6] Michal R.Schweiger,[1,4] Ulrike Krüger,[1] Götz Frommer,[7] Björn Fischer,[1,4] Uwe Kornak,[1,4] Ricarda Flöttmann,[1] Amin Ardeshirdavani,[8] Yves Moreau,[8] Suzanna E. Lewis,[9] Melissa Haendel,[10] Damian Smedley,[11] Denise Horn,[1] Stefan Mundlos,[1,4,5] Peter N Robinson,[1,4,5,12,*]

Affiliations

1) Institute for Medical and Human Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany
2) Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
3) Labor Berlin – Charité Vivantes GmbH, Humangenetik, Föhrer Straße 15,13353 Berlin, Germany
4) Max Planck Institute for Molecular Genetics, Ihnestr. 63–73, 14195 Berlin, Germany
5) Berlin Brandenburg Center for Regenerative Therapies (BCRT), Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany
6) Max Delbrück Center for Molecular Medicine, Robert-Rössle-Str. 10, 13125 Berlin, Germany
7) Agilent Technologies, Hewlett-Packard-Straße 8, 76337 Waldbronn, Germany
8) Department of Electrical Engineering, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium.
9) Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
10) University Library and Department of Medical Informatics and Epidemiology, Oregon Health & Sciences University, Portland, OR, USA.
11) Mouse Informatics group, Wellcome Trust Sanger Institute, Hinxton, UK
12) Institute for Bioinformatics, Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany
§) equal contribution
*) Correspondence to.

Phenotype-driven bioinformatic prioritization of candidate genes was applied for the diagnosis of rare, genetic diseases using targeted next-generation sequencing of the disease-associated genome.

Abstract

Less than half of patients with suspected genetic disease receive a molecular diagnosis. We have therefore integrated next-generation sequencing, bioinformatics, and clinical data into an effective diagnostic workflow. We used variants in the 2741 established Mendelian disease genes (the disease-associated genome (DAG)) to develop a targeted enrichment DAG panel (7.1 Mb), which achieves a coverage of 20-fold or better for 98% of bases. Furthermore, we established a computational method (Phenotypic Interpretation of eXomes (PhenIX)) that evaluated and ranked variants based on pathogenicity and semantic similarity of patients' phenotype described by Human Phenotype Ontology (HPO) terms to those of 3991 Mendelian diseases. In computer simulations, ranking genes based on the variant score put the true gene in first place less than 5% of the time; PhenIX placed the correct gene in first place over 86% of the time. A retrospective test of PhenIX on 52 patients with previously identified mutations and known diagnoses, achieving a mean rank of 2.1 for the correct gene. In a prospective study on 40 individuals without a diagnosis, PhenIX analysis enabled a diagnosis in 11 cases (28%, at a mean rank of 2.4). Thus, the combination of targeted next generation sequencing (NGS) investigation of the DAG followed by phenotype-driven bioinformatic analysis allows quick and effective differential diagnostics in medical genetics.

Introduction

At the time of this writing, roughly 7,000 Mendelian diseases are recognized (*1-3*). Although these diseases are individually rare, up to 8% of the population is affected by a specific genetic disorder (*4*). Because of the vast number of diseases, many of which have a broad and incompletely understood phenotypic spectrum, and the high genetic heterogeneity of many clinical syndromes such as intellectual disability, the diagnostic process in medical genetics is often challenging, even for experienced and expert clinicians. The traditional medical genetics evaluation relies upon recognizing a characteristic pattern of signs or symptoms to guide targeted genetic testing for confirmation of the diagnosis, with the major diagnostic methods including karyotyping, array comparative genomic hybridization (CGH), biochemical testing, and Sanger sequencing of individual genes. However, the diagnostic yield remains less than 50% even after extensive workups (*5*), with the costs of clinical and molecular genetic analysis for patients whose diagnosis is not clear after the first visit reaching 25,000 US dollars or more (*5*).

The term “diagnostic odyssey” has been used to describe the experience of patients and families affected by rare diseases that cannot be diagnosed; for instance, the average time between the onset of symptoms and the correct diagnosis is currently 14 years for patients with type 2 myotonic dystrophy (*6*). The lack of a diagnosis can mean missed opportunities for tailored approaches to clinical management and treatment strategies, a substantial burden of guilt and uncertainty for families, and the inability to make accurate statements on recurrence risk and prognosis, not to mention the economic costs of unnecessary diagnostic procedures.

Whole-exome sequencing (WES), first used in 2010 to identify the cause of a Mendelian disease (*7*), is rapidly becoming attractive as a tool for diagnostic testing in general medical genetics (*8*). Additionally, NGS-panel, WES, and whole genome sequencing (WGS) approaches have been introduced for carrier screening (*9*) as well as in neonatal intensive care units (*10*). However, medical interpretation of WES results remains challenging, and the successes have for the most part been limited to single cases or small groups of patients

(*11*). Identifying the one or two causative mutations amongst the myriad of variants present in the WES findings of an individual has been compared to finding a needle in a haystack (*12*). A typical exome contains well over 30,000 variants when compared to the human reference sequence, with about 10,000 of them representing nonsynonymous amino acid substitutions, alterations of conserved splice site residues, or small insertions or deletions (*13, 14*). Although the community has developed numerous bioinformatic tools to filter out common variants and predict their pathogenicity (*15, 16*), each human genome harbors about 100 genuine loss of function variants with ~20 genes completely inactivated (*17*). Therefore, purely sequenced-based evaluation of genes in diagnostic WES typically identifies tens or hundreds of candidates. While this is acceptable in a research context, in which other strategies such as genetic linkage or comparison with a study group of individuals thought to have the same disease can often reduce the search space, extensive evaluation of long lists of candidate genes does not scale well to the diagnostic setting.

Depth and uniformity of coverage have a major influence on the performance of targeted capture for next-generation sequencing. For instance, at a mean on-target read depth of 20x, up to 15% of heterozygous single nucleotide variants will be missed (*18*). Although initial WES studies aimed for a coverage of 20-fold, deeper coverage is needed for accurate detection of heterozygous variants (*19*), and current studies typically employ a coverage of 50-70-fold (*20, 21*) or higher. This has led to debate in the community as to the relative value of various NGS approaches for diagnostics, with proponents of targeted panel sequencing (*22*), WES (*23*), and whole genome sequencing (WGS) (*24*).

In this work, we explore a different approach towards the translation of NGS-based diagnostics into clinical diagnostics in a medical genetics clinic. We contend that WES is not optimal in a purely diagnostic setting, since we can currently offer a confident interpretation of variants only in ~2740 known Mendelian disease genes; the identification of a potentially pathogenic variant in a gene regarded as a good candidate because of biochemical or model-organism data often represents the starting point for a good research project, but is more likely to engender confusion in a diagnostic setting. Therefore, by enriching for genes

known to be associated with Mendelian disease, we shift the focus from the whole exome to that part of the exome/genome that is clinically interpretable in a diagnostic setting. We refer to this portion of our genome as the disease-associated genome (DAG). A pathogenic variant in one of these genes is, in principle, interpretable in the context of the presenting clinical phenotype and our knowledge of the diseases associated with the gene in question. We have previously shown that phenotypically driven genomic data fusion (*25*) and comparison of human to model organism phenotypes (*26*) dramatically improves the ability to correctly identify candidate disease-causing mutations in WES studies. Here, we use the Human Phenotype Ontology (HPO) and associated data to develop a computational procedure for differential diagnosis with the DAG panel. The HPO provides a structured, comprehensive and well-defined set of over 10,000 terms describing human phenotypic abnormalities. It provides annotations of nearly 7,300 human hereditary syndromes that yield computable representations of the diseases, associated disease genes, as well as the signs, symptoms, laboratory findings, and other phenotypic abnormalities that characterize the diseases (*3, 27*). Here, we adapt our semantic similarity approach towards differential diagnosis, using terms and annotations from the HPO (*28*), to rank candidate genes in a diagnostic setting. Our algorithm is freely available for academic use through the website http://compbio.charite.de/PhenIX/.

Results

Here we present an approach to Mendelian disease diagnostics that involves the targeted sequencing of the DAG panel combined with a phenotype-driven computational analysis strategy (PhenIX) that ranks candidate genes on the basis of the presence of rare, predicted pathogenic variants and the clinical relevance of the genes with associated disease phenotypes. Our algorithm first filters the variants according to rarity, target region location, and predicted pathogenicity. Next, the remaining candidate genes are evaluated for clinical relevance on the basis of the semantic similarity of the patient's phenotypic abnormalities to the phenotypic spectrum of diseases associated with each candidate gene. In brief, our method aims to identify and rank disease genes by combining potential clinical relevance with deleterious variants found within those genes (see Methods section).

Design and Validation of the Disease-Associated Genome Panel

We established a comprehensive catalog of Mendelian disease genes using data from the Human Phenotype Ontology project (*3*), part of which is derived from information in the Online Mendelian Inheritance in Man (OMIM) (*1*) and Orphanet (*2*) resources. The HPO project, which was initiated in 2007, has grown to include over 10,000 terms describing individual phenotypic abnormalities that have been used to generate over 110,000 annotations to over 7000 mainly Mendelian disease entries (*3, 27*). The data in the HPO thus provides a powerful curated resource for translational research by providing the means to capture, store, and exchange phenotypic information about human disease and has been used to integrate phenotypic information into computational analysis (*25, 26, 28-32*). We additionally surveyed the recent literature to obtain additional information about plausible candidate disease genes from recent publications describing large-scale WES studies were also included (*8, 33-37*), for a total of 2741 genes (genes and references are included in Table S6).

Since our aim was to obtain nearly complete coverage of the DAG, we designed enrichment probes for the DAG using SureSelect technology (*38*). In total 96 samples were sequenced (six samples per lane of an Illumina HiSeq 1500 sequencer), resulting in an average coverage of 361.7±81.6 reads (135.6 ±10.6 after removal of duplicates), with 98% of the DAG target region being covered by at least 20 reads (Figure S1; Tables S1 and S2).

In order to estimate the advantage of the high coverage of the DAG panel with respect to comprehensive variant calling, we randomly sampled reads from the Binary Alignment/Map (BAM) files from the DAG target region (twice over each of the 96 sequenced DAG samples) to a target average coverage of 100-fold to simulate the coverage expected from typical exome sequencing. After this, the down-sampled BAM files were processed in the same way as the original BAM files, and the distribution of called variants was compared (Table S3). A substantial number of variants called from the original BAM file were not called from the files simulated to have exome or genome coverage, including an average of 5.2 ± 2.0 variants listed in the Human Gene Mutation Database (HGMD) (*39*).

Phenotypic Interpretation of eXomes: PhenIX

We developed a computational algorithm to filter and rank candidate genes according to variant rarity and pathogenicity and potential clinical relevance of the gene harboring the variants. As input, PhenIX requires (i) a variant call format (VCF) file representing the results of sequencing the DAG target region (or an exome or genome), and (ii) a list of Human Phenotype Ontology (HPO) terms representing the clinical features of the individual being sequenced. Each variant is scored on the basis of rarity and predicted pathogenicity; after this, all variants mapping to a given gene are combined. The genes harboring predicted pathogenic variants are assigned a phenotype score by using the semantic similarity between associated disease phenotypes and the patient's phenotype. However, the gene is down-weighted if the distribution of variants in a gene is incompatible with the mode of inheritance of the associated disease, e.g., if a single heterozygous variant is observed in a

gene associated with an autosomal recessively inherited disease. Finally, a rank is calculated based on the combined variant and phenotype scores.

To estimate the performance of our method, we conducted extensive computational simulations using mutation data from the HGMD. Sample datasets were simulated for a given disease and inheritance model by spiking with mutations from HGMD into a VCF file generated with the DAG panel. Appropriate HPO terms were chosen from the annotations of the corresponding disease. Several test scenarios were considered. The performance of the method was near 100% when all the HPO terms annotating the disease (e.g., Greig cephalopolydactyly syndrome is annotated with 44 HPO terms representing individual signs and symptoms of that disease). In another, more realistic, test scenario, up to five terms were chosen, of which two were made imprecise by exchanging them with the more general parent term, and two unrelated confounder ("noise") terms were added at random. Here, the correct gene was ranked in first place in 86.5% of 8504 simulations**,** corresponding to a 32.5-fold improvement over pure variant filtering (Figure 1, Figure S2).

Retrospective analysis

We then tested the performance of our method with the generated DAG data from 52 individuals with a diagnosis of Mendelian disease that had been confirmed by Sanger sequencing (Table 1). HPO terms were entered and filtering was performed at a frequency threshold of 1%. The average rank of the correct gene amongst the 2741 disease genes in the DAG panel was 2.1. The mean rank of the autosomal recessive genes was 5, substantially lower than for the autosomal dominant genes (1.7). The lower rank for the recessive genes was partially related to results for an individual with eczematoid acrodermatitis enteropathica, who had a missense mutation in *SLC39A4* that was correctly flagged as pathogenic as well as a synonymous mutation that had been shown to cause a splice defect. The latter mutation was not identified as deleterious by PhenIX, resulting in a final rank of 14 for *SLC39A4*.

Prospective analysis

To further validate our methodology, we investigated 40 individuals who, after extensive clinical genetic evaluation (physical exam by medical geneticist, array CGH, and often targeted Sanger gene sequencing), remained without a diagnosis (clinical features summarized in Table 2). We designed a standard evaluation procedure in which deep phenotyping (*40*) with the selection of representative HPO terms (*3, 27*) was followed by targeted NGS of the DAG panel. Computational analysis was performed as described above to generate a ranked list of candidates based on the combined variant and clinical relevance scores. Since our computational simulations almost always placed the true disease gene in the top 10 candidates, we limited our evaluation to the top 20 ranked genes as well as any gene with a pathogenic mutation at the same nucleotide position listed in HGMD (*39*) or ClinVar (*41*) for each patient. Initial clinical evaluation was performed by one of the authors, and a short list of the most likely candidates was presented to the entire group in clinical rounds, where up to the best two candidate genes were chosen based on clinical experience. These genes were subjected to Sanger validation and cosegregation studies. If the variants in the selected genes cosegregated as expected and the clinical manifestations of the patient were sufficiently explained by a disease associated with the gene, then a positive diagnosis was made. Otherwise, the short list was re-examined for additional candidates (Fig 2). We estimate an experienced clinical geneticist would spend a total of one hour in the initial evaluation of the patient and decision whether to perform DAG panel sequencing and an additional one hour studying the list of top 20 candidates, evaluating the results of Sanger validation and cosegregation studies before being able to decide whether a definitive diagnosis can be made.

By applying this procedure to 40 individuals, we identified a definitive diagnosis in eleven (28%) cases. Table 2 shows a clinical summary of these cases, and Tables S4 and S5 include a full list of HPO terms used to search in PhenIX. PhenIX analysis was performed

according to the flow chart in Figure 2, and the top 20 genes were inspected. Discussion at clinical genetics rounds flagged one (n=16 only one) or two (n=6) genes as being likely candidates. These genes were then subjected to Sanger validation, cosegregation studies and close examination. This led to definitive diagnoses being made in 11 of 40 patients (28%) (Table 3).

Discussion

Genomic medicine, including WES and WGS, is poised to transform clinical practice in many fields (*42*). Here, we present a phenotype-driven computational and clinical workflow for the efficient diagnosis of rare Mendelian diseases. Our approach uses the results of clinical analysis to substantially improve the ranking of candidate genes, and provides a clear pathway to integrate the results of bioinformatic analysis into the clinical workflow by clinical evaluation of phenotypic matching amongst the best candidates.

In this work, we have shown how to use a computable representation of clinical phenotypes to prioritize candidate genes in diagnostic sequencing with a target panel of 2741 known Mendelian disease genes. Our workflow represents a tight integration of clinical and bioinformatic analysis (Fig. 2). Clinical expertise is required to perform deep phenotyping and choose representative HPO terms to describe the clinical features of the patient being investigated. Experience is necessary to realize whether a given phenotypic abnormality is likely to be characteristic of a disease or an incidental finding, e.g. a feature such as low-grade myopia may not be related to the genetic disease being sought and adding this feature to PhenIX analysis may lower the score of the actual disease-causing gene. Following sequencing, alignment, and variant calling, PhenIX analysis is used to generate a list of the top 20 candidates. Additional candidates can be listed if desired. Clinical expertise is required to examine this list for promising candidates based on additional information from original publications and databases, such as OMIM. To assist with this process, the PhenIX webpage provide links to a number of useful resources including OMIM, the UCSC Genome Browser, ClinVar, and HGMD. We suggest that a presentation of the case together with a

description of the best PhenIX candidates at clinical genetics rounds should be performed, followed by validation of the most plausible candidate(s) by Sanger sequencing and cosegregation studies. In our experience, discussions on the differential diagnosis proceed quickly when organized in this fashion and fit well into a typical clinical workflow. We chose to limit our NGS analysis only to the sample from the affected individual, because in the diagnostic setting family samples (trios) may not be available initially. In addition, the cost of sequencing may be a factor. However, trio sequencing could easily be adapted into our workflow.

On the basis of our results, we suggest that targeting all known disease genes, that is a DAG, rather than the whole exome or genome, is advantageous in terms of target coverage, cost per sample, and the ability to provide quick and accurate clinical interpretation of the variants. Cases that remain unsolved after PhenIX analysis of the DAG Panel can be considered for more time-intensive clinical research WES/WGS studies, as these approaches are able to search for potential mutations in previously undescribed disease genes.

There are several areas in which our approach can be improved and extended. The phenotypic analysis based on semantic similarity depends on an annotated corpus of information about the phenotypic features that characterize various diseases. The HPO currently has over 110,000 annotations to over 7000 diseases listed in the Online Mendelian Inheritance in Man (*3*). Increasing the depth of annotation to these diseases would improve the performance (*43*). A number of challenges remain in the ontological modeling of certain classes of diseases and phenotypes in areas such as neurobehavioral abnormalities (*44*). The DAG panel as presented here currently contains baits only for protein coding genes. However, other medically relevant sequences of the genome could be captured in a similar way, such as enhancers of the sonic hedgehog gene, in which point mutations can cause characteristic skeletal malformations (*45*). Hand in hand with this, future bioinformatics research will be

required to confidently identify medically relevant variants in non-coding sequences as well as presumptive synonymous variants that actually lead to a deleterious effect such as defective splicing in the case of the “silent” *SLC39A4* mutation mentioned above. Our approach concentrates on known disease genes, and is thus not designed or intended to identify novel disease genes; other computational tools such as the Exomiser (*26*) and eXtasy (*25*) have been presented for this purpose.

In summary, we have presented a diagnostic tool for genetics professionals that combines targeted enrichment and next-generation sequencing of a comprehensive panel of genes known to be associated with Mendelian disease; bioinformatics analysis of sequencing results is tightly coupled to the expertise and workflow of genetics professionals, allowing a complete workup of NGS results in roughly two hours per patient. A recent study on the use of diagnostic exome sequencing of 250 unselected, consecutive cases achieved a diagnostic yield of 25% (*8*), and another larger scale exome-based study on persons with intellectual disability reached a diagnostic yield of 16% (*46*). Although it is hard to compare the diagnostic yield between different studies, the results presented here are competitive, with an average rank of the correct gene of 2.1 in a retrospective study on representative diseases and a yield of 28% in prospective study with cases chosen for the fact that a diagnosis could not be achieved. Additionally, our method requires less sequencing than high coverage WES or WGS which may translate into cost benefits. Our bioinformatic and clinical workflow could be completed in roughly two hours per patient, and PhenIX analysis is easy to use, requiring only a VCF file and a list of HPO terms. Our method thus provides the means for quick and effective differential diagnostics in medical genetics.

MATERIALS AND METHODS

Consent

This study was approved by the Institutional Review Board of the Charité Universitätsmedizin Berlin. Informed written consent was obtained from adult subjects and parents of children.

Case selection

The control group consisted of 52 individuals with suspected genetic diagnoses seen at the Institute of Medical Genetics and Human Genetics of the Charité university hospital between 2010 and 2013, and who received an etiological diagnosis based on clinical findings and the identification of mutations in the genes indicated in Table 1. In addition, 38 patients seen during this time frame who remained without an etiological diagnosis were investigated in this study. Patients were chosen on the basis of availability of DNA samples from parents (for validation of cosegregation by Sanger sequencing), consent for research, and the inability to identify a genetic diagnosis despite a high index of suspicion of an underlying genetic cause. Two additional cases were referred from external clinics and were not seen in our department (P6 and P10 in Table 3).

Capture of the targeted disease-related genome and Next-Generation Sequencing

A SureSelectXT Automation Custom Capture Library (Agilent) target enrichment panel was generated using the coordinates given in Table S6. The enrichment panel comprised all coding exons of 2741 genes associated with at least one Mendelian disease as well as 133 control genes. Capture was performed according to the manufacturer's instructions using an NGS Workstation Option B (Agilent) for automated library preparation starting with 3 µg DNA per sample. Then, sequencing of 100 bp paired-end reads was carried out on a HiSeq 1500 (Illumina). Sequence reads were mapped to the haploid human reference genome (hg19) with Novoalign (Novocraft Technologies). Single nucleotide variants (SNVs) and short insertions and deletions (indels) were called using GATK version 2.8 (*47*). Variant annotation was performed with Jannovar (*48*). In total, 96 samples were sequenced on two HiSeq 1500 flowcells.

PhenIX: Bioinformatic ranking of candidate genes. Ranking of candidate genes was performed in two steps. First, off-target and synonymous variants were removed, and the remaining variants were analyzed with respect to population frequency by using data from dbSNP (*49*) and from the Exome Variant Server (NHLBI GO Exome Sequencing Project 2014, http://evs.gs.washington.edu/EVS/). For the purposes of analysis, we assumed the minor allele frequency of each variant to be the maximum frequency reported by dbSNP or that of the African American or European American populations represented in the Exome Variant Server. A frequency score is calculated as $\max(0{,}1 - 0.13533e^{100*f})$, and variants with no frequency data ($f = 0$) were assigned a score of 1.0, and results in values between 1.0 and 0.0 for variants with frequencies of up to 2%. Predicted pathogenicity of missense variants was derived from dbNSFP version 2.4 (*50*) using the fields for MutationTaster (*16*), polyphen-2 (*15*), and SIFT (*51*). Scores from these three prediction tools were normalized to be between 0.0 (benign) and 1.0 (pathogenic), and the single most pathogenic score was taken for each variant. For classes of variants other than missense mutations, a pathogenicity score was calculated as described (*26*). Finally, the overall variant score was calculated as the product of the frequency and pathogenicity score. A clinical relevance score was calculated using the semantic similarity between phenotypic abnormalities entered by the user and 2741 disease genes in our database. The phenotypic abnormalities of all diseases associated with a given gene were assigned to the gene, since our method ranks candidate genes rather than individual diseases. For instance, the *FBN1* gene is mutated in Marfan syndrome, acromicric dysplasia, and a number of other diseases, and the phenotypic abnormalities of each of those diseases were assigned to *FBN1*. Then, the semantic similarity score of the Phenomizer algorithm (*28*) was calculated for each of the genes. The maximum score was set to 1.0, and the other scores were normalized accordingly. The final score was calculated as the average of the variant and the gene-relevance score. However, if the variant distribution for a gene was not compatible with the mode of inheritance of the associated diseases (e.g., a gene has only a single heterozygous mutation but the

associated disease is autosomal recessive, or the gene has only a single homozygous mutation but the disease is autosomal dominant), then the gene relevance score was divided by 2 before calculating the final score. The final score was calculated as the mean of the variant score and the gene relevance score. The major distinction between PhenIX and our previously published algorithm PHIVE, which is implemented in the Exomiser (*26*) is thus the restriction of the analysis to variants in clinically interpretable disease genes using only human phenotype information rather than model organism phenotype data, the analysis of sequencing results for previously reported mutations in ClinVar and the public version of HGMD, and the use of prioritization based on the modes of inheritance of diseases associated with candidate genes compared with the distribution of sequenced variants.

Computational evaluation of PhenIX prioritization

To test the performance of PhenIX prioritization with DAG panel sequencing, we used a simulation approach based on known disease-causing mutations from the Human Gene Mutation Database (HGMD). A total of 28,516 mutations were selected on the basis of being assigned as a disease-causing, single-nucleotide mutations (including indels) by HGMD and with HPO annotations available for the disease in question. For the simulations, 10,000 variants were randomly selected from this set. We first removed the causative mutations from the 52 VCF files generated from the retrospective cohort with known mutations. Then, we added an additional mutation to one of these files. For autosomal dominant diseases, one heterozygous mutation was added; and for autosomal recessive diseases, either one homozygous mutation or two heterozygous mutations were added. The phenotypic (HPO) annotations for the corresponding disease were then compared to the HPO annotations associated with the 2741 disease genes (if a disease gene was associated with multiple diseases, all annotations were merged). There were three test scenarios. In the first case, all HPO annotations for the disease in question were used. In order to simulate incomplete phenotyping, we performed the simulations with up to five HPO

terms chosen at random from the annotations of the disease. Finally, in order to simulate the effects of noise, we randomly chose 2 or the 5 terms and promoted them to their less specific parent terms, and finally 2 new terms were chosen randomly from the whole of HPO and added to the annotations

A rank was determined for the original disease gene following PhenIX analysis. In all the analysis, an ordinal ranking method was used in which equal scoring genes are resolved arbitrarily but consistently by assigning a unique rank to each of the ties. In our case, we simply sorted the equally scored genes alphabetically and assign the ranks. We recorded the number of times the correct disease gene was ranked in first place, as well as the total recall (correct gene listed at any rank). For each simulation, one of the 52 DAG panel VCF files was chosen.

Clinical evaluation and validation of NGS results

We clinically evaluated the NGS results using the PhenIX server, which implements the algorithm described above. PhenIX presents a ranked gene list together with links to various other resources such as the UCSC browser (*52*), Entrez Gene (*53*), OMIM (*1*), Orphanet (*2*), ClinVar (*41*), MutationTaster (*16*), and HGMD (*39*). Evaluation was performed by trained genetics professionals. For each unsolved case, the top 20 ranked candidates were examined by comparison with the above mentioned data sources and as appropriate with the original literature. An initial assessment of these 20 candidates was possible in about two hours, and resulted in a short list of candidates thought to be potential matches. These were discussed at clinical rounds by a team of clinicians and researchers including LM, TZ, LGM, SD, NE, MS, NCØ, MRS, UK, PK, PNR, SM, and DH. A consensus decision was reached on candidates to be validated by Sanger sequencing and cosegregation studies. We considered a case to be solved after clinical analysis and cosegregation studies if a degree of certainty was reached that led to reporting of the mutation and diagnosis in our clinical setting.

TABLES

Mode of inheritance	Genes	Average rank
AD	ACVR1, ATL1, BRCA1, BRCA2, CHD7 (4), CLCN7, COL1A1, COL2A1, EXT1, FGFR2 (2), FGFR3, GDF5, KCNQ1, MLH1 (2), MLL2/KMT2D, MSH2, MSH6, MYBPC3, NF1 (6), P63, PTCH1, PTH1R (2), PTPN11 (2), SCN1A, SOS1, TRPS1, TSC1, WNT10A	1.7
AR	ATM, ATP6V0A2, CLCN1 (2), LRP5, PYCR1, SLC39A4	5
X	EFNB1, MECP2 (2), DMD, PHF6	1.8

Table 1. 52 control patient cases with known mutations. The number of patients with a mutation in the given gene is indicated in parentheses.

Clinical presentation	N
Intellectual disability + multiple congenital anomalies (= more than 2 other organ systems affected)	13
Intellectual disability + other neuropsychological features	7
Intellectual disability + musculoskeletal abnormalities	5
Intellectual disability + eye abnormalities	1
Intellectual disability + dysmorphic features	1
Multiple congenital anomalies (more than 2 organ systems affected) without intellectual disability	6
Skeletal phenotype	5
Eye and/or Ear phenotype	2

Table 2. Summary of clinical signs and symptoms in 40 patients with unknown diagnosis

ID	Age, Sex	Presentation	Gene	Rank	Diagnosis	MoI
P1	3y (f)	Intellectual disability + multiple congenital anomalies	*MLL*	2	Wiedemann-Steiner syndrome (*54*)	AD
P2	5y (f)	Intellectual disability + multiple congenital anomalies	*SYNGAP1*	4	Mental retardation, MRD5 (*55*)	AD
P3	6y (f)	Skeletal phenotype	*FGFR2*	1	Pfeiffer syndrome (*56*)	AD
P4	Death at 5.5m (f)	Multiple congenital anomalies without intellectual disability	*SH3PXD2B*	6	Frank-ter Haar syndrome (*57*)	AR
P5	6m (f)	Intellectual disability + neurological abnormalities	*SLC6A3*	1	Parkinsonism-dystonia (*58*)	AR
P6	Fetus (m) Death at 22w of gestation	Skeletal phenotype	*ALPL*	2	Infantile hypophosphatasia (*59*)	AR
P7	7y (m)	Eye phenotype	*NHS*	2	Nance-Horan Syndrome / Cataract 40, X-linked (*60*)	XR
P8	14y (m)	Intellectual disability + multiple congenital anomalies	*MLL*	1	Wiedemann-Steiner syndrome (*54*)	AD
P9	6y (f)	Intellectual disability + multiple congenital anomalies	*DYRK1A*	4	Mental retardation, MRD7 (*61*)	AD
P10	4 children between 1 ½ and 7y	Intellectual disability + multiple congenital anomalies	*MCOLN1*	1	Type IV mucolipidosis (*62*)	AR
P11	3y (m)	Intellectual disability + multiple congenital anomalies	*RBM10*	3	TARP syndrome (*63*)	XR

Table 3. Clinical category and final diagnosis of 11 patients whose diagnosis was identified by PhenIX analysis. Additional information, including complete lists of HPO terms used to describe the phenotypic abnormalities seen in these patients is available in Table S4. Patients P6 and P10 were referred from external centers. The “Rank” column shows the rank after PhenIX analysis before clinical evaluation. The average rank for all 11 cases was 2.5. MoI: mode of inheritance, AD: autosomal dominant, AR: autosomal recessive, XR: X-linked recessive.

FIGURE LEGENDS

Fig.1. Computational evaluation of PhenIX**.** HGMD mutations were inserted into variant files from DAG panels from which the causative mutations had been removed and phenotypic annotations of the corresponding diseases were extracted from the HPO database. The genes were ranked using PhenIX. Results were simulated either on the entire disease set (All), or by filtering for known autosomal dominant (AD) or autosomal recessive (AR) diseases (see Figure S2). A total of 8504 (All), 3471 (AD), and 5006 (AR) simulations were performed. Data are shown as the percentage of simulations in which the correct genes was ranked in first place. Variant, only variant scores used to rank candidate genes. All terms, All HPO terms used to annotate a disease were used for PhenIX analysis. ≤5 terms, Up to 5 HPO terms were chosen at random from the terms used to annotation the disease. ≤5 terms & noise, Up to 5 annotations are used, 2 of which are made imprecise by exchanging them with a more general parent term; additionally, two random “noise” terms were added. Results are shown for the correct gene being ranked as the single top hit, or being among the top 5, 10, or 20 hits for the three test scenarios.

Fig 2. PhenIX workflow showing the clinical and bioinformatic analysis steps. After initial clinical evaluation, a decision is made to perform PhenIX analysis if no clinical diagnosis can be found. After sequencing and computational analysis, clinical evaluation of the top 20 gene candidates identifies genes for validation by Sanger sequencing and cosegregation studies.

Figure S1. Distribution of the coverage fraction for all sequenced 96 samples.

Figure S2. Computational evaluation of PhenIX.

Table S1. Percentage of target bases that exceed coverages of 10, 20, ..., 100 reads.

Table S2. Read alignment and coverage summary statistics for the 96 DAG panels sequenced for this project (40 patients with unknown diagnosis P1-P40, 52 patients with known diagnosis C41-C92, 4 control samples R93-R96).

Table S3. Average number of variants called only from the original BAM files from the DAG panels but not in simulated BAM files generated down-sampling reads to a typical exome coverage (100x).

Table S4. Detailed clinical and molecular findings for the 11 individuals in whom a previously unknown diagnosis was clarified by PhenIX analysis. HPO terms shown in bold match with the disease profiles in the HPO database for these diseases.

Table S5. Clinical presentation of 29 patients for whom PhenIX analysis failed to reveal a molecular diagnosis.

Table S6. List of genes (with references) present in the DAG panel.

References

1. J. Amberger, C. Bocchini, A. Hamosh, A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). *Hum Mutat* **32**, 564 (May, 2011).
2. A. Rath *et al.*, Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum Mutat* **33**, 803 (May, 2012).
3. S. Köhler *et al.*, The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res* **42**, D966 (Jan 1, 2014).
4. P. A. Baird, T. W. Anderson, H. B. Newcombe, R. B. Lowry, Genetic disorders in children and young adults: a population study. *Am J Hum Genet* **42**, 677 (May, 1988).

5. V. Shashi *et al.*, The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. *Genet Med* **16**, 176 (Feb, 2014).
6. J. E. Hilbert *et al.*, Diagnostic odyssey of patients with myotonic dystrophy. *J Neurol* **260**, 2497 (Oct, 2013).
7. S. B. Ng *et al.*, Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30 (Jan, 2010).
8. Y. Yang *et al.*, Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* **369**, 1502 (Oct 17, 2013).
9. C. J. Bell *et al.*, Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* **3**, 65ra4 (Jan 12, 2011).
10. C. J. Saunders *et al.*, Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci Transl Med* **4**, 154ra135 (Oct 3, 2012).
11. K. A. Johansen Taber, B. D. Dickinson, M. Wilson, The promise and challenges of next-generation genome sequencing for clinical care. *JAMA Intern Med* **174**, 275 (Feb 1, 2014).
12. G. M. Cooper, J. Shendure, Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**, 628 (Sep, 2011).
13. K. Pelak *et al.*, The characterization of twenty sequenced human genomes. *PLoS Genet* **6**, e1001111 (Sep, 2010).
14. M. X. Li *et al.*, Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet* **9**, e1003143 (2013).
15. I. A. Adzhubei *et al.*, A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248 (Apr, 2010).
16. J. M. Schwarz, C. Rodelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575 (Aug, 2010).
17. D. G. MacArthur *et al.*, A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823 (Feb 17, 2012).
18. A. M. Meynert, L. S. Bicknell, M. E. Hurles, A. P. Jackson, M. S. Taylor, Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics* **14**, 195 (2013).
19. S. S. Ajay, S. C. Parker, H. O. Abaan, K. V. Fajardo, E. H. Margulies, Accurate and comprehensive sequencing of personal genomes. *Genome Res* **21**, 1498 (Sep, 2011).
20. K. E. Lohmueller *et al.*, Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes. *Am J Hum Genet* **93**, 1072 (Dec 5, 2013).
21. F. M. Williams *et al.*, Genes contributing to pain sensitivity in the normal population: an exome sequencing study. *PLoS Genet* **8**, e1003095 (2012).
22. H. L. Rehm, Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet* **14**, 295 (Apr, 2013).
23. B. O. Choi *et al.*, Exome sequencing is an efficient tool for genetic screening of Charcot-Marie-Tooth disease. *Hum Mutat* **33**, 1610 (Nov, 2012).
24. M. P. Ball *et al.*, A public resource facilitating clinical use of genomes. *Proc Natl Acad Sci U S A* **109**, 11920 (Jul 24, 2012).
25. A. Sifrim *et al.*, eXtasy: variant prioritization by genomic data fusion. *Nat Methods* **10**, 1083 (Nov, 2013).

26. P. N. Robinson *et al.*, Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* **24**, 340 (Feb, 2014).
27. P. N. Robinson *et al.*, The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* **83**, 610 (Nov, 2008).
28. S. Köhler *et al.*, Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* **85**, 457 (Oct, 2009).
29. S. Bauer, S. Kohler, M. H. Schulz, P. N. Robinson, Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics* **28**, 2502 (Oct 1, 2012).
30. S. C. Doelken *et al.*, Phenotypic overlap in the contribution of individual genes to CNV pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish. *Dis Model Mech* **6**, 358 (Mar, 2013).
31. R. Hoehndorf, P. N. Schofield, G. V. Gkoutos, PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res* **39**, e119 (Oct, 2011).
32. T. Hwang *et al.*, Co-clustering phenome-genome for phenotype classification and disease gene discovery. *Nucleic Acids Res* **40**, e146 (Oct, 2012).
33. P. S. Tarpey *et al.*, A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet* **41**, 535 (May, 2009).
34. G. Cho, Y. Lim, J. A. Golden, XLMR candidate mouse gene, Zcchc12 (Sizn1) is a novel marker of Cajal-Retzius cells. *Gene Expr Patterns* **11**, 216 (Mar-Apr, 2011).
35. H. Najmabadi *et al.*, Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* **478**, 57 (Oct 6, 2011).
36. M. G. Kapetanaki *et al.*, The DDB1-CUL4ADDB2 ubiquitin ligase is deficient in xeroderma pigmentosum group E and targets histone H2A at UV-damaged DNA sites. *Proc Natl Acad Sci U S A* **103**, 2588 (Feb 21, 2006).
37. A. Rauch *et al.*, Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674 (Nov 10, 2012).
38. A. Gnirke *et al.*, Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182 (Feb, 2009).
39. P. D. Stenson *et al.*, The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1 (Jan, 2014).
40. P. N. Robinson, Deep phenotyping for precision medicine. *Hum Mutat* **33**, 777 (May, 2012).
41. M. J. Landrum *et al.*, ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**, D980 (Jan, 2014).
42. S. F. Kingsmore, C. J. Saunders, Deep sequencing of patient genomes for disease diagnosis: when will it become routine? *Sci Transl Med* **3**, 87ps23 (Jun 15, 2011).
43. M. Oti, M. A. Huynen, H. G. Brunner, The biological coherence of human phenome databases. *Am J Hum Genet* **85**, 801 (Dec, 2009).
44. P. N. Robinson, C. Webber, Phenotype ontologies and cross-species analysis for translational research. *PLoS Genet* **10**, e1004268 (Apr, 2014).
45. M. M. Al-Qattan, I. Al Abdulkareem, Y. Al Haidan, M. Al Balwi, A novel mutation in the SHH long-range regulator (ZRS) is associated with preaxial polydactyly, triphalangeal thumb, and severe radial ray deficiency. *Am J Med Genet A* **158A**, 2610 (Oct, 2012).

46. J. de Ligt *et al.*, Diagnostic exome sequencing in persons with severe intellectual disability. *N Engl J Med* **367**, 1921 (Nov 15, 2012).
47. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297 (Sep, 2010).
48. M. Jäger *et al.*, Jannovar: A Java library for Exome Annotation. *Hum Mutat* **in press**, (2014).
49. S. T. Sherry *et al.*, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308 (Jan 1, 2001).
50. X. Liu, X. Jian, E. Boerwinkle, dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat* **34**, E2393 (Sep, 2013).
51. P. C. Ng, S. Henikoff, SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812 (Jul 1, 2003).
52. D. Karolchik *et al.*, The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42**, D764 (Jan, 2014).
53. NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **42**, D7 (Jan, 2014).
54. W. D. Jones *et al.*, De novo mutations in MLL cause Wiedemann-Steiner syndrome. *Am J Hum Genet* **91**, 358 (Aug 10, 2012).
55. M. H. Berryer *et al.*, Mutations in SYNGAP1 cause intellectual disability, autism, and a specific form of epilepsy by inducing haploinsufficiency. *Hum Mutat* **34**, 385 (Feb, 2013).
56. S. Jay *et al.*, The fibroblast growth factor receptor 2 p.Ala172Phe mutation in Pfeiffer syndrome--history repeating itself. *Am J Med Genet A* **161A**, 1158 (May, 2013).
57. Z. Iqbal *et al.*, Disruption of the podosome adaptor protein TKS4 (SH3PXD2B) causes the skeletal dysplasia, eye, and cardiac abnormalities of Frank-Ter Haar Syndrome. *Am J Hum Genet* **86**, 254 (Feb 12, 2010).
58. M. A. Kurian *et al.*, Homozygous loss-of-function mutations in the gene encoding the dopamine transporter are associated with infantile parkinsonism-dystonia. *J Clin Invest* **119**, 1595 (Jun, 2009).
59. M. J. Weiss *et al.*, A missense mutation in the human liver/bone/kidney alkaline phosphatase gene causing a lethal form of hypophosphatasia. *Proc Natl Acad Sci U S A* **85**, 7666 (Oct, 1988).
60. S. P. Brooks *et al.*, Identification of the gene for Nance-Horan syndrome (NHS). *J Med Genet* **41**, 768 (Oct, 2004).
61. J. B. Courcet *et al.*, The DYRK1A gene is a cause of syndromic intellectual disability with severe microcephaly and epilepsy. *J Med Genet* **49**, 731 (Dec, 2012).
62. M. Sun *et al.*, Mucolipidosis type IV is caused by mutations in a gene encoding a novel transient receptor potential channel. *Hum Mol Genet* **9**, 2471 (Oct 12, 2000).
63. J. J. Johnston *et al.*, Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am J Hum Genet* **86**, 743 (May 14, 2010).
64. W. D. Jones *et al.*, De novo mutations in MLL cause Wiedemann-Steiner syndrome. *Am J Hum Genet* **91**, 358 (Aug 10, 2012).
65. F. F. Hamdan *et al.*, Mutations in SYNGAP1 in autosomal nonsyndromic mental retardation. *N Engl J Med* **360**, 599 (Feb 5, 2009).

66. S. Jay *et al.*, The fibroblast growth factor receptor 2 p.Ala172Phe mutation in Pfeiffer syndrome--history repeating itself. *Am J Med Genet A* **161A**, 1158 (May, 2013).
67. Z. Iqbal *et al.*, Disruption of the podosome adaptor protein TKS4 (SH3PXD2B) causes the skeletal dysplasia, eye, and cardiac abnormalities of Frank-Ter Haar Syndrome. *Am J Hum Genet* **86**, 254 (Feb 12, 2010).
68. M. A. Kurian *et al.*, Homozygous loss-of-function mutations in the gene encoding the dopamine transporter are associated with infantile parkinsonism-dystonia. *J Clin Invest* **119**, 1595 (Jun, 2009).
69. J. M. Schwarz, C. Rodelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575 (Aug, 2010).
70. C. Draguet, Y. Gillerot, E. Mornet, [Childhood hypophosphatasia: a case report due to a novel mutation]. *Arch Pediatr* **11**, 440 (May, 2004).
71. M. Spentchian *et al.*, Characterization of missense mutations and large deletions in the ALPL gene by sequencing and quantitative multiplex PCR of short fragments. *Genet Test* **10**, 252 (Winter, 2006).
72. I. Brun-Heath *et al.*, Delayed transport of tissue-nonspecific alkaline phosphatase with missense mutations causing hypophosphatasia. *Eur J Med Genet* **50**, 367 (Sep-Oct, 2007).
73. S. P. Brooks *et al.*, Identification of the gene for Nance-Horan syndrome (NHS). *J Med Genet* **41**, 768 (Oct, 2004).
74. J. B. Courcet *et al.*, The DYRK1A gene is a cause of syndromic intellectual disability with severe microcephaly and epilepsy. *J Med Genet* **49**, 731 (Dec, 2012).
75. M. Sun *et al.*, Mucolipidosis type IV is caused by mutations in a gene encoding a novel transient receptor potential channel. *Hum Mol Genet* **9**, 2471 (Oct 12, 2000).
76. M. K. Raychowdhury *et al.*, Molecular pathophysiology of mucolipidosis type IV: pH dysregulation of the mucolipin-1 cation channel. *Hum Mol Genet* **13**, 617 (Mar 15, 2004).
77. K. W. Gripp *et al.*, Long-term survival in TARP syndrome and confirmation of RBM10 as the disease-causing gene. *Am J Med Genet A* **155A**, 2516 (Oct, 2011).
78. M. G. Reese, F. H. Eeckman, D. Kulp, D. Haussler, Improved splice site detection in Genie. *J Comput Biol* **4**, 311 (Fall, 1997).

Acknowledgements: The authors thank the patients and their families for taking part in this study. Funding: The study was supported by grants from the Bundesministerium für Bildung und Forschung (BMBF project number 0313911), core infrastructure funding from the Wellcome Trust, NIH 1R24OD011883-02, and by the Director, Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under contract no. DE-AC02-05CH11231, and a grant to SM by the Max Planck Foundation (MPF).

Author contributions: TZ, SM, NCØ, DH, PNR participated in drafting/or revising the manuscript; TZ, DH, DS, SM, PNR designed the study; TZ, SK, LM, MJ, JH, PK, LGN, SD, NE, MS, NCØ, MRS, UK, GF, BF, UK, RF, AA, YM, SEL, MH, DS, DS, SM, PNR participated

in the acquisition and/or analysis of data; TZ, DH, SM, PNR, GF provided administrative, technical or supervisory support.

Competing interests: None.

Data and materials availability

The PhenIX server is freely available for academic use at http://compbio.charite.de/PhenIX/

The HPO is freely available for all users at: http://www.human-phenotype-ontology.org

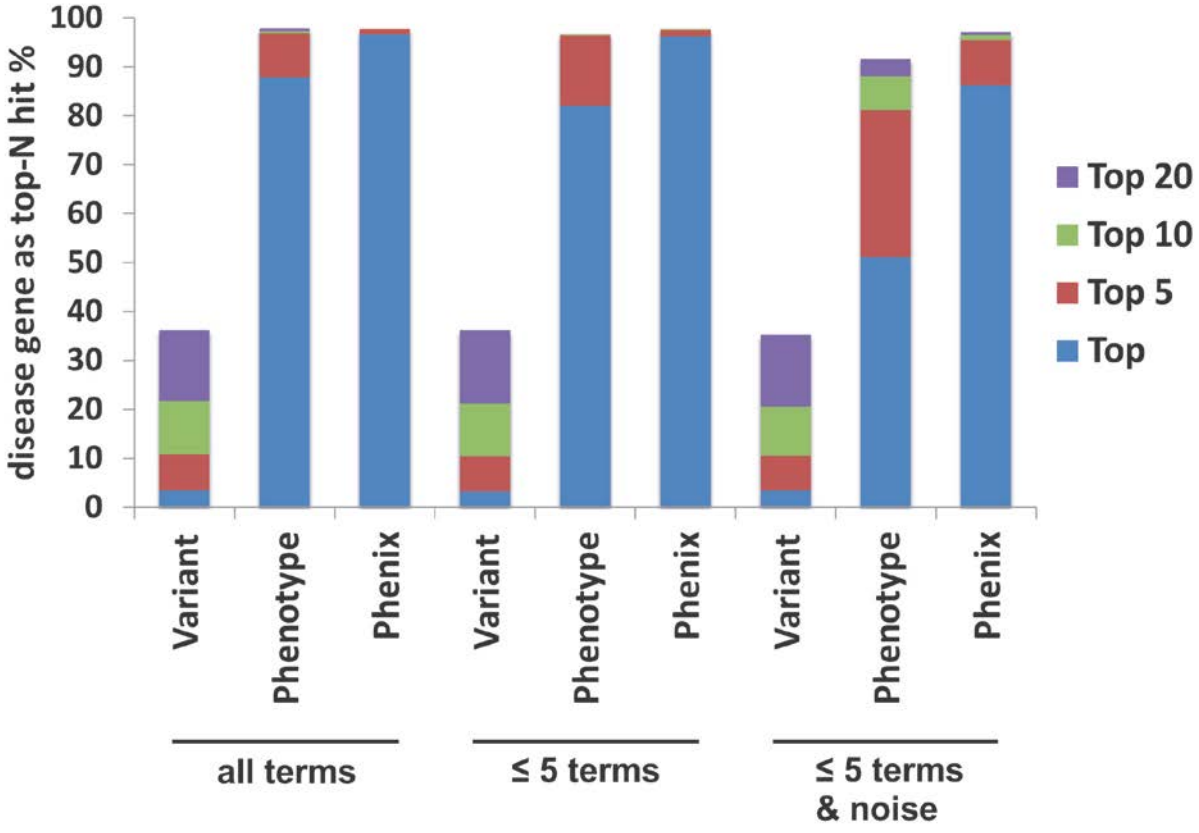
disease gene as top-N hit %
100
90
80
70
60
50
40
30
20
10
0
Variant
Phenotype
Phenix
Variant
Phenotype
Phenix
Variant
Phenotype
Phenix
all terms
≤ 5 terms
≤ 5 terms & noise
Top 20
Top 10
Top 5
Top

Figure 1

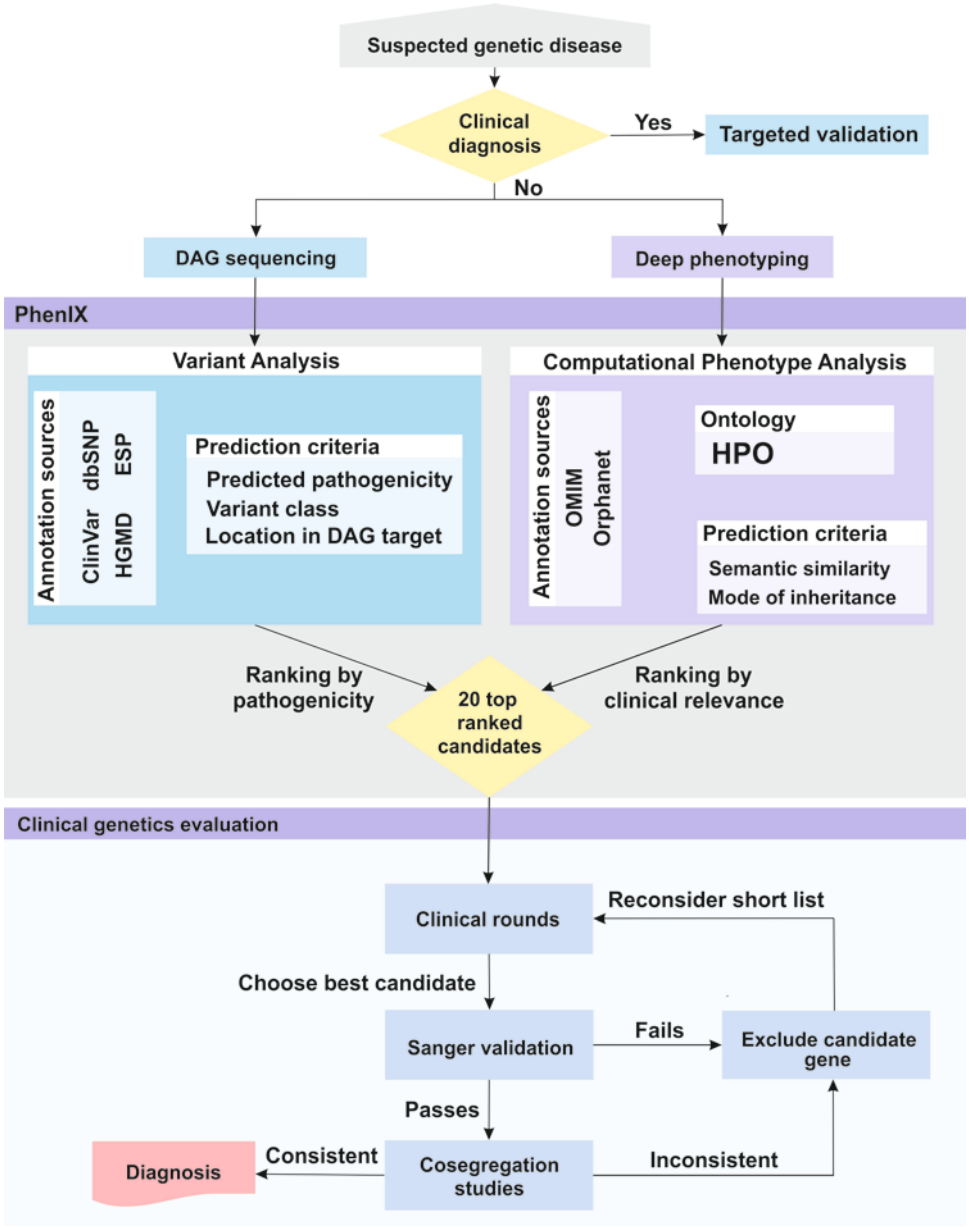
Suspected genetic disease
Clinical diagnosis
Yes
Targeted validation
No
DAG sequencing
Deep phenotyping
PhenIX
Variant Analysis
Annotation sources
dbSNP
ESP
ClinVar
HGMD
Prediction criteria
Predicted pathogenicity
Variant class
Location in DAG target
Computational Phenotype Analysis
Annotation sources
OMIM
Orphanet
Ontology
HPO
Prediction criteria
Semantic similarity
Mode of inheritance
Ranking by pathogenicity
20 top ranked candidates
Ranking by clinical relevance
Clinical genetics evaluation
Clinical rounds
Reconsider short list
Choose best candidate
Sanger validation
Fails
Exclude candidate gene
Passes
Diagnosis
Consistent
Cosegregation studies
Inconsistent

Figure 2