[BIOINFO 2025]

Accelerating whole-genome analysis: benchmarking GPU- and CPU-based variant calling pipelines

Hye-Yeong Jo[1], Jaehoo Kim[2], Ga Young Park[1], Minseok Kwon[2], and Sang Cheol Kim[1,*]

[1]*Division of Healthcare and Artificial Intelligence, Department of Precision Medicine, National Institute of Health, Cheong-Ju, 28159, South Korea*
[2]*RexSoft Corp., Seoul, South Korea.*
**Corresponding author: sckim.knih@korea.kr*

Next-generation sequencing (NGS), including Whole-genome sequencing (WGS), has advanced rapidly and become increasingly cost-effective, leading to the generation of large-scale genomic datasets, which in turn, necessitates high-performance computational solutions for timely analysis. As sequencing throughput continues to grow, the speed and accuracy of downstream bioinformatics pipelines have become critical bottlenecks. Conventionally, most genomic analysis workflows have been implemented on central processing units (CPUs). With the advent of graphics processing units (GPUs) coupled with the Parabricks software suite, GPU-accelerated pipelines offer the potential to significantly reduce computational burden and improve efficiency in large-scale genomic datasets.

In this study, we assessed computational performance and variant calling consistency by implementing the CPU-based gold-standard best practice GATK pipeline and the GPU-accelerated Parabricks pipeline (dual NVIDIA H100 accelerators), in the context of WGS-based single nucleotide variant (SNV) and indel calling. The GPU-based analysis markedly outperformed the CPU-based approach, reducing total runtime by approximately 18 hours. In the meantime, variant calls between CPU and GPU workflows were highly concordant, with about 99.7% of variants consistently identified by both CPU and GPU workflows. Following variant quality score recalibration (VQSR), only a small fraction of variants remained unique to each platform—294 detected exclusively by the GPU-based workflow and 239 by the CPU-based workflow—together representing just 0.3% of total calls. We further performed functional annotation of variants to compare biological interpretability across workflows.

We provide in-depth analysis of the resource usage, suggesting practical insights into GPU and other system requirements. This comprehensive evaluation of accelerated NGS pipelines highlights their potential to enhance the scalability, efficiency, and accessibility of WGS analysis, ultimately guiding researchers in selecting optimized workflows for large-scale genomics studies.