Whole Genome Sequencing of A Healthy Sundanese Individual As an Initial Study for Sundanese Population Genetic Reference Foundation

Richa Wijayanti Jamilah[1], *Dimas Andrianto*[1]*,*Rahadian Pratama*[1], *Vincentius Simeon Weo Budhyanto*[2], and *Nabila Tsoerayya Gustia Pudjas*[2]

[1]Biochemistry, IPB University, 16680 Bogor, Indonesia
[2]Satriabudi Dharma Setia Foundation, Tangerang, Indonesia

Abstract.

Sundanese is the second-largest tribe in Indonesia. Despite its substantial and diverse population, the application and database of whole-genome sequencing in Indonesia, especially among Sundanese individuals, are limited. This study aims to establish a genetic reference for Indonesia by presenting phenotypic and clinical data from Whole-Genome Sequencing of Sundanese individuals. Method selection of the subject based on physical health and three-generation Sundanese ancestry. Results show that clinical assessments revealed a body weight of 55.3 kg, a height of 171.8 cm, a BMI of 18.7, a fasting blood glucose level of 69 mg/dL, a cholesterol level of 141 mg/dL, and a uric acid level of 3.3 mg/dL. Anthropometric measurements, including nasal index, classified the nose as broad. DNA extraction using proteinase K yielded 28 ng/µL, with mechanical shearing recovering 40.04% of the extracted DNA. A total of 2,215,390 reads with an average coverage of 33.9x were obtained from the alignment process and genotype data, which can be used for downstream whole-genome sequencing analysis. These genotype data support downstream WGS analysis, contributing to a foundational genomic resource for Indonesia's Sundanese population and can be used as a basis for studying unique genetic variants and clinical implementation in the future.
Keywords: dna extraction, genome variants, healthy sundanese, whole genome sequencing

1 Introduction

Indonesia is the fourth most populous country in the world after India, China, and the United States. Based on its area and population, Indonesia is a country with a highly diverse

* Corresponding author: dimasandrianto@apps.ipb.ac.id

population [1]. This diversity is not limited to culture, language, race, and ethnicity. According to Statistics Indonesia, the number of ethnic groups in Indonesia has reached 1,340. The Sundanese tribe is one of the tribes with the largest population in Indonesia, after the Javanese tribe (40%), with a population of 15.5%.

Research on the human genome has expanded rapidly in developed countries worldwide. Data on genome variation in human populations is helpful for further studies, such as genome-wide association studies, which include disease, human-specific characteristics, migration patterns, and natural selection [2]. In 2022, Ministry of Health Republic of Indonesia launched the Biomedical & Genome Science Initiative that focuses on the Whole Genome Sequencing method with the target of collecting a genomic database of Indonesian people to be used as a genomic characterization of disease profiles, prediction of individual susceptibility to a disease based on the genome, and the potential discovery of personalized medicine.

Research on the human genome in Indonesia remains limited, with insufficient databases or reference data on population genomic variations, particularly for the Sundanese, a major ethnic group with a significant population. The condition is in stark contrast to the situation in developed countries, where a population genome database already exists, including genome banks from various individuals. The United Kingdom is one of the countries that already has a genome database of its population. It is estimated that there are more than 1 million individual genome data points available in the UK Biobank, with data such as genome sequences, health records, and lifestyle information used for reference in health and pharmacogenomics research [3]. Several other countries also have similar projects and databases, such as China (China Kadoorie Biobank), Iceland (deCODE Genetics), the United States (All of Us Project), and Sweden (Swedish Biobank).

Today, data from the Asia and Oceania region, including Indonesia, are available in the Indonesia Genome Diversity Project database [4]. However, this data only includes 161 individuals from 14 tribes' populations in Asia and Oceania. As of 2021, the population in Indonesia according to Statistics Indonesia had reached 280 million individuals from 1340 tribes. The study aims to screen subjects and analyze the quality of whole-genome sequencing data on healthy Sundanese individuals and determine the genomic variations. This study contributes significantly to the understanding of human genetic diversity by providing whole-genome sequencing (WGS) data from a healthy individual of Sundanese ethnicity and supports global efforts in genomic databases.

2 Methods

This research has been approved by the Human Research Ethics Commission of IPB University with the number: 1412/IT3.KEPMSM-IPB/SK/2024.

2.1 Respondents Screening

The screening form, which contains information on identity, health record, and ancestry, was distributed to 60 students of IPB University, with inclusion and exclusion criteria outlined in the informed consent form and approved by the ethics commission.

Subjects were then given informed consent and a respondent form that had been approved by the Research Ethics Commission of IPB University, and a physical examination was conducted. The process was followed by biological sampling.

Body Weight and height were measured using the Bfla One Station Body Check (Zhongshan Jinli Electronic), and blood pressure was measured using Omron (Suzhou Thriving Medical). Glucose, cholesterol, and uric acid concentrations in the blood were measured using

EasyTouch GCU (Bioptik Technology0. The Subject that matched all criteria had a blood peripheral sample taken intravenously using an EDTA Tube (OneMed).

2.2 DNA Extraction

Blood samples were extracted according to the FavorGen FavorPrep Blood Genomic DNA Extraction Mini (FavorGen Biotech Corp) product protocol. A total of 200 µL of blood sample was transferred into a microcentrifuge tube. Then, Proteinase K Solution and FABG Buffer were added, and the mixture was vortexed until homogeneous. The sample was incubated and then transferred to the FABG mini purification column in the tube. The process of washing and drying the column was continued. The sample was transferred to the elution tube, and elution buffer was added. Quantification of extracted DNA was carried out using the QuantiFluor(R) ONE dsDNA System, 500rxn Promega (Promega Corporation).

2.3 Ligation Sequencing Method

2.3.1 DNA Shearing

The extracted DNA was transferred entirely into the g-TUBE and then subjected to the DNA shearing stage using mechanical techniques with a G25 micro syringe needle, pulled up and down manually 70 times. The remaining volume was determined using a micropipette, and concentration measurements were performed using a Quantus.

2.3.2 Library Preparation

Library Preparation was performed using the Ligation kit V14 DNA Sequencing SQK-LSK114 (Oxford Nanopore Technology), which involves three main processes: DNA repair and End-Prep, Adapter Ligation and Clean-up, and Priming and loading. In the DNA repair and End-Prep process, the fragmented, sheared DNA was repaired and prepared for adapter ligation by adding DNA Repair Buffer, DNA Repair Mix, and Ultra II end-prep enzyme mix, and (v=1:1) AMPure XP Beads (AXP). Washing of the pellets attached to the beads with Ethanol in Nuclease-Free Water was also performed.

Adapter Ligation and Clean-Up are performed to attach specific DNA adapters to the selected fragments and facilitate the downstream process. End-prep DNA was mixed with Ligation Buffer reagent, Quick T4 DNA ligase, Ligation Adapter, and AMPure XP Beads (AXP). Beads were washed with Fragment Buffer, and qubit measurements were made on the Ligation DNA Adapter. Clean-Up aims to remove unwanted adapters. Next, the R.10.41 flow cell (FLO-MIN114) was prepared by opening the flow cell and removing a small volume with a micropipette to remove air bubbles. After 5 minutes, the sequencing buffer solution, library beads, and library solution were thoroughly mixed with the DNA library results. The priming mixture was put into the flow cell hole, and then the DNA library mixture was carefully added. The sequencing process was performed using PromethION P2 Solo (Oxford Nanopore Technology) sequencer.

2.4 Data Analysis

Raw FASTQ data from sequencing were analyzed using the "WGS Human" workflow on the EPI2ME platform from Oxford Nanopore Technology, starting with sequence quality control using FASTQC. Alignment with the human genome reference GRCh38

(PRJNA31257) was performed using minimap2, resulting in a BAM file. Variant calling was performed using Clair3 [5].

3 Result and Discussion

3.1 Sample Screening

Data from the distributed forms included personal data, disease history, tribal origin up to the grandparents' generation, and regional origin. Of the 60 respondents, 38 were willing to join the research, and 38 were screened for the ethnic aspect. Respondents who had a difference of one tribe in their family were excluded from the exclusion criteria. A total of 13 respondents were obtained, including 6 Sundanese respondents, 3 Javanese respondents from the Central Java region, and 4 Javanese respondents from the East Java region, with a total of 5 male respondents and 8 female respondents. The use of three generations of descendants can also determine mutations and epigenetics in the family tree [6]. According to the examination results and medical history collected from 6 Sundanese tribes, it is evident that 2 respondents, specifically samples with codes S01 and S04, meet the criteria for normal clinical parameters (Table 1).

Table 1. Clinical Data of Sundanese Respondents

Sample Code	Tribes	Sex	Weight (kg)	Height (cm)	Pulse (times/minute)	Body Mass Index	Fasting Glucose (mg/dL)	Cholesterol (mg/dL)	Uric Acid (mg/dL)
S01*	**Sunda**	M	**55.30**	**171.80**	**67**	**18.70**	**69**	**141**	**3.3**
S02	Sunda	F	74.30	140.90	95	37.40	73	182	6.1
S03	Sunda	F	60.50	158.20	103	24.20	91	140	3.0
S04	Sunda	M	53.50	170.10	73	18.50	80	102	3.0
S05	Sunda	M	76.70	177.70	83	24.30	67	177	17.0
S06	Sunda	F	72.20	156.80	72	29.40	84	124	4.3

*sample

: Does not meet blood chemistry profile screening standards

3.2 Phenotype of Sundanese Sample

The Sundanese subject (S01) is a 21-year-old male from Sukabumi, West Java. Based on the screening and interview results, the subject is a descendant of the Sundanese tribe, tracing back to three generations, including the grandparents of both fathers and mothers. The subject had no history of degenerative diseases and no family history of autosomal mutations. Autosomal mutations occur due to gene variations on one of the 22 autosomes, which can cause disease or disorder. Autosomal mutation diseases include polydactyly, in which patients clinically experience the addition of fingers and toes, thalassemia, which causes hemoglobin gene disorders, and hemophilia, a blood clotting disorder [7].

Health screening and anamnesis by medical doctors were carried out with examinations including Weight (BB), Height (TB), Body Mass Index (BMI), blood pressure, pulse rate, fasting blood glucose levels cholesterol levels, and uric acid, as well as anamnesis of complaints, history of drug consumption, and history of illness. Parameters of blood chemical levels, such as fasting blood glucose, cholesterol levels, and uric acid levels, are one of the efforts in screening for the detection of degenerative diseases, including diabetes mellitus, hypertension, cardiovascular disease, and other non-communicable diseases. Subjects who participated in the whole genome sequencing (S01) analysis were known to

have fasting blood sugar, cholesterol, and uric acid levels within the normal range. Health indicators, such as blood glucose levels, cholesterol levels, and uric acid levels, can be influenced by age, lifestyle, physical activity, and dietary habits [8].

The International Obesity Task Force has published a BMI assessment reference for Asian populations. Individuals with a BMI less than 18.5 are categorized as underweight. Individuals with a BMI ranging from 18.5 to 22.9 are categorized as usual, 23 to 24.9 as at risk of obesity, and a BMI ranging from 25 to 29.9 as grade I obesity; all other BMI values are considered grade II obesity.

The fasting blood glucose level indicates the overall glucose balance after a person has fasted for 8 hours. Normal levels of fasting blood glucose are <100 mg/dL. Blood cholesterol levels can also serve as a marker for metabolic disorders. People with high total cholesterol levels are prone to cardiovascular disease and have higher mortality. Normal total cholesterol levels are no more than 200 mg/dL [9].

In addition to clinical examination, ethnic anthropometric measurements were taken on 49 body parts, and nasal index calculations were also conducted. The nasal index measurement is 123.23, calculated from the ratio of nasal width and nasal height, and belongs to the type of nasal platyrrhine or broad nose with a nasal angle size of 45°. Nasal Index is a form of human anatomical identification that is determined based on the ratio of nose width and nose height. The nasal index is generally dependent on climatic conditions. There are four main types of Nasal Index categories, namely (i) Leptorrhine (fine nose) with a value of 70.00 or more and is generally found in populations with cold and dry climates, (ii) Messorrhine (medium Nose) index values ranging from 70.0 to 84.9, and (iii) Platyrrhine (broad nose) with index values greater than 85 and found in populations with warmer and humid climates. The nasal index of the Sundanese sample is categorized as Platyrrhine or broad nose. However, previous studies have indicated that there is no dominant nasal shape associated with a particular ethnicity [10].

3.3 **Data Quality**

Biological samples, in the form of blood, were collected through a vein and extracted using the whole blood method, as outlined in the Favorgen kit protocol, with the Proteinase K lysis agent. The blood was chosen as the biological sample due to considerations of yield and purity. Blood samples are known to produce lower microbial contamination than saliva samples in sequencing alignment results with human oral microbes, although the difference is not significant. The reading results from blood samples also show higher coverage than saliva samples [11].

The quantification result of DNA sample extraction was 28 ng/µL with a total eluent volume of 48 µL and 49 µL. Extraction was done in duplicate. Before the library preparation stage, the DNA Shearing stage was first carried out mechanically with a 25G syringe needle. This method was carried out as a modification to break high-molecular-weight DNA into fragments of 200 bp to 10 kb, depending on the needs of the downstream analysis. Mechanical DNA shearing techniques are based on breaking DNA fragments using physical (non-enzymatic) forces, such as sonication with ultrasonic waves, nebulization with high-pressure, slight friction, hydrodynamics, and needle shearing [12].

Mechanical DNA shearing is known to reduce sample loss during transfer and simplify the workflow. Mechanical techniques also show even coverage in Whole Genome Sequencing and produce a similar insert size distribution between libraries. In this study, the mechanical technique of the G25 syringe needle method was employed, and the results of qubit measurement of DNA shearing concentration were obtained at 18 ng/µL with an eluent volume of 61 µL. Thus, the total DNA recovery after and before the DNA shearing process reached 40.04% and the volume of eluent was reduced by 37 µL. The reduction in eluent

volume is due to the remaining eluent taken using a syringe left on the needle column or attached to the needle body wall.

Furthermore, after completing the adapter ligation and clean-up stages, the concentration of DNA, as measured by Qubit, was 30 ng/μL with a total eluent volume of 25 mL. This concentration is the final concentration of the library preparation, and then it is pushed into the flow cell for the priming process. Based on these results, the total DNA for the priming process was 750 ng, which is within the minimum amount recommended by the kit.

Sequencing process using promethION P2Solo flow cell PAY03998 for 72 hours, then FASTQ sequencing data was carried out quality control using FASTQC and continued alignment against the human genome reference GRCh38 or hg38 using minimap2 and produced BAM files which were further analyzed to obtain quality metrics such as read length, read quality, and mean coverage. The GRCh38 reference is the latest version of the human genome reference, derived from donor genomes, and was based on the findings of the Human Genome Project.

Based on the alignment results of the sample on the reference sequence, the total reads and other data are obtained in Table 2. Total reads Alignment is the number of reads or reads from sequencing results that can be recognized and matched against the genome reference. From the alignment results, it was found that the Sunda Tribe sample mapped 100% to the hg38 reference genome. The alignment results also show a mean coverage value of 33.9x, which means that each sample genome position was read an average of 34 times (Table 2).

Table 2. Alignment Result Matrix

Matrix	Value
Total Reads	2.215.390
N50	14.715 bp
% Mapped read	100%
Mean Coverage	33,908x

3.4 Variants Analysis

According to the variant calling results, the Sundanese S01 sample was found to have a total of 596,799 genetic variants, comprising 473,269 single-nucleotide variants (SNVs) and 124,094 insertion-deletions (INDELs) (Figure 1). This distribution aligns with the general characteristics of human genomic variants. SNVs were the most common type of variant found. This proportion of variants indicates a high level of genetic variation within an individual. The transition to transversion ratio (Ti/Tv ratio) of 2.07 indicates good data quality and is consistent with the ratio value commonly found in human genome data, which is around 2.0-2.1 [13]. This value suggests that the identified variants are likely to be biologically relevant.

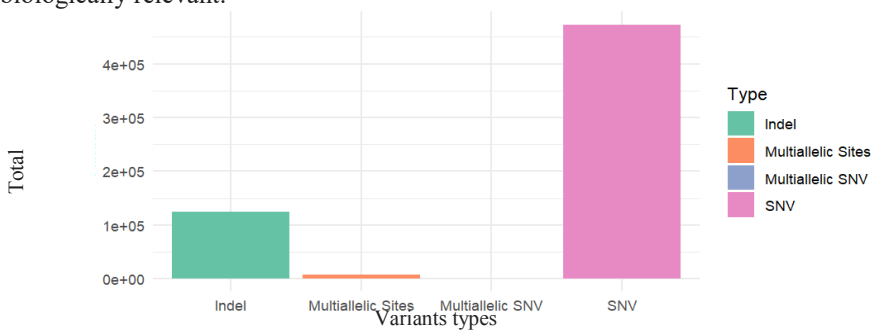

Fig. 1. Distribution of Total variants.

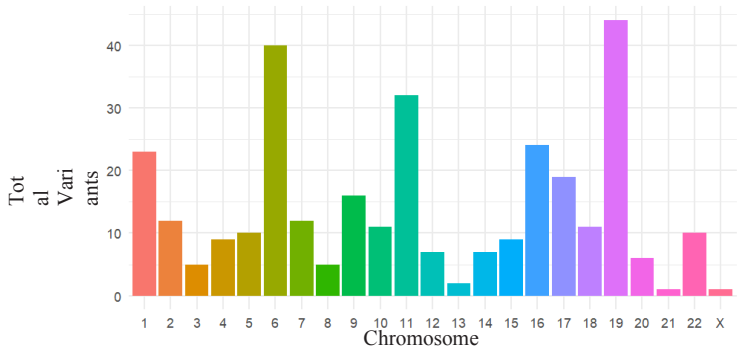

Fig. 2. Total Variants per chromosome in the ClinVar Database

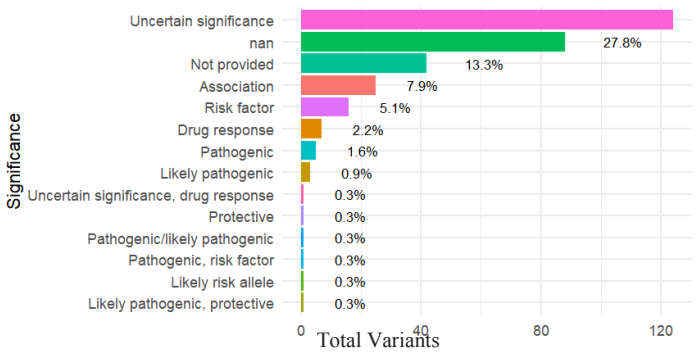

Fig. 3. Variants significance Distribution in the Clinvar Database

Through annotation with the ClinVar database, the variant significance distribution of the genomic variants of sample S01 was obtained (Figure 2). The most variants, with a percentage of 27.8%, are of uncertain significance. This means that the data in the database was not enough to conclude that a variant belongs to the common or pathogenic group. From the results of the distribution of variant significance, there are at least 1.6% pathogenic variants and 0.9% likely pathogenic (Figure 3). One of the pathogenic variants is located on chromosome 1 at position 976,215 and is classified as a Single Nucleotide Variant (SNV) in the PERM1 gene, resulting in a base change from Thymine (T) to Cytosine (C) at nucleotide position 2330 in the gene transcript. The consequence of this variant results in the inclusion of a different amino acid in the protein (missense). The PERM1 gene encodes proteins that regulate muscle and heart function, as well as energy metabolism [14].

4 Conclusion

Whole-genome sequencing analysis was performed on clinically screened healthy Sundanese individuals (S01). The anthropometric nasal index of the Sundanese sample is a wide nose type. Biological blood samples were extracted, and DNA fragmentation was performed using the needle shearing method, resulting in a recovery rate of 40.04% of the extracted DNA. Library preparation was performed using a total of 750 ng of DNA, and the

alignment process on the GRCh38 reference genome yielded 2,215,390 reads with a mean coverage of 33.9x. The Sundanese sample genome mapped 100% to the reference genome. Sample S01 has a total genomic variation of 473,269 SNVs and 124,094 INDELS. The significance of pathogenic variants detected based on ClinVar annotation was 1.6%.

The future work of this research is to determine the genetic variation pattern and the unique variants of the Sundanese tribe, thereby establishing a genetic reference for the Sundanese population. This study will necessitate the recruitment of additional subjects.

Authors express sincere gratitude to Yayasan Satriabudi Dharma Setia and Panin Bank for their support and valuable contributions to this work using grant number 0927/YSDS/PKS/VIII/2023 with Dimas Andrianto, PhD as the Recipient.

References

1. W. Anggraini, "Keanekaragaman hayati dalam menunjang perekonomian masyarakat kabupaten oku timur," *Jurnal Aktual STIE Trisna Negara*, vol. 16, no. 2, pp. 99–106, (2018). https://doi.org/10.47232/aktual.v16i2.24.
2. T. K. Clarke *et al.*, "Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK biobank (N=112117)," *Mol Psychiatry*, vol. 22, no. 10, pp. 1376–1384, Oct. (2017). https://doi.org/10.1038/mp.2017.153.
3. B. V. Halldorsson *et al.*, "The sequences of 150,119 genomes in the UK Biobank," *Nature*, vol. 607, no. 7920, pp. 732–740, Jul. (2022). https://doi.org/10.1038/s41586-022-04965-x.
4. N. Brucato *et al.*, "Chronology of natural selection in Oceanian genomes," *iScience*, vol. 25, no. 7, Jul. (2022). https://doi.org/10.1016/j.isci.2022.104583.
5. Z. Zheng, J. Su, L. Chen, Y.-L. Lee, T.-W. Lam, and R. Luo, "ClairS: a deep-learning method for long-1 read somatic small variant calling". https://doi.org/10.1101/2023.08.17.553778.
6. A. Kanzi *et al.*, "Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance," *Front Genet*, vol. 11, p., (2020). https://doi.org/10.3389/fgene.2020.544162.
7. P. Xie, F.-Q. Yuan, H.-H. Zhou, X. Li, and Z.-Q. Liu, "The molecular genetics related to polydactyly: an updated review," *Pharmacogenomics Research and Personalized Medicine*, vol. 0, pp. 0–0, Jan. (2021). https://doi.org/10.21037/prpm-20-2.
8. A. Mustika *et al.*, "Correlation Glucose, Uric Acid, and Cholesterol Levels Towards Health Conditions in the Highlands: POCT Approach," *Indonesian Journal of Clinical Pathology and Medical Laboratory*, vol. 3, no. 30, pp. 206–211, (2024). https://doi.org/10.24293/ijcpml.v30i3.2203.
9. M. Essa *et al.*, "Abstract P178: Prevalence of Elevated LDL-C and Non-HDL-C in Adults With Total Serum Cholesterol in the Normal Range," *Circulation*, vol. 149, no. 1, pp. 1–7, Mar. (2024). https://doi.org/10.1161/circ.149.suppl_1.P178.
10. A. Karolina, A. A. Rusman, and Y. F. Syukriani, "The Similarity Between Chinese-Indonesian, Sundanese and Bataknese Based on Facial and Nasal Index," *Indonesian Facial and Nasal Index eJKI*, vol. 8, no. 2, (2020). https://doi.org/10.23886/ejki.8.12013.

11. R. A. Yao, O. Akinrinade, M. Chaix, and S. Mital, "Quality of whole genome sequencing from blood versus saliva derived DNA in cardiac patients," *BMC Med Genomics*, vol. 13, no. 1, Jan. (2020). https://doi.org/10.1186/s12920-020-0664-7.
12. T. Ribarska, P. M. Bjørnstad, A. Y. M. Sundaram, and G. D. Gilfillan, "Optimization of enzymatic fragmentation is crucial to maximize genome coverage: a comparison of library preparation methods for Illumina sequencing," *BMC Genomics*, vol. 23, no. 1, Dec. (2022). https://doi.org/10.1186/s12864-022-08316-y.
13. M. Dash, P. Meher, A. Kumar, S. S. Satapathy, and N. D. Namsa, "High frequency of transition to transversion ratio in the stem region of RNA secondary structure of untranslated region of SARS-CoV-2," *PeerJ*, vol. 12, no. 4, (2024). https://doi.org/10.7717/peerj.16962.
14. S. I. Oka *et al.*, "Perm1 regulates cardiac energetics as a downstream target of the histone methyltransferase Smyd1," *PLoS One*, vol. 15, no. 6, Jun. (2020). https://doi.org/10.1371/journal.pone.0234913.