


Evaluation of phenotype-driven gene prioritization methods for Mendelian diseases

Xiao Yuan, Jing Wang, Bing Dai, Yanfang Sun, Keke Zhang, Fangfang Chen, Qian Peng, Yixuan Huang, Xinlei Zhang, Junru Chen, Xilin Xu, Jun Chuan, Wenbo Mu, Huiyuan Li, Ping Fang, Qiang Gong and Peng Zhang 

Corresponding authors: Peng Zhang, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China. Tel.: +86-10-59616161; E-mail: zhangpengdyx@163.com; Qiang Gong, Changsha KingMed Center for Clinical Laboratory, Changsha 410000, China. Tel.: +86-139-7580-6193; E-mail: hn-gongqiang@kingmed.com.cn

Abstract

It's challenging work to identify disease-causing genes from the next-generation sequencing (NGS) data of patients with Mendelian disorders. To improve this situation, researchers have developed many phenotype-driven gene prioritization methods using a patient's genotype and phenotype information, or phenotype information only as input to rank the candidate's pathogenic genes. Evaluations of these ranking methods provide practitioners with convenience for choosing an appropriate tool for their workflows, but retrospective benchmarks are underpowered to provide statistically significant results in their attempt to differentiate. In this research, the performance of ten recognized causal-gene prioritization methods was benchmarked using 305 cases from the Deciphering Developmental Disorders (DDD) project and 209 in-house cases via a relatively unbiased methodology. The evaluation results show that methods using Human Phenotype Ontology (HPO) terms and Variant Call Format (VCF) files as input achieved better overall performance than those using phenotypic data alone. Besides, LIRICAL and AMELIE, two of the best methods in our benchmark experiments, complement each other in cases with the causal genes ranked highly, suggesting a possible integrative approach to further enhance the diagnostic efficiency. Our benchmarking provides valuable reference information to the computer-assisted rapid diagnosis in Mendelian diseases and sheds some light on the potential direction of future improvement on disease-causing gene prioritization methods.

Keywords: benchmarking, gene prioritization, HPO, Mendelian diseases

Introduction

Mendelian diseases, or so-called genetic disorders, impact about 8% of the population in the world [1]. The successful deciphering of the human genome accelerates

human cognition of the association between genes and Mendelian diseases. The next-generation sequencing technology brought a new dawn to the field of Mendelian disease molecular diagnosis due to its high throughput

Xiao Yuan is a senior bioinformatician of Changsha KingMed Center for Clinical Laboratory. He works in high-throughput sequencing data analysis of genetic diseases and cancer.

Jing Wang is a certified senior genetic analyst of Changsha KingMed Center for Clinical Laboratory. She focuses on molecular diagnosis of inherited neuromuscular disorders.

Bing Dai is the general manager of the Diagnostic Laboratory of Changsha KingMed Center for Clinical Laboratory. She is in charge of the management and use of laboratory data.

Yanfang Sun is a certified genetic analyst of Changsha KingMed Center for Clinical Laboratory. She focuses on molecular diagnosis of inherited pediatric neurological disorders.

Keke Zhang is a certified genetic analyst of Changsha KingMed Center for Clinical Laboratory. She focuses on molecular diagnosis of inherited endocrine disorders.

Fangfang Chen is a certified genetic analyst of Changsha KingMed Center for Clinical Laboratory. She focuses on molecular diagnosis of inherited reproductive disorders.

Qian Peng is a junior bioinformatician of Changsha KingMed Center for Clinical Laboratory. He works in high-throughput sequencing data analysis of infectious diseases.

Yixuan Huang is a senior bioinformatician of Beijing Geneworks Technology Co., Ltd.

Xinlei Zhang is the CEO of Beijing Geneworks Technology Co., Ltd.

Junru Chen is a senior bioinformatician of Reproductive & Genetics Hospital of Citic&Xiangya.

Xilin Xu is a senior bioinformatician of Reproductive & Genetics Hospital of Citic&Xiangya.

Jun Chuan is the leader of bioinformatics of Genetalks Biotech. Co., Ltd. He works in big data analysis.

Wenbo Mu is the leader of bioinformatics of Guangzhou Kingmed Center for Clinical Laboratory. His team is currently working on an intelligent analysis system for sequencing data interpretation and visualization.

Huiyuan Li is the vice president of Guangzhou Kingmed Center for Clinical Laboratory. His team focuses on translational medicine in the field of molecular diagnosis.

Ping Fang is the vice president of Guangzhou Kingmed Center for Clinical Laboratory, and a certified senior genetic analyst. She is a fellow of the American College of Medical Genetics and Genomics (ACMG). She is in charge of R&D of molecular diagnostic technology.

Qiang Gong is the director of Diagnosis Center for Genetic Diseases and Cancer of Changsha KingMed Center for Clinical Laboratory, and a certified senior genetic analyst. His team focuses on molecular diagnosis of genetic diseases and cancer.

Peng Zhang is a professor at the Beijing Key Laboratory for Genetics of Birth Defects, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health.

Received: November 15, 2021. **Revised:** January 10, 2022. **Accepted:** January 13, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

in sequencing DNA fragments. Ng SB and his colleagues used whole-exome sequencing (WES) to first discover the causal gene of a Mendelian disease called Miller syndrome in 2010 [2]. Since then, WES and whole-genome sequencing (WGS) are widely used to seek or identify disease-causing genes (variants) in patients with congenital abnormalities. However, the average positive diagnosis rates of these two approaches are 36 and 41% respectively [3], mainly because the molecular basis of a considerable part of the known Mendelian diseases is still unclear [4]. In the meantime, novel disease-gene associations are being identified every year due to the perseverance of the whole scientific community [5]. To date, about 5800 single-gene disorders caused by roughly 4000 genes are diagnosable at a molecular level according to the statistics in OMIM (<https://omim.org/statistics/geneMap>).

In NGS-based molecular diagnosis, finding the true causal variants among hundreds of thousands of variants is time-consuming. Researchers have developed numerous algorithms or software to predict the pathogenicity of the variants [6]. These genotype-based computational predictive tools could be classified into three categories [7]: (i) function-prediction methods that predict the probability of a given missense variant leading to dysfunction of the protein. SIFT [8], PolyPhen-2 [9] and MutationTaster [10] could be representative of this kind of method; (ii) conservation methods including PhyloP [11], SiPhy [12] and GERP++ [13] measure the conservativeness of a given nucleotide site across multiple species; (iii) ensemble methods that create multiple models and then combine them to produce results which are usually more accurate than that of a single model. CADD [14], M-CAP [15], and REVEL [16] all belong to this kind of method. Until now, genotype-based predictions and annotations have been an indispensable procedure for common variants filtration and candidate variants selection.

Initiated in 2007 [17], the Human Phenotype Ontology (HPO) is a standardized vocabulary of phenotypic abnormalities containing about 13 000 terms associated with more than 7000 diseases and is increasingly being adopted as a standard for clinical synopsis by international organizations, worldwide clinical labs especially those focusing on genetic disorders and numerous biomedical resources, guidelines, and software [18]. The expansion and optimization of HPO terms directly promoted the emergence and development of phenotype-based gene prioritization methods. These methods usually compare the phenotype of a patient with a curated knowledge base that consists of associations between phenotypes, genes and diseases, and then give lists of prioritized candidate genes. The phenotype-based gene prioritization method could be divided into two kinds according to the input. Those using both genotype and phenotype information of a patient as input include eXtasy [19], Phevor [20], Phen-Gen [21], PHIVE [22], PhenIX

[23], Exomiser [24], OVA [25], VarElect [26], Omimexplorer [27], QueryOR [28], PDR [29], PhenoVar [30], DeepPVP [31], PhenoPro [32], Phenoxome [33], Xrare [34], eDiVA [35], MutationDistiller [36], AMELIE [37] and LIRICAL [38]. Tools accepting only HPO (s) include Phenomizer [39], Phenolyzer [40], Phrank [41], PhenoRank [42], HANRD [43], GADO [44] and Phen2Gene [45]. Driven by phenotypic features, these *in silico* prioritization methods substantially improve the performance of NGS analytic pipelines in the identification of disease-causing genes [46].

Data for benchmarking on phenotype-driven gene prioritization methods should contain as many real patient cases as possible, each with a single diagnosed causal gene through a stringent assessment procedure for gene pathogenicity and expert-curated HPO terms from the medical record or other clinical information [45]. Moreover, for methods using both phenotype and genotype as input a VCF file produced by a bioinformatic pipeline is required. Methods to be evaluated are run using these real cases and whether a method can properly rank the causal genes highly could simply be regarded as an indication of performance for the evaluation. Bone *et al.* [47] evaluated the performance of Exomiser using simulated and real data from the National Institutes of Health Undiagnosed Diseases Program (UDP) and found that Exomiser ranked the causative variants within the top 10 variants for all 11 previously diagnosed cases. Pengelly *et al.* [48] evaluated the performance of PhenIX, Exomiser, OVA and eXtasy using 21 exomes with known causal variants identified by traditional clinical evaluation and concluded that PhenIX was the most effective method, ranking the causal variant within the top 10 in 85% of total cases. Ebiki *et al.* [49] evaluated the performance of PhenIX, hiPHIVE, Phen-Gen and eXtasy using both simulated data of 100 diseases and real in-house data of 20 Japanese patients, and showed the detection rates of the top most causal variant were 71.4% for PhenIX and 65.0% for hiPHIVE. Cipriani *et al.* [50] evaluated the performance of Exomiser using 134 cases with a range of rare retinal disorders and known causal variants and reported that Exomiser ranked the causal variants as the top candidate in 74% of total cases and top five in 94%.

These pioneering benchmarking works have limitations in the amount of both methods for benchmarking and samples for the evaluation. Thus, these results may miss some up-to-date competitive methods and also may not reflect the true performance of the evaluated methods because of the limited sample size. Based on this situation, we conducted the most comprehensive benchmarking research that evaluated 10 carefully selected methods by using 305 cases from the DDD project [51] and 209 in-house cases from Changsha KingMed Center for Clinical Laboratory (CKCCL) via a relatively unbiased methodology, aiming at providing valuable reference information to the computer-assisted rapid molecular diagnosis in Mendelian diseases.

Materials and method

Datasets curation

For this benchmarking research, phenotypic and genotypic data of 305 positive-diagnosed patients with neurodevelopmental disorders and congenital anomalies in the DDD project [51] and an in-house cohort representing a wide range of phenotypic abnormalities were compiled respectively. DDD project dataset with accessions of EGAD00001001355 (VCF files of 1133 trios) and EGAD00001001413 (HPO terms of 1133 trios) was downloaded via European Genome-Phenome Archive (EGA) [52] download client application with authorization. The list file, containing a single causal gene of each proband, was obtained by additional application. Each input VCF file in our in-house dataset was generated via BWA-GATK bioinformatic pipeline after the exonic DNA fragments of the blood sample were captured using xGen™ Exome Research Panel kit and sequenced by Illumina Nextseq 550 sequencer.

Performance evaluation

Each method was run using default parameters described in its users manual or official website. For each case in the two datasets, the exact rank of the known causal gene was recorded for each method. The proportions of cases with causal genes ranked in top-1, and within top-5, -10, -20, -30, -40 and -50 were calculated respectively for each method.

Statistical visualization

The cumulative distribution function (CDF) curve, bar plot, lollipop chart and pie plot presented in this work were performed by using the 'ggplot2' package in R software (version 4.0). The Venn diagrams were plotted using the 'venneuler' package in R software. The distribution plots illustrate the percentage of cases with causal genes ranked in top-1, and within top-5, -10, -20, -30, -40 and -50 by each method. The CDF plots illustrate the percentage of cases with causal genes ranked within the top k by each method. k could be any integer between 1 and 50 (inclusive) and this cumulative display is better for the visualization of results generated under continuously changing conditions than the regular distribution plots. The bar plots illustrate the relative proportion of each group involving cases with causal genes ranked within a designated range.

Results

Overview of curated datasets and selected methods

We benchmarked the gene prioritization performance of a total of 10 phenotype-based methods for Mendelian diseases based on two curated datasets, including the DDD dataset ($N=305$) and an in-house KingMed Changsha Genetic Diseases (KMCGD, $N=209$) dataset (Figure 1). DDD dataset is recognized as the gold standard data of developmental disorder research. A total of 305 proband

cases were selected for this benchmarking and each of them carried a single nucleotide variant (SNV) or insertion and deletion (INDEL) causal variant which was rated as 'definitely pathogenic' by the DDD project. On average, each case in the final set was characterized by 7.5 phenotypic terms and the corresponding VCF file contained 100 033 variants (Table 1). In addition to the DDD dataset, we built a real-world clinical dataset by collecting 209 patients with a wide range of syndromes as an additional cohort for benchmarking. The recruited patients who received molecular diagnoses of the WES approach at CKCCL between 2018 and 2021 were chosen according to these three criteria: (i) the patient consented to providing his/her genotype and phenotype information for research purposes; (ii) the sequence variant interpretation procedure was performed stringently under the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) guidelines [53] and its refinement and updated recommendations [54–57] and a single Sanger-confirmed pathogenic gene (Supplementary Table 1) was reported in the final clinical interpretive report; (iii) all the HPO terms of the case (Supplementary Table 1) were assigned unambiguously after a manual investigation of the patient's medical record by one of the four well-trained and certified genetic analysts and reviewed independently by another two senior genetic analysts. Any ambiguous term was removed and this 'putting quality before quantity' criterion made the average number of HPO terms per case at 2.0. The average number of variants in the original input VCF files of the KMCGD dataset was 83 587 (Table 1).

About 30 software or algorithms for causal-gene prioritization of Mendelian diseases developed in the last decade were evaluated preliminarily (Supplementary Table 2), and 10 of them were selected for this benchmarking (Table 2) according to whether the approach is available currently, updated periodically and free for academic use. The designated software used for this benchmarking work could be divided into two categories: those using a VCF file and HPO(s) as input ('HPO + VCF' methods) including PhenIX [23], Exomiser [24], DeepPVP [31], Xrare [34], AMELIE [37] and LIRICAL [38]; and software accepting only HPO(s) ('HPO-only' methods) including Phenolyzer [40], HANRD [43], GADO [44] and Phen2Gene [45]. It should be pointed out that AMELIE could be run using HPO + VCF mode or HPO only mode (hereinafter referred to as 'AMELIE_HPO') and the performance of both modes was evaluated respectively in this research.

PhenIX [23] filtered and ranked the candidate's genes according to the combination of a variant score which indicated variant rarity, pathogenicity and phenotype score representing the potential clinical relevance of the gene harboring the variants. Exomiser [24] comprised a suite of algorithms including PhenIX mentioned above. As with the default prioritization called hiPHIVE, the phenotypic similarity was calculated with human phenotypic data as well as mouse and zebra-fish data,

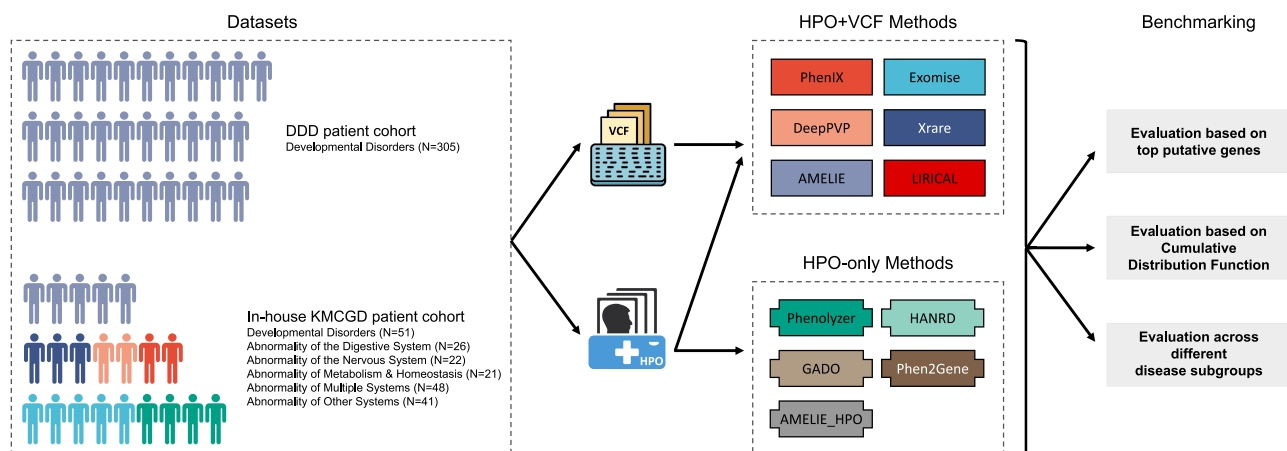


Figure 1. Illustration of study workflow. Flowchart of data collection and method implementation in this work. DDD patient cohort includes 305 cases with developmental disorders (represented as light blue) while the in-house KMCGD patient cohort involves a total of 209 cases with a wide range of syndromes (represented as various colors). Then, curated HPO terms and a VCF file of each case in both cohorts are imported into six ‘HPO + VCF’ prioritization methods. Additionally, curated HPO terms of each case are imported into five ‘HPO-only’ prioritization methods. In particular, AMELIE is run in both ‘HPO + VCF’ mode and ‘HPO-only’ mode (AMELIE_HPO). Finally, for each case, the ranking position of the known causal gene in the gene list output by each method is recorded, based on which the performance of each method is evaluated.

Table 1. Overview of the datasets used in this work; the basic information, gender composition, age composition, frequent causal genes and disease subgroup composition of the DDD and KMCGD dataset

| | | DDD | KMCGD |
|--|------------------------|--------------------------------|-------------|
| Basic information | Size | 305 | 209 |
| | Average HPO amount | 7.5 | 2.0 |
| | Average variant amount | 100 033 | 83 587 |
| Gender | Male | 141 (46.2%) | 125 (59.8%) |
| | Female | 164 (53.8%) | 84 (40.2%) |
| Age (years) | 0–1 | 3 | 61 |
| | 1–7 | 148 | 30 |
| | 7–18 | 154 | 51 |
| | 18–65 | – | 65 |
| | 65+ | – | 2 |
| Gene (frequency) | >9 | ARID1B (11) | ATP7B (37) |
| | 9 | – | SRD5A2 |
| | 8 | MED13L | – |
| | 7 | ANKRD11, SYNGAP1 | – |
| | 6 | KCNQ2, SATB2, SCN2A | UGT1A1 |
| | 5 | CTNNB1, DDX3X, PPP2R5D, STXBP1 | PAH |
| Developmental disorder | | 305 | 51 |
| Abnormality of Metabolism and Homeostasis | | – | 21 |
| Abnormality of the Digestive System | | – | 26 |
| Abnormality of the Nervous System | | – | 22 |
| Abnormality of Multiple Systems | | – | 48 |
| Abnormality of Other Systems | | – | 41 |

allowing novel candidate genes discovery. LIRICAL [38] required the library files of Exomiser. It calculated the likelihood ratio (LR) for each phenotype and genotype input and measured how much any individual phenotypic observation had contributed to the prioritization result. DeepPVP combined automated inference with deep neural networks to classify and identify the likely causative variants [31]. Xrare was developed based on a phenotype-similarity scoring method. The Xrare model was trained and validated using known pathogenic variants from ClinVar. AMELIE was an online application helping Mendelian diagnosis by matching patient phenotype and genotype to primary literature [37]. Via parsing a huge amount of abstracts and full-text articles in

PubMed, AMELIE used natural language processing (NLP) to construct a homogeneous knowledge base for causal gene prioritization. Phenolyzer [40] and Phen2Gene [45] came from the same lab. The former used prior information to implicate genes involved in diseases with a machine learning model. The latter was an enhanced version of the former and calculated a prioritized gene list based on a probabilistic model. HANRD was a prioritization approach using heterogeneous networks in the context of rare diseases [43]. Researchers of this software developed a graph convolution-based technique to infer new phenotype-gene associations. GADO [44] predicted which genes cause specific phenotypes based on the public RNA sequencing dataset.

Table 2. Overview of the methods evaluated in this work; the brief features, running time, version numbers and the released time of the 10 methods

| Input | Method | Feature | Time | Version | Year |
|-----------|------------|--|-------|-------------|-----------|
| HPO + VCF | PhenIX | Computational phenotype analysis | 103 s | 1.16 | 2014 [23] |
| | Exomiser | Cross-species phenotype comparison | 35 s | 12.1.0 | 2015 [24] |
| | DeepPVP | Deep learning | 280 s | 2.1 | 2019 [31] |
| | Xrare | Machine learning | 260 s | pub:2015 | 2019 [34] |
| | AMELIE | Text mining and natural language processing | 94 s | Oct 5, 2020 | 2020 [37] |
| | LIRICAL | Likelihood ratio framework | 31 s | 1.3.0 | 2020 [38] |
| HPO only | Phenolyzer | Machine learning | 107 s | 0.4.0 | 2015 [40] |
| | HANRD | Heterogeneous networks and graph convolution | 628 s | – | 2018 [43] |
| | GADO | Gene network based on transcriptome data | 6 s | 1.0.1 | 2019 [44] |
| | Phen2Gene | Probabilistic model | 6 s | 1.2.3 | 2020 [45] |

The running time is tested under the default setting of each software using a VCF file (29 968 variants, size: 23.87 MB) of case NA12878 from the Genome in a Bottle project and a set of randomly chosen HPO terms including HP:0000002 (Abnormality of body height), HP:0003020 (Enlargement of the wrists), HP:0006089 (Palmar hyperhidrosis), HP:0009023 (Abdominal wall muscle weakness) and HP:0012047 (Hemeralopia).

Benchmarking phenotype-driven gene prioritization methods for clinical data sets

We employed the selected 10 methods to estimate prioritized genes based on the two curated datasets and evaluated the accuracy used the top 1, -5, -10, -20, -30, -40 and -50 putative genes on all the phenotype-driven gene prioritization tasks. As shown in Figure 2, the performance of each method varied greatly in our benchmarking experiments. In general, three methods including LIRICAL, AMELIE and Xrare outperformed all the other algorithms and showed a better performance across all putative gene percentage experiments. In both DDD and KMCGD datasets, AMELIE stood out at the top-1, -5 and -10 experiments, while the LIRICAL method ranked the best at the top-30, -40 and -50 experiments. Specifically, AMELIE correctly assigned the causal gene at the very top in 47.9% of the total 305 DDD cases and 57.9% of the total 209 KMCGD cases (Supplementary Table 3). Besides, AMELIE identified the causal gene within the top-5 and -10 candidates in about 81.3 and 86.2% of the total DDD cases and about 73.7 and 78.9% of the total KMCGD cases. While LIRICAL captured the causal genes within the top-30, -40 and -50 for about 94.4, 95.4 and 96.1% of the DDD cohort, and about 86.6, 88.5 and 90.4% of the KMCGD cohort (Supplementary Table 3). As with the top-20 experiment, AMELIE outperformed LIRICAL in DDD dataset (90.5 versus 89.8%) but LIRICAL did better in KMCGD dataset (84.2 versus 81.3%). Xrare ranked third across all the different settings in the DDD dataset and three of seven experiments in the KMCGD dataset. The performance of Exomiser and PhenIX software followed that of these three methods under most circumstances but exceeded that of both AMELIE and Xrare at the top-30, -40 and top-50 settings in the KMCGD dataset. The other six methods including DeepPVP and all five ‘HPO-only’ methods (Phenolyzer, HANRD, GADO, Phen2Gene, and AMELIE_HPO) had an overall lower performance on all the benchmarking settings. The hits percentage of these six methods were all lower than 50% even at the top 50 putative genes based on the two curated datasets.

Next, we evaluated the performance of all methods using a random synthetic dataset. A total of 50 artificial cases were synthesized for this simulation experiment (Supplementary Table 4). Each case consisted of (i) a VCF file randomly chosen from the healthy parents of the ‘original’ DDD dataset which contains a totally of 1133 trios [51]; (ii) fake phenotypic features with 3~12 HPO terms randomly chosen from the official HPO pool; (iii) a randomly chosen low-frequency (<1% in gnomAD v2.1.1 Exomes) nonsynonymous mutation which was inserted into the corresponding VCF file as the hypothetical causal variant. In the evaluation result for these 50 negative control cases, all methods showed a very poor performance (Supplementary Figure 1), with some methods capturing few hits by chance.

Performance evaluation based on the cumulative distribution function

The expected value of performance alone was insufficient to properly evaluate an algorithm as it discarded information about the variance in performance. A more useful representation of performance could be obtained by the cumulative frequency analysis of the performance of overall sources of variance, so we investigated the performance of phenotype-driven gene prioritization methods based on the cumulative distribution function (CDF) next. Consistent with the above results, five ‘HPO + VCF’ methods including LIRICAL, AMELIE, Xrare, Exomiser and PhenIX were far more effective than DeepPVP and all ‘HPO-only’ methods in prioritizing disease-causing genes (Figure 3A and B). LIRICAL and AMELIE were the two best performers, and then we made a prioritizing genes comparison between them (Figure 3C). At the top-1 setting, the causal genes of 146 DDD cases and 121 KMCGD cases were correctly put at the very top of the final gene lists generated by AMELIE. In LIRICAL, the corresponding numbers were 125 and 107 respectively. However, the common case of the two software at this setting were only 67 and 78 in DDD and KMCGD cohort respectively (Figure 3C). AMELIE provided 79 and 43 additional cases

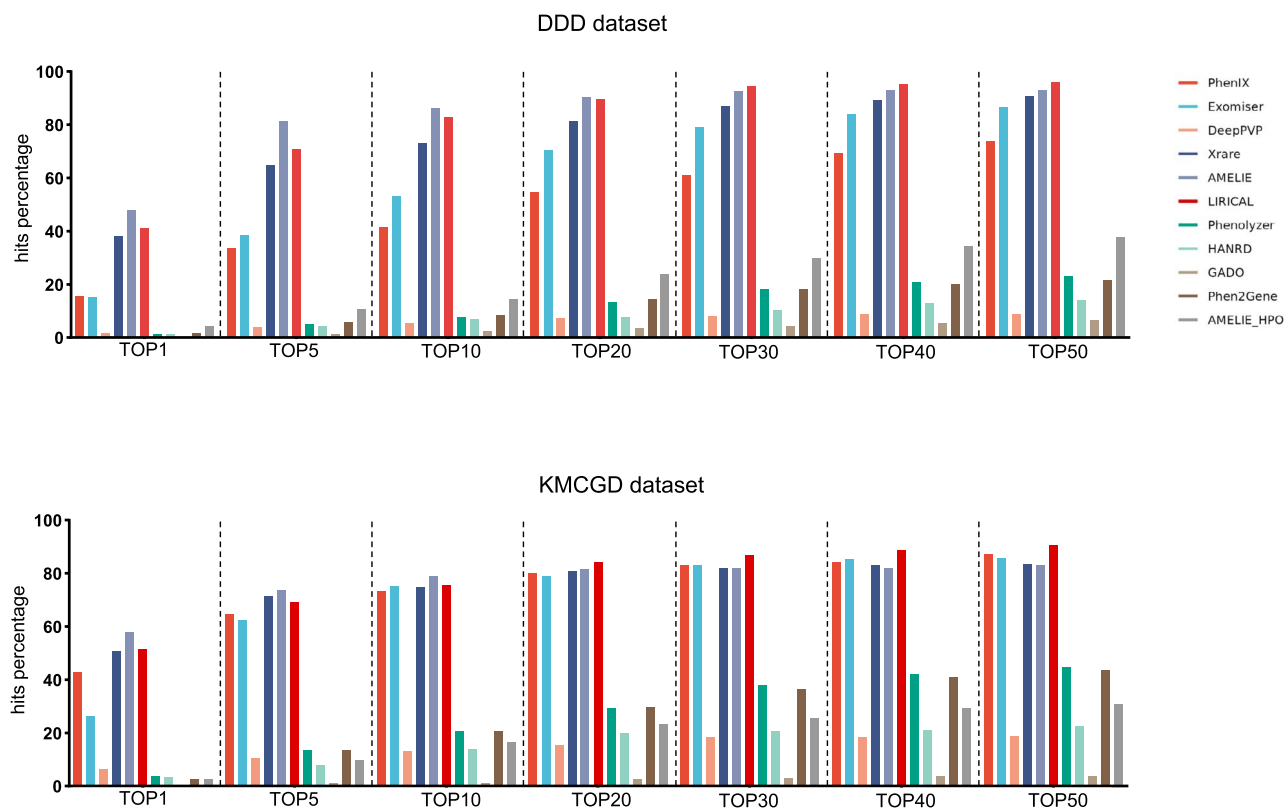


Figure 2. Distribution plots of performance evaluation results. Distribution plots of performance evaluation results of 10 phenotype-driven gene prioritization methods on the DDD (A) and KMCGD (B) datasets. The distribution plots illustrate the percentage of the cases with causal genes ranked in top-1 and within the top-5, -10, -20, -30, -40 and -50 by each method. Each method is represented by a different color.

than the intersections with casual genes ranked at the very top, while LIRICAL presented 58 and 29. The symmetric differences were still considerable at top-5 and top-10 settings but became inconspicuous in the follow-up settings. In the final top-50 settings, the cases of the two software almost overlapped (Figure 3C). These findings suggested that LIRICAL and AMELIE could complement each other in the performance to rank causal genes highly.

Methods using only phenotypic features based on HPO terms had an overall poorer performance than 'HPO + VCF' approaches, suggesting the disadvantage of the 'HPO-only' scoring method in evaluating causal genes. This was in line with expectations because the VCF file provided genotype information to greatly reduce the ranking scope. AMELIE_HPO outshone other methods of the same category at all ranking levels in the DDD dataset. It puts the true gene in top-1, and within top-5, -10, -20, -30, -40 and -50 for about 4.3, 10.8, 14.4, 23.9, 29.8, 34.4 and 37.7% of the total cases (Supplementary Table 3). For the KMCGD dataset, Phenolyzer and Phen2Gene, whose CDF curves twisted each other, had an outstanding performance in phenotype-only methods. This intertwined situation also happened in the DDD dataset, suggesting similar performance of the two methods.

Performance evaluation across different disease subgroups

In the final step, we explored whether the complexity and variability of disease patterns could affect the performance of each prioritization method. According to the HPO constitutions per case and the official HPO hierarchy, KMCGD cases were further divided into five disease subgroups including Abnormality of Metabolism & Homeostasis ($N=21$, 10.0%), Abnormality of the Digestive System ($N=26$, 12.4%), Abnormality of the Nervous System ($N=22$, 10.5%), Abnormality of Multiple Systems ($N=48$, 23.0%), Abnormality of Other Systems ($N=41$, 19.6%) and an unofficially-defined subgroup called Developmental Disorder ($N=51$, 24.4%) (Figure 4A, Supplementary Table 1). The subgroup of Abnormality of Multiple Systems involved cases with HPO terms classified into at least two systems, and the subgroup of Abnormality of Other Systems was a collection of those small groups (less than 20 cases) of other systems such as the cardiovascular system. The whole frequency distribution of the HPO parent classes (defined by the official HPO hierarchy) for the KMCGD dataset is displayed in Figure 4B. As seen in Figure 4C, these performance results showed the five best methods mentioned above (LIRICAL, AMELIE, Xrare, Exomiser and PhenIX) still had an overall superior performance

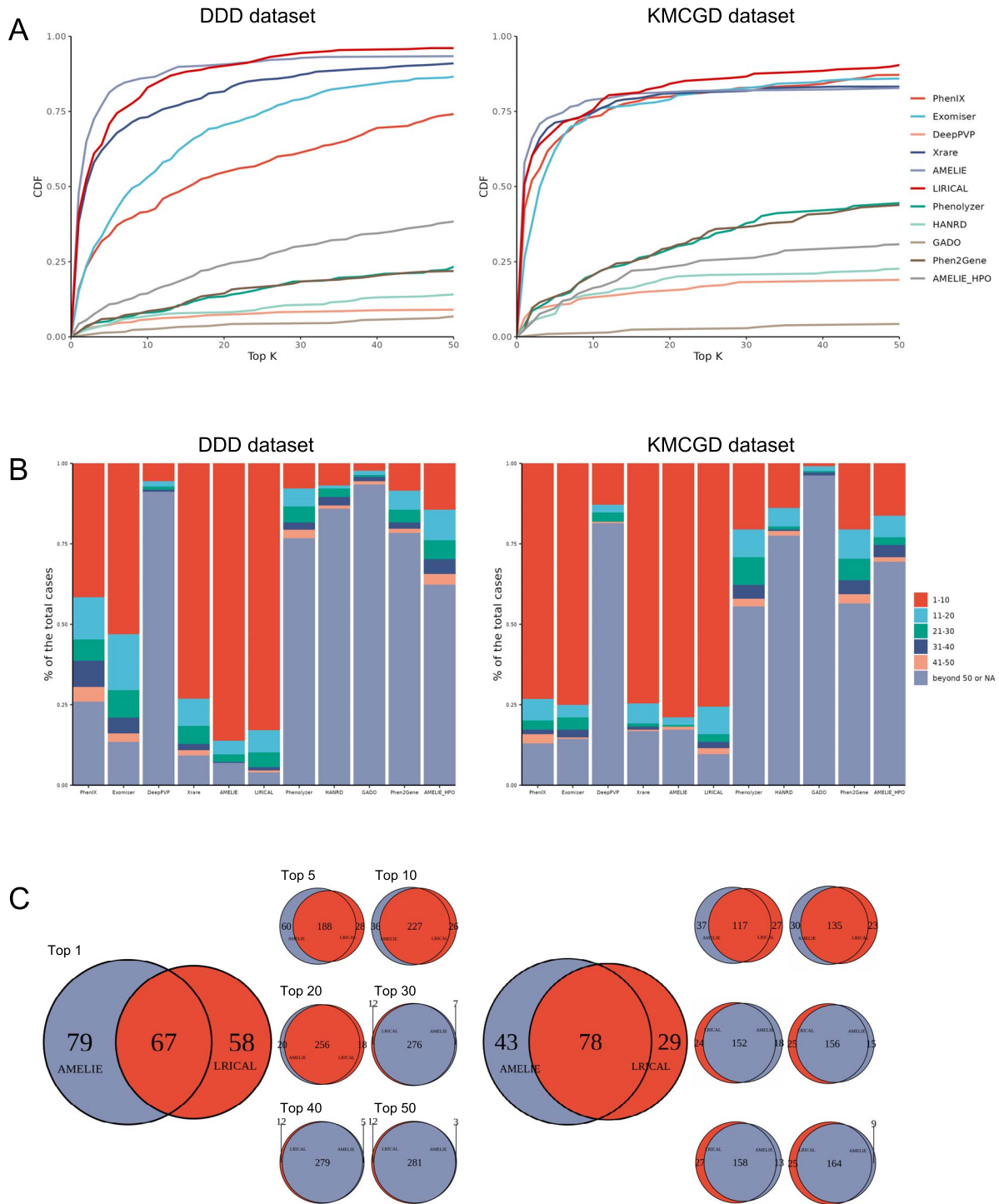


Figure 3. CDF and bar plots of performance evaluation results. CDF plots (A) and bar plots (B) of performance evaluation results of 10 phenotype-driven gene prioritization methods on the DDD (left) and KMCGD (right) datasets. The CDF plots illustrate the percentage of the cases with causal genes ranked within the top k by each method. k could be any integer between 1 and 50 (inclusive). Each method is represented by a different color. The bar plots illustrate the relative proportion of each group involved cases with causal genes ranked within a designated range. Each group is represented by a different color. (C) The overlapping set of cases with causal genes ranked in top-1 and within top-5, -10, -20, -30, -40 and -50 by LIRICAL and AMELIE in DDD (left) and KMCGD (right) dataset.

for the various disease subgroups when compared with other methods, but the prioritization accuracy of each method differed among disease subgroups and some methods even showed a preference. LIRICAL performed best in the subgroup of Development Disorder, Abnormality of the Nervous System and Abnormality of

Other Systems. Especially in the Developmental Disorder phenotypic group, LIRICAL held a significant lead over all other methods. These findings, combined with the performance evaluation in the DDD dataset, suggested that LIRICAL was good at identifying the causal gene in developmental disorders. For the disease subgroups of

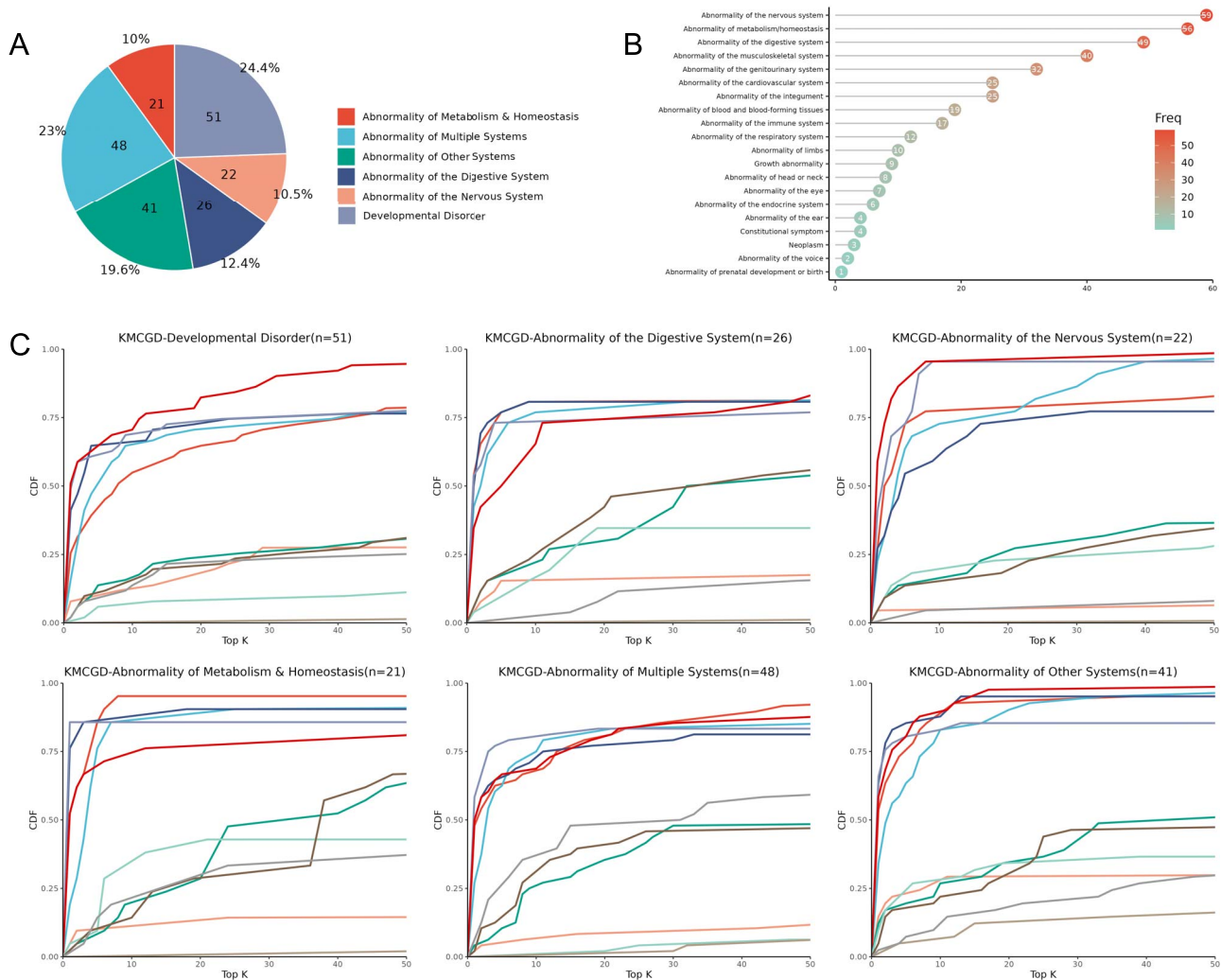


Figure 4. Performance evaluation across different disease subgroups. (A) Frequency distribution of the HPO parent classes for the KMCGD dataset. The HPO terms of each case in the KMCGD dataset are assigned to HPO parent classes according to the official HPO hierarchy and some cases involve more than one kind of HPO parent class. (B) Disease subgroup composition of KMCGD dataset. Case amount and proportion are tagged for each subgroup. (C) CDF plots of performance evaluation results of 10 phenotype-driven gene prioritization methods on each subgroup of the KMCGD dataset.

Abnormality of the Digestive System and Abnormality of Multiple Systems, the performance of the best five was very close. Interestingly, the evaluation result for the Abnormality of Metabolism & Homeostasis subgroup had a distinct pattern when compared with other disease subgroups. The PhenIX method ranked best for this subgroup while the performance of the LIRICAL method was not prominent.

Discussions

Timely and effective diagnosis for patients with Mendelian diseases will bring chances for appropriate approaches to clinical management and treatment. Computer-assisted prioritization methods could substantially improve the performance of NGS-based analysis pipelines to identify disease-causing genes [46]. Objective evaluation of these methods could provide valuable reference information to practitioners, providing convenience

for choosing an appropriate tool for their workflows. Previously published reports [47–50] benchmarked different prioritization methods or a single approach based on real and/or simulated data, but the overall scales of these researches were limited. Here we provided more comprehensive benchmarking analyses using two large-scale datasets with more than 500 real-world patient cases to investigate the performance of as many as 10 methods including some prestigious software and newly developed approaches. Patients in the DDD dataset are undiagnosed children and adults in the UK with developmental disorders while patients in KMCGD dataset are Chinese with a wide range of genetic abnormalities. Thus, we believe that this broader evaluation based on the abundant number of cases could minimize or avoid bias derived from the possible preference for a population or disease-type in a certain method. Three remarkable findings were observed in our results firstly, the methods using HPO terms and

VCF files as input performed better overall than those using phenotypic data alone, which indicated that providing both genotype and phenotype information of a patient to causal-gene prioritization methods would probably get a more effective diagnosis result. Secondly, as the two best performers, LIRICAL and AMELIE had obvious different sets to each other in cases with the causal genes ranked highly in both DDD and KMCGD datasets, suggesting that an integrative approach may further facilitate the pinpoint of targets. Thirdly, disease preference was observed in some methods during the evaluation of disease subgroup data, but more data and further sophisticated experiments are required to support this finding.

In our benchmarking work, the difference between the shapes of the CDF-curve bunch of the five 'HPO + VCF' methods in the two datasets was significant. The bunch was more compact in the DDD dataset. This phenomenon is probably due to the more detailed phenotypic records in the DDD dataset which increase the discriminative power of the experiments, and more evidence is needed to support this hypothesis. When comparing the results of each method, we found that some software failed in a few of the provided cases, which might be due to the untimely update of the internal HPO database, hence these software could not profit from parts of the regularly updated HPO. Future updates of the HPO database of these software may fix this issue. Besides, we found that causal genes including CPLANE1, H4C3, CERT1 and NEXMIF related to six cases of the DDD dataset were represented as their alias (C5orf42, HIST1H4C, COL4A3BP and KIAA2022) in AMELIE and Xrare when checking the gene ranking lists. Improper uses of gene names might slightly influence the exact performance of these two methods and we've contacted the authors about this issue. Moreover, we evaluated all methods by using the default settings and minor modifications, while adjusting these settings for a specific dataset might improve the performances. Note that the HPO terms were manually curated by curators from different institutions. Thus, the differences in both curation criteria, methods and styles, and the background experience and levels of expertise of the curators between the DDD project and CKCGL could bring unavoidable deviations to the evaluation results. For example, some phenotypic descriptions might introduce a few noisy HPO terms due to the relatively looser curation style, and these might eventually affect the ranking performance of some software.

To prioritize disease-causing genes, AMELIE parses a huge number of primary literature to construct a knowledge-based storing the relationship between genes, variants and phenotypes [37]. Thus, AMELIE tends to have an upper hand in identifying targets for published cases (DDD dataset) over unpublished ones (KMCGD dataset). Therefore, we investigated whether the published cases could lead to bias to the performance evaluation of AMELIE. The professional version of the

Human Gene Mutation Database (HGMD) provides up-to-date information on human inherited gene mutations and mutations in the database are manually curated from the scientific literature [58]. So, we removed 156 HGMD (version: Pro-2021.2) included cases from the DDD dataset and used the rest of 149 unpublished DDD cases to re-evaluate the performance of each method. The result showed an almost-unchanged overall trend for the performances of all methods but a slightly lower performance of AMELIE in the top-20 experiment (Supplementary Figure 2). Specifically, it was consistent with the result using the whole DDD dataset (305 cases), AMELIE ranked first at the top-1, -5, and -10 experiments and second at the top-30, -40 and -50 experiments. However, when compared with the result using the whole DDD dataset, AMELIE was surpassed by LIRICAL in the top-20 experiment. According to the results of these extensive analyses, we believe that whether or not the input DDD case is published, it would not bring substantial bias to the performance evaluation of the 10 methods including AMELIE for benchmarking.

Currently, accurately and efficiently extracting phenotype from patients' medical records remains labor-intensive. Artificial intelligence (AI) especially NLP has been applied to automatically extract and normalize HPO terms from electronic health records of patients and shows great power [59–64]. Impressively, PEDIA [65] has tried NGS data interpretation with portrait photographs of patients. This research uses deep-learning-based facial analysis to quantify the phenotypic similarity and proves the possibility to use image analysis to obtain HPO terms from patients' photographs and/or medical images. The evaluation of automatic HPO extractors will be one of our possible research directions in the future. In addition, the interpretation of non-coding variants still constitutes a major challenge in the application of NGS in Mendelian disease. Genomiser [66] as the Exomiser's extended version, is an analysis framework to associate variants in the non-coding genome to specific Mendelian diseases, setting an example to meet the challenge. As the disease-gene knowledge base of human beings expands and bioinformatic technology develops, more tools will emerge to rank the causal variants in both coding and non-coding regions for genetic disease diagnosis. We will compare the prioritization performance of the top methods in our benchmarking work with those well-designed new tools in our future works. We believe our comprehensive evaluation work could provide a guideline for researchers to select and apply suitable prioritization methods in their analysis frameworks and shed some light on the potential direction of future improvement on disease-causing gene prioritization methods.

Key Points

- Comprehensive benchmarking analyses using two large-scale datasets with totally more than 500 real-world cases have been conducted to evaluate the performance

of as many as 10 phenotype-driven gene prioritization methods.

- Methods using HPO terms and VCF files as input have shown better performance than those using phenotypic data alone in causal gene prioritization.
- The two best performers LIRICAL and AMELIE had obvious different sets to each other in cases with the causal genes ranked highly, and this complementarity suggests a possible integrative approach to further enhance the diagnostic efficiency.

Availability of data or materials

The materials of the DDD dataset used for the current study are deposited in the European Genome-phenome Archive(EGA) [52] hosted by the EBI (Study ID EGAS00001000775), and accessible through the application. Other materials are available upon reasonable request from the corresponding author. The analysis codes can be downloaded from: <https://github.com/yuanxiao1988/Benchmarking-of-Phenotype-driven-Gen-Prioritization-Methods>.

Authors' contributions

X.Y. performed the main analysis and prepared the manuscript. J.W., Y.S., K.Z. and F. C. curated the KMCGD dataset. Q.P., J.C. and W.M. performed partial bioinformatic analysis. X.Z., J.C. and X.X. helped to review the HPO terms of each case of the KMCGD dataset. B.D., Y.H., H.L. and P.F. helped to supervise the studies. P.Z. and Q.G. supervised the studies, designed the analysis, and revised the manuscript. All authors reviewed and approved the manuscript.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Acknowledgments

We thank the Deciphering Developmental Disorders(DDD) project [51] for their generous sharing of data. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [HICF-1009-003], and makes use of DECIPHER (<http://www.deciphergenomics.org>), which is funded by Wellcome. X.Y., J.W., B.D., Y.S., K.Z., F.C., Q.P., H.L. and Q.G. are all employees of Changsha Kingmed Center for Clinical Laboratory. X.Y., W.M., H.L., P.F. and Q.G. are all employees of Guangzhou Kingmed Center for Clinical Laboratory. Y.H. and X.Z. are all employees of Beijing Geneworks Technology Co., Ltd.. J.C. is an employee of Genetalks Biotech. Co., Ltd.. All other authors declare no conflict of interest.

Funding

Innovation and Entrepreneurship Technology Investment project of Hunan Province (2019GK5019);

Science and Technology Program of Guangzhou, China (201802030010).

References

1. Baird PA, Anderson T, Newcombe H, et al. Genetic disorders in children and young adults: a population study. *Am J Hum Genet* 1988;**42**:677.
2. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* 2010;**42**:30–5.
3. Clark MM, Stark Z, Farnaes L, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom Med* 2018;**3**:1–10.
4. Chong JX, Buckingham KJ, Jhangiani SN, et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* 2015;**97**:199–215.
5. Boycott KM, Rath A, Chong JX, et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am J Hum Genet* 2017;**100**:695–705.
6. Umlai U-KI, Bangarusamy DK, Estivill X, et al. Genome sequencing data analysis for rare disease gene discovery. *Brief Bioinform* 2022;**23**:bbab363.
7. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* 2015;**24**: 2125–37.
8. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;**4**:1073.
9. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**: 248–9.
10. Schwarz JM, Rödelsperger C, Schuelke M, et al. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;**7**:575–6.
11. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005;**15**:901–13.
12. Garber M, Guttman M, Clamp M, et al. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 2009;**25**:i54–62.
13. Davydov EV, Goode DL, Sirota M, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010;**6**:e1001025.
14. Kircher M, Witten DM, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5.
15. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* 2016;**48**:1581.
16. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;**99**:877–85.
17. Robinson PN, Köhler S, Bauer S, et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet* 2008;**83**:610–5.
18. Köhler S, Vasilevsky NA, Engelstad M, et al. The human phenotype ontology in 2017. *Nucleic Acids Res* 2017;**45**:D865–76.
19. Sifrim A, Popovic D, Tranchevent LC, et al. eXtasy: variant prioritization by genomic data fusion. *Nat Methods* 2013;**10**:1083–4.
20. Singleton MV, Guthery SL, Voelkerding KV, et al. Phevor combines multiple biomedical ontologies for accurate identification of

- disease-causing alleles in single individuals and small nuclear families. *Am J Hum Genet* 2014;**94**:599–610.
21. Javed A, Agrawal S, Ng PC. Phen-gen: combining phenotype and genotype to analyze rare disorders. *Nat Methods* 2014;**11**:935–7.
 22. Robinson PN, Köhler S, Oellrich A, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res* 2014;**24**:340–8.
 23. Zemojtel T, Köhler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med* 2014;**6**:252ra123–3.
 24. Smedley D, Jacobsen JO, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nat Protoc* 2015;**10**:2004–15.
 25. Antanaviciute A, Watson CM, Harrison SM, et al. OVA: integrating molecular and physical phenotype data from multiple biomedical domain ontologies with variant filtering for enhanced variant prioritization. *Bioinformatics* 2015;**31**:3822–9.
 26. Stelzer G, Plaschkes I, Oz-Levi D, et al. VarElect: the phenotype-based variation prioritizer of the GeneCards suite. *BMC Genomics* 2016;**17**(Suppl 2):444.
 27. James RA, Campbell IM, Chen ES, et al. A visual and curatorial approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. *Genome Med* 2016;**8**:13.
 28. Bertoldi L, Forcato C, Vitulo N, et al. QueryOR: a comprehensive web platform for genetic variant analysis and prioritization. *BMC Bioinformatics* 2017;**18**:225.
 29. Kramer A, Shah S, Rebres RA, et al. Leveraging network analytics to infer patient syndrome and identify causal genes in rare disease cases. *BMC Genomics* 2017;**18**:551.
 30. Thuriot F, Buote C, Gravel E, et al. Clinical validity of phenotype-driven analysis software PhenoVar as a diagnostic aid for clinical geneticists in the interpretation of whole-exome sequencing data. *Genet Med* 2018;**20**:942–9.
 31. Boudellioua I, Kulmanov M, Schofield PN, et al. DeepPVP: phenotype-based prioritization of causative variants using deep learning. *BMC Bioinform* 2019;**20**:1–8.
 32. Li Z, Zhang F, Wang Y, et al. PhenoPro: a novel toolkit for assisting in the diagnosis of Mendelian disease. *Bioinformatics* 2019;**35**:3559–66.
 33. Wu C, Devkota B, Evans P, et al. Rapid and accurate interpretation of clinical exomes using Phenoxome: a computational phenotype-driven approach. *Eur J Hum Genet* 2019;**27**:612–20.
 34. Li Q, Zhao K, Bustamante CD, et al. Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. *Genet Med* 2019;**21**:2126–34.
 35. Bosio M, Drechsel O, Rahman R, et al. eDiVA-classification and prioritization of pathogenic variants for clinical diagnostics. *Hum Mutat* 2019;**40**:865–78.
 36. Hombach D, Schuelke M, Knierim E, et al. MutationDistiller: user-driven identification of pathogenic DNA variants. *Nucleic Acids Res* 2019;**47**:W114–20.
 37. Birgmeier J, Haeussler M, Deisseroth CA, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci Transl Med* 2020;**12**:eaau9113.
 38. Robinson PN, Ravanmehr V, Jacobsen JO, et al. Interpretable clinical genomics with a likelihood ratio paradigm. *Am J Hum Genet* 2020;**107**:403–17.
 39. Köhler S, Schulz MH, Krawitz P, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet* 2009;**85**:457–64.
 40. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods* 2015;**12**:841–3.
 41. Jagadeesh KA, Birgmeier J, Guturu H, et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet Med* 2019;**21**:464–70.
 42. Cornish AJ, David A, Sternberg MJE. PhenoRank: reducing study bias in gene prioritization through simulation. *Bioinformatics* 2018;**34**:2087–95.
 43. Rao A, Saipradeep V, Joseph T, et al. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Med Genet* 2018;**11**:1–12.
 44. Deelen P, van Dam S, Herkert JC, et al. Improving the diagnostic yield of exome-sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. *Nat Commun* 2019;**10**:1–13.
 45. Zhao M, Havrilla JM, Fang L, et al. Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom Bioinform* 2020;**2**:lqaa032.
 46. Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med* 2015;**7**:81.
 47. Bone WP, Washington NL, Buske OJ, et al. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet Med* 2016;**18**:608–17.
 48. Pengelly RJ, Alom T, Zhang Z, et al. Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. *Sci Rep* 2017;**7**:13509.
 49. Ebiki M, Okazaki T, Kai M, et al. Comparison of causative variant prioritization tools using next-generation sequencing data in Japanese patients with Mendelian disorders. *Yonago Acta Med* 2019;**62**:244–52.
 50. Cipriani V, Pontikos N, Arno G, et al. An improved phenotype-driven tool for rare Mendelian variant prioritization: benchmarking exomiser on real patient whole-exome data. *Gen* 2020;**11**:460.
 51. Deciphering Developmental Disorders S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* 2015;**519**:223–8.
 52. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European genome-phenome archive of human data consented for biomedical research. *Nat Genet* 2015;**47**:692–5.
 53. Richards S, Aziz N, Bale 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**:405–24.
 54. Abou Tayoun AN, Pesaran T, DiStefano MT, et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum Mutat* 2018;**39**:1517–24.
 55. Ghosh R, Harrison SM, Rehm HL, et al. Updated recommendation for the benign stand-alone ACMG/AMP criterion. *Hum Mutat* 2018;**39**:1525–30.
 56. Biesecker LG, Harrison SM. The ACMG/AMP reputable source criteria for the interpretation of sequence variants. *Genet Med* 2018;**20**:1687–8.
 57. Brnich SE, Abou Tayoun AN, Couch FJ, et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med* 2020;**12**:1–12.
 58. Stenson PD, Ball EV, Mort M, et al. Human gene mutation database (HGMD): 2003 update. *Hum Mutat* 2003;**21**:577–81.

59. Son JH, Xie G, Yuan C, et al. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. *Am J Hum Genet* 2018;**103**:58–73.
60. Deisseroth CA, Birgmeier J, Bodle EE, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet Med* 2019;**21**:1585–93.
61. Mishra R, Burke A, Gitman B, et al. Data-driven method to enhance craniofacial and oral phenotype vocabularies. *J Am Dent Assoc* 2019;**150**:933–939 e932.
62. Liu C, Ta CN, Rogers JR, et al. Ensembles of natural language processing systems for portable phenotyping solutions. *J Biomed Inform* 2019;**100**:103318.
63. Han Q, Yang Y, Wu S, et al. Cruxome: a powerful tool for annotating, interpreting and reporting genetic variants. *BMC Genomics* 2021;**22**:407.
64. Havrilla JM, Zhao M, Liu C, et al. Clinical phenotypic spectrum of 4095 individuals with down syndrome from text mining of electronic health records. *Genes (Basel)* 2021;**12**:1159.
65. Hsieh TC, Mensah MA, Pantel JT, et al. PEDIA: prioritization of exome data by image analysis. *Genet Med* 2019;**21**:2807–14.
66. Smedley D, Schubach M, Jacobsen JO, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. *Am J Hum Genet* 2016;**99**:595–606.