Yonago Acta Medica 2019;62(3):244–252 doi: 10.33160/yam.2019.09.001 *Original Article*

Comparison of Causative Variant Prioritization Tools Using Next-generation Sequencing Data in Japanese Patients with Mendelian Disorders

Mitsutaka Ebiki,*† Tetsuya Okazaki,‡§ Masachika Kai,‖ Kaori Adachi,¶ and Eiji Nanba§**

**The Development of Innovative Future Medical Treatment, Graduate School of Medical Sciences, Tottori University, Yonago 683-8504, Japan, †KUSUNOKI SCALE INC., Yonago 683-0832, Japan, ‡Division of Child Neurology, Department of Brain and Neurosciences, School of Medicine, Tottori University Faculty of Medicine, Yonago 683-8504, Japan, §Division of Clinical Genetics, Tottori University Hospital, Yonago 683-8504, Japan, ‖Technical Department, Tottori University, Yonago 683-8503, Japan, ¶Research Initiative Center, Organization for Research Initiative and Promotion, Tottori University, Yonago 683-8503, Japan, and **Research Strategy Division, Organization for Research Initiative and Promotion, Tottori University, Yonago 683-8503, Japan*

ABSTRACT

Background During the investigation of causative variants of Mendelian disorders using next-generation sequencing, the enormous number of possible candidates makes the detection process complex, and the use of multidimensional methods is required. Although the utility of several variant prioritization tools has been reported, their effectiveness in Japanese patients remains largely unknown.

Methods We selected 5 free variant prioritization tools (PhenIX, hiPHIVE, Phen-Gen, eXtasy-order statistics, and eXtasy-combined max) and assessed their effectiveness in Japanese patient populations. To compare these tools, we conducted 2 studies: one based on simulated data of 100 diseases and another based on the exome data of 20 in-house patients with Mendelian disorders. To this end we selected 100 pathogenic variants from the "Database of Pathogenic Variants (DPV)" and created 100 variant call format (VCF) files that each had pathogenic variants based on reference human genome data from the *1000 Genomes Project*. The later "in-house" study used exome data from 20 Japanese patients with Mendelian disorders. In both studies, we utilized 1-5 terms of "Human Phenotype Ontology" as clinical information.

Results In our analysis based on simulated disease data, the detection rate of the top 10 causative variants was 91% for hiPHIVE, and 88% for PhenIX, based on 100 sets of simulated disease VCF data. Also, both software packages detected 82% of the top 1 causative variants. When we used data from our in-house patients instead, we found that these two programs (PhenIX and hiPHIVE) produced higher detection rates than the other three systems in our study. The detection rate of the top 1 causative variant was 71.4% for PhenIX, 65.0% for hiPHIVE.

Conclusion The rates of detecting causative variants in two Exomizer software packages, hiPHIVE and PhenIX, were higher than for the other three software systems we analyzed, with respect to Japanese patients.

Key words computational biology; databases; genetic; whole exome sequencing

Next-generation sequencing (NGS) can exhaustively analyze genes in one sequencing process, and this new capability has dramatically altered the fields of genomic research and medical genetics. In particular, many undiagnosed patients have been receiving benefit from genetic diagnoses by exhaustive gene analysis. Our previous study showed the utility of genetic diagnoses for patients with Mendelian disorders using NGS.[1] Research methods of nationwide large-scale genomic studies, such as Genomics England and the Initiative of Rare and Undiagnosed Diseases project in Japan, have also been based on NGS technology.

Variant prioritization plays a central role in the genetic diagnosis of patients with Mendelian disorders when using NGS techniques, including whole genome sequencing and whole exome sequencing (WES). WES can detect some 30,000 more variants compared to human reference sequences, and approximately 10,000 of these represent nonsynonymous amino acid substitutions, alterations of conserved splice site residues, or small insertions or deletions.[2] Therefore, to detect the causative variant among an enormous number of possibilities, subsequent prioritization steps are required. For example, to interpret variants, we utilize several reference databases and software systems, including a common variant database, pathogenic variant databases, and in-silico prediction tools. Ultimately, detailed consideration by clinicians and bioinformaticians is needed for detecting the causative variant.[1] In such

Corresponding Author: Tetsuya Okazaki, MD
t-okazaki@tottori-u.ac.jp
Received 2019 June 20
Accepted 2019 July 17
Online published 2019 September 13
Abbreviations: DPV, Database of Pathogenic Variants; HPO, human phenotype ontology; NGS, next-generation sequencing; SNPs, single nucleotide polymorphisms; SNVs, single nucleotide variants; VCF, variant call format; WES, whole exome sequencing

Table 1. Comparison of phenotype-based variant detection tools[8, 10, 12]

Software	Availability	Population,Disease-Specific, and Sequence Databases	In Silico Predictive Algorithms	Framework Algorithm
eXtasy-order statistics eXtasy-combined max	Website and Command line	1000 Genomes Project dbNSFP database HGMD	Polyphen2 SIFT MutationTaster CAROL LRT PhastCons Phylop	Phenomiser algorithm Endeavour algorithm Random Forest learning Haploinsufficiency prediction score
Phen-Gen	Website and Command line	HGMD	Polyphen2 SIFT	Bayesian framework Unifying framework Genomewide approach Phenomiser algorithm Random-walk-with-restart algorithm
PhenIX	Website and Command line	1000 Genomes Project ESP 6500	Polyphen2 SIFT MutationTaster	Phenomiser algorithm
hiPHIVE	Website and Command line	1000 Genomes Project ESP 6500 MGD IMPC	Polyphen2 SIFT MutationTaster	Phenomiser algorithm PhenoDigm algorithm Random-walk analyses Random-walk-with-restart algorithm

CAROL, Calculated Combined Annotation Scoring Tools; dbNSFP, database for nonsynonymous SNPs' functional predictions; ESP 6500, Exome Server Project; HGMD, Human Gene Mutation Database; IMPC, International Mouse Phenotyping Consortium; LRT, Likelihood-Ratio Test; MGD, Mouse Genome Database; SIFT, Sorting Intolerant from Tolerant. Software version: eXtasy (Sifrim et al. 2013)[8] ver.0.1, Phen-Gen (Javed et al. 2014)[10] ver.1.0, PhenIX and hiPHIVE (Smedley et al. 2015)[12] ver. 10.0.1.

variant prioritization processes, clinical information is essential. Specifically, we use gene lists that have been made by clinicians for each specific analysis. Such records can help select related variants for a given patient's symptoms, but in those cases where the causative gene is omitted from such a list, we would not be able to detect the causative variant. Identification of disease-causing variants for patients with Mendelian disorders is very complicated and easily qualifies as a "needle in a haystack" challenge.[3]

Several phenotype-driven software tools are now available to help select differential diagnoses.[4] Additionally, to rank the candidate variants in the context of the enormous number of variants that can be detected by NGS, several software systems use both the patient's phenotypic information and the NGS-derived genotypic data.[5–12] These software packages are also able to directly reference a variety of diverse databases and ancillary software systems (Table 1). Comparative evaluations of these software products using patient data from Western countries has already been published,[13] and their utility for causative variant detection is well described. However, the utility of these variant detection software systems using data from Japanese patients' data is unknown.[11] The genetic basis of differences among ethnic groups has been under-investigated. However, we cannot rule out that differences in individual single nucleotide variants (SNVs) in each ethnic group can affect genetic testing results.[14] Therefore, we proceeded to undertake the first effort to evaluate their capabilities with regard to Japanese patients.

MATERIALS AND METHODS

To our knowledge, 11 software packages have been established as variant prioritized tools using Human Phenotype Ontology (HPO).[15] In this research, software was selected using the following criteria using exome sequencing data and available as downloaded packages. We finally selected five variant prioritization software packages PhenIX,[12] hiPHIVE,[12] Phen-Gen,[10] eXtasy-order statistics,[8, 16] and eXtasy-combined max.[8] These products, except for the Phen-Gen package, have already been compared in a previous report.[13] To test these software packages, we selected 1–5 HPO terms corresponding to specific variant call format (VCF) files.[15] The HPO provides a common lexicon of phenotypic abnormalities found in human diseases, and has been utilized in many databases and NGS projects.[17] For our analyses using PhenIX and hiPHIVE, we settled on a cut-off of allele frequency at 1%. We ran all the software programs in the form we received them via downloads. In this research, we performed two different

analyses, one using artificially simulated data and the other using actual data from in-house Japanese patients (Fig. 1). We retained the default parameter settings in all the software products. In addition to the comparison of variant detection rates across the various software products, we also compared variant types and detection rates. To classify each variant type, we utilized the guidelines of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology, which were developed in 2015. In that statement, null variants including nonsense, frameshift and splice sites variants, were classified as very strong pathogenic criteria.[18]

Simulated disease data analysis

For this part of our study, we created 100 VCF files using reference genome bam data (1000 Genomes Project,[19] http://www.internationalgenome.org/home). In addition, we randomly selected 100 pathogenic variants from the "Database of Pathogenic Variants (DPV), http://dpv.cmg.med.keio.ac.jp/dpv-pub/variants)." This database contains germline pathogenic variants in Japanese patients with Mendelian disorders. We then created 100 simulated patient VCF files using each pathogenic variant.

The HPO of each disease was randomly selected using the "Phenomizer[15, 20] system (http://compbio.charite.de/phenomizer)." With this tool, we can pick up causative disorders using the HPO. Additionally, the HPO terms with regard to each disease are described on this site. We analyzed data on 100 simulated patients with the five software packages, using each VCF file and between 1 to 5 selected HPO terms.

In-house patient data analysis

In the second part of this research project, we analyzed 20 exome samples collected from Japanese patients with Mendelian disorders. All samples were sequenced using the TruSight One sequencing panel (Illumina, San Diego, CA). Sequencing and variant detection methods have already been described in our previous research.[1] We performed WES using the Ion AmpliSeq TM Exome RDY kit (Thermo Fisher Scientific, Waltham, MA) for undiagnosed patients despite having performed TruSightOne sequencing. All causative variants were validated by Sanger sequencing. Clinical geneticists selected 1 to 5 HPO of each patient based upon a review of their medical records. Using both the HPO and VCF information, we evaluated the detection rate of each software program. This study was approved by the ethics committee at Tottori University (dated September 22, 2014, approval number G152).

RESULTS

Simulated disease data analysis

We analyzed 100 simulated disease VCF data using the five different software packages (Fig. 2a). The detection rates of causative variants revealed that the best two systems at detecting the top ten variants were hiPHIVE at 91%, and PhenIX at 88%. In addition, both products detected 82% of the top 1 variant. We also note that both of the eXtasy software packages (eXtasy-order statistics and eXtasy-combined max) have limitations on their HPO terms available for analysis and that they could not analyze 20 VCF data items. The detection rate of the top 10 causative variants in eXtasy-order statistics was 19.0%, and in eXtasy-combined max product was 21.0%. In addition, the percentage of variants that could be detected as the chief cause was 6.0% in the order statistics system and 10.0% in combined max product. In Phen-Gen software, the detection rate of the top 5 causative variants was 29%, the top 10 was 33%, and the top 1 was zero percent. Detailed results from the simulated disease data analysis are presented in Supplementary Table S1.

PhenIX and hiPHIVE could detect a total of 82 causative variants as being the most critical. They also produced higher detection rates than the other three software products (Phen-Gen, eXtasy-order statistics and eXtasy-combined max), which correctly detected a total of only 12 variants as being the most critical. We confirmed that the difference in productivity between PhenIX or hiPHIVE and the three other products was quite statistically significant ($P < 0.001$, by Fisher's exact test). Similarly, PhenIX and hiPHIVE detected 87 causative variants as being the first to fifth main factors and showed a higher detection rate than the other three systems (Phen-Gen, eXtasy-order statistics and eXtasy-combined max), which only detected 40 variants as being in the first to fifth factors. We also confirmed the significance of this difference between the PhenIX or hiPHIVE systems and the three other software packages using the detection rates regarding the top first to fifth data ($P < 0.001$, by Fisher's exact test).

Subsequently, we compared the results of PhenIX and hiPHIVE against each other. A total of 79 variants were detected as the top causative factors by both products. The hiPHIVE system detected 7 causative variants as having a higher priority than did the PhenIX product (ID 3, 7, 30, 60, 72, 88, 96), but the PhenIX software detected 6 causative variants as higher priority than did the hiPHIVE product (ID 29, 38, 59, 70, 71, 73).

In this simulated disease aspect of our study, the analyzed variants consisted of 51 missense mutation variants (Fig. 2b) and 44 null variants (Fig. 2c). Because

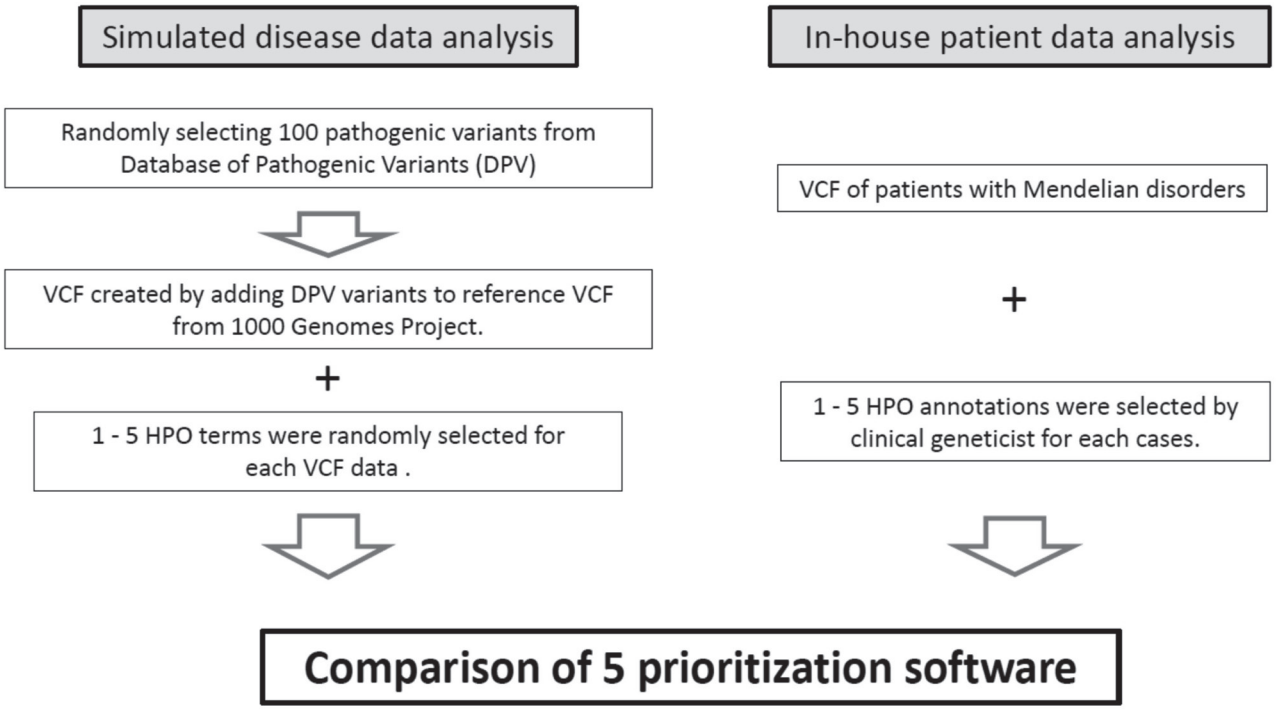

Fig. 1. Overview of the prioritization software comparison study.
Two separate investigations were performed, one using data from in-house patients, and the other based on simulated data. We compared 5 different software products (PhenIX, hiPHIVE, Phen-Gen, eXtasy-order statistics and eXtasy-combined max). These systems use variant call format (VCF) files and the human phenotype ontology (HPO). In simulated disease analyses, we created 100 virtual patient VCF files, and added pathogenic variants based on reference to the VCF files in the 1000 Genome Project. We selected 100 pathogenic variants from the Japanese pathogenic variant database (http://dpv.cmg.med.keio.ac.jp/dpv-pub/variants). Additionally, we selected several HPO files for each VCF from the Phenomizer site (http://compbio.charite.de/phenomizer). For the in-house patient analysis, we used VCF files and HPO for specific patients with Mendelian disorders.

five variants were deleted from the DPV site, we excluded these from our comparison of the detection rates and variant types. We found that Exomiser's two kinds of software (PhenIX and hiPHIVE) correctly detected 80% of the VCF files as the top 1 through 5 for both the missense mutations and null variants. For Phen-Gen, no remarkable difference was found in the detection rate in distinguishing between these variant types. For the eXtasy-order statistics package, the "not ranked" detection rate was 64% (28/44) of the null variants and only 11.8% (6/51) for the missense variants. "Not ranked" means that the software did not remarkably detect any causative variants within the top 1–100. This tendency was also observed in the eXtasy-combined max program, where 64% (28/44) of the null variants and 11.8% (6/51) of the missense variants were also not detected.

In-house patient data analysis

We also analyzed the data from our 20 in-house patients with Mendelian disorders (Fig. 3) using each of the five software products. The detection rate of the top 10 causative variants for the hiPHIVE system was 85.7%, and for the PhenIX product this was 76.2%. Also, the percentage of variants that could be detected as the chief cause was 61.9% in hiPHIVE and 71.4% in PhenIX. In our analysis of eXtasy, we found that it discovered the principal causative variant in only 5% of the cases, and its finding of the top 10 factors was only 10%. Remarkably, we also found that, for eXtasy, the combined max analysis could not detect any of the main causative variants, nor any within the top 10. Likewise, Phen-Gen also generated poor results, identifying zero of the top five causative variants, only 5% of the top 10, and zero of the chief causes. Detailed results of variant data analysis are described in Supplementary Table S2.

PhenIX could detect 15 of the top causative variants and showed a higher detection rate than the other three software products (Phen-Gen, order statistics, and combined max eXtasy), each of which only detected one top variant. We confirmed the significance of the differences between PhenIX and these three other software products ($P < 0.001$, by Fisher's exact test).

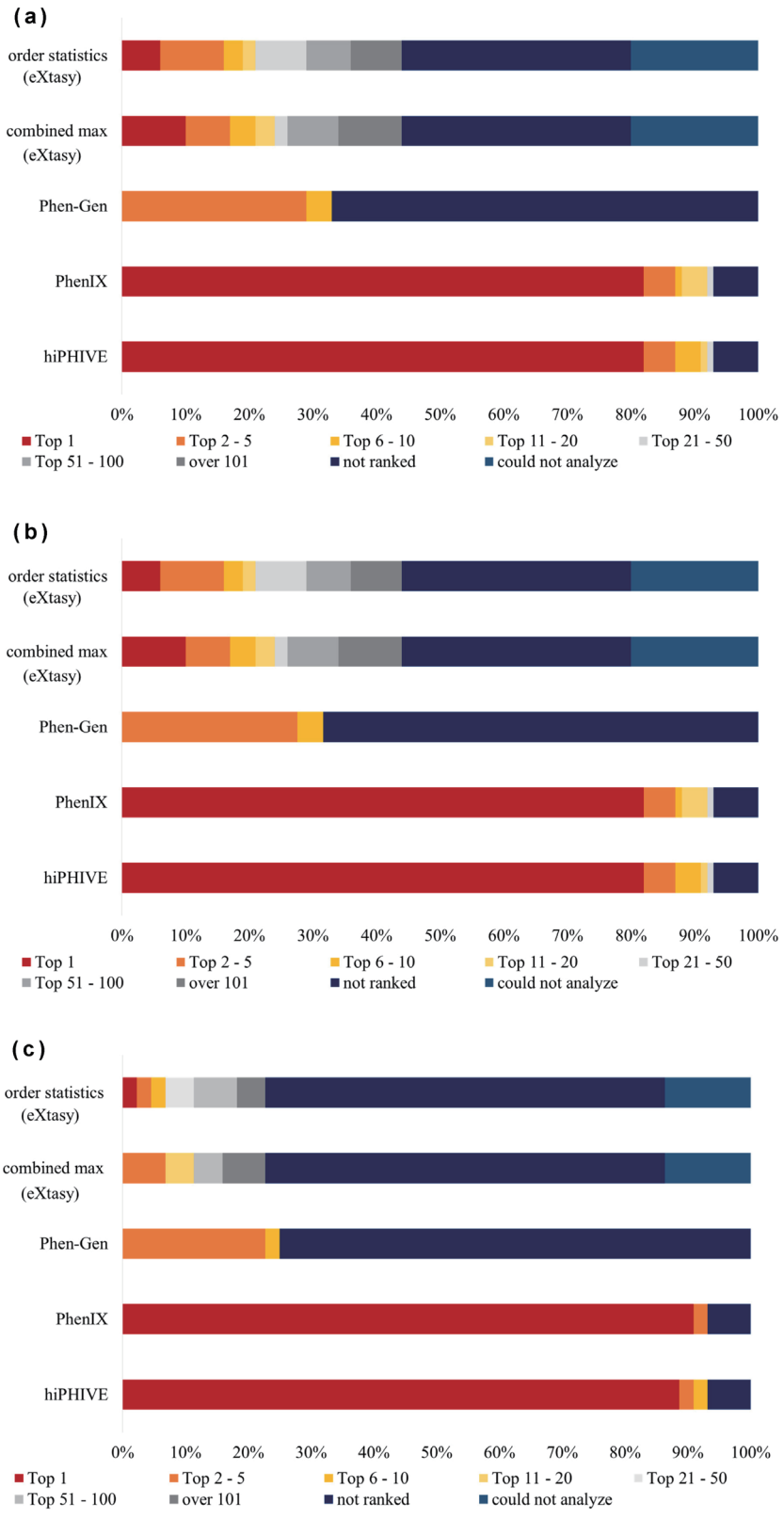

Fig. 2. Simulated disease analysis data ranked by category.
(**a**) Ranking of causative variant detection in simulated disease data analysis (n = 100). (**b**) Ranking of missense causative variants detection in simulated disease data analysis (n = 51). (**c**) Ranking of null causative variants detection in simulated disease data analysis (n = 44). The term "not ranked" means software could not detect causative variants; "could not analyze" means the software could not perform the analysis.

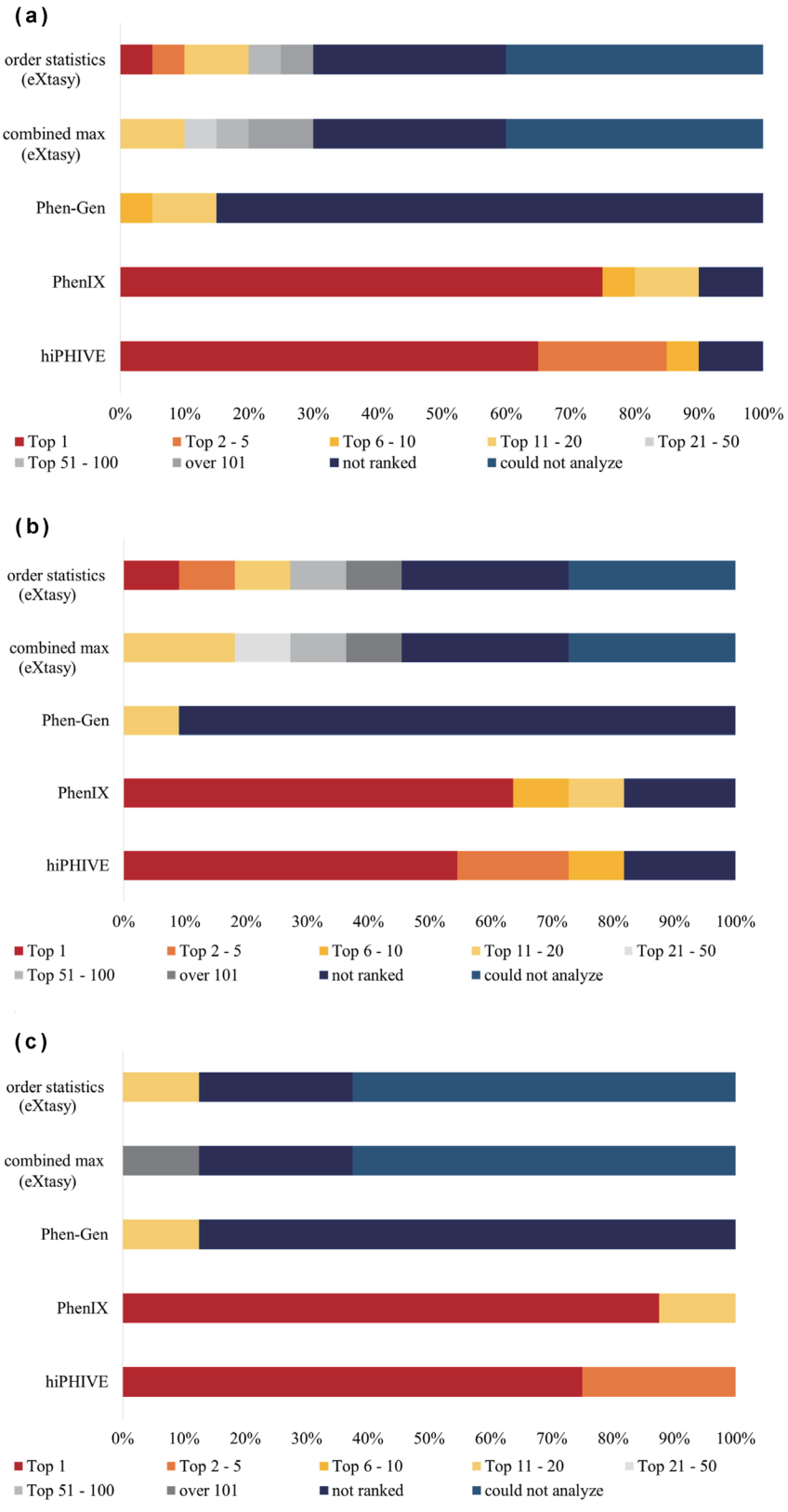

Fig. 3. In-house patient data ranked by category.
(**a**) Data from 20 patients was used in the analysis of each software product. (**b**) Ranking of missense causative variants detection regarding in-house patient data analysis (n = 11). (**c**) Ranking of null causative variants detection of in-house patient data analysis (n = 8). The term "not ranked" means software could not detect causative variants; "could not analyze" means the software could not perform the analysis.

Similarly, PhenIX detected 15 of the top 1–5 causative variants and in this regard also showed a higher detection rate than the other three software systems (Phen-Gen, order statistics, and combined max eXtasy), which only detected two variants, respectively. We again confirmed this statistical difference between PhenIX and three software ($P < 0.001$, by Fisher's exact test).

In a similar pattern, hiPHIVE could detect 13 causative variants as being the top 1, and 17 variants as being in the top 1–5. This software also produced a higher detection rate than the three other software systems (Phen-Gen, order statistics, and combined max eXtasy), which detected only one variant as a top 1 and 2 variants as being among the top 1–5 causes. We likewise confirmed this difference in top 1 and top 1–5 detection rates between hiPHIVE and three software as being statistically significant ($P < 0.001$, by Fisher's exact test).

We subsequently compared the results produced by the PhenIX and hiPHIVE systems against each other and found that the same 13 variants were detected as the top causes by both products. For three causative variants (patients 1, 16, 19), hiPHIVE identified them as having a higher priority than did PhenIX. In contrast, PhenIX detected two causative variants as having a higher priority than hiPHIVE (patients 6, 10).

Over the course of this in-house study, we analyzed 11 missense mutation variants (Fig. 3b) and 8 null variants (Fig. 3c). Because patient 3's variant types were missense mutation and null variant, we excluded the case from our comparison of the detection rates and variant types. With the Phen-Gen product, we found no remarkable difference in the detection rate despite the distinction of the variant types. In Exomiser's two types of software (PhenIX and hiPHIVE), the detection rate of causative variants within the top 10 was 100% in the analysis of the null variants, and also quite high for the missense variants: PhenIX at 72.7% (8/11) and hiPHIVE at 81.8% (9/11). In addition, the detection rates of "not ranked" for both software were 18.2% (2/11) in the missense variants analysis. With the two eXtasy software products (order statistics and combined max) the proportion of "not ranked" results was 25% (2/8) for the null variants data and 27% (3/11) for the missense variants.

DISCUSSION

Our simulated disease data analysis indicated that the detection rates with PhenIX and hiPHIVE were substantially higher than those with the remaining 3 tools; this trend was corroborated by the previously reported results.[13] However, the utility of these two software packages, specifically regarding Japanese patients, had previously been unknown. Our present research indeed showed that the two software systems generated high detection rates when using simulated Japanese patient data and also when using the data from our in-house patients. We provide detailed information available to us regarding each software product in Table 1, but only looking at this information, we do not understand why there is such a substantial difference in the detection rates across the various products. Software update frequency might be one of the most important factors in improving detection rates. In fact, frequent updates are done for Exomiser. For example, integration of a usable database and adding the most recent algorithms are done continuously.[17] The detection rate of the top 10 causative variants in hiPHIVE was 20.0% in the previous report,[13] whereas the present research detection rate of this software was 90.0%. Certain mechanisms are unknown, but frequent software updates in Exomizer might be one of the key elements of improving the identification rate.

The detection rate of the causative variants for hiPHIVE and PhenIX was higher than for the other software except for the data related to ID 71 (Supplementary Table S1). In this case, with c.755G > C, and p Arg252Pro in the *PAH* gene, we found that Phen-Gen and eXtasy-order statistics detected the causative variant as in the top 3, while eXtasy-combined max detected it among the top 7. In contrast, hiPHIVE recognized it as being among the top 10, and PhenIX identified it as among the top 14.

Similarly, other variants of the *PAH* gene in ID 70 and 72 also had relatively low detection results in hiPHIVE and PhenIX. In comparison, the detection rate of ID 73 in the *PAH* gene for hiPHIVE and PhenIX was higher than in the two eXtasy software programs. Also, Phen-Gen could detect more than could the two eXtasy products. We considered many possible reasons for this particular result regarding the *PAH* gene, but we could not reach a conclusion about the mechanism involved due to a lack of more detailed information about each software product and the methods they use. Nonetheless, this result suggests that the detection rates can differ depending on the gene.

Associations between the detection rate and the variant type were not found for hiPHIVE, PhenIX, and Phen-Gen products, except for the eXtasy software systems. Notably, in the two types of eXtasy software, usable HPOs were limited, and we could not analyze 20/100 (20%) of the VCF files. Although statistical analysis could not be performed because of our small

sample size, this result suggests that the detection rate of the 2 eXtasy software products might be inadequate with regard to the analysis of null variants.

In this simulated study, we created VCF files to simulate patient genomic data. The 100 bam data files registered in the 1000 Genome Project were randomly selected in this study. These bam data files contain not only Japanese data files but also non-Japanese data files, and we cannot deny that specific SNVs related to non-Japanese data affected the analysis results. In our subsequent analyses using this simulated data, we obtained results similar to what we found using data from actual in-house patients. As a result, we speculate that this approach of using simulated patient data might indeed be more broadly useful for the comparisons of other variant prioritization tools.

From in-house patient data analysis results involving ranked comparisons, it is clear that the causative variant detection rate of hiPHIVE and PhenIX was higher than that of the other software products. We particularly note that hiPHIVE was able to detect all the causative variants within the top 10 except for only the two *ECHS1* items. Indeed, these two *ECHS1* gene variants from patients 7 and 12 were missed by all five software products. The *ECHS1* gene was not included in our TruSight One sequencing panel, and we only detected these variants in the *ECHS1* gene using WES. We suggest that since the mitochondrial short-chain enoyl-CoA hydratase 1 deficiency due to *ECHS1* mutations was first reported only in 2014,[21] it is possible that the software packages we examined might be unable to readily detect such relatively new disorders (Fig. 3). All five software products use the Phenomizer program described in Table 1 of this paper, and we hypothesize that this relatively new gene was not listed in it. However, since these software packages and their related databases are updated daily, we have to assume that Phenomizer was updated at the time of our analyses.

Several variants cause specific phenotypes in Mendelian disorders. N540K in the *FGFR3* of hypochondroplasia and S2G in the *SHOC2* of Noonan-like syndrome with loose anagen hair are well-known variants related to specific phenotypes among clinical geneticists. We found that hiPHIVE and PhenIX could detect these variants, but the other software products remarkably could not. This failure is noteworthy because we would expect that purpose-built causative variant prioritization tools should be easily able to detect these well-known variants.

Exomiser, including its use in hiPHIVE and PhenIX, recently received the approval of the International Rare Diseases Research Consortium as a recognized resource.[17] Overall, the detection rates of causative variants with hiPHIVE and PhenIX was higher than with the other software products in our study. This result is the same found in previous research.[13]

As a technical note, we should mention that although one can find many diagnostic software tools online using HPO and VCF files,[11] in our present study, this would have required uploading sensitive patient data. Therefore, to fully adhere to the protection of personal information, in this present research, we operated all five software as downloaded packages.

HPO provides a comprehensive bioinformatic resource for the analysis of human diseases and phenotypes.[15] It has therefore been adopted as the standard for phenotypic terms in international rare affected tissues, registries, clinical laboratories, biomedical resources, and clinical software tools.[17, 22–27] The description of phenotypic variations has become critical for genomic medicine and translational research,[28–31] and our having "computable" descriptions of human diseases using HPO phenotypic profiles is a key element in the use of Phenotype-based exome analysis tools. Further studies of HPO setting methodologies for these tools are needed to enable us to detect causative variants more efficiently.

In conclusion, we confirmed the utility of causative variant prioritization tools with regard to Japanese patients. In particular, the detection rate of causative variants in two Exomizer software products, hiPHIVE and PhenIX, was higher than in that of the three other systems we analyzed.

Acknowledgments: I would like to thank M.Shimada, Y.Endo, H.Sunada, A.Koga, K.Uehara and M.Ueki at Advanced Medicine of Innovation and Clinical Research Center of Tottori University Hospital for their support. I also thank K.Yoshida, H.Noma and K.Higaki for your advice and cooperation. Finally, we are grateful to the referees for their useful comments.

The authors declare no conflict of interest.

REFERENCES

1 Okazaki T, Murata M, Kai M, Adachi K, Nakagawa N, Kasagi N, et al. Clinical Diagnosis of Mendelian Disorders Using a Comprehensive Gene-Targeted Panel Test for Next-Generation Sequencing. Yonago Acta Med. 2016;59:118-25. PMID: 27493482

2 Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014;6:252ra123. PMID: 25186178, DOI: 10.1126/scitranslmed.3009262

3 Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12:628-40. PMID: 21850043, DOI: 10.1038/nrg3046

4 Jalali Sefid Dashti M, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. Biotechniques. 2017;62:18-30. PMID: 28118812, DOI: 10.2144/000114492

5 Robinson PN, Krawitz P, Mundlos S. Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet. 2011;80:127-32. PMID: 21615730, DOI: 10.1111/j.1399-0004.2011.01713.x

6 Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur J Hum Genet. 2012;20:490-7. PMID: 22258526, DOI: 10.1038/ejhg.2011.258

7 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al.; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011;27:2156-8. PMID: 21653522, DOI: 10.1093/bioinformatics/btr330

8 Sifrim A, Popovic D, Tranchevent LC, Ardeshirdavani A, Sakai R, Konings P, et al. eXtasy: variant prioritization by genomic data fusion. Nat Methods. 2013;10:1083-4. PMID: 24076761, DOI: 10.1038/nmeth.2656

9 Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet. 2012;13:523-36. PMID: 22751426, DOI: 10.1038/nrg3253

10 Javed A, Agrawal S, Ng PC. Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat Methods. 2014;11:935-7. PMID: 25086502, DOI: 10.1038/nmeth.3046

11 Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. Genome Med. 2015;7:81. PMID: 26229552, DOI: 10.1186/s13073-015-0199-2

12 Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10:2004-15. PMID: 26562621, DOI: 10.1038/nprot.2015.124

13 Pengelly RJ, Alom T, Zhang Z, Hunt D, Ennis S, Collins A. Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. Sci Rep. 2017;7:13509. PMID: 29044180, DOI: 10.1038/s41598-017-13841-y

14 Burchard EG, Ziv E, Coyle N, Gomez SL, Tang H, Karter AJ, et al. The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med. 2003;348:1170-5. PMID: 12646676, DOI: 10.1056/NEJMsb025007

15 Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017;45:D865-76. PMID: 27899602, DOI: 10.1093/nar/gkw1039

16 Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006;24:537-44. PMID: 16680138, DOI: 10.1038/nbt1203

17 Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47:D1018-27. PMID: 30476213, DOI: 10.1093/nar/gky1105

18 Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al.; ACMG Laboratory Quality Assurance Committee. 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;17:405-23. PMID: 25741868, DOI: 10.1038/gim.2015.30

19 Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al.; 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526:68-74. PMID: 26432245, DOI: 10.1038/nature15393

20 Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85:457-64. PMID: 19800049, DOI: 10.1016/j.ajhg.2009.09.003

21 Peters H, Buck N, Wanders R, Ruiter J, Waterham H, Koster J, et al. ECHS1 mutations in Leigh disease: a new inborn error of metabolism affecting valine metabolism. Brain. 2014;137:2903-8. PMID: 25125611, DOI: 10.1093/brain/awu216

22 Ramoni RB, Mulvihill JJ, Adams DR, Allard P, Ashley EA, Bernstein JA, et al.; Undiagnosed Diseases Network. The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. Am J Hum Genet. 2017;100:185-92. PMID: 28157539, DOI: 10.1016/j.ajhg.2017.01.006

23 Taruscio D, Groft SC, Cederroth H, Melegh B, Lasko P, Kosaki K, et al. Undiagnosed Diseases Network International (UDNI): white paper for global actions to meet patient needs. Mol Genet Metab. 2015;116:223-5. PMID: 26596705, DOI: 10.1016/j.ymgme.2015.11.003

24 Gahl WA, Mulvihill JJ, Toro C, Markello TC, Wise AL, Ramoni RB, et al.; UDN. The NIH Undiagnosed Diseases Program and Network: applications to modern medicine. Mol Genet Metab. 2016;117:393-400. PMID: 26846157, DOI: 10.1016/j.ymgme.2016.01.007

25 Gall T, Valkanas E, Bello C, Markello T, Adams C, Bone WP, et al. Defining Disease, Diagnosis, and Translational Medicine within a Homeostatic Perturbation Paradigm: The National Institutes of Health Undiagnosed Diseases Program Experience. Frontiers in Medicine. 2017;4:62. PMID: 28603714, DOI: 10.3389/fmed.2017.00062

26 Thompson R, Johnston L, Taruscio D, Monaco L, Béroud C, Gut IG, et al. RD-Connect: an integrated platform connecting databases, registries, biobanks and clinical bioinformatics for rare disease research. J Gen Intern Med. 2014;29(suppl 3):780-7. PMID: 25029978, DOI: 10.1007/s11606-014-2908-8

27 Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users. Hum Mutat. 2012;33:803-8. PMID: 22422702, DOI: 10.1002/humu.22078

28 Biesecker LG. Phenotype matters. Nat Genet. 2004;36:323-4. PMID: 15054484, DOI: 10.1038/ng0404-323

29 Robinson PN, Webber C. Phenotype ontologies and cross-species analysis for translational research. PLoS Genet. 2014;10:e1004268. PMID: 24699242, DOI: 10.1371/journal.pgen.1004268

30 Robinson PN. Deep phenotyping for precision medicine. Hum Mutat. 2012;33:777-80. PMID: 22504886, DOI: 10.1002/humu.22080

31 Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, et al. Finding our way through phenotypes. PLoS Biol. 2015;13:e1002033. PMID: 25562316, DOI: 10.1371/journal.pbio.1002033