IMPLEMENTATION OF AN AI-ASSISTED AUTOMATED PIPELINE FOR COMPREHENSIVE QUALITY CONTROL, ENRICHMENT ANALYSIS, AND REPORT GENERATION OF EXOME SEQUENCING DATA

GodayalageMalshaSandeepani Gunasinghe[1], NilakshaFreeson Neththikumara[2]*, Therese Lakshmi Shamala Tirimanne[1], Vajira Harshadeva Weerabaddana Dissanayake[2]

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

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

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

Abstract

Genetic variant identification plays a central role in human medical research. Whole Exome Sequencing (WES), a Next-Generation Sequencing (NGS) technology, targets disease-associated protein-coding regions to enable variant calling. The reliability of these analyses depends on sequencing quality, alignment accuracy, and target enrichment efficiency. Although many bioinformatics tools exist, they remain fragmented and require significant expertise to interpret quality control (QC) metrics. In the present study, ExomeQC was developed to address this gap by providing an integrated framework for quality assessment and interpretation of human WES secondary analysis data. The pipeline integrates well-established tools, includingSAMtools, BEDtools, GATK, and mosdepth to compute a comprehensive set of sequencing and target enrichment quality metrics. The gene-level analysis module was developed to assess coverage and sequencing quality for user-selected genes, to calculate per-target metrics by integrating GATK and novel methods. The output of the primary pipeline is subsequently analyzed using the LlaMA 3.1 (8B) large language model (LLM) integrated with a Retrieval-Augmented Generation (RAG) framework to generate automated interpretations of quality metrics. The knowledge base for the RAG system was constructed using a Facebook AI Similarity Search (FAISS) vector database, with embeddings generated using a BAAI/bge-base-en-v1.5 embedding model. The pipeline was executed against secondary WES data analyzed at the Center for Genetics and Genomics, Faculty of Medicine, University of Colombo. Output evaluation against industry benchmarks and senior bioinformatics expert review confirmed the production of expected values and patterns. Evaluation using the RAGAs framework, the LLM-generated interpretations achieved a faithfulness score of 0.9815, a context precision score of 0.8041 and an answer relevancy score of 0.6282, indicating minimal hallucination and unsupported content, while achieving moderately relevant interpretation. To enhance usability, a graphical user interface (GUI) was developed, while a command-line interface (CLI) ensures seamless integration into existing bioinformatics workflows. The pipeline eliminates the technical bottlenecks traditionally associated with multi-tool QC workflows and delivers a reliable and reproducible automated reporting mechanism significantly supporting clinical and genomic research data interpretation.

Keywords: Next-Generation Sequencing (NGS), Whole Exome Sequencing (WES), Large Language Model (LLM), Retrieval-Augmented Generation (RAG)