A REPRODUCIBLE FULL-STACK FRAMEWORK FOR NATIONAL AND ETHNIC GENOMIC DATABASES: ENABLING LABORATORIES TO ARCHIVE LOCALLY-DERIVED SNVS, INDELS, AND CNVS DIRECTLY FROM VCF FILES

Savindu Lathika Weerathunga[1],Nilaksha Freeson Neththikumara[2*] , Therese Lakshmi Shamala Tirimanne[1] , Vajira Harshadeva Weerabaddana Dissanayake[2]

[1]Department of Plant Sciences, Faculty of Science, University of Colombo, Colombo, Sri Lanka

[2]Center of Genetics and Genomics, Faculty of Medicine, University of Colombo, Colombo, Sri Lanka

*Corresponding author: nilaksha@genetics.cmb.ac.lk

Abstract

National and Ethnic Genomic Databases (NEGDs) serve as centralized repositories for archiving human genetic variants from specific local populations while Reference Population Databases archive global population variants greatly skewed towards European populations. While laboratories increasingly generate sequencing data to the Variant Calling Format (VCF) level, many regions lack standardized infrastructure to ingest these files, compute local allele frequencies, and archive the full spectrum of variant classes especially for exome-derived Copy Number Variants (CNVs), for which dedicated archival architecture remains scarce, limiting informed clinical interpretation. This study aimed to develop a reproducible, full-stack NEGD framework capable of ingesting de-identified VCF files already validated and received ethical approval and participant consent for publishing, to store Single Nucleotide Variants (SNVs), indels and CNVs, allowing secure submission, browsing, and download of variants with population specific metrics, deployable by any laboratory or country as a local resource. Initially, existing NEGD frameworks were reviewed to identify common architectural principles. This was integrated into a three-tier modular system: a relational database (MySQL) optimized for multi-scale variant storage; a React-based frontend; and Django backend featuring an automated ingestion pipeline that parses VCF files and dynamically calculates local allele frequencies, with external API integration for sourcing global annotations. The framework was validated through a reference deployment of the Sri Lanka Genetic Variation Database containing SNV and CNV data from over 1,000 exomes and which is maintained by the Centre for Genetics and Genomics, Faculty of Medicine, University of Colombo. The architecture enables querying of variants via simple and complex searches, with dedicated variant pages depicting local allele frequency, variant type, and consequences alongside global frequencies and links to external databases. A secure portal allows users to upload, edit, and manage submitted data. This framework addresses a critical gap by enabling archival of variants for underrepresented populations while supporting equitable diagnostics and personalized medicine.

Keywords: National and Ethnic Genomic Database, Human genetic variants, Biomedical Informatics