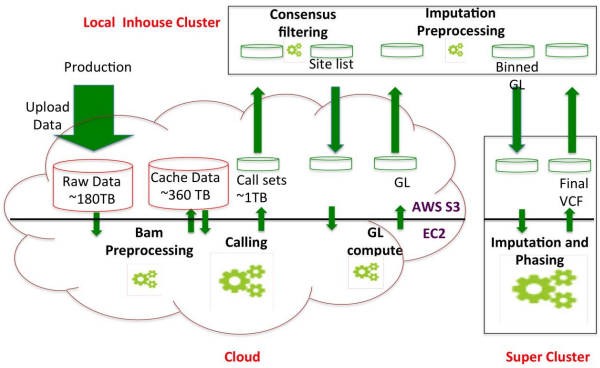
A Hybrid Computational Strategy to Address WGS Variant Analysis in >5000 Samples

Achievement: Developed scalable hybrid computational strategy to conduct analysis of large-scale whole genome datasets.

Significance and Impact: A hybrid computation strategy involving multiple HPC systems of different architectures is required for current and future WGS datasets analysis. The approach mentioned in the publication can be scaled to over 10,000 samples and turnaround time can be reduced from months to weeks.

Research Details: The hybrid computational strategy involved mapping various stages of the genomics pipeline (implemented with goSNAP) onto different computation resources optimizing for cost and turnaround time, while not sacrificing sensitivity or specificity. Figure shows the mapping of goSNAP pipeline. Variant calling (Stage A) and genotype likelihood (Stage C) calling is done on the AWS cloud, consensus filtering and imputation preprocessing is accomplished in the LHPC (Stage B) and Imputation and Phasing (Stage D) is done at the supercomputers at Rice and Oak Ridge National Laboratory



Sponsor/Facility: Oak Ridge Leadership Computing Facility

ORNL PI and affiliation: Manjunath Gorentla Venkata, CSMD – Oak Ridge National Laboratory

Team: Zhuoyi Huang (Baylor College of Medicine), Navin Rustagi (BCM), Narayanan Veeraraghavan (BCM), Andrew Carroll (DNAnexus), Richard A. Gibbs (BCM), Eric Boerwinkle (BCM and U. Texas Health Science Center), Manjunath Gorentla Venkata (ORNL), Fuli Yu (BCM)

Publication: Zhuoyi Huang, Navin Rustagi, Narayanan Veeraraghavan, Andrew Carroll, Richard A. Gibbs, Eric Boerwinkle, Manjunath Gorentla Venkata, Fuli Yu, A hybrid computational strategy to address WGS variant analysis in >5000 samples. BMC Bioinformatics 17: 361 (2016)

Overview: The decreasing costs of sequencing are driving the need for cost effective and real time variant calling of whole genome sequencing data. The scales of these projects are far beyond the capacity of typical computing resources available with most research labs. Other infrastructures like the cloud AWS environment and supercomputers also have limitations due to which large scale joint variant calling becomes infeasible, and infrastructure specific variant calling strategies either fail to scale up to large datasets or abandon joint calling strategies. In this paper, we present a high throughput framework including multiple variant callers for single nucleotide variant (SNV) calling, which leverages hybrid-computing infrastructure consisting of cloud AWS, supercomputers and local high performance computing infrastructures. We present a novel binning approach for large scale joint variant calling and imputation which can scale up to over 10,000 samples while producing SNV callsets with high sensitivity and specificity. As a proof of principle, we present results of analysis on Cohorts for Heart And Aging Research in Genomic Epidemiology (CHARGE) WGS freeze 3 dataset in which joint calling, imputation and phasing of over 5300 whole genome samples was produced in under 6 weeks using four state-of-the-art callers.

HPCWire:
https://www.hpcwire.com/2016/09/22/new-genomics-pipeline-combines-aws-local-hpc-supercomputing/

Baylor College of Medicine News:
https://www.bcm.edu/news/genome-sequencing/new-scalable-whole-genome-data-analysis

ORNL News:

https://www.ornl.gov/news/ornl-helps-develop-hybrid-computational-strategy-efficient-sequencing-massive-genome-datasets