Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Clinical phenotype and trio whole exome sequencing data from a patient with glycogen storage disease IV in Indonesia

Ivan William Harsono[a], Yulia Ariani[b,*], Beben Benyamin[c,d,e], Fadilah Fadilah[f,g], Dwi Ari Pujianto[b], Cut Nurul Hafifah[h], Titis Prawitasari[h]

[a] *Faculty of Medicine, Doctoral Program in Biomedical Sciences, Universitas Indonesia, Jakarta 10430, Indonesia*
[b] *Department of Medical Biology, Faculty of Medicine, Universitas Indonesia, Jakarta 10430, Indonesia*
[c] *Australian Centre for Precision Health, University of South Australia, Adelaide, SA 5000, Australia*
[d] *UniSA Allied Health and Human Performance, University of South Australia, Adelaide, SA 5000, Australia*
[e] *South Australian Health and Medical Research Institute (SAHMRI), University of South Australia, Adelaide, SA 5000, Australia*
[f] *Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia, Jalan Salemba Raya number 4, Jakarta 10430, Indonesia*
[g] *Faculty of Medicine, Universitas Indonesia, Bioinformatics Core Facilities - IMERI, Jalan Salemba Raya number 6, Jakarta 10430, Indonesia*
[h] *Department of Child Health, Faculty of Medicine, Dr. Cipto Mangunkusumo Hospital, University of Indonesia, Jakarta 10430, Indonesia*

ARTICLE INFO

Article history:
Received 9 September 2024
Revised 6 December 2024
Accepted 10 December 2024
Available online 16 December 2024

Dataset link: Trio Whole Exome Sequencing of Rare Disease (Original data)

Keywords:
Homo sapiens
Rare disease
Whole exome sequencing
Cipto mangunkusumo national hospital
Jakarta
Indonesia

ABSTRACT

Glycogen storage disease type IV (GSD IV) is a rare disease caused by a defect in glycogen branching enzyme 1 (GBE1), which played a crucial role in glycogen branching. GSD IV occurs once in approximately 1 in every 760,000 to 960,000 live births and is inherited in an autosomal recessive pattern. Early diagnosis of GSD IV is challenging due to non-specific symptoms, such as liver and spleen enlargement, which can overlap with other hematologic and hepatobiliary disorders. The non-specific clinical finding (phenotype) and identification of novel mutation adds the complexity of diagnosing and confirming rare disease. This often results in delayed diagnosis, typically 5.6 to 7.6 years later, with only 50% of

* Corresponding author.
E-mail address: yulia.ariani@ui.ac.id (Y. Ariani).
Social media: @ivanwilliammd (I.W. Harsono)

https://doi.org/10.1016/j.dib.2024.111231

cases being diagnosed, while the remaining cases are classified as undiagnosed rare diseases due to either the absence of identifiable potential variants or the presence of novel variants requiring further functional studies to confirm their pathogenicity. Proband and trio whole exome sequencing analysis remains a cost-effective and widely available method for diagnosing rare diseases detecting between 21 and 40% of cases. We present a trio (familial) exome sequences data from a patient with Glycogen Storage Disease IV from Indonesia. The clean and adapter trimmed FASTQ files of these sequences are available under BioProject accession number PRJNA1077459 with Sequence Read Archive accession numbers SRR27997290-SRR27997292.

Specifications Table

Subject	Clinical Genetics.
Specific subject area	Genomics, Phenomics, Rare Disease, Trio Samples.
Type of data	Processed / Cleaned Sequences File (FASTQ).
Data collection	DNA was extracted from the buffy coat of whole blood sample, followed by quality check of DNA quantification, and preparation of DNA libraries. After sequencing BGI DNBSeq sequencing platform. Raw sequences, filtered with SOAPnuke, including removal of adaptor sequences, contamination, and low-quality reads from raw reads.
Data source location	Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia – Ciptomangunkusumo National Referral Hospital, Jakarta, Indonesia.
Data accessibility	Clinical phenotypes attached in this paper. Clean and adapter-trimmed data (FASTQ) files have been deposited to National Center for Biotechnology Information (NCBI), https://www.ncbi.nlm.nih.gov/, under BioProject database: https://www.ncbi.nlm.nih.gov/bioproject/1077459, with BioSample database of SAMN39972817-SAMN39972819. and SRA database: with accession number: SRR27997290-SRR27997292 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP490127). Repository name: National Center for Biotechnology Information (NCBI) Data identification number: BioProject database: https://www.ncbi.nlm.nih.gov/bioproject/1077459 Direct URL to data: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP490127
Related research article	Harsono IW, Ariani Y, Benyamin B, Fadilah F, Pujianto DA, Hafifah CN: IDeRare: a lightweight and extensible open-source phenotype and exome analysis pipeline for germline rare disease diagnosis. JAMIA Open 2024, 7(2).

1. Value of The Data

- This dataset includes clinical phenotypes and trio exome sequences from a patient and both of his parents, with the proband carrying a germline recessive mutation.
- This dataset is valuable for future population analysis of patients with rare diseases, particularly in developing countries considering the establishment of rare disease biobanks.
- This dataset can provide novel insight into mutation discovery, enrich the dbSNP, and enhance the Indonesian population database for patients with functional rare diseases.
- This dataset will serve as a valuable asset for training purposes to validate new diagnostic pipeline utilizing phenotype-genotype correlations in rare diseases.

2. Background

Rare diseases (RD) are commonly defined as diseases with a prevalence of less than 1 in 2,000 people in a population [1]. There are approximately 5,000 to 8,000 types of RDs globally, affecting an estimated 400 million people [2]. These diseases cause up to 35% of deaths in the first year of life, and as many as 3 in 10 do not survive beyond the age of 5 years [1], or they survive with major disabilities and a decreased quality of life [3]. Approximately 72% of rare diseases are of genetic origin [4]. Glycogen storage disease (GSD) is one of functional rare diseases, causing abnormalities in the glycogen cycle [5]. GSD IV is inherited in an autosomal recessive pattern, caused by a defect in the *GBE1* gene, and occurs in 1 out of every 760,000 to 960,000 live births [5,6]. Diagnosing RDs remained a challenge, with delayed diagnosis and referral loops lasting 5.6 to 7.6 years in developed countries [7]. Currently, proband (patient) and trio (patient and parents) whole exome sequencing remains a cost-effective and widely available modality for diagnosing 21–40% of rare diseases [8]. Common approaches to diagnosing rare disease include analysis of clinical phenotypes alongside genetic sequencing data using phenotype-variant prioritization. This data has been shown to facilitate the development of open-source rare disease pipelines [9], and enhance the understanding of the pathogenicity of mutations through functional mutational analysis.

3. Data Description

We presented 3 exome sequences of trio exome sequencing conducted of a male patient suspected of having functional rare disease with initial symptoms of progressively hepatomegaly at the age of 2 years old, which later resulted in liver failure and death before reaching 5 years old, with working diagnosis of suspected glycogen storage disease type IV. Coded phenotype and differential diagnoses during patient care were presented in Table 1. Isolated DNA was purified and QC-ed using spectrophotometer, before cryopreserved at -20 °C in May 2021 (Table 2). DNA was thawed in March 2023, and re-quantified using Qubit 3.0 before sequencing (Table 2). Paired ends were obtained after exome sequencing runs. FASTQ raw data files have been deposited in the NCBI database under BioProject database: https://www.ncbi.nlm.nih.gov/bioproject/1077459, with BioSample database of SAMN39972817-SAMN39972819 and SRA database with accession number: SRR27997290-SRR27997292 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP490127). The detail of the cleaned sequencing data is presented in Table 3. Validation of variant calling result and Mendelian inheritance shown in Tables 4 and 5. This data will be useful for building rare disease registry and biobank, further downstream bioinformatics analysis, creating phenotype-genotype pipeline, enrichment of dbSNP, population genomics, and potential large scale familial studies for functional rare disease pathomechanism discovery through integrative multiomics and functional study.

4. Experimental Design, Materials and Methods

4.1. Experimental workflow

This subchapter describes the comprehensive workflow and experimental procedures to create and use the dataset in a practical use-case as training data for rare disease mutational analysis pipeline. Further details on the experimental procedures and workflow will be provided in the subsequent subchapter.

1. **Gather Phenotype Data**
    - 1.1. **Patient Consultation:** An initial clinical interview was conducted to gather patient history and clinical data.
    - 1.2. **Further laboratory/radiology workup:** Positive workup results were noted.

Table 1
Patient's phenotype and working differential diagnoses data.

Clinical information	Description	Terminology Code
Phenotype		
Family history	Disease inherited in autosomal recessive pattern	SNOMEDCT:258211005
Physical examination	Liver and spleen enlargement	SNOMEDCT:36760000
Physical examination	Anemia	SNOMEDCT:271737000
Physical examination	Fluid build up in abdomen (ascites)	SNOMEDCT:389026000
Problem list	Inadequate red blood cell production in bone marrow	SNOMEDCT:70730006
Problem list	Abnormal morphology of bone marrow cell	SNOMEDCT:12703.5006
Problem list	Slowed flow of bile from liver to small intestine (cholestasis)	SNOMEDCT:33688009
Problem list	Abnormal liver function	SNOMEDCT:75183008
Problem list	Impending hepatic failure	SNOMEDCT:59927004
Problem list	Lower bone density (osteopenia)	SNOMEDCT:312894000
Problem list	Mitral regurgitation	SNOMEDCT:48724000
Problem list	Metabolic alkalosis	SNOMEDCT:1388004
Routine laboratory workup	Low albumin serum level	LOINC:1751-7|L
Routine laboratory workup	Low high density lipoprotein (HDL) level	LOINC:2085-9|L
Routine laboratory workup	Low platelet count	LOINC:777-3|L
Routine laboratory workup	Increased lactate level	LOINC:2519-7|H
Routine laboratory workup	Increased alanine aminotransferase (ALT) level	LOINC:1742-6|H
Routine laboratory workup	Increased aspartate aminotransferase (AST) level	LOINC:1920-8|H
Disorder group	Abnormal lower motor neuron function	HP:0002366
Biopsy liver result	Increase hepatic glycogen content	HP:0006568
Bone marrow biopsy	Foam cells identified in bone-marrow biopsy	HP:0004333
Growth and development	Failure to thrive during infancy	HP:0001531
Differential diagnoses		
Differential diagnosis	Beta thalassemia	SNOMEDCT:65959000
Differential diagnosis	Gaucher disease	SNOMEDCT:190794006
Differential diagnosis	Niemann pick disease type C	SNOMEDCT:66751.000
Differential diagnosis	Glycogen storage diseases spectrum	ICD-10:E74.0

This table provides a comprehensive overview of the patient's clinical phenotypes and differential diagnoses, coded according to various medical terminology standards. Each entry includes a description and the corresponding terminology code, utilizing ICD-10 (International Classification of Diseases, 10th Revision), SNOMEDCT (Systematized Nomenclature of Medicine Clinical Terms), LOINC (Logical Observation Identifiers Names and Codes), and HP (Human Phenotype Ontology). The table is divided into two main sections: Phenotype and Diagnoses. The phenotype section captures key clinical findings from the patient, such as family history, physical examination results, problem lists, routine laboratory workups, biopsy results, and growth and development observations. The diagnoses section lists potential differential diagnoses relevant to the patient's condition.

Table 2
Double-stranded DNA concentration and absorbance.

Sample	Sample Name	Spectrophotometer			Qubit 3.0
		Conc. (µg/mL)	A260/230	A260/280	dsDNA Conc. (µg/mL)
V350145665_L04_B5EHOMdmhwXAAAA-515	Proband	273.4	1.883	2.828	75.9
V350145665_L04_B5EHOMdmhwXAABA-517	Mother	336.6	1.856	2.626	63.9
V350145665_L04_B5EHOMdmhwXAACA-519	Father	86.24	1.955	3.102	108

This table presents the concentration and absorbance measurements of double-stranded DNA (dsDNA) samples from the proband, mother, and father. The samples were quantified using both spectrophotometry and Qubit 3.0 Fluorometer methods. The table includes data on the concentration measured by spectrophotometry (in µg/mL), absorbance ratios A260/230 and A260/280, and the dsDNA concentration measured by the Qubit 3.0 Fluorometer (in µg/mL). Desirable DNA purity absorbance ratio: A260/230 1.80–2.00, A260/280≥2.00.

Table 3
Descriptive information of cleaned sequencing data.

Sample	BioSample accession number	SRA accession number	Clean Reads	Clean Base	Read Length	Q20(%)	Q30(%)	GC(%)
Proband	SAMN39972817	SRR27997292	48,121,571	14,436,471,300	PE150	97.68	93.55	51.27
Mother	SAMN39972818	SRR27997291	48,165,644	14,449,693,200	PE150	97.53	93.13	50.95
Father	SAMN39972819	SRR27997290	48,117,318	14,435,195,400	PE150	97.66	93.50	50.44

This table provides detailed information on the cleaned sequencing data for three samples: Proband, Mother, and Father. **Sample:** Identifier of the sample. **BioSample Accession Number:** Unique identifier assigned to each sample in the BioSample database. **SRA Accession Number:** Unique identifier for each sample in the Sequence Read Archive. **Clean Reads:** Number of sequencings reads after cleaning. **Clean Base:** Total number of bases in clean reads. **Read Length:** Length of each sequencing read. **PE150:** paired-end 150 base pairs. **Q20 (%):** Percentage of bases with a quality score of 20 or higher. **Q30 (%):** Percentage of bases with a quality score of 30 or higher. **GC (%):** Percentage of guanine and cytosine bases in the sequencing reads.

Table 4
Sequence alignment and variant calling statistics.

Metrics	Proband	Mother	Father
Sequence alignment			
Total reads (QC-passed + QC-failed)	46,571,627	57,063,310	43,655,733
Secondary alignments	0	0	0
Supplementary alignments	261,587	265,556	234,341
Duplicates	0	0	0
Mapped reads	46,560,885 (99.98%)	57,052,295 (99.98%)	43,644,291 (99.97%)
Paired in sequencing	46,310,040	56,797,754	43,421,392
Read 1	23,155,020	28,398,877	21,710,696
Read 2	23,155,020	28,398,877	21,710,696
Properly paired	45,580,370 (98.42%)	56,012,376 (98.62%)	42,707,274 (98.36%)
With itself and mate mapped	46,290,124	56,777,386	43,400,548
Singletons	9,174 (0.02%)	9,353 (0.02%)	9,402 (0.02%)
Mate mapped to different chr	519,754	556,436	500,090
Mate mapped to different chr (mapQ $\geq$ 5)	434,509	465,748	417,530
% mapQ $\geq$ 5	83.6%	83.7%	83.5%
Variant Calling			
SNPs	355,041	381,531	342,916
MNPs	0	0	0
Insertions	25,433	27,049	24,069
Deletions	29,301	29,068	25,752
Indels	237	233	191
Same as reference	269,513	238,966	284,094
Missing genotype	627	8,884	8,729
Partial genotype	174	238	218
SNP transitions/transversions (Ts/Tv)	2.22	2.26	2.26
Total Het/Hom ratio	0.79	0.71	0.7
SNP Het/Hom ratio	0.78	0.69	0.68
MNP Het/Hom ratio	-	-	-
Insertion Het/Hom ratio	0.78	0.79	0.76
Deletion Het/Hom ratio	0.96	0.95	0.91
Indel Het/Hom ratio	78	57.25	94.5
Insertion/Deletion ratio	0.87	0.93	0.93
Indel/SNP+MNP ratio	0.15	0.15	0.15

This table summarizes the sequence alignment and variant calling statistics for the proband, mother, and father. The key metrics include the total number of reads, mapped reads, duplicates, and various quality indicators for sequence alignment, as well as counts for SNPs, indels, and other variant statistics. **QC**: Quality Control. **SNP:** Single Nucleotide Polymorphism. **MNP**: Multiple Nucleotide Polymorphism. **Indel**: Insertion-Deletion. **Het/Hom**: Heterozygous/Homozygous ratio. **Ts/Tv**: Transitions/Transversions ratio. **Chr**: Chromosome.

Table 5
Sample concordance.

Concordance result	Value
Trio variant calling with DP >10 (A)	119,598
Mendelian consistency status cannot be determined (B)	1.490
Variant violates Mendelian inheritance constraints (C)	1.254
Overall Concordance with all DP>10 (A-B-C)/(A-B)	**98.9%**
Trio variant calling with DP >20 (E)	59,527
Mendelian consistency status cannot be determined (F)	747
Variant violates Mendelian inheritance constraints (G)	584
Overall Concordance with all DP>20 (E-F-G)/(E-F)	**99.0%**

This table provides concordance statistics for Mendelian inheritance based on trio variant calling data. The metrics include the number of variants with a read depth (DP) greater than 10 and 20, the number of variants with undetermined Mendelian consistency status, and the number of variants violating Mendelian inheritance constraints. Overall concordance percentages are calculated for each read depth threshold.

1.3. **Code the phenotype in FHIR HL7 terminology standard**: The data were manually coded using ICD-10, SNOMEDCT, LOINC, and/or HPO formats.

2. **Gather Genotype Data**

2.1. **Sample Collection:**

2.1.1. Blood samples were collected from the proband and both parents after obtaining informed consent.

2.1.2. The blood samples were processed to extract DNA according to the manufacturer's guidelines (e.g. Geneaid^{TM} DNA Isolation Kit).

2.1.3. DNA quantification was performed to ensure sample quality and concentration (e.g. Varioskan microplate reader and/or Qubit® 3.0 Fluorometer).

2.2. **Library Preparation and Sequencing:**

2.2.1. Library preparation and exome sequencing were performed according to the manufacturer sequencing protocol. An example of DNBSEQ procedure:

2.2.1.1. **DNA library construction:** DNA library construction was carried out using the SureSelect Human All Exon V6 kit, which includes steps such as gDNA shearing, size selection, end repair, A tailing, adaptor ligation, pre-PCR and hybrid capture, washing streptavidin beads, and post-PCR amplification.

2.2.1.2. **Sequencing:** The DNA libraries were sequenced using the BGI DNBSeq system, following standard procedures to prepare the library, reagents, and DNBs, and finally loading and running the sequencing chip.

2.2.2. Post-sequencing, raw reads were filtered to remove adaptor sequences, contamination, and low-quality reads using software (e.g. SOAPnuke software). Quality control metrics and variant statistics were calculated, including read depth (DP), quality scores (Q20, Q30), and GC content.

2.3. **Publish the genotype data:** Genotype data could be uploaded to national biodatabank or international repository (e.g. NCBI BioProject).

3. **Utilizing the Dataset for Rare Disease Variants Diagnosis Pipeline:**

3.1. **Data preparation:** Phenotype data can be obtained from this paper, and genotype data from NCBI BioProject.

3.2. **Analysis:**

3.2.1. Phenotype data were analyzed using phenotyping software or pipelines (e.g. IDeRare, Phenomizer, or other phenotype analysis tools).

3.2.2. Sequencing data were aligned to the GRCh38.p14 reference genome using either complete genotype analysis pipeline (e.g. IDeRare) or separate tools for sequence alignment (e.g. bwa, bwa-mem), duplicate removal and sorting (e.g. samtools, sambamba), variant calling (e.g. DeepVariant, DeepTrio, tiddit), variant annotation (e.g. SnpEff, SnpSift), phenotype-based gene prioritization (e.g. Exomiser).

3.3. **Variant Interpretation:**

3.3.1. Variants are filtered based on Mendelian inheritance patterns, with a focus on detecting pathogenic variants responsible for the proband's condition.
3.3.2. Identified variants are further classified into known and novel mutations with respective pathogenicity status according to The American College of Medical Genetics and Genomics (ACMG) classification.
3.3.3. The findings are compiled into a detailed clinical report, highlighting key phenotypic features and genetic variants by manually sort the relevant data or automatically (using IDeRare, Exomiser, or other reporting pipeline or software).

4.2. Phenotype data

Clinical finding phenotype data were gathered through clinical interviews with the patient's attending physician. All significant phenotypes and differential diagnoses were coded according to Fast Healthcare Interoperability Resource (FHIR) Health Level Seven International (HL7) terminology standards by a clinical informatician. Table 1 shows the coded phenotypes and differential clinical diagnoses using the following standards:

- **International Classification of Diseases, 10th Revision (ICD-10):** Used to represent diagnosis groups or disorders spectrum.
- **Systematized Nomenclature of Medicine Clinical Terms (SNOMEDCT):** Used to represent clinical findings, clinical problems, and clinical diagnoses.
- **Logical Observation Identifiers Names and Codes (LOINC):** Used to represent laboratory work-up results indicated as high (H) or low (L), separated by a pipe (|).
- **Human Phenotype Ontology (HPO/HP):** Used to represent specific clinical findings or problems for rare diseases that are not covered in SNOMEDCT terminology set.

4.3. Sample collection and DNA isolation

Blood samples were collected from the affected male child and both of his parents. Informed consent for research and publication was obtained from the parents. Purified DNA was extracted from the blood buffy coat using reagents from Geneaid™ DNA Isolation Kit (Blood) according to the manufacturer's recommendation. Spectrophotometer quantification after isolation was perfomed using a Varioskan microplate reader (Thermo Scientific) prior to cryopreservation. Quantification of double-stranded DNA (dsDNA) was done using a Qubit®3.0 Fluorometer (Thermo Fisher Scientific) with the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific). Table 2 provides the quantification results of dsDNA of each sample.

4.4. Library preparation

The DNA library construction was carried out using SureSelect Human All Exon V6, 60Mb (Agilent, Santa Clara, CA, USA) according to the manufacturer's guidelines, with the following detailed steps: (1) **gDNA shear:** The qualified genomic DNA sample was randomly fragmented; (2) **Size selection:** the size of the library fragments was primarily distributed between 150bp and 250bp; (3) **End repair, A tailing:** The end repair of DNA fragments was performed, and an "A" base was added at the 3'-end of each strand; (4) **Adaptor ligation:** Adapters were ligated to both ends of the end-repaired/dA-tailed DNA fragments for amplification and sequencing; (5) **Pre-PCR and Hybrid Capture:** Size-selected DNA fragments were amplified, purified, and hybridized to the exome array; (6) **Wash streptavidin beads:** Non-hybridized fragments were washed out; (7) **Post-PCR:** Captured fragments were circularized, followed by the sequencing process.

4.5. Whole exome sequencing data

The DNA libraries were sequenced using BGI DNA Nanoball Sequencing (DNBSEQ) system according to the manufacturer recommendations: (1) preparing the library, (2) preparing the reagents for DNB master mix, (3) creating the DNB by rolling circle amplification (RCA), (4) quantifying the DNB, (5) adding DNB loading buffer to DNB product and placing it on the DNBs loading machine, (6) installing the sequencing chip and loading it, (7) removing the sequencing chip out and installing it in sequencing machine, (8) loading the sequencing reagent kit, opening the DNBSeq software, and running the sequencing process.

After sequencing, the raw reads were filtered to remove adaptor sequences, contamination, and low-quality reads from raw reads with read statistics shown in Table 3. Filtering was conducted using SOAPnuke [10] software developed by BGI with the filter parameters of *"-n 0.001 -l 10 –adaMR 0.25 –minReadLen 150",* which removes entire reads if sequencing reads matches ≥ 25% of adapter sequences, filter out sequencing read less than 150bp, remove entire reads if N content accounts for 0.1% of entire read, and filter base quality values less than 10.

4.6. Sequence alignment and variant calling

Sequence alignment and variant calling were conducted according to the IDeRare pipeline [9], which complies with germline genomic analysis best practice. Sequence alignment was compared with GRCh38.p14 reference sequence, duplicates were removed, and variants were called using DeepVariant and DeepTrio. Table 4 shows alignment statistics from *sambamba flagstat* command and variant calling statistics. Table 5 shows mendelian inheritance concordance by manually hard filtering the read depth (DP) from the trio variant calling file, removing undetermined (./.) variant, and counting variants violating Mendelian inheritance constraints.

Limitations

None.

Ethics Statement

Ethical clearance was obtained from the Ethics Committee of the Faculty of Medicine, University of Indonesia – Cipto Mangunkusumo Hospital (approval number: KET-1395/UN2.F1/ETIK/PPM.00.02/2022). Written informed consent was obtained from patient's parents for all experiments described here, biological sample usage, and publication.

Data Availability

Trio Whole Exome Sequencing of Rare Disease (Original data) ((NCBI)).

CRediT Author Statement

Ivan William Harsono: Conceptualization, Methodology, Software, Validation, Writing – original draft; **Yulia Ariani:** Conceptualization, Methodology, Writing – review & editing, Supervision; **Beben Benyamin:** Methodology, Software, Validation, Writing – review & editing; **Fadilah Fadilah:** Methodology, Software, Validation, Writing – review & editing; **Dwi Ari Pujianto:** Methodology, Writing – review & editing; **Cut Nurul Hafifah:** Writing – review & editing; **Titis Prawitasari:** Writing – review & editing.

Acknowledgments

This research was supported by grants from Directorate of Research and Development, Universitas Indonesia under Hibah PUTI 2023 (Grant No. NKB-162/UN2.RST/HKP.05.00/2023).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. European, Directorate-general for R, innovation: Rare diseases: a major unmet medical need. 2017.
[2] C.M.W. Gaasterland, M.C.J. van der Weide, M.J. du Prie-Olthof, M. Donk, M.M. Kaatee, R. Kaczmarek, C. Lavery, K. Leeson-Beevers, N. O'Neill, O. Timmis, et al., The patient's view on rare disease trial design - a qualitative study, Orphanet. J. Rare Dis. 14 (1) (2019) 31.
[3] I. Pearson, B. Rothwell, A. Olaye, C. Knight, Economic modeling considerations for rare diseases, Value Health 21 (5) (2018) 515–524.
[4] S. Nguengang Wakap, D.M. Lambert, A. Olry, C. Rodwell, C. Gueydan, V. Lanneau, D. Murphy, Y. Le Cam, A. Rath, Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database, Eur. J. Hum. Genet. 28 (2) (2020) 165–173.
[5] Z. Beyzaei, F. Ezgu, B. Geramizadeh, M.H. Imanieh, M. Haghighat, S.M. Dehghani, N. Honar, M. Zahmatkeshan, A. Jassbi, M. Mahboubifar, et al., Clinical and genetic spectrum of glycogen storage disease in Iranian population using targeted gene sequencing, Sci. Rep. 11 (1) (2021) 7040.
[6] T. Sandhu, M. Polan, Z. Yu, R. Lu, A. Makkar, Case of neonatal fatality from neuromuscular variant of glycogen storage disease type IV, JIMD Rep 45 (2019) 51–55.
[7] C.P. Austin, C.M. Cutillo, L.P.L. Lau, A.H. Jonker, A. Rath, D. Julkowska, D. Thomson, S.F. Terry, B. de Montleau, D. Ardigo, et al., Future of rare diseases research 2017-2027: An IRDiRC Perspective, Clin. Transl. Sci. 11 (1) (2018) 21–27.
[8] C.F. Wright, D.R. FitzPatrick, H.V. Firth, Paediatric genomics: diagnosing rare disease in children, Nat. Rev. Genet. 19 (5) (2018) 253–268.
[9] I.W. Harsono, Y. Ariani, B. Benyamin, F. Fadilah, D.A. Pujianto, C.N. Hafifah, IDeRare: a lightweight and extensible open-source phenotype and exome analysis pipeline for germline rare disease diagnosis, JAMIA Open 7 (2) (2024).
[10] Y. Chen, Y. Chen, C. Shi, Z. Huang, Y. Zhang, S. Li, Y. Li, J. Ye, C. Yu, Z. Li, et al., SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data, GigaScience 7 (1) (2017).