JAMIA Open, 2024, **7(2)**, ooae052
https://doi.org/10.1093/jamiaopen/ooae052
Application Notes

Application Notes

IDeRare: a lightweight and extensible open-source phenotype and exome analysis pipeline for germline rare disease diagnosis

Ivan William Harsono, MD, MCS[1], Yulia Ariani, MD, PhD[2,*], Beben Benyamin, PhD[3,4,5], Fadilah Fadilah, PhD[6,7], Dwi Ari Pujianto, PhD[2], Cut Nurul Hafifah, MD[8]

[1]Doctoral Program in Biomedical Sciences, Faculty of Medicine, Universitas Indonesia, Jakarta 10430, Indonesia, [2]Department of Medical Biology, Faculty of Medicine, Universitas Indonesia, Jakarta 10430, Indonesia, [3]Australian Centre for Precision Health, University of South Australia, Adelaide 5000, Australia, [4]UniSA Allied Health and Human Performance, University of South Australia, Adelaide 5000, Australia, [5]South Australian Health and Medical Research Institute (SAHMRI), University of South Australia, Adelaide 5000, Australia, [6]Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia, Jakarta 10430, Indonesia, [7]Bioinformatics Core Facilities—IMERI, Faculty of Medicine, Universitas Indonesia, Jakarta 10430, Indonesia, [8]Department of Child Health, Dr Cipto Mangunkusumo Hospital, Faculty of Medicine, University of Indonesia, Jakarta 10430, Indonesia

*Corresponding author: Yulia Ariani, MD, PhD, Gedung Fakultas Kedokteran UI, Jl. Salemba Raya No.6, PO Box 1358, Jakarta 10430, Indonesia (yulia.ariani@ui.ac.id)

Abstract

Objectives: Diagnosing rare diseases is an arduous and challenging process in clinical settings, resulting in the late discovery of novel variants and referral loops. To help clinicians, we built IDeRare pipelines to accelerate phenotype-genotype analysis for patients with suspected rare diseases.

Materials and Methods: IDeRare pipeline is separated into phenotype and genotype parts. The phenotype utilizes our handmade Python library, while the genotype part utilizes command line (bash) and Python script to combine bioinformatics executable and Docker image.

Results: We described various implementations of IDeRare phenotype and genotype parts with real-world clinical and exome data using IDeRare, accelerating the terminology conversion process and giving insight on the diagnostic pathway based on disease linkage analysis until exome analysis and HTML-based reporting for clinicians.

Conclusion: IDeRare is freely available under the BSD-3 license, obtainable via GitHub. The portability of IDeRare pipeline could be easily implemented for semi-technical users and extensible for advanced users.

Lay Summary

Diagnosing rare diseases is challenging and has a debilitating impact if diagnosed late. The lack of national clinical guidelines, advanced omics laboratories, and an integrated phenotype-genotype bioinformatic pipeline forces clinicians, especially in developing countries, to manually curate patient phenotypes and annotate patient genotype variants. Recently, efforts to standardize clinical data collection have used ICD-10, SNOMED-CT, and LOINC terminologies through national programs such as the SATUSEHAT ecosystem in Indonesia. However, OMIM and HPO are more specialized for rare diseases and are not commonly collected in standard medical records. Automatic conversion of these common terminologies to OMIM and HPO, along with linkage analysis and phenotype-based recommendations, can benefit clinicians and geneticists in diagnosing rare diseases. To ease their burden, we built IDeRare, a lightweight, extensible phenotyping and variant analysis pipeline. This pipeline integrates existing tools and runs on a mid-tier personal computer with NVIDIA 8GB GPU memory. The IDeRare pipeline is available on GitHub under the BSD-3 license.

Key words: phenotype; genotype; terminology; rare disease; open-source software.

Introduction

Diagnosing rare diseases

Diagnosing rare diseases is challenging due to diverse mutations, the potential for silenced variants, and their atypical presentation. Functional rare diseases pose greater diagnostic difficulties than structural syndromes, especially in developing archipelago countries, where the lack of national guidelines and limited access to specialized facilities contribute to delayed diagnoses averaging 5.6-7.6 years.[1,2] Approximately 72% of rare diseases are genetically inherited, highlighting the importance of gene and phenotype analysis.[3] Whole exome sequencing (WES) is pivotal for diagnosing rare diseases, offering cost-effective and adequate potential to detect novel mutations. Despite a 28% diagnostic rate, WES may still identify variants of uncertain significance. Trio WES, involving both the proband and parents' sequences, enhances diagnostic yield by 40%.[4]

Interoperability and the EMR implementation landscape in Indonesia

SATUSEHAT, the Indonesian national health interoperability ecosystem, consists of both SATUSEHAT platform and

Received: April 11, 2024; **Revised:** May 20, 2024; **Editorial Decision:** May 23, 2024; **Accepted:** May 27, 2024

SATUSEHAT mobile. Initiated by the Ministry of Health of the Republic of Indonesia (MoH Indonesia) in mid-2022, SATUSEHAT platform serves as a national Fast Healthcare Interoperability Resources (FHIR) R4 interoperability server accessible by healthcare facilities (HCF) and healthcare providers (HCP), promoting standardization at the national level. SATUSEHAT mobile, a counterpart of SATUSEHAT platform, functions as a national personal health record (PHR) displaying medical resume data aggregated through SATUSEHAT platform. The use-cases of SATUSEHAT platform provides state-of-the-art guidelines for clinical data capture, documenting patient journeys (outpatient, inpatient, emergency visits), or specific diseases requiring continuity of care, such as mother and child health, dental health, nutrition, and cancer. Current terminologies supported by SATUSEHAT include ICD-10 for diagnoses, ICD-9 CM for procedures, SNOMED-CT for disorders and clinical findings, LOINC for observable and laboratory findings, and *Kamus Farmasi & Alkes* (KFA) for drugs and medical devices. Data sent from HCF's electronic medical record (EMR) to SATUSEHAT platform are validated based on FHIR R4 structure and allowable terminology subsets, ensuring that EMR implementation follows SATUSEHAT predefined standards.[5]

The Biomedical and Genome Science Initiative (BGSI), focusing on building a research ecosystem of genomic biodatabanks and registries, was initiated by MoH Indonesia in mid-2022 and is still in its early phases.[6] BGSI has recently adopted SATUSEHAT FHIR R4 standards for collecting cancer registry data since 2023 and will soon follow with other registry use-cases in the future. However, to date, neither SATUSEHAT nor BGSI have released a genetic disorders and variant reporting standard.[5] This situation indirectly impacts rare disease diagnoses by restricting the transfer of Human Phenotype Ontology (HPO)[7] and Online Mendelian Inheritance in Man (OMIM)[8] terminology from the EMR to SATUSEHAT platform, despite both being the main terminology standards for rare disease phenotyping.

Before the existence of SATUSEHAT, all clinical data input at each HCF was done in an unstructured manner, either digitally or on paper. However, this paradigm has progressively shifted towards structured EMR implementation, driven by MoH Indonesia initiatives in 2023 to enforce all EMR vendors and HCFs to be accredited for their compliance and implementation of SATUSEHAT FHIR R4 profile use-cases, specifically Encounter and Condition resources. The existence of SATUSEHAT has positively influenced traditional unstructured clinical data collection towards internationally acceptable, structured data collection via FHIR-compliant EMR system.

Phenomics and genomics of rare diseases

To date, rare disease phenotyping is primarily conducted by healthcare or research institutions and heavily relies on manual curation to narrow down possible diagnoses before referral to centers with genetic sequencing capabilities. Since HPO7 and OMIM8 haven't been included as acceptable SATUSEHAT terminology inputs, there is an unmet need for a flexible, lightweight, and extensible library to: (1) automate the conversion of terminologies to HPO directly from EMR; (2) facilitate differential diagnoses similarity scoring and linkage analysis of diagnoses compared to the patient's phenotype; and (3) provide recommendations of potential diagnoses for geneticists.

While numerous premium software suites for clinical exome analysis exist, such as VarSome Clinical,[9] Geneyx,[10] Golden Helix,[11] and SOPHiA DDM platform,[12] lack of independence for customization, integration with preexisting EMRs, and affordability are major issues in low- to middle-income countries. Other rare disease genetic analysis pipelines utilize multiple open-source programs, such as SIMPLEX[13] and nf-core/raredisease.[14] However, these pipelines do not consider integrating phenotype data input sourced from clinical records, and their VCF output requires manual curation by expert bioinformaticians and/or geneticists. To the author's knowledge, there has been no integrative open-source solution for a lightweight, extensible end-to-end phenotype-genotype translational bioinformatics tool easily integrated into preexisting EMR or information systems adopting FHIR R4 with ICD-10, SNOMED-CT, LOINC, HPO, and/or OMIM.

Objectives

Our primary objectives were to design and build a pipeline that enables streamlined clinical workflow utilizing data provided by clinicians through EMR. Specifically, the pipeline should have the minimum capability to: (1) automatically convert common EMR terminology (ICD-10, SNOMED-CT, and LOINC) to specialized rare disease terminology (HPO and OMIM); (2) perform phenotype-based linkage analysis and dendrogram between disorders; (3) provide potential recommendations of working diagnoses based on phenotype; (4) offer an easy setup exome analysis pipeline for germline variant discovery for proband or trio exome analysis; (5) align, variant call, and annotate exome data; and (6) return summarized clinically significant variants based on gene-phenotype prioritization. Additionally, this pipeline design should: (1) adhere to structured electronic data capture and terminology for phenotype and genotype analysis; (2) support lightweight, portable deployment in mid-tier computing settings to anticipate the scarcity of high performance computing clusters in rural areas; and (3) be extensible to be interconnected with any information system supporting standardized data capture.

Methods

Phenotype software dependencies

To enhance the extensibility of the phenotype pipeline, we developed a Python3 library with the minimum prerequisites of Python 3.8+. This library utilized the pandas,[15] hpo3,[16] PyYAML,[17] matplotlib,[18] and scipy[19] library packages. Data dependencies include subsets from SNOMED-CT Indonesia MLDS,[20] LOINC2HPO annotation,[21] Orphanet Rare Disease Ontology,[22] and HPO7 ontology and annotation. The script used to build subsets is accessible on GitHub.

Genotype software dependencies

The core workflow of the genotype pipeline consists of five main steps: (1) sequence alignment; (2) duplicate removal and sorting; (3) variant calling for single-nucleotide variants; small insertions and deletions for both proband and trio; along with structural variant calling for the proband; (4) variant annotation; and (5) gene-phenotype prioritization. An optional initial step for adapter trimming and QC reporting is also available. Complete instructions for prerequisites and

Table 1. IDeRare steps, software, and database used.

Steps		Software/database /library	Link
Phenotype pipeline (iderare_phenotyping.sh)			
1	Phenotype pipeline	iderare-pheno	https://github.com/ivanwilliammd/iderare-pheno
Genotype pipeline (iderare.sh)			
1	**Adapter trimming and QC report**[a]	fastp[23]	https://github.com/OpenGene/fastp
2	**Sequence Alignment**	bwa-mem2 2.2.1[24]	https://github.com/bwa-mem2/bwa-mem2
	Reference Sequence	GRCh38.p14	http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/p14/hg38.p14.fa.gz
3	**Duplicate removal and sorting**	sambamba 1.0.0[25]	https://github.com/biod/sambamba
4	**Variant calling**		
	Proband SNV and small indel	DeepVariant 1.5.0[26]	https://github.com/google/deepvariant
	Trio SNV and small indel	DeepTrio 1.5.0[27]	https://github.com/google/deepvariant
		GLnexus 1.2.7[28]	https://github.com/dnanexus-rnd/GLnexus
	Proband Structural variant calling	tiddit 3.6.0[29]	https://github.com/SciLifeLab/TIDDIT
5	**Variant annotation**	SnpEff 5.1d[30] SnpSift[31]	https://github.com/pcingola/SnpEff
	Annotation database	dbSNP b156[32]	https://ftp.ncbi.nih.gov/snp/latest_release/VCF/
		ClinVar20230514[33]	https://ftp.ncbi.nih.gov/pub/clinvar/vcf_GRCh38/
		dbNSFP4.4a[34]	https://sites.google.com/site/jpopgen/dbNSFP
6	**Gene-phenotype prioritization**	Exomiser 13.3.0[35], 2302 database	https://github.com/exomiser/Exomiser
7	**Supporting tools**	bcftools 1.17[36]	https://github.com/samtools/bcftools

[a] Optional, if fastq data have not been trimmed or QC checked.

code execution are provided on GitHub. The detailed dependencies, versions, and reference databases used are listed in Table 1.

Hardware dependencies

The phenotype pipeline runs smoothly on multiple operating systems and hardware specifications, including MacOS 14, PopOS 22.04, and Windows 11, using the Anaconda virtual environment with Python 3.8+ and RAM ranging from 8 to 64 GB. The genomic pipeline has been evaluated on a personal computer equipped with an Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz, 64 GB RAM with 48 GB swap memory, 2TB SATA SSD storage, NVIDIA RTX2080 8 GB, PopOS 22.04 operating system, Anaconda environment with Python 3.8.8, and Docker support. By default, the genomic pipeline utilizes all CPU threads with adjustable maximum memory usage. The minimum hardware requirement for the IDeRare genomic pipeline should include an NVIDIA RTX workstation with 8 GB of GPU memory, 32 GB of RAM to smoothly process intermediary exome data, and an SSD to prevent file I/O bottlenecks.

Phenotype and genotype data examples

We conducted phenotype and genotype analysis based on previous rare disease cases reported in 2021 by the attending physician, who manually analyzed the cases for months to provide working diagnoses of Glycogen Storage Disease IV. We obtained raw, unstructured phenotype data and transcribed all phenotype findings and differential diagnoses to SNOMED-CT, LOINC, and ICD-10 as primary terminology choices, in compliance with SATUSEHAT acceptable terminology. Additionally, we used HPO phenotype as secondary terminology to supplement disease pathology findings or spectrum occurring at specific onsets (such as infancy or toddlerhood), which may not be well-defined in the primary terminologies. This combination of SNOMED-CT, LOINC, and ICD-10 with HPO ensured complete representation of phenotype data for linkage analysis, differential diagnosis recommendation, and gene-phenotype-based prioritization. The transcribing processes were conducted by a clinical terminologist and validated by a senior pediatrician. Genotype data is accessible from the NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject/1077459) and phenotype data is provided at https://github.com/ivanwilliammd/IDeRare#clinical-information-example. All phenotype and genotype data have been anonymized, and patient consent has been obtained.

Results and discussion

Complete IDeRare phenotype and genotype pipelines are distributed under the open-source 3-Clause BSD License and can be accessed on the IDeRare GitHub repository (https://github.com/ivanwilliammd/IDeRare). The phenotype part (IDeRare-Pheno) is also available as a Python package downloadable on PyPi (https://pypi.org/project/iderare-pheno/) and is distributed under the same open-source 3-Clause BSD License on the IDeRare-Pheno GitHub repository (https://github.com/ivanwilliammd/iderare-pheno). Web UI implementations of the phenotype pipeline can be found at https://bioinformatics-ivanwilliamharsono.streamlit.app/IDeRare_Pheno. The genotype part is only available as a command-line script for on-premises installation. Figure 1 depicts the schematic diagram of the IDeRare pipeline.

IDeRare inputs

As illustrated in Figure 1, the pipeline requires phenotype, genotype, and pedigree information. The phenotype part accepts SNOMEDCT, LOINC, ICD-10, OMIM, HP, and ORPHA terminologies, which should be provided as newline-separated text files, tsv files, or FHIR R4 outputs generated by EMR. Additionally, genotype filenames and pedigrees (for trio mode) should be provided in ***iderare.yml***.

IDeRare-pheno as part of an analysis pipeline, a standalone data exploration project, or web services

IDeRare-pheno library was imported and integrated with the argparse library to provide robust customizable parameters such as *threshold*, *differential*, and *recommendation*.

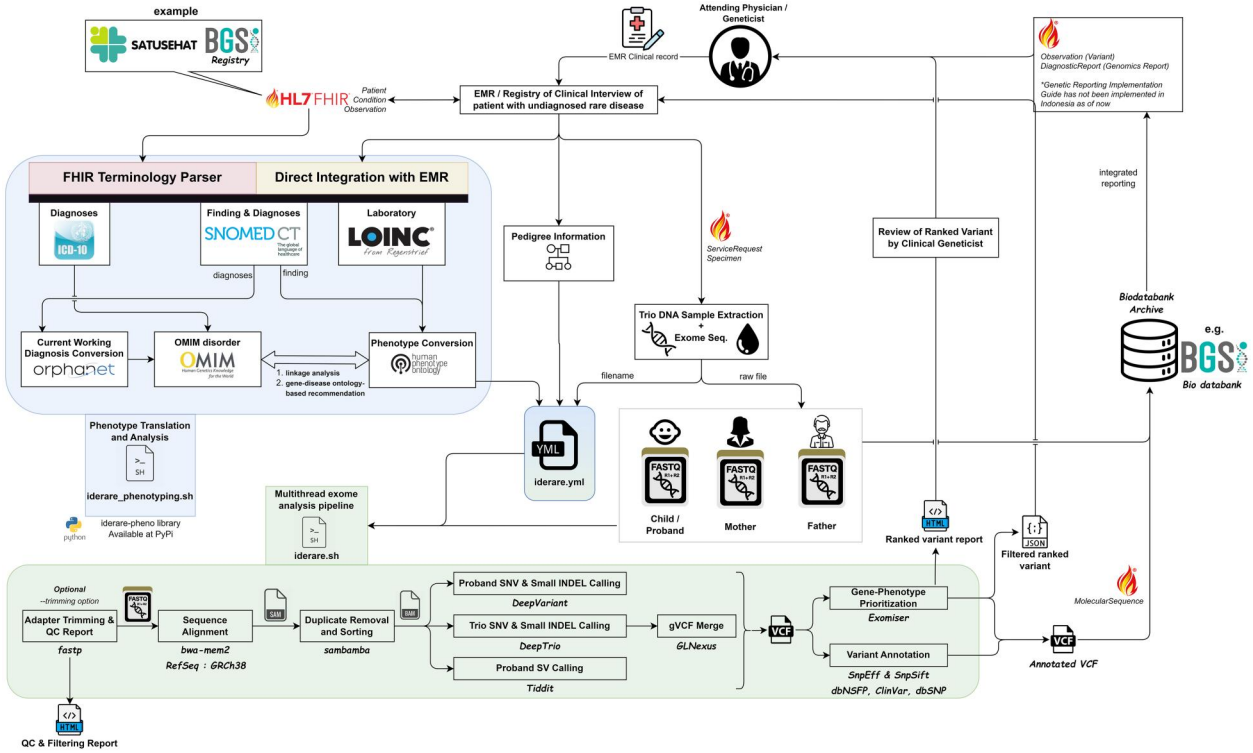

Figure 1. Schematic diagram of the IDeRare complete pipeline (blue = phenotype, green = genotype).

Table 2. IDeRare-pheno function and action.

Function	Action
term2orpha(str)	Convert specific input to ORPHA code list. Input accepted: SNOMEDCT:[code]
term2hpo(str)	Convert specific input to HPO code list. Input accepted: SNOMEDCT:[code], LOINC:[code]|[interpretation].
term2omim(str)	Convert specific input to OMIM code list. Input accepted: ORPHA:[code] and ICD-10:[code].
batchconvert(list)	Automatically convert respective input of mixed terminology code of SNOMEDCT:[code], LOINC:[code], ICD-10:[code], ORPHA:[code], OMIM:[code], HP:[code], OMIM:[code] returning HPO lists and OMIM lists.
omim2object(str)	Objectify OMIM string terminology OMIM:[code] to OMIM PyHPO object.
hpos2set(list)	Objectify HP:[code] string list to HPOSet PyHPO object.
hpo2name(list)	Acquire the HPO name given the list of HP:[code] List.
omim2name(list)	Acquire the OMIM name given the list of OMIM:[code] List.
similarity_linkage(omim_sets, hpo_sets, threshold, min_n, linkage)	Run the similarity score and linkage analysis between provided HPOSet PyHPO object, and OMIM HPOSet list. Returning the full, filtered (by threshold, or falling back of min_n) data sorted by similarity score in descending manner. Linkage analysis run utilizing graph based information coefficient (graph IC)[37] method and best-match average (BMA)[37] combination method.
hpo2omim_similarity(omim_sets, hpo_sets, threshold, differential)	Wrapper function to automatically objectify HP:[code] and OMIM:[code] string list and run similarity_linkage function.
omim_recommendation(hpo_set, type=[gene/disease], threshold, recommendation)	Wrapper function to get all available gene and disease from OMIM dictionary, objectify HPO code to HPOSet, and run through full similarity score, and linkage analysis of filtered OMIM gene/disease based on threshold and recommendation.
generate_yml(hpo_sets, filename)	Create template iderare.yml file required by IDeRare genotype analysis given the HP:[code] string List.
linkage_dendrogram(linkage_data, labels, title, threshold, path_to_save)	Draw a dendrogram based on linkage analysis data, labels provided, title with threshold red line, and path_to_save.
list2tsv(term_id, name, sim_score, filename)	Convert, and generate tab separated file given ID, name, and/or similarity score. Input could be any kind of terminology.

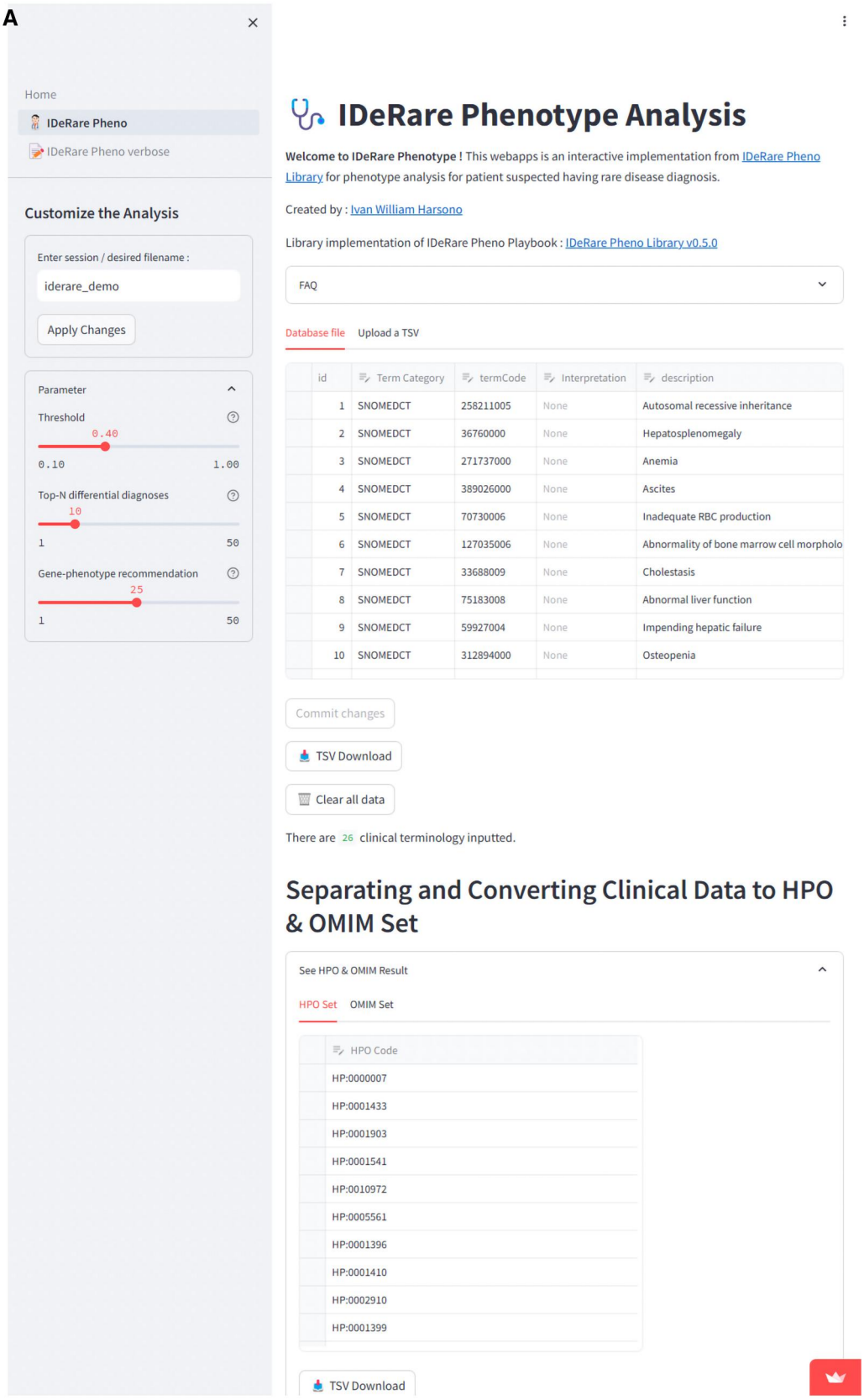

Figure 2. Example of IDeRare analysis using clinical data (A-C) and trio clinical report of exome analysis (D).

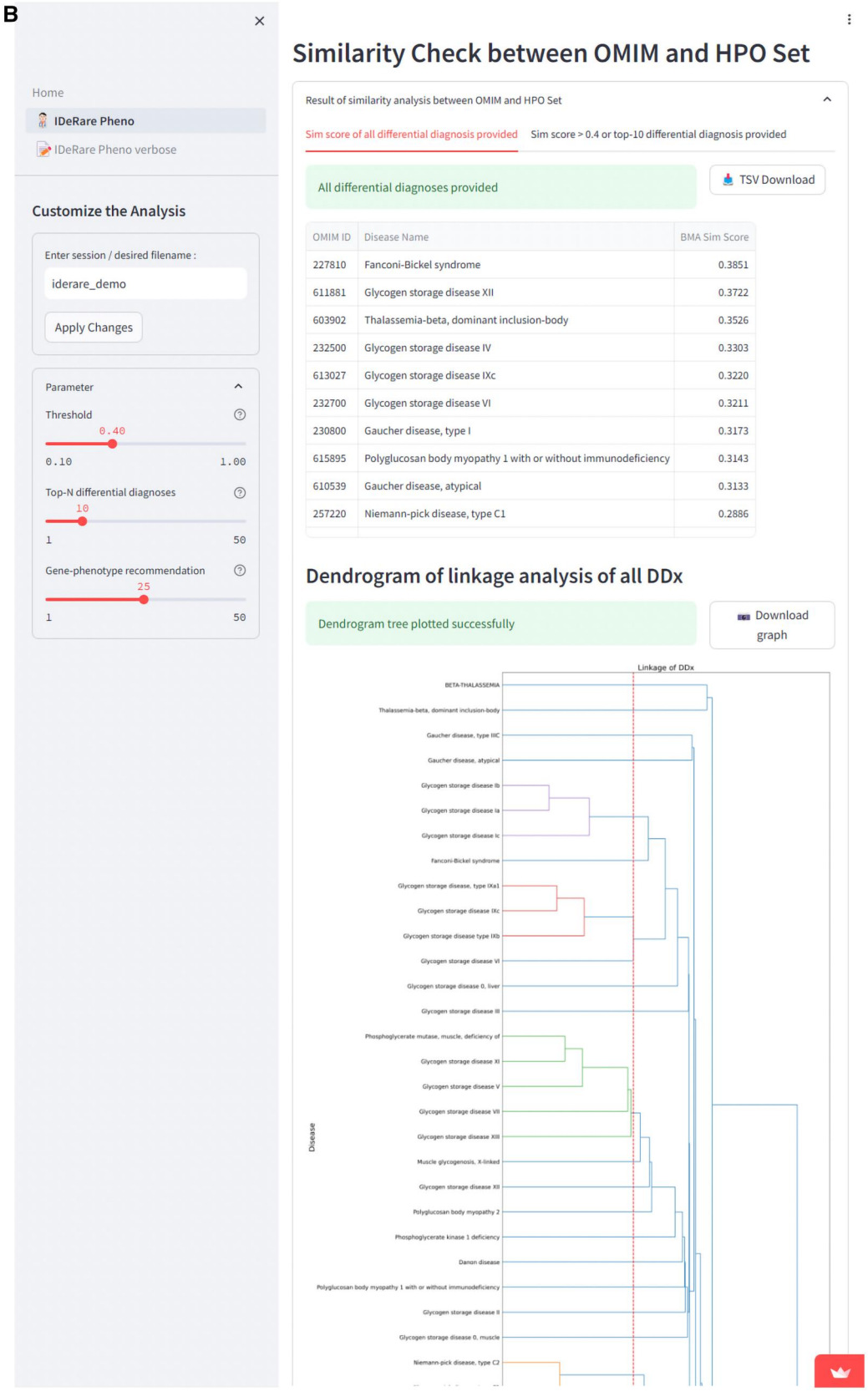

Figure 2. (Continued).

C

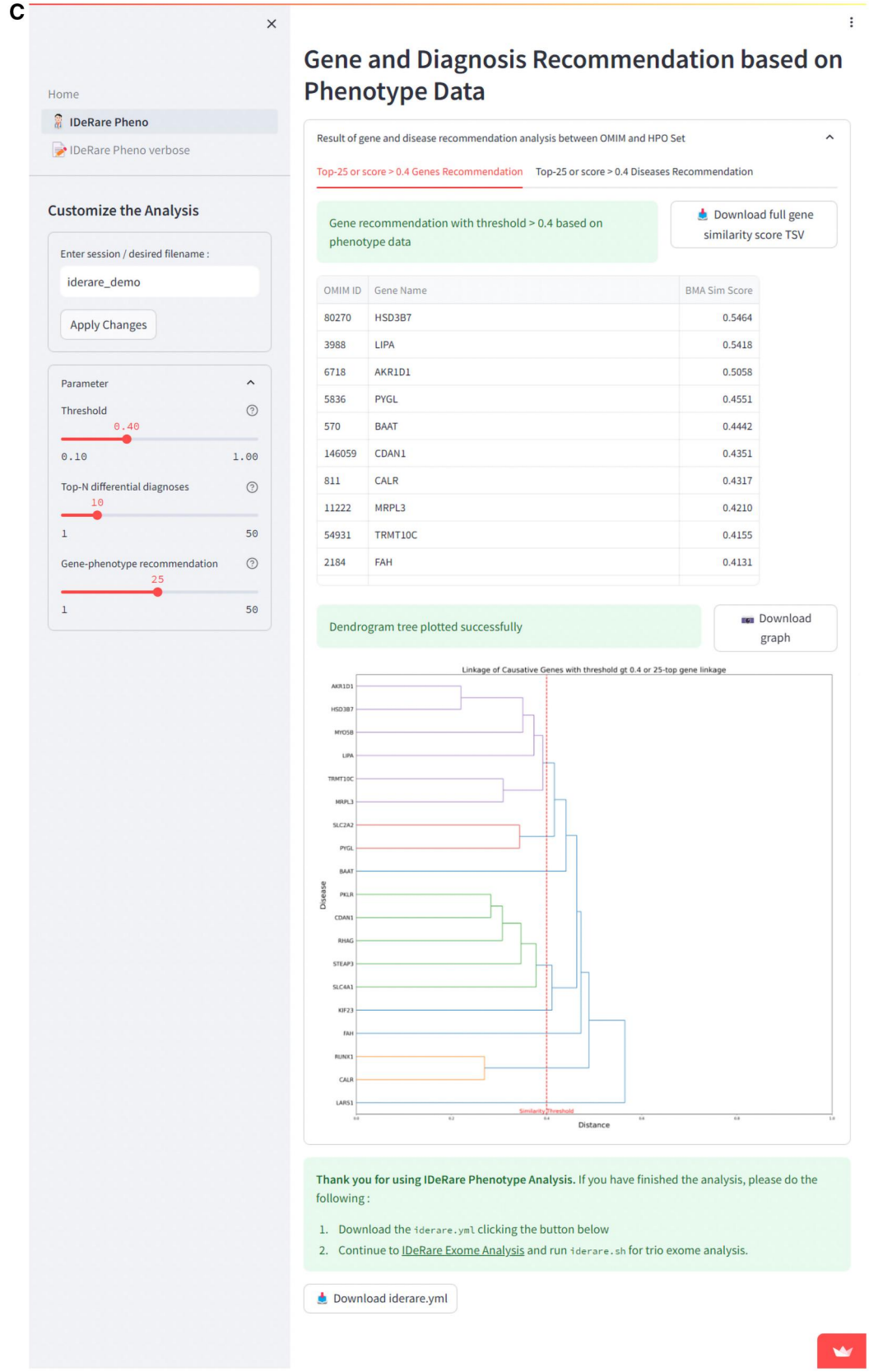

Figure 2. (Continued).

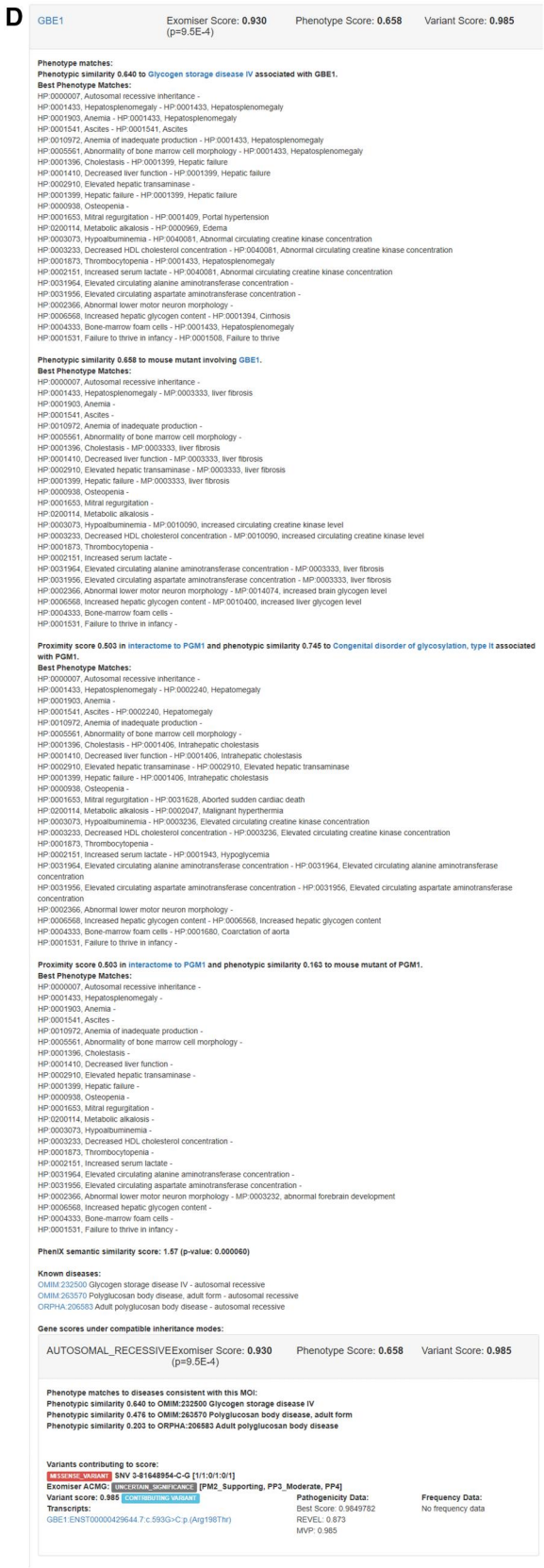

Figure 2. (Continued).

Examples of usage and Jupyter notebook examples can be found on the IDeRare-pheno GitHub repository. Additionally, web services to parse FHIR data input are available for testing via Postman Workspace: https://www.postman.com/ivanwilliamharsono/workspace/iderare-pheno/overview. The full functionality of IDeRare-pheno is listed in Table 2.

IDeRare genotype pipeline configuration

The IDeRare genotype pipeline supports analysis in ***proband***, ***trio***, and ***both*** modes, with an optional parameter to enable ***fastq*** trimming. By default, the analysis will be conducted for ***both*** proband and trio modes, and the trimming process will be skipped. This can be adjusted by inputting parameters alongside the iderare.sh command, such as *iderare.sh—mode [both/proband/trio]—trimming [false/true]*.

Demonstration of IDeRare-pheno web application implementation

The example of the Streamlit phenotype application showcases a streamlined UI with an interactive table-like input for mixed phenotype, laboratory, and clinical findings results, along with responsive interactive sliders in the left sidebar, as depicted in Figure 2A-C. All graph and table results can be directly downloaded from the application. The UI is interactive enough to be explored by non-technical users.

Demonstration of the IDeRare complete pipeline using real-case clinical and exome data

The testing results utilizing real clinical data can be observed on Streamlit pages by accessing the session of ***iderare_demo***. The diagnosis obtained was Glycogen Storage Disease IV, with a similarity score of 0.33, ranking 4 out of 37 OMIM differential diagnoses. This diagnosis was provided using a combination of SNOMEDCT and ICD-10 codes. In terms of exome variant prioritization, the top-ranked result obtained from both proband and trio approaches using IDeRare was the GBE1(NM_000158.4):c.593G>C(p.Arg198Thr) mutation inherited in a compound heterozygote manner. This variant had a p-value of 9.5×10^{-4} with Exomiser score of 0.930 and was classified as VUS fulfilling ACMG Criteria PM2, PP3, and PP4 (Figure 2D). The processing time for phenotype and genotype analysis was approximately 5 min for the phenotype part and 18 h for the proband and trio genotype analysis.

Extensibility and usability of IDeRare

Given the availability of structured clinical and genomics data input outlined in the IDeRare inputs section, IDeRare can be seamlessly extended, customized, and integrated with existing EMR supporting FHIR R4 outputs. These systems should adhere to ICD-10, SNOMED-CT, and LOINC as primary clinical terminology standards, complemented by secondary rare disease-specific terminologies such as HPO, OMIM, or ORPHA to encapsulate disease spectrum nuances or pathology findings not covered by primary terminologies. This integration process is facilitated by the modular nature of the IDeRare pipeline, which comprises the IDeRare-pheno Python library, which can be utilized as-is, integrated into web services, or incorporated into web apps, along with the genotype command line pipeline, consolidated into a unified pipeline.

Limitation and solution

The main limitation of the IDeRare pipeline is its reliance on available rare disease patient medical records and exome files for the diagnosis of germline variant mutations. This reliance is constrained by the limited availability of complete clinical information with exome files, particularly from institutions like Cipto Mangunkusumo, and restricted access to international rare disease registries. To address this limitation, future validation of the IDeRare pipeline could be conducted using a cohort of rare disease registries, such as those provided by BGSI registries.

Conclusion

In conclusion, we have developed a portable and extensible phenotype-genotype analysis pipeline tailored for both proband and trio-based analyses. This pipeline encompasses essential features for clinicians, including seamless terminology conversion, phenotype-diagnoses similarity scoring, visualization of phenotype linkage analysis, processing of exome sequencing files, and the generation of meaningful clinical reporting. Our demonstration with real-case clinical and exome data showcases the effectiveness and utility of the IDeRare pipeline in rare disease diagnosis and analysis.

Author contributions

Ivan William Harsono (Conceptualization, Methodology, Software, Validation, Visualization, Writing—Original Draft), Yulia Ariani (Conceptualization, Methodology, Writing—Review & Editing, Supervision), Beben Benyamin and Fadilah Fadilah (Conceptualization, Methodology, Validation, Writing—Review & Editing), Dwi Ari Pujianto (Methodology, Writing—Review & Editing), and Cut Nurul Hafifah (Investigation, Writing—Review & Editing).

Funding

This work was supported by Directorate of Research and Development, Universitas Indonesia under Hibah PUTI 2022 grant number NKB-1425/UN2.RST/HKP.05.00/2022.

Conflicts of interest

None declared.

Data availability

The genotype data sample used have been deposited to National Center for Biotechnology Information (NCBI) under BioProject accession number PRJNA1077459 with BioSample database of SAMN39972817-SAMN39972819, and Sequence Read Archive accession numbers SRR27997290-SRR27997292. Phenotype data sample available at IDeRare GitHub pages and accessible from https://github.com/ivanwilliammd/IDeRare#clinical-information-example.

Ethic statements

Ethical clearance was obtained from The Ethics Committee of the Faculty of Medicine, University of Indonesia—Cipto Mangunkusumo Hospital and approved on 12 December

2022 (KET-1395/UN2.F1/ETIK/PPM.00.02/2022). Written informed consent was obtained from patient's parents for all experiments described here for biological sample usage, and publication.

References

1. Wiseman V, Thabrany H, Asante A, et al. An evaluation of health systems equity in Indonesia: study protocol. *Int J Equity Health*. 2018;17(1):138.
2. Austin CP, Cutillo CM, Lau LPL, et al. Future of rare diseases research 2017-2027: an IRDiRC perspective. *Clin Transl Sci*. 2018;11(1):21-27.
3. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *Eur J Hum Genet*. 2020;28(2):165-173.
4. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet*. 2018;19(5):253-268.
5. Ministry of Health of Republic of Indonesia. 2024. SATUSEHAT Platform. Accessed May 01, 2024. https://satusehat.kemkes.go.id/platform/docs/id/playbook/
6. Ministry of Health of the Republic of Indonesia. *Blueprint for Digital Health Transformation Strategy 2024*. Ministry of Health of the Republic of Indonesia; 2021.
7. Gargano MA, Matentzoglu N, Coleman B, et al. The human phenotype ontology in 2024: phenotypes around the world. *Nucleic Acids Res*. 2024;52(D1):D1333-D1346.
8. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*. 2019;47(D1):D1038-D1043.
9. Saphetor SA. 2024. Varsome Clinical. Accessed May 01, 2024. https://landing.varsome.com/varsome-clinical
10. Geneyx Genomex Ltd. 2024. NGS Data Analysis | Clinical Genome Research. Accessed May 01, 2024. https://geneyx.com/
11. Golden Helix I. 2024. Genomic Data Analysis Software—Golden Helix. Accessed May 01, 2024. https://www.goldenhelix.com/
12. SOPHiA GENETICS. 2024. Technology—SOPHiA GENETICS. Accessed May 01, 2024. https://www.sophiagenetics.com/technology/
13. Fischer M, Snajder R, Pabinger S, et al. SIMPLEX: cloud-enabled pipeline for the comprehensive analysis of exome sequencing data. *PLoS One*. 2012;7(8):e41948.
14. Stranneheim H, Lagerstedt-Robinson K, Magnusson M, et al. Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med*. 2021;13(1):40.
15. The pandas development t. pandas-dev/pandas: Pandas. Accessed April 11, 2024. https://github.com/pandas-dev/pandas
16. Marcello J. hpo3. Accessed April 11, 2024. https://github.com/anergictcell/hpo3
17. Simonov K. PyYAML. Accessed April 11, 2024. https://github.com/yaml/pyyaml
18. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90-95.
19. Jones E, Oliphant T, Peterson P. OthersSciPy: open source scientific tools for Python. 2001. Accessed April 11, 2024. https://scipy.org/
20. International Health Terminology Standards Development Organisation. *Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)*. SNOMED International; 2024.
21. The Jackson Laboratory. loinc2hpoAnnotation. Accessed April 11, 2024. https://github.com/TheJacksonLaboratory/loinc2hpoAnnotation
22. ORPHANET. Orphanet Rare Disease Ontology. 2024. Accessed April 11, 2024. https://bioportal.bioontology.org/ontologies/ORDO
23. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i890.
24. Vasimuddin M, Misra S, Li H, Aluru S, eds. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Rio de Janeiro, Brazil; 2019; 20–24 May 2019.
25. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015;31(12):2032-2034.
26. Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36(10):983-987.
27. Kolesnikov A, Goel S, Nattestad M, et al. DeepTrio: variant calling in families using deep learning. bioRxiv, 2021, 2021.04.05.438434.
28. Lin MF, Rodeh O, Penn J, et al. GLnexus: joint variant calling for large cohort sequencing. *bioRxiv*. 2018:343970.
29. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. *F1000Res*. 2017;6:664. https://doi.org/10.12688/f1000research.11168.2
30. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92.
31. Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet*. 2012;3:35. https://doi.org/10.3389/fgene.2012.00035
32. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311.
33. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067.
34. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med*. 2020;12(1):103.
35. Smedley D, Jacobsen JOB, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the exomiser. *Nat Protoc*. 2015;10(12):2004-2015.
36. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993.
37. Deng Y, Gao L, Wang B, Guo X. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *PLoS One*. 2015;10(2):e0115692.

JAMIA Open, 2024, 7, 1–10
https://doi.org/10.1093/jamiaopen/ooae052
Application Notes