HBVgenomer: A comprehensive workflow for mixed-strain hepatitis B virus genomic surveillance and haplotype reconstruction

Authors:

Choga WT[1], Ratsoma-Sethibe T[1,2], Gaongalelwe FG[1], Phakedi BK[1,3], Phinius BB[1], Anderson M[1,4,5], Moyo S[1,3,6,7,8], Gaseitsiwe S[1]

[1] Botswana Harvard Health Partnership, Gaborone, Botswana, [2] Department of Biological sciences, Faculty of Science, University of Botswana, [3] Department of Medical Sciences, Faculty of Allied Health Professions, University of Botswana, Gaborone, Botswana, [4] Africa Health Research Institute, Durban, South Africa, [5] The Francis Crick Institute, London, UK, [6] Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA, [7] Division of Medical Virology, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa, [8] School of Health Systems and Public Health, University of Pretoria, Pretoria, South Africa

Background: Recent advances in next-generation sequencing (NGS) have facilitated high-throughput viral genome sequencing for surveillance of epidemic pathogens such as hepatitis B virus (HBV). However, standardized bioinformatics workflows specific to HBV genomics remain limited. Most currently available pipelines assume a single-strain HBV infection and collapse mixed-strain reads into a single consensus sequence dominated by the major variant. This approach masks minor populations, misclassifies genotypes and drug resistance variants, and may contribute to suboptimal clinical management decisions.

Methods: We developed HBVgenomer, a comprehensive end-to-end workflow incorporating a five-detector orthogonal mixed-infection framework designed to sensitively identify infections containing multiple HBV strains. HBVgenomer accepts FASTQ data generated from both short-read and long-read sequencing platforms. Following mixed-infection screening, a 14-stage haplotype reconstruction framework is triggered to recover candidate haplotypes for each detected strain. Per-strain read recovery was enhanced using a hybrid assembly strategy that combines reference-guided assembly with best-fit reference selection and de novo rescue of unmapped HBV reads using integrated nucleotide and amino acid similarity scoring. HBVgenomer additionally performs per-strain genotyping, recombination screening, mutation profiling (drug resistance, vaccine escape, immune escape, covarying mutations, and epitope decay dynamics), and in silico functional analyses. We successfully generated 200 near-full-length HBV genomes using Illumina and Oxford Nanopore Technologies (ONT) platforms and analysed all genomes using HBVgenomer. The workflow was further benchmarked against widely used commercial and open-source HBV analysis pipelines using published datasets from 2019–2024.

Results: HBVgenomer demonstrated superior sensitivity and specificity for mixed-infection detection, strain discrimination, and consensus sequence reconstruction, particularly for complex quasispecies, recombinant strains, and low-frequency variants. We developed and optimized the DiploCirc engine, which further reduced

false-positive strain assignments while maintaining high sensitivity across both sequencing platforms.

Conclusion: HBVgenomer applies empirically optimized thresholds to support standardized NGS analysis and enables robust, scalable, and near real-time HBV genomic surveillance. The workflow provides a comprehensive framework for accurate HBV strain resolution, genomic characterization, and translational clinical interpretation.

Disclosure of Interest Statement: None