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

Trupti Joshi[1,2,3], Babu Valliyodan[2,3,4], Saad M. Khan[2,5], Yang Liu[2,5], Joao V. Maldonado dos Santos[4], Yongqing Jiao[3,4], Dong Xu[1,2,3] , Henry T. Nguyen[2,3,4] , Nicole Hopkins[6], Mats Rynge[7] , 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, MO. [6] iPlant Collaborative, University of Arizona

[7] Information Sciences Institute, University of Southern California

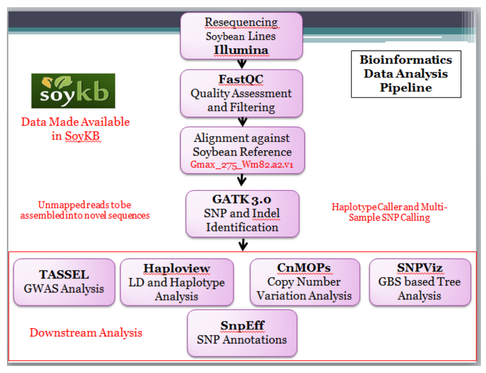
**Introduction**

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.

We have conducted resequencing of 1000+ soybean germplasm lines 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 conducted initial bioinformatics analysis of the NGS data from these 1000+ lines. The analysis identified SNPs and insertion, deletions by comparisons against the soybean reference genome, Williams 82 using GATK software. We have also conducted structural variation analysis in terms of copy number variations (CNV) using CNV-seq software. As a case study, we have selected 25 out of the 1000+ resequenced genomes for the in depth analysis of SCN resistance and further classified them into four different categories of resistance and susceptibility levels. The GWAS analysis helped identify major SNPs associated with the phenotypic changes between these lines. We have performed analysis using Haploview for linkage disequilibrium, haplotype identification and Cladogram tree generation. We have also applied 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.

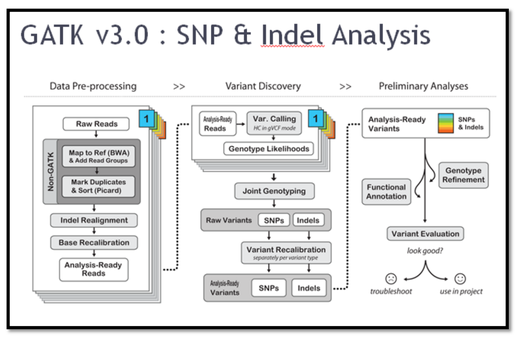
The informatics analysis results help set a strong foundation for handling and analysis of subsequent large scale resequencing efforts in future. The NGS resequencing data represents a rich source of information and can lead to significant discoveries when it comes to mining genotypic data for phenotypic inferences. All data including GWAS, SNP and genome structure information generated from this resequencing project can be accessed through Soybean Knowledge Base (SoyKB) at http://soykb.org.

## Bioinformatics Analysis Pipeline




We have developed a bioinformatics pipeline for the analysis of the resequencing data as above. The analysis involves quality assessment and filtering steps with FastQC and identifies SNPs and insertions/deletions by comparison against the soybean reference genome Williams 82 using GATK analysis pipelines. The SNPs generated from GATK analysis are used for the subsequent downstream analysis as described below.

## GATK Analysis




| Sample | Total Reads | % Mapped | % Mapped unique | Coverage | SNP_HC filtered | Indel_HC filtered |
|---|---|---|---|---|---|---|
| HN001 | 178012181 | 99.39930459 | 85.66 | 16.46 | 921174 | 139546 |
| HN002 | 189725050 | 99.28611957 | 86.19 | 17.543 | 1747381 | 305884 |
| HN003 | 197817579 | 99.33175049 | 85.55 | 18.291 | 1867448 | 300665 |
| HN004 | 175095378 | 99.31592198 | 85.59 | 16.19 | 1600603 | 259031 |
| HN005 | 200314390 | 99.3604139 | 85.83 | 18.522 | 1801279 | 303319 |
| HN006 | 185557717 | 99.23813732 | 87.06 | 17.158 | 1370865 | 230198 |
| HN007 | 195461975 | 99.23788911 | 85.81 | 18.073 | 1543388 | 253697 |
| HN008 | 165493144 | 99.10996253 | 85.8 | 15.302 | 1619672 | 313594 |
| HN009 | 188816175 | 99.18097589 | 86.97 | 17.459 | 1602653 | 304931 |
| HN010 | 149643309 | 99.36637261 | 85.43 | 13.837 | 1268596 | 181983 |
| HN011 | 161364750 | 98.94882866 | 84.73 | 14.921 | 1236654 | 225272 |
| HN012 | 161134565 | 99.31436747 | 84.64 | 14.899 | 1654295 | 250173 |
| HN013 | 189992517 | 99.34214778 | 86.65 | 17.568 | 1062485 | 178084 |
| HN014 | 167022679 | 99.36191599 | 84.95 | 15.444 | 1079994 | 153715 |
| HN015 | 195359928 | 99.20419401 | 84.33 | 18.064 | 1840714 | 279665 |
| HN018 | 174490395 | 99.26196797 | 85.6 | 16.134 | 1678565 | 282060 |
| HN019 | 197618994 | 99.40807157 | 86.4 | 18.273 | 1850267 | 328124 |
| HN020 | 132029402 | 99.33248505 | 84.69 | 12.208 | 1019364 | 155568 |
| HN021 | 189123447 | 99.41646527 | 85.52 | 17.487 | 1622514 | 253096 |
| HN022 | 179241418 | 99.50280744 | 84.75 | 16.574 | 1740524 | 266309 |
| HN023 | 151464063 | 99.43830108 | 85.95 | 14.005 | 1126029 | 189553 |
| HN024 | 211827511 | 99.45493105 | 85.53 | 19.587 | 1672883 | 264496 |
| HN026 | 182524198 | 99.38366802 | 85.35 | 16.877 | 1458105 | 232437 |
| HN027 | 202205881 | 99.37399793 | 86.11 | 18.697 | 972851 | 145152 |

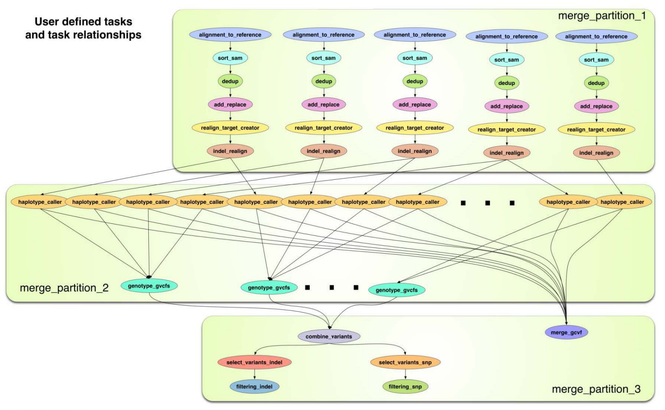
Table 1. GATK statistics for analysis of 25 soybean resequencing germplasm.

## Processing on XSEDE and iPlant Infrastructures

In collaboration with iPlant and XSEDE Extended Collaborative Support Service (ECSS) the analysis pipeline was implemented in Pegasus WMS. Primarily XSEDE compute resources are used for executing the pipeline, and iPlant provides infrastructure for managing the overall analysis using Pegasus appliance (VM) in iPlant cloud platform Atmosphere and the storage of, and access to, data products using the iPlant Data Store. Making this pipeline easy to use by multiple research groups with similar analytics needs.

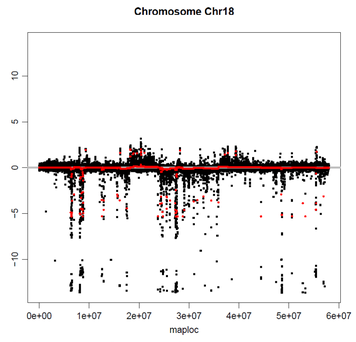
The NGS resequencing data (~25 TB) is housed in the iPlant Data Store (iDS) developed using iRODS, this allows rapid data provisioning and geo-replication. Data can be replicated from the primary data store at U. Arizona data center for iPlant to the dedicated iDS resource server at TACC, allowing low latency data access to the inputs and output when running workflow steps on TACC Stampede. iDS supports the pre and post analysis data management tasks such as integration with Genome browsers, metadata tagging and searching.

The use of Pegasus to manage workflow, Github for handling versioned workflow steps, iPlant Atmosphere and iDS for coordinating execution and data management tasks has proved to be a reproducible, reliable and scalable solution. This platform allows marshaling of distributed computing resources (XSEDE, USC/ISI, U. Arzona) allowing teams of biologists to effectively manage computational tasks associated with large scale genome analysis.



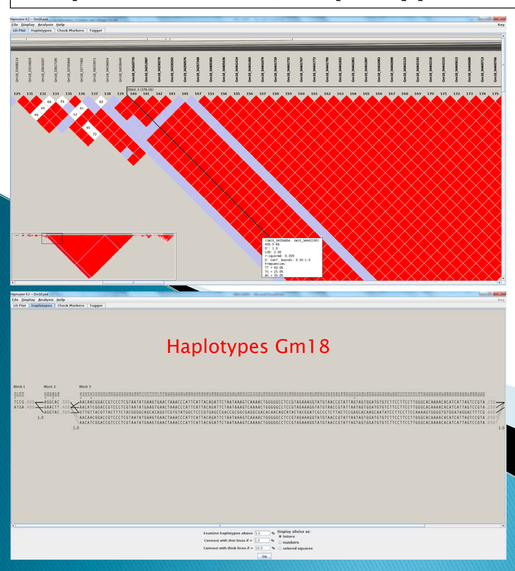


## Copy Number Variation Analysis




We also conducted CNV analysis using CnMOPs tool to identify the copy number gain and loss between genomes.

## Haploview : LD, Haplotype Analysis




We have also performed linkage disequilibrium (LD) and haplotype analysis using Haploview using chromosome Gm18 data. Analysis for all other soybean chromosomes and continued analysis for identification on QTL regions for specific traits is currently on-going.

## TASSEL: GWAS Analysis

We performed GWAS analysis on the SNP data generated by GATK using TASSEL. We performed Generalized linear models (GLM) and Mixed linear models (MLM) analysis on the datasets for several chromosomes and identified the significant SNPs based on a pvalue cutoff of 1e-8, using the SCN resistance/susceptibility categories assigned to our 25 genomes in SCN case study as below. We identified total 3337 significant SNPs using GLM amongst the 13 studied chromosomes that differentiate the SCN resistant and susceptible genomes. Significant SNPs were only identified using GLM. Cladogram using NJ method was also generated.

**Generalized linear models (GLM) :**

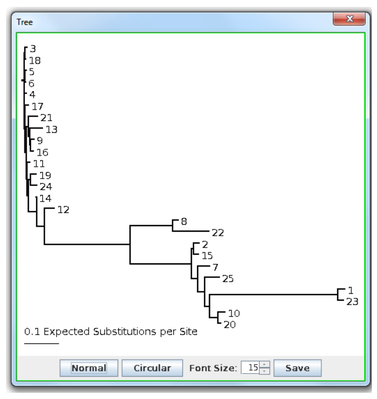
Population structure + Marker effect + residual

**Mixed linear models (MLM) :**

Population structure + Marker effect + Individuals + residual

| Sample | Country of Origin | RACE 1 | RACE 2 | RACE 3 |
|---|---|---|---|---|
| Ref | -- | Susceptible | Susceptible | Susceptible |
| H001 | U.S. | Susceptible | Susceptible | Resistant |
| H002 | China | Resistant | Susceptible | Resistant |
| H003 | China | Resistant | Mod. Resistant | Resistant |
| H004 | China | Resistant | Resistant | Resistant |
| H005 | China | Resistant | Mod. Resistant | Resistant |
| H006 | South Korea | -- | -- | Susceptible |
| H007 | South Korea | -- | -- | Mod. Susceptible |
| H008 | China | Resistant | Susceptible | Resistant |
| H009 | China | Mod. Susceptible | Mod. Resistant | Resistant |
| H010 | Japan | Resistant | Mod. Resistant | Resistant |
| H011 | China | Resistant | Mod. Resistant | Resistant |
| H012 | China | -- | -- | Mod. Susceptible |
| H013 | U.S. | Mod. Resistant | Resistant | Resistant |
| H014 | U.S. | -- | -- | Susceptible |
| H015 | China | Resistant | Mod. Resistant | Resistant |
| H018 | China | Resistant | Mod. Resistant | Resistant |
| H019 | China | Resistant | Mod. Susceptible | Resistant |
| H020 | China | -- | Susceptible | Resistant |
| H021 | Japan | Mod. Susceptible | Susceptible | Resistant |
| H022 | China | Resistant | -- | Resistant |
| H023 | South Korea | -- | -- | Susceptible |
| H024 | China | Mod. Susceptible | Susceptible | Mod. Susceptible |
| H026 | North Korea | Resistant | -- | Mod. Susceptible |
| H027 | U.S. | Susceptible | Mod. Resistant | Mod. Resistant |

Table 4. SCN Resistance / Susceptible Categories for Race 1, 2 and 3 for 25 genomes in case study.

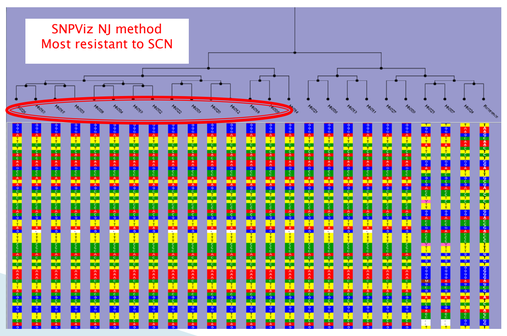



| Chromosome | Number of SNPs | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gm01 (D1a) | Gm04 (C1) | Gm05 (A1) | Gm06 (C2) | Gm08 (A2) | Gm10 (O) | Gm11 (B1) | Gm12 (H) | Gm14 (B2) | Gm17 (D2) | Gm18 (G) | Gm19 (L) | Gm20 (I) | Total |
| All | 391865 | 465305 | 225871 | 477299 | 355100 | 426873 | 302612 | 294649 | 294649 | 412418 | 834040 | 435679 | 341950 | 5265510 |
| SNPs/SCN | 65 | 364 | 68 | 748 | 819 | 127 | 70 | 78 | 106 | 157 | 628 | 54 | 53 | 3337 |
| Race 1 | 9 | 216 | 7 | 174 | 742 | 33 | 5 | 9 | 16 | 11 | 204 | 8 | 8 | 1442 |
| Race 2 | 4 | 6 | 1 | 13 | 5 | 4 | 4 | 7 | 3 | 94 | 38 | 13 | 4 | 196 |
| Race 3 | 52 | 142 | 60 | 561 | 72 | 90 | 61 | 62 | 87 | 52 | 386 | 33 | 41 | 1699 |

Table 5. Significant SNPs identified using GLM analysis in TASSEL for thirteen soybean chromosomes.

## SNPViz : Tree Clustering

We also generated tree using NJ method in our in house developed SNPViz tool, which is incorporated into SoyKB for automatic retrieval of the 25 genomes GBS datasets.




## Soybean Knowledge Base (SoyKB)

The data is available for access in SoyKB at http://soykb.org