Next Generation Resequencing of Soybean Germplasm for Trait Discovery on XSEDE using Pegasus Workflows and iPlant Infrastructure

Trupti Joshi[1,2,3]
Yang Liu[2,5]
Dong Xu[1,2,3]
Nicole Hopkins[6]
Babu Valliyodan[2,3,4]
Joao V. Maldonado dos Santos[4]
Henry Nguyen[2,3,4]
Mats Rynge[7]
Saad M. Khan[2,5]
Jeuxin Wang[1,2]
Nirav Merchant[6]

[1] Department of Computer Science; [2] Christopher S. Bond Life Sciences Center; [3] National Center of Soybean Biotechnology; [4] Division of Plant Sciences; [5] Informatics Institute; University of Missouri, Columbia
[6] iPlant Collaborative, University of Arizona
[7] Information Sciences Institute, University of Southern California

ABSTRACT

This poster presents an effort for conducting bioinformatics analysis using XSEDE as the computing infrastructure, iPlant as the data and cloud infrastructure, and the Pegasus workflow systems to control and coordinate the data management and computational tasks. It outlines best practices for efficient utilization of distinct and unique Cyberinfrastructure (CI) resources available through multiple providers, with emphasis on creating extensible and scalable workflows that can be easily modified and deployed.

With the advances in next generation sequencing (NGS) technology and significant reduction in sequencing costs it is now possible to sequence large sets of crop germplasm and generate whole genome scale structural variations and genotypic data. In-depth informatics analysis of the genotypic data can provide better understanding of the links with the observed phenotypic changes. This approach can be used to further understand and study different traits for the improvement of crops by design. The NGS resequencing data represents a rich source of information and can lead to significant discoveries when it comes to mining genotypic data for phenotypic inferences.

We have conducted resequencing of more than 1000 soybean germplasm lines from multiple projects, selected for major traits including oil, protein, soybean cyst nematode resistance (SCN), abiotic stress resistance (drought, heat and salt) and root system architecture. .

We have data for 108 soybean genomes from Phase I and part of data for 110 genomes from Phase II. We expect remaining data from Phase II to be available beginning June 2014 and Phase III later by Dec 2014. We have outlined the steps in bioinformatics analysis pipeline for SNP and Indel identification using GATK 3.0 haplotype caller and multi sample SNP calling against soybean Gmax_275_Wm82.a2.v1 reference genome.

The NGS reseqeuncing data (25TB) is housed in iDS (iPlant Data Store), which is based on iRODS (Integrated Rule Oriented Data-Management System), key features utilized for this analysis include the ability to replicate data closer to computational resources utilizing parallel data transfer capabilities, while maximizing on available network bandwidth. Data is transparently replicated from primary data servers at iPlant (University of Arizona, Tucson) to resource servers at TACC, which allows low latency data access to the inputs and output while running the workflows on Stampede or other TACC systems.

The Pegasus workflow system is used to define and control the required computational tasks. These include user-defined tasks, such as BWA, Picard, and GATK, as well as Pegasus added tasks such as data staging between the iDS and Stampede scratch filesystem. Pegasus also adds data cleanup tasks to maintain and minimize the workflow footprint on the Lustre scratch filesystem as the workflow progresses.

Key challenge has been that the application from BWA, Picard, and GATK suite, each require different configurations and have upper limits in terms of number of cores (basic Linux threads) and memory that they can efficiently utilize. These requirements vary from a single core and 6 GBs of RAM, to 16 cores and 22 GBs of RAM. The Pegasus workflow is fully defined with all the requirements. In order to execute the tasks efficiently on Stampede, Pegasus clusters the tasks together into pegasus-mpi-cluster jobs. Pegsus-mpi-cluster jobs are self-contained sub-DAGs of the workflow, and at runtime pegasus-mpi-cluster will execute the tasks in the sub-DAG by considering the core/memory requirements of the task, and tracking available cores/memory on the compute nodes. Each workflow gets mapped into three pegasus-mpi-cluster jobs; one with BWA going to the normal compute nodes, one for large memory Picard and GATK tasks mapped to the large memory nodes available on Stampede, and the last one lower memory GATK tasks. The number of nodes used by each pegasus-mpi-cluster job varies based on the number of input samples the user provides to the workflow.

To submit and manage workflows, we have created a publicly accessible Virtual Appliance hosted in the iPlant Atmosphere cloud platform, allowing anyone to launch an instance to manage their own workflow; these instances includes all required software and application dependencies. We plan to develop additional workflows for copy number variations (CNV) analysis using CNV-seq, SNP annotations using SnpEff, and genotype to phenotype predictions., Haploview for linkage disequilibrium, haplotype identification and Cladogram tree generation. We will also generate generalized linear models (GLM) and mixed linear models (MLM) using TASSEL for identifying SNPs significant for phenotypic changes between the various resistance and susceptibility categories of this trait. All data GWAS, SNP and genome structure information generated from this resequencing project can be accessed through Soybean Knowledge Base (SoyKB) at http://soykb.org and workflow documentation at http://goo.gl/kaaFgt , hosted at the iPlant Collaborative cyber-infrastructure.