Accelerating insights into plant genomics: The vital role of bioinformatics in High Performance Computing (HPC)

Nagarajan Kathiresan[1,$], Yong Zhou[2,3,$], Doreen Ware[6,7], Jianwei Zhang[4], Kenneth L. McNally[5], Rod A. Wing[2,3,5,#]

[1]KAUST Supercomputing Laboratory (KSL), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia
[2]Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia
[3]Arizona Genomics Institute (AGI), School of Plant Sciences, University of Arizona, Tucson, Arizona 85721, USA
[4]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China
[5]International Rice Research Institute (IRRI), Los Baños, 4031 Laguna, Philippines
[6]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[7]USDA ARS NEA Plant, Soil & Nutrition Laboratory Research Unit, Ithaca, NY, 14853, USA

[$]These authors contributed equally: Nagarajan Kathiresan and Yong Zhou
[#]The corresponding author supervised this work: Rod A. Wing, rwing@ag.arizona.edu

Keywords: High-Performance Computing (HPC), Genome Analysis Toolkit (GATK), Genome Index splitter (GIS), Single-nucleotide polymorphisms (SNPs), INDELs, Rice Genome

Abstract: Bioinformatics on high-performance computing (HPC) platforms is crucial for improving agricultural research, especially for rice crops which is essential for global food security. Using HPC systems, scientists study the complex genetic makeup of rice to learn more about its traits and how it adapts to different environments. In our presentation, we introduce an automated method for finding genetic variations in rice, designed specifically for HPC platform called as "High-Performance Computing Genome Variant Calling Workflow" (HPC-GVCW, https://doi.org/10.1186/s12915-024-01820-5). This method can efficiently process data from 3,000 rice samples in just 1-3 days and it also suitable for other plant genomes like maize, soybean and sorghum. The process includes mapping the rice genome, discovering genetic variations, refining the results for accuracy, and putting together a complete picture of the variations findings. Our method is designed to work smoothly on different types of computers, like clusters, cloud systems, and high-end workstations. We use advanced tools like Genome Analysis Toolkit (GATK) and a special Genome Index splitter (GIS) to speed up the process by running multiple tasks of disjoint genome intervals across the different nodes in the HPC platform. Each step of our method works independently, making the whole process faster and more efficient. During our presentation, we will demonstrate how our method finds genetic variations in real-time across 3,000 rice samples at Shaheen III as an example HPC platform. This will show how our approach can be practically used to improve genomic research in agriculture, leading to new innovations that help farmers grow better crops.