Check for updates

Sequencing whole genomes of the West Javanese population in Indonesia reveals novel variants and improves imputation accuracy

Edwin Ardiansyah[1], Anca-Lelia Riza[2], Sofiati Dian[1,3],
Ahmad Rizal Ganiem[1,3], Bachti Alisjahbana[1,4],
Todia P. Setiabudiawan[5], Arjan van Laarhoven[5],
Reinout van Crevel[5] and Vinod Kumar[5,6]* on behalf of the
ULTIMATE consortium

[1]Research Center for Care and Control of Infectious Diseases, Universitas Padjadjaran, Bandung, Indonesia, [2]Laboratory of Human Genomics, University of Medicine and Pharmacy of Craiova, Craiova, Romania, [3]Department of Neurology, Hasan Sadikin Hospital, Faculty of Medicine, Universitas Padjadjaran, Bandung, Indonesia, [4]Department of Internal Medicine, Hasan Sadikin Hospital, Faculty of Medicine, Universitas Padjadjaran, Bandung, Indonesia, [5]Department of Internal Medicine and Radboud Center of Infectious Diseases (RCI), Radboud University Medical Center, Nijmegen, Netherlands, [6]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, Netherlands

Existing genotype imputation reference panels are mainly derived from European populations, limiting their accuracy in non-European populations. To improve imputation accuracy for Indonesians, the world's fourth most populous country, we combined Whole Genome Sequencing (WGS) data from 227 West Javanese individuals with East Asian data from the 1,000 Genomes Project. This created three reference panels: EAS 1KGP3 (EASp), Indonesian (INDp), and a combined panel (EASp + INDp). We also used ten West-Javanese samples with WGS and SNP-typing data for benchmarking. We identified 1.8 million novel single nucleotide variants (SNVs) in the West Javanese population, which, while similar to the East Asians, are distinct from the Central Indonesian Flores population. Adding INDp to the EASp reference panel improved imputation accuracy ($R2$) from 0.85 to 0.90, and concordance from 87.88% to 91.13%. These findings underscore the importance of including West-Javanese genetic data in reference panels, advocating for broader WGS of diverse Indonesian populations to enhance genomic studies.

Introduction

Genome-wide association studies (GWAS) have played a pivotal role in advancing our understanding of the genetic underpinnings of diseases over the past decade (Uffelmann et al., 2021; Sollis et al., 2023). Genotype imputation, has gained paramount importance in GWAS, by inferring unobserved genotypes (Marchini and Howie, 2010). To facilitate

imputation, the 1,000 Genomes Project (1KGP3) has furnished the genome sequencing data required for constructing reference panels (Auton et al., 2015). However, the current repository of genotypes is dominated by genotype information from the European populations, thereby inadequately representing global diversity. Consequently, imputing genotypes for individuals from non-European populations can result in reduced accuracy due to differences in genetic variation between populations (Schurz et al., 2019). Prior research has also indicated significant disparities in imputation performance when employing common reference panels across diverse populations (Huang et al., 2008; Lin et al., 2018). Such inaccuracies particularly limit the possibility of detecting genetic association at low-frequency variants, leading to incomplete understanding of the genetic architecture underlying complex diseases in ethnically diverse populations.

Indonesia, the world's fourth most populous country, with over 17,000 islands and a diverse population (ICF, 2018) is currently underrepresented in human genomic studies. The closest super-population in the 1KGP3 (Auton et al., 2015) database to Indonesia is the East Asian (EAS) panel, composed of populations from China, Japan, and Vietnam. However, this East Asian panel may inadequately represent the genetic landscape specific to Indonesia. Addressing this shortfall, the GenomeAsia 100 K (GAsP) consortium has sequenced 1,739 individuals from Asia (Wall et al., 2019). GAsP included 68 individuals from the central parts of Indonesia, but does not capture the rich genetic diversity present throughout the archipelago.

In the present study, we examined the largest cohort of genomes from individuals originating from West Java, the most populous province of Indonesia with 50 million inhabitants. Our primary objective was to identify novel genetic variants specific to this region. Furthermore, we investigated the added value of incorporating our WGS data in the current Asian reference panels to improve the accuracy of imputation.

Results

Whole genome sequencing of West Java population identifies 1.8 million novel SNVs

We performed whole genome sequencing (WGS) on 239 tuberculous meningitis patients from West Java using DNBSeq with 30× coverage. We excluded 3 individuals with inconsistent sex between genotype and phenotype data, 4 individuals visually identified as outliers in the principal component analysis (PCA) plot, and 5 individuals with high discordancy (>30%) between the WGS genotype count and the SNP-array genotype, leaving 227 samples for analysis. Within these genomes (Supplementary Table 1), we identified 14,283,158 single nucleotide variants (SNVs) and 2,449,610 insertions and deletions (InDels). Among these, 6,616,414 (46.32%) SNVs and 941,524 (38.44%) InDels had a minor allele frequency (MAF) of less than 1%; 2,250,980 (15.76%) and 506,675 (20.68%) between 1% and 5%; and 5,415,764 (37.92%) and 1,001,411 (40.88%) more than 5%. All of the variants were then annotated using dbSNP build 153 as the reference. Based on the annotation, 1,867,419 (13.07%) SNVs and 432,345 (17.65%) InDels with MAF less than 1%; 22,122 (0.15%) and
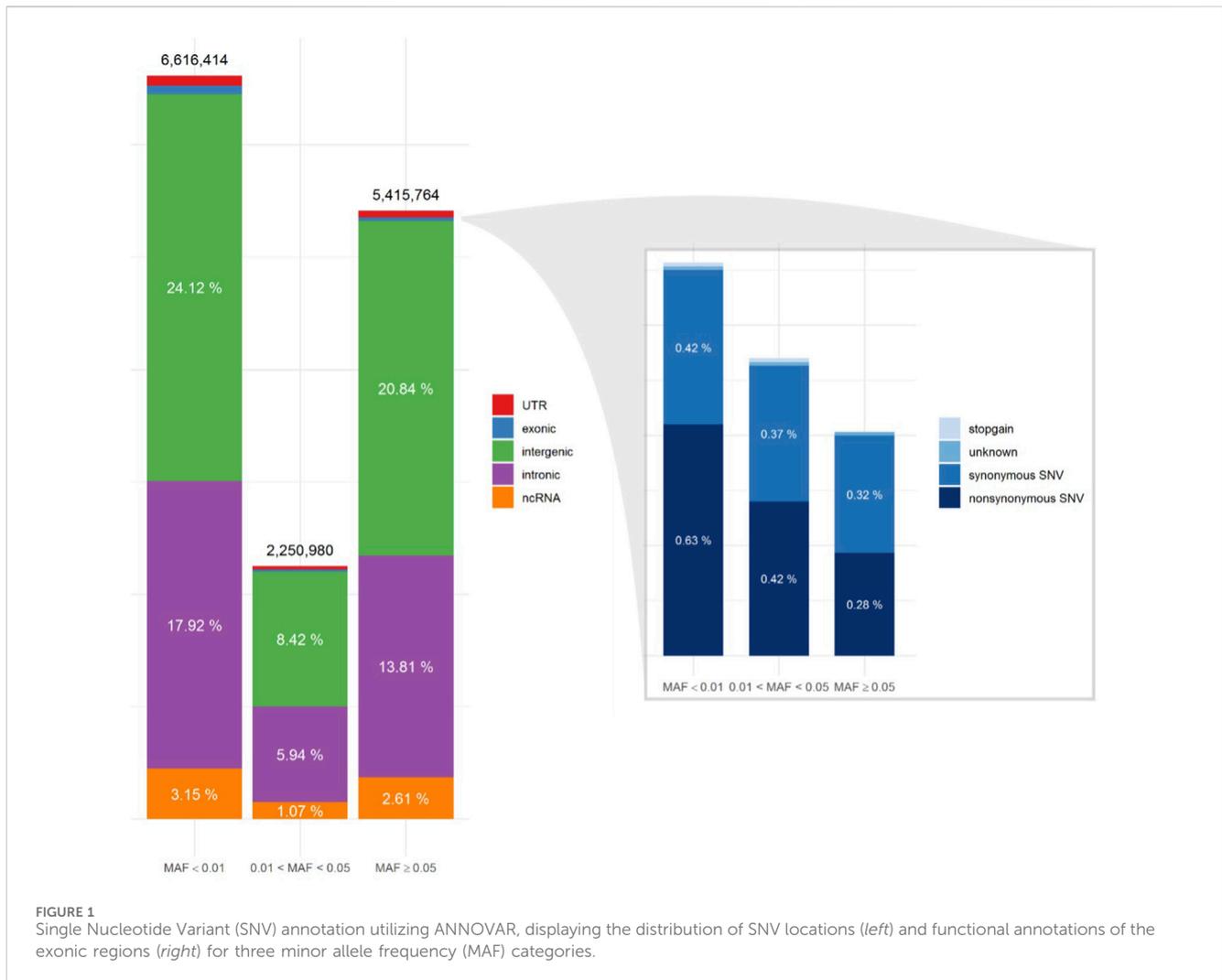
200,711 (8.19%) with MAF between 1% and 5%; and 137 (0.001%) and 449,936 (18.37%) with MAF more than 5% were novel.

To understand the genomic region and function of the variants, we performed region-based and functional annotation using ANNOVAR. The majority of the variants was located in intronic and intergenic regions of the genome, consistent in all minor allele frequency (MAF) categories (Figure 1). The proportion of nonsynonymous SNV (nsSNV) among the total 14,283,158 SNVs, increased from 0.28% in the common (MAF >5%) to 0.42 in the intermediate (MAF 1%–5%, $p < 0.001$) and 0.63% in the rare (MAF <1%) SNV categories (Figure 1).

West Javanese population genetic architecture is distinct from other East Asian genomes

To evaluate the population structure of the West Javanese population, we performed Principal Component Analysis (PCA) including worldwide population reference data from the 1,000 Genomes Project. The West Java population was found to be genetically closer to the East Asian (EAS) cluster than to other 1,000 Genomes populations (Figure 2A; Supplementary Figure 1), but did not overlap with the ethnicities (Japanese, Chinese and Vietnamese individuals) in the EAS (Figure 2B; Supplementary Figure 2). We also compared the genetic architecture of our West Javanese population to the individuals represented in the Genome Asia 100 K (GAsP) project. These 68 individuals came from Flores, another island located approximately 1,500 km east from West Java (Figure 3), and comprises of four different ethnicities: Flores Bena, Flores Cibal, Flores Rampasasa, and Austronesian. The West Javanese individuals were shown to be largely different from both the Austronesian population as well as the Flores Bena, Flores Cibal, and Flores Rampasasa ethnicities which together form a third cluster (Figure 2C; Supplementary Figure 3).

Next, we performed two unsupervised admixture analyses: one combining the West Javanese population with reference populations from the 1,000 Genomes Project (African, East Asian, Caucasian, Hispanic, and South Asian populations), and the other combining the West Javanese population with populations from the Genome Asia 100 K dataset (Austronesian, Flores Bena, Flores Cibal, and Flores Rampasasa populations). The results revealed that the West Javanese population's ancestry patterns closely align with those of the East Asian populations from the 1,000 Genomes Project. In contrast, the West Javanese population displayed distinct genetic differences from the African, Caucasian, Hispanic, and South Asian populations (Figure 4A). Comparisons with the Genome Asia 100 K populations further demonstrated that, despite originating from the same country, the West Javanese population exhibits a distinct ancestry pattern compared to other Indonesian populations within the Genome Asia 100 K dataset (Figure 4B). To quantify the number of shared SNVs found in West Java population, with the Flores population and the populations in 1KGP3, we counted shared SNVs and categorized them by their MAF (Table 1). Of the variants with a MAF <0.01 in IND, less than half had been identified in both the Flores or the 1KGP3 populations. With increasing MAF, the proportion of known variants increased to approximately 90%

FIGURE 1
Single Nucleotide Variant (SNV) annotation utilizing ANNOVAR, displaying the distribution of SNV locations (*left*) and functional annotations of the exonic regions (*right*) for three minor allele frequency (MAF) categories.

(GAsP) and 98% (1KGP3). The larger representation in 1KGP3 pEAS could be explained because of a larger sample size, but also emphasizes the unexplored genetic diversity present within different ethnicities in Indonesia.

Comparison of imputed genotype and whole genome sequencing data reveal reduced accuracy

In addition to whole genome sequencing, 219 out of the 227 West Java individuals were also genotyped using a SNP-chip (HumanOmniExpressExome-8 v1.0; Illumina; San Diego, CA, United States). After imputation against the East Asian population in the 1KGP3 reference and QC check ($R^2 > 0.3$ and MAF >0.1), a final set of 4,751,257 SNVs was obtained.
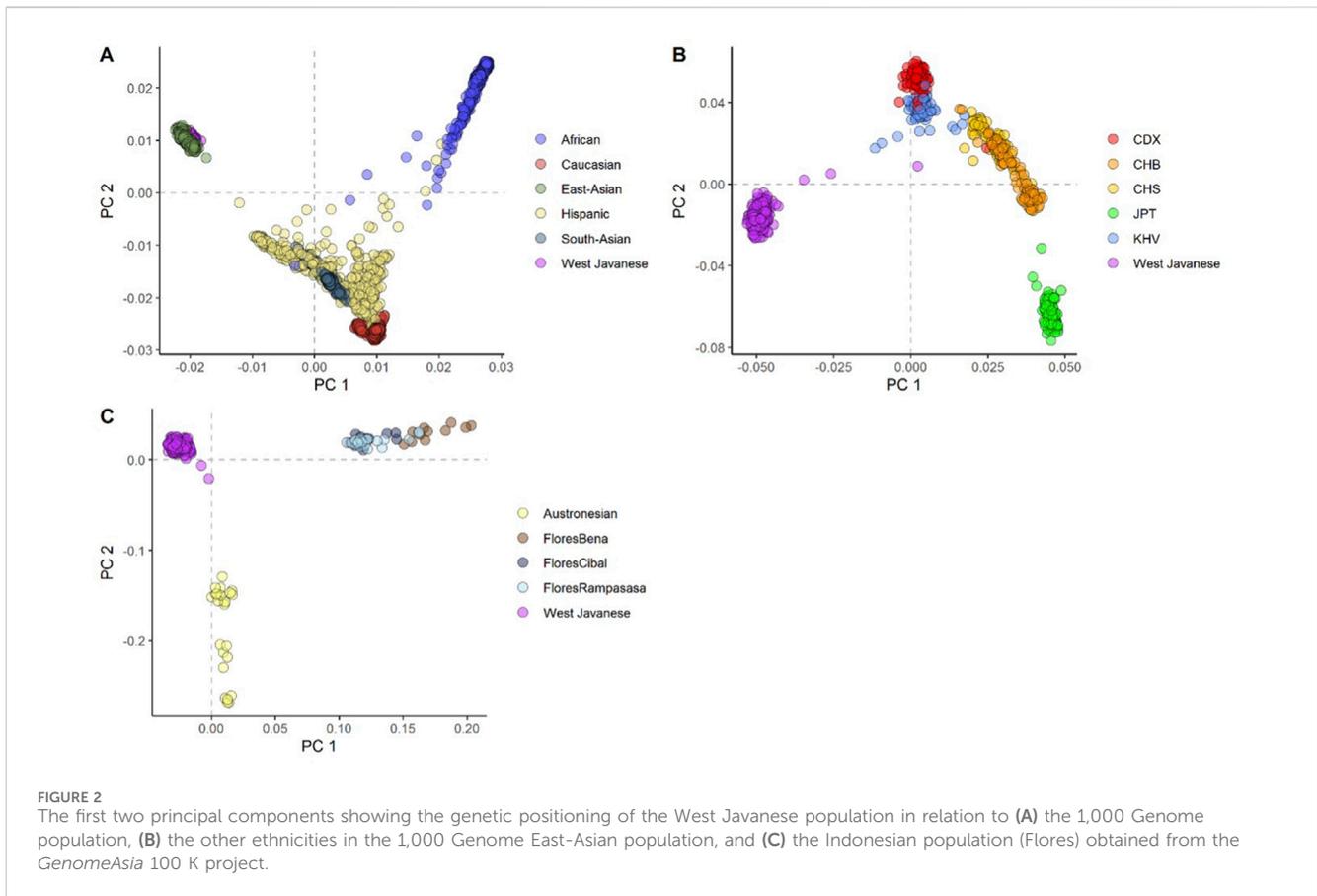
To assess the accuracy of the imputed genotypes, treating our whole genome sequencing as the truth set, we compared the genotypes measured from the 2 different platforms for all 219 individuals. To visualize, we plotted the heterozygous allele count between the imputed SNVs and the whole genome sequencing (Figure 5). As expected, SNVs with low R^2 had larger discrepancies

in heterozygous allele count between the imputed and whole genome sequencing. Importantly, even a cut-off of R^2 greater than 0.8, leaves relevant discrepancies.

Imputation evaluation against EAS, IND, and EAS-IND panel

Inclusion of population-specific genome sequencing data into a reference panel has shown to increase the imputation accuracy (O'Connell et al., 2021; Ritari et al., 2020). We aimed to assess the extent of accuracy improvement achievable by incorporating the West Java population into the existing 1KGP3 EAS panel. To accomplish this, we constructed three distinct reference panels: the 1KGP3 EAS panel (EASp), the Indonesian panel derived from West Java whole-genome sequencing (INDp), and the merged panel combining EAS and West Java datasets (EASp + INDp). Subsequently, we compared the imputation results of SNVs across all three reference panel configurations.

Out of 227 West Java WGS samples, 217 samples were utilized to create the reference panel consisting of INDp and EASp + INDp, and accuracy of imputation was evaluated for the three panels using

FIGURE 2
The first two principal components showing the genetic positioning of the West Javanese population in relation to **(A)** the 1,000 Genome population, **(B)** the other ethnicities in the 1,000 Genome East-Asian population, and **(C)** the Indonesian population (Flores) obtained from the *GenomeAsia* 100 K project.

two approaches: firstly, by comparing the R2 values (equivalent to the "info" metric in IMPUTE2). Ten remaining samples were used for benchmarking imputed SNVs against the WGS data by assessing the concordance between imputed SNVs and the WGS truth set in the 10 benchmark samples.

Imputation against EASp + INDp results in a higher count of SNVs with a total of 14,617,245 SNVs, compared to imputation to EASp (12,266,600 SNVs) or INDp (10,144,296 SNVs) individually. This outcome is in line with expectations, as merging two distinct panels increases the overall number of SNVs in the reference. Among the imputed SNVs, a subset of 7,792,202 SNVs is found in all three reference panels.

The EASp + INDp panel exhibited the highest mean R2 values, followed by the INDp and EASp panels (Figure 6A), across the entire minor MAF spectrum. This finding suggests an enhanced imputation accuracy achieved by incorporating the Indonesian population. Subsequently, we assessed the genotype concordance between the imputed genotypes and the whole-genome sequencing (WGS) dataset using the 10 benchmarking samples. To ensure a precise evaluation of the concordance, we excluded genotypes that were expected to be imputed as homozygous reference. This step was crucial to prevent the misclassification of rare variants with low MAF as having high concordance, as the majority of the population is expected to have the homozygous reference genotype. We observed that the concordance of the imputed genotypes from the INDp and EASp + INDp panels were comparable, and both exhibited higher concordance rates compared to the EASp panel (Figure 6B).

Discussion

In this study, by examining the largest Indonesian cohort of genomes from individuals originating from West Java, we not only identified >1.8 million novel variants but also highlight the rich genetic diversity present throughout the archipelago. Furthermore, by combining West-Javanese and East Asian reference panels, we built an imputation reference panel to demonstrate significant improvement in SNP imputation accuracy and the concordance between imputed and genotyped variants. The implications of our findings are two-fold. Firstly, given the limited representation of non-European populations in genetic studies, combining the genetic architecture of this population can potentially help in understanding East-Asian population genetics. Additionally, the utilization of an imputation reference panel that incorporates specific populations becomes crucial to improve the imputation accuracy and in turn facilitate GWAS in diverse populations.

The high imputation accuracy observed among European GWA studies can be attributable to diverse representation of European populations in 1,000 Genomes reference data. On the other hand, East Asian genomes from the 1,000 Genomes data as well as the GenomeAsia data (Auton et al., 2015; Wall et al., 2019) contains populations that are genetically closest to our study cohort. Although the GenomeAsia contains data from 219 population groups and 64 countries across Asia, our study clearly demonstrates that the Indonesian population possesses a
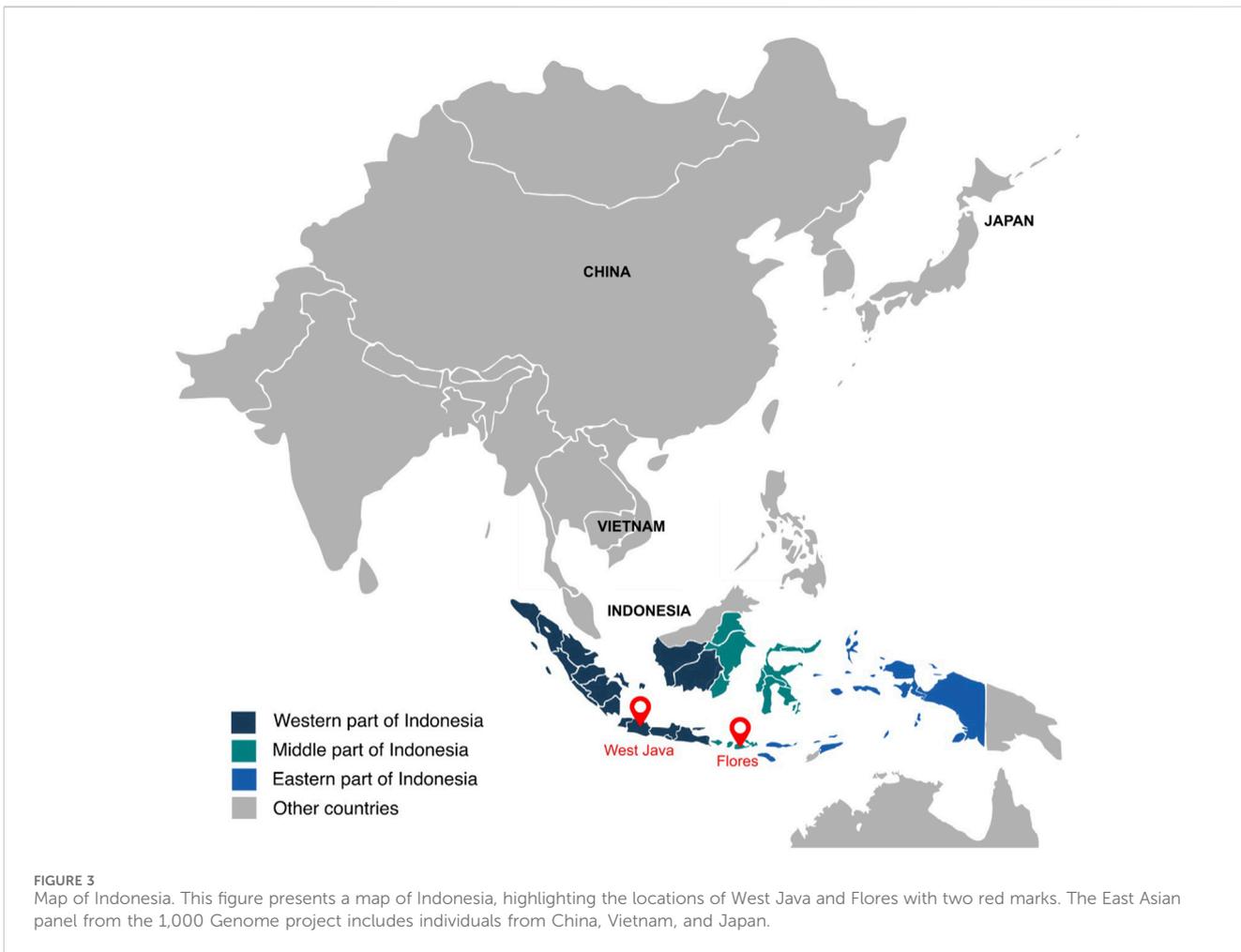
FIGURE 3
Map of Indonesia. This figure presents a map of Indonesia, highlighting the locations of West Java and Flores with two red marks. The East Asian panel from the 1,000 Genome project includes individuals from China, Vietnam, and Japan.
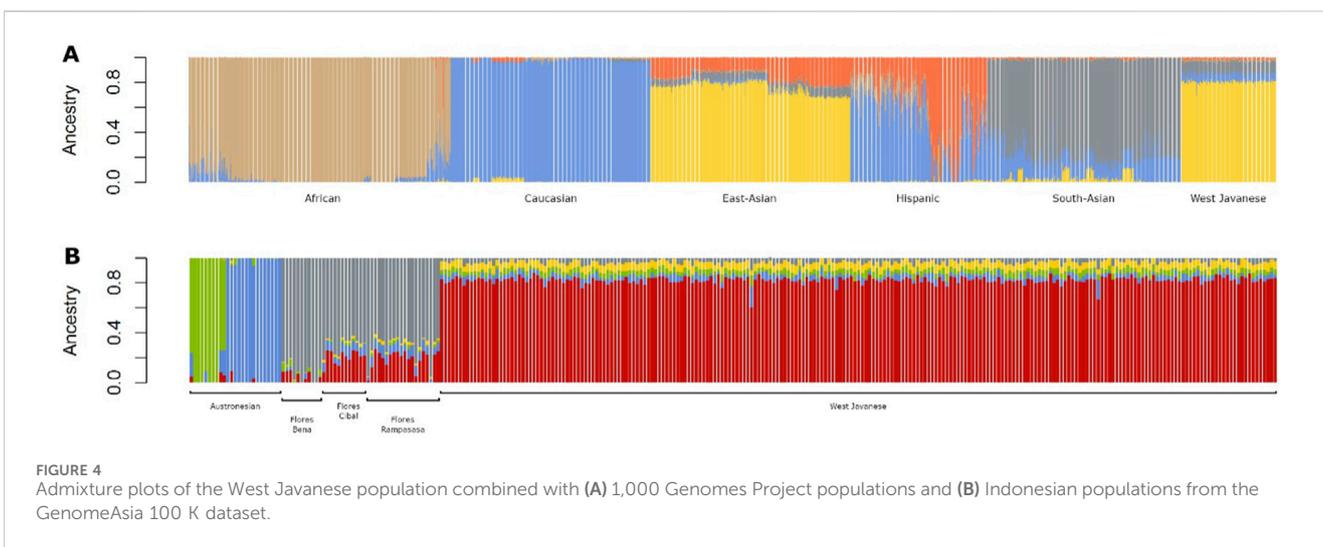

FIGURE 4
Admixture plots of the West Javanese population combined with **(A)** 1,000 Genomes Project populations and **(B)** Indonesian populations from the GenomeAsia 100 K dataset.

distinctive genetic architecture compared to neighboring Asian countries. Furthermore, we identified a substantial number of novel rare variants that were absent in the 1,000 Genomes East Asian population. These findings underscore the unique genetic profile of the Western Javanese population. Moreover, the results

demonstrated a clear distinction between the genetic makeup of the Western Javanese population and populations originating from the central region of Indonesia.

It is not surprising that adding the Western Javanese population WGS data to the reference panel improved the imputation accuracy

TABLE 1 Number of Single Nucleotide Variants (SNVs) identified in the genomes of West Javanese population, compared to the 100 K Asia and 1,000 G populations.

MAF	SNV (n)	Present in 100 K Asia	Present in 1,000 G
<0.01	6,818,363	1,093,308 (16.03%)	2,975,954 (43.65%)
0.01–0.05	2,263,232	1,445,656 (63.88%)	1,962,389 (86.71%)
0.05–0.1	1,054,398	922,880 (87.53%)	1,023,669 (97.09%)
0.1–0.2	1,393,726	1,248,343 (89.57%)	1,358,766 (97.49%)
0.2–0.3	1,109,694	999,800 (90.10%)	1,088,726 (98.11%)
0.3–0.4	966,199	869,030 (89.94%)	950,003 (98.32%)
0.4–0.5	922,765	829,212 (89.86%)	908,100 (98.41%)

FIGURE 5
The scatter plots display the heterozygosity count of Single Nucleotide Variants (SNVs) from individuals who underwent both whole genome sequencing and SNP-array genotyping, imputed against the East Asian panel of the 1,000 Genome project. Each dot on the plot represents an individual SNV, while the color indicates the imputation accuracy (R^2) of the SNP-array data. The plots are presented based on three R2 thresholds: **(A)** $R^2 > 0.3$, **(B)** $R^2 > 0.6$, and **(C)** $R^2 > 0.8$.

FIGURE 6
Imputation accuracy measurements. The figure presents a comparison of imputation accuracy across three panels using **(A)** internal "INFO" metrics of IMPUTE2 and **(B)** the actual genotype concordance between imputed single nucleotide variants (SNVs) and the whole-genome sequencing (WGS) dataset.

of common variants in our study. In fact, several previous studies have already shown the positive impact on imputation performance when additional populations were combined in reference panels (Jostins et al., 2011; Nelson et al., 2016; Vergara et al., 2018). However, this improvement in accuracy also varied depending on the allele frequencies of the variants, where much stronger

improvement was seen for common variants (MAF >0.1). It is shown that in addition to the use of large number of samples to build reference panels (Chou et al., 2016; Halldorsson et al., 2022), using population-specific reference panels (Mitt et al., 2017; Pistis et al., 2015) strongly benefit rare variant imputation. As Genome Asia currently has a small sample size of 68 individuals from Indonesia, our study helps to enrich the Asian reference datasets further to assist in rare variant imputation.

A notable strength of our study lies in the substantial sample size employed to construct the reference panel, consisting of 217 individuals. The inclusion of a large number of individuals allows for a more comprehensive representation of haplotypes within the population, consequently improving the accuracy of imputation. A potential limitation of our study is that the population in our study comprises of patients diagnosed with tuberculous meningitis, and this selection may impact the distribution and representation of SNPs within the population. Therefore, we cannot completely rule out the possibility that some of the variants could be associated with tuberculous meningitis. This may especially be relevant for rare variants that would put individuals at risk for tuberculous meningitis but less so for variants that are highly polymorphic (MAF>0.1). Given that the number of disease-affecting loci is small compared to the entire genome, capturing population-specific aspects should not be significantly impacted (Andrews et al., 2020). As cost of WGS is reducing, future studies should ideally include healthy individuals to mitigate the potential selection bias effects and further validate our findings.

In conclusion, our study provides evidence that the genetic architecture of the Indonesian population exhibits distinct characteristics when compared to other Asian countries. Our study also serves as an important resource to improve the GWAS of complex phenotypes in the West Javanese population. Given the rich ethnic diversity in the country, it is crucial for future genetic research to encompass a wider range of Indonesian ethnicities to capture the full extent of genetic diversity present in the country. By expanding our knowledge of the genetic architecture within Indonesia, we can pave the way for more targeted and effective precision medicine strategies tailored to the diverse needs of different ethnic groups.

STAR methods

Whole genome sequencing and data processing

As part of the ULTIMATE project (van Crevel et al., 2021), we sequenced DNA from 239 tuberculous meningitis patients. The genomic DNA sample was fragmented randomly by Covaris technology, resulting in fragments of 350 bp after selecting the appropriate size range. The fragmented DNA was subjected to end repair, followed by the addition of an "A" base at the 3′-end of each strand. Adapters were then ligated to both ends of the DNA fragments and subjected to amplification by ligation-mediated PCR (LM-PCR), followed by single-strand separation and cyclization. Rolling circle amplification (RCA) was used to generate DNA Nanoballs (DNBs) from the qualified DNBs,

which were loaded onto patterned nanoarrays and pair-end reads were obtained on the DNBseq platform. High-throughput sequencing was performed for each library to ensure adequate sequencing coverage. The raw image files generated during sequencing were processed by DNBseq base calling software using default parameters to obtain sequence data in the form of paired-end reads, which is defined as "raw data" and stored in FASTQ format.

FASTQ files of each sample was first mapped to the human reference genome (NCBI Build 37, hg19) (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/) using BurrowsWheeler Aligner (BWA) (Li and Durbin, 2009). To ensure reliable results, we followed the recommended Best Practices for variant analysis with the Genome Analysis Toolkit (GATK) (Schmidt et al., 2010; Van der Auwera et al., 2013; Depristo et al., 2011). Local realignment around InDels and base quality score recalibration was performed using GATK v3.7-0 (Schmidt et al., 2010; Van der Auwera et al., 2013; Depristo et al., 2011), with duplicate reads removed by Picard tools (Institute B, 2024). GVCF files for each sample were then created using HaplotypeCaller. All the individual GVCFs were then jointly genotyped using GATK GenotypeGVCFs, after first combining them with GATK CombineGVCFs. Variants were further recalibrated using GATK Variant Quality Score Recalibration (VQSR), and variants not passing the VQSR filtering and have missingness >5% were removed. Region-based and functional annotation were done using ANNOVAR version 7 June 2020 (Wang et al., 2010).

Whole genome sequencing data of Genome Asia 100 K

After receiving approval from the 100 K GenomeAsia consortium (Wall et al., 2019), VCF files of the Indonesian population were obtained, consisting of 68 individuals. In the final VCF files containing variant sites for Indonesian population, non-variant sites with an allele count of 0 were excluded.

Admixture analysis

Admixture analysis was conducted using the ADMIXTURE v1.3.0 tool (Alexander et al., 2009). Sequence data from West Javanese individuals were intersected separately with datasets from the 1,000 Genomes Project and the GenomeAsia 100 K. The combined datasets were pruned for linkage disequilibrium (LD) using PLINK v1.9 (Purcell et al., 2007). The optimal number of clusters (k) was determined based on cross-validation results.

Imputation reference panel creation

Using IMPUTE2 (Howie et al., 2009), we created three imputation reference panels: The East Asian panel (EASp), Indonesian panel (INDp), and the combined East Asian and Indonesian panel (EASp + INDp). Out of 227 Indonesian whole genome sequences (WGS) that passed QC, 10 Indonesian WGS were

taken out for benchmarking purpose, and the remaining 217 were used to create the reference panel. To obtain the EASp, the publicly available reference panel The 1,000 Genomes phase 3 (1KGP3) (Auton et al., 2015) was downloaded and the East Asian population was extracted. For both EAS and IND WGS data, multiallelic sites were split into biallelic sites, and variants with missingness >5%, Hardy Weinberg Equilibrium (HwE) $< 1 \times 10^{-10}$, and allele count <2 were removed. The 217 IND WGS were then phased using SHAPEITv2 (Delaneau et al., 2012) and converted to a reference panel using IMPUTE2 (Howie et al., 2009). The EAS and IND panel (EASp + INDp) was constructed by merging the EAS and IND panel using the "-merge_ref_panels_output_ref" option in IMPUTE2 (Howie et al., 2009), which performed the merging in two steps: first, imputing variants that were specific to one panel to the other panel and vice versa, and second, combining the two panels by taking the union of variants from both panels.

Imputation against IMPUTE2 reference panel

Genotypes of 509 TBM patients established with SNP-typing (HumanOmniExpressExome-8 v1.0; Illumina; San Diego, CA, United States) were used as data input for imputation. Variants with >5% missingness or HwE <0.00001 were removed, and then pre-phased using SHAPEITv2 (Delaneau et al., 2012). To align the strand against each of the reference panel, we used GenotypeHarmonizer (Deelen et al., 2014), which exploit linkage disequilibrium pattern to solves the unknown strand issue by aligning ambiguous A/T and G/C SNPs to a specified reference, thus eliminating the need of prior knowledge of the used strand.

Imputation was performed against the 3 panels using IMPUTE2 (Howie et al., 2009) by first dividing the input genome into 5 Mb chunks to increase computation and memory efficiencies. After completion of imputation, all chunks were re-combined to obtain the final imputed genotype, in IMPUTE2 haplotype format. The haplotype files were converted to VCF for further analysis using SHAPEITv2 (Delaneau et al., 2012).

Evaluation of imputation accuracy

To assess the imputation accuracy, we first utilized the internal quality metrics obtained from IMPUTE2, specifically the INFO score. Subsequently, we calculated the genotype concordance of the overlapping single nucleotide variants (SNVs) between the imputed SNVs and the whole-genome sequencing (WGS) dataset using *vcfcompare* tools. For an accurate evaluation of concordance, we excluded SNVs expected to be homozygous reference and calculated the concordance solely based on the alternate allele.

Data availability statement

The variant data for this study have been deposited in the European Variation Archive (EVA) at EMBL-EBI under accession number PRJEB85629 (https://www.ebi.ac.uk/eva/?eva-study=PRJEB85629).

Ethics statement

Author contributions

EA: Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing–original draft, Writing–review and editing, Software. A-LR: Data curation, Investigation, Methodology, Validation, Writing–review and editing. SD: Data curation, Methodology, Resources, Writing–review and editing. AG: Project administration, Resources, Writing–review and editing. BA: Project administration, Resources, Writing–review and editing, Funding acquisition. TS: Methodology, Validation, Writing–review and editing. AvL: Project administration, Resources, Visualization, Writing–review and editing, Conceptualization. RvC: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing–original draft, Writing–review and editing. VK: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing–original draft, Writing–review and editing.

Group member of the ULTIMATE consortium

Le Thanh Hoang Nhat, Hoang Thanh Hai, Nguyen Tran Binh Minh, Thai Minh Triet, Trinh Thi Bich Tram, Joseph Donovan, Dorothee Heemskerk, Tran Thi Hong Chau, Nguyen Thuy Thuong Thuong, Guy Thwaites (Oxford University Clinical Research Unit, Ho Chi Minh City, Viet Nam; Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom); Kirsten C. J. van Abeelen, Mihai G. Netea (Department of Internal Medicine and Radboud Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, Netherlands); Rovina Ruslami (Research Center for Care and Control of Infectious Diseases, Universitas Padjadjaran, Bandung, Indonesia); Julian Avila-Pacheco, Amy Deik, Jesse Krejci, Jeff Pruyne, Lucas Dailey, Clary B. Clish (Broad Institute of MIT and Harvard. Cambridge, Massachusetts, United States); Nguyen Duc Bang (Pham Ngoc Thanh Hospital, Ho Chi Minh City, Viet Nam); Riwanti Estiasari, Raph L. Hamers, Darma Imran, Kartika Maharani (Faculty of Medicine, Universitas Indonesia, Jakarta, Indonesia).

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1492602/full#supplementary-material

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi:10.1101/gr.094052.109

Andrews, S. J., Fulton-Howard, B., and Goate, A. (2020). Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *Lancet Neurol.* 19, 326–335. doi:10.1016/S1474-4422(19)30435-1

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi:10.1038/nature15393

Chou, W. C., Zheng, H. F., Cheng, C. H., Yan, H., Wang, L., Han, F., et al. (2016). A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. *Sci. Rep.* 6, 39313. doi:10.1038/srep39313

Deelen, P., Bonder, M. J., Van Der Velde, K. J., Westra, H. J., Winder, E., Hendriksen, D., et al. (2014). Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* 7 (1), 901. doi:10.1186/1756-0500-7-901

Delaneau, O., Marchini, J., and Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi:10.1038/nmeth.1785

Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi:10.1038/ng.806

Halldorsson, B. V., Eggertsson, H. P., Moore, K. H. S., Hauswedell, H., Eiriksson, O., Ulfarsson, M. O., et al. (2022). The sequences of 150,119 genomes in the UK Biobank. *Nature* 607, 732–740. doi:10.1038/s41586-022-04965-x

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5, e1000529. doi:10.1371/journal.pgen.1000529

Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., et al. (2008). Genotype-imputation accuracy across worldwide human populations. *Am. J. Hum. Genet.* 84, 235–250. doi:10.1016/j.ajhg.2009.01.013

ICF (2018). Indonesia demographic and Health survey 2017. Available at: http://dhsprogram.com/pubs/pdf/FR342/FR342.pdf.

Institute B (2024). *Picard toolkit*. Cambridge, United States: Broad Institute. Available at: http://broadinstitute.github.io/picard/.

Jostins, L., Morley, K. I., and Barrett, J. C. (2011). Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur. J. Hum. Genet.* 19 (6), 662–666. doi:10.1038/ejhg.2011.10

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Lin, Y., Liu, L., Yang, S., Li, Y., Lin, D., Zhang, X., et al. (2018). Genotype imputation for han Chinese population using haplotype reference consortium as reference. *Hum. Genet.* 137, 431–436. doi:10.1007/s00439-018-1894-z

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511. doi:10.1038/nrg2796

Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., et al. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* 25, 869–876. doi:10.1038/ejhg.2017.51

Nelson, S. C., Stilp, A. M., Papanicolaou, G. J., Taylor, K. D., Rotter, J. I., Thornton, T. A., et al. (2016). Improved imputation accuracy in Hispanic/Latino populations with larger and more diverse reference panels: applications in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Hum. Mol. Genet.* 25, 3245–3254. doi:10.1093/hmg/ddw174

O'Connell, J., Yun, T., Moreno, M., Li, H., Litterman, N., Kolesnikov, A., et al. (2021). A population-specific reference panel for improved genotype imputation in African Americans. *Commun. Biol.* 4, 1269. doi:10.1038/s42003-021-02777-9

Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., et al. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur. J. Hum. Genet.* 23, 975–983. doi:10.1038/ejhg.2014.216

Purcell, S., Neale, B., Todd-brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (September), 559–575. doi:10.1086/519795

Ritari, J., Hyvärinen, K., Clancy, J., FinnGen, P. J., and Koskela, S. (2020). Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. *Nar. Genomics Bioinforma.* 2, lqaa030. doi:10.1093/nargab/lqaa030

Schmidt, S., McKenna, A., Hanna, M., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

Schurz, H., Müller, S. J., Van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., et al. (2019). Evaluating the accuracy of imputation methods in a five-way admixed population. *Front. Genet.* 10, 34. doi:10.3389/fgene.2019.00034

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* 51, D977–D985. doi:10.1093/nar/gkac1010

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., et al. (2021). Genome-wide association studies. *Nat. Rev. Methods Prim.* 1, 59. doi:10.1038/s43586-021-00056-9

van Crevel, R., Avila-Pacheco, J., Thuong, T. T. N., Ganiem, A. R., Imran, D., Hamers, R. L., et al. (2021). Improving host-directed therapy for tuberculous meningitis by linking clinical and multi-omics data. *Tuberculosis* 128 (February), 102085. doi:10.1016/j.tube.2021.102085

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., et al. (2013). From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43, 1–11. doi:10.1002/0471250953.bi1110s43

Vergara, C., Parker, M. M., Franco, L., Cho, M. H., Valencia-Duarte, A. V., Beaty, T. H., et al. (2018). Genotype imputation performance of three reference panels using African ancestry individuals. *Hum. Genet.* 137, 281–292. doi:10.1007/s00439-018-1881-4

Wall, J. D., Stawiski, E. W., Ratan, A., Kim, H. L., Kim, C., Gupta, R., et al. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 576 (7785), 106–111. doi:10.1038/s41586-019-1793-z

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603