

July 14, 2014

Soybean Knowledge Base (SoyKB) Pipeline on XSEDE

Mats Rynge <rynge@isi.edu>

XSEDE ECSS

XSEDE

Extreme Science and Engineering
Discovery Environment



ECSS Workflows

We provide expertise and assistance with scientific workflow software tools that make challenging workflow problems more efficiently executed, easier to manage, and easier to reproduce.

<https://www.xsede.org/ecss-workflows>

XSEDE specific workflow tutorials and submit node:

<https://sites.google.com/site/xsedeworkflows/>

The XSEDE logo is rendered in a large, bold, white sans-serif font. It is positioned on a dark blue background that features a grid of glowing blue squares and a faint image of a planet's horizon.

XSEDE Allocation

PI: Dong Xu

Trupti Joshi, Saad Kahn,
Yang Liu, Juexin Wang,
Badu Valliyodan, Jiaojiao
Wang

<http://soykb.org>

SOYBEAN KNOWLEDGE BASE (SoyKB)

A web resource for Soybean Translational Genomics



SoyKB Home



A hallmark of modern biology is tremendous amounts of complex omics data, which require large-scale data management, comprehensive computational analyses, and efficient integration, for better understanding of the data and hypothesis generation. For soybean with a newly sequenced genome, there is an increasing need from the soybean community to have a one-stop interactive, web-based portal to browse, access and share knowledge about soybean.

Towards this, we developed the Soybean Knowledge Base (SoyKB), a comprehensive all-inclusive web resource for soybean. SoyKB is designed to handle the storage and integration of the gene, genomics, EST, microarray, transcriptomics, proteomics, metabolomics, pathway and phenotype data.

SoyKB provides an informatics-based social network system to build connections among soybean researchers, producers and consumers.



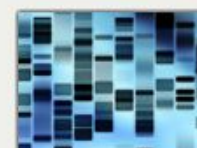
Gene Card



Latest News



SoyKB: a powerful tool at the junction of plant biology and computer science



University of Missouri Leads Soybean Sequencing Effort



SoyKB: Leading the convergence of wet and dry science in the era of Big Data



Data made available via
SoyKB

Unmapped reads assembled into
novel sequences

NGS Resequencing
Soybean Germplasm
Illumina

FastQC
Quality Assessment
and Filtering

Alignment against
Soybean Reference
Gmax_275_Wm82.a2.v1

GATK 3.0
SNP and Indel
Identification

XSEDE

SnpEff - SnpSift
SNP Annotations

LDexplorer
LD and Haplotype
Analysis

CnMOPs
Copy Number
Variation Analysis

SNPViz 2.0
GWAS Data Tree
Analysis

Downstream Analysis



XSEDE



iPlant Collaborative™

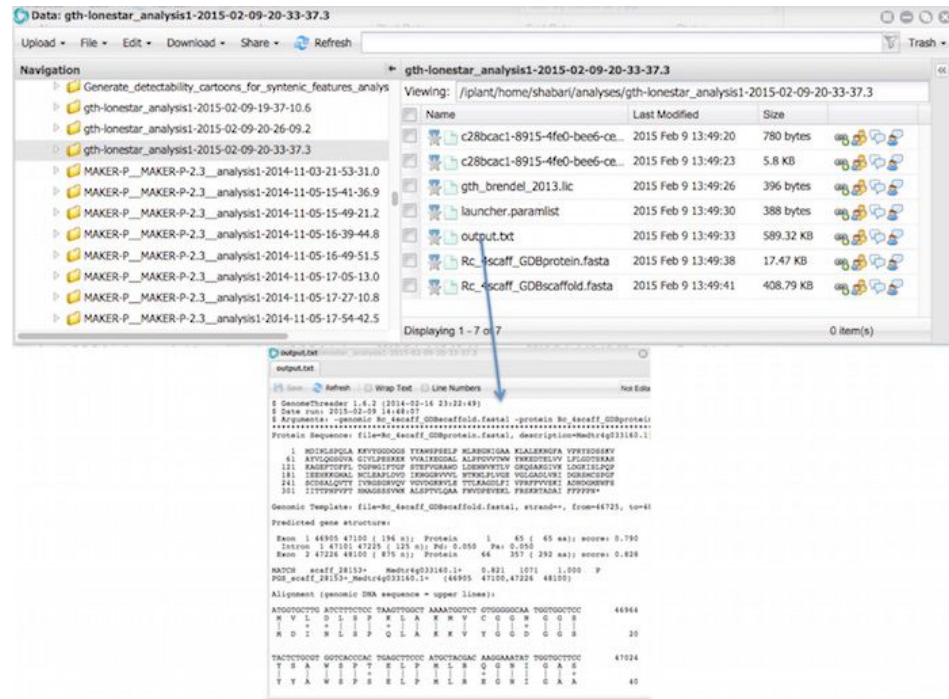
Data Store

iRods - Workflow input and output data

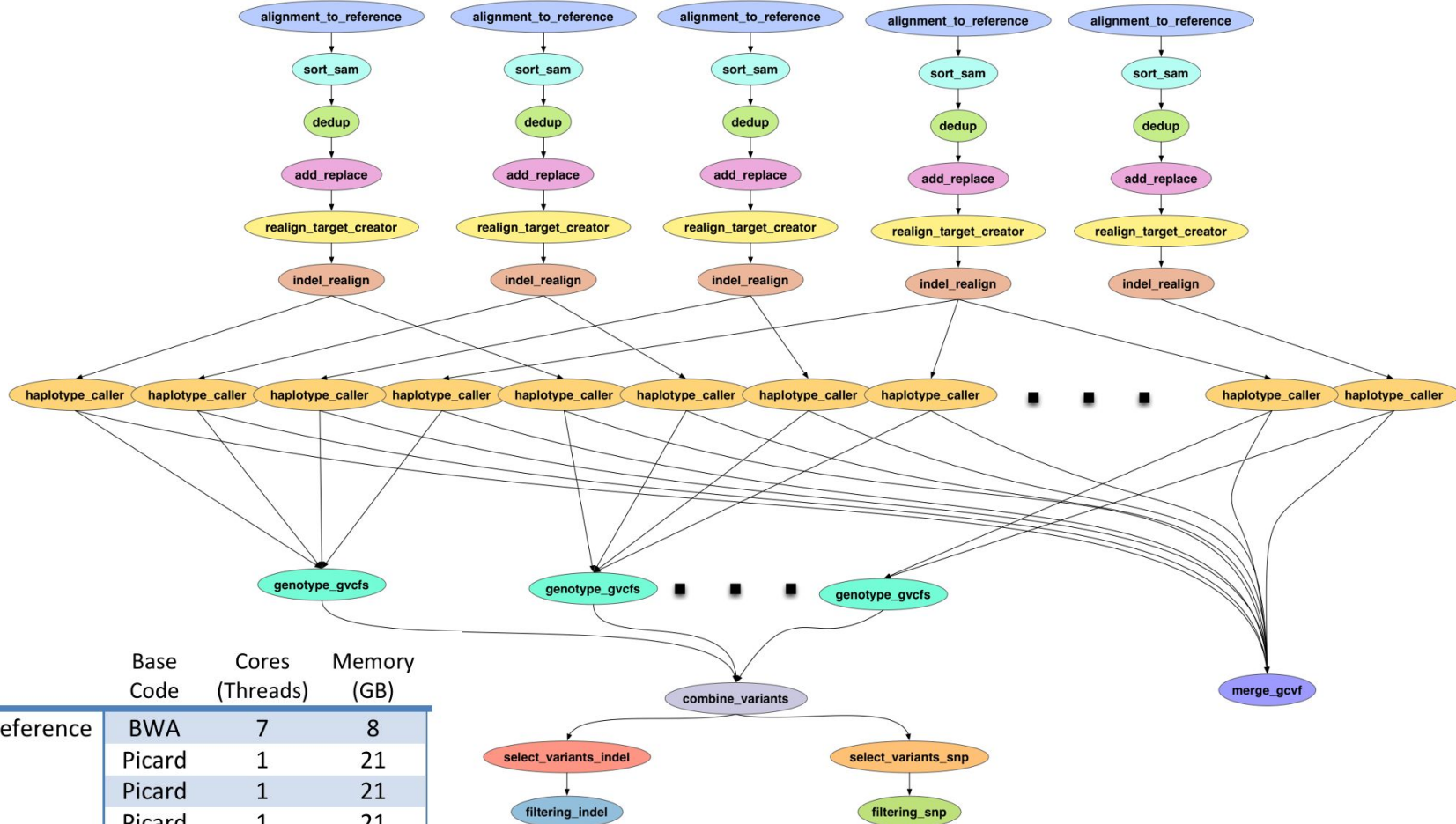
Can replicate data to servers at TACC

Discovery Environment

Atmosphere



XSEDE



| Task | Base Code | Cores (Threads) | Memory (GB) |
|------------------------|-----------|-----------------|-------------|
| Alignment_to_reference | BWA | 7 | 8 |
| Sort_sam | Picard | 1 | 21 |
| Dedup | Picard | 1 | 21 |
| Add_replace | Picard | 1 | 21 |
| Realign_target_creator | GATK | 15 | 10 |
| Indel_realign | GATK | 1 | 10 |
| Haplotype_caller | GATK | 1 | 3 |
| Genotype_gvcfs | GATK | 1 | 10 |
| Merge_gvcf | GATK | 10 | 20 |
| Combine_variants | GATK | 1 | 10 |
| Select_variants | GATK | 14 | 10 |
| Filtering | GATK | 1 | 10 |

TACC Stampede as Execution Environment

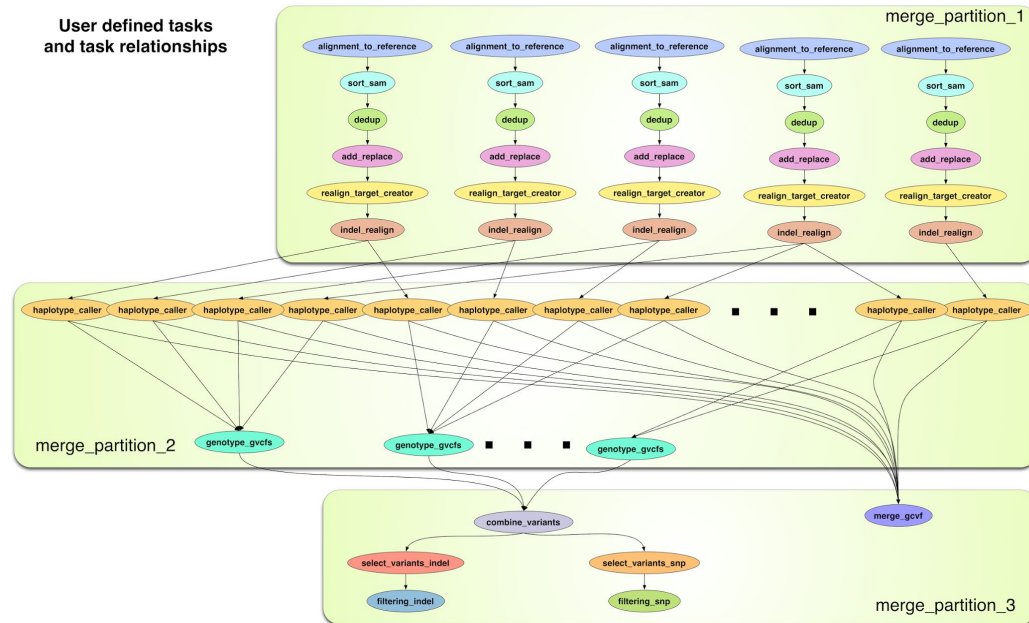
48 hour job runtime

Width of job 1 and 2 is determined by the number of input genomes.
3rd job is only 1 node

Pegasus MPI Cluster

<https://pegasus.isi.edu/wms/docs/latest/cli-pegasus-mpi-cluster.php>

User defined tasks
and task relationships



Executable workflow



XSEDE

TACC Wrangler as Execution Environment

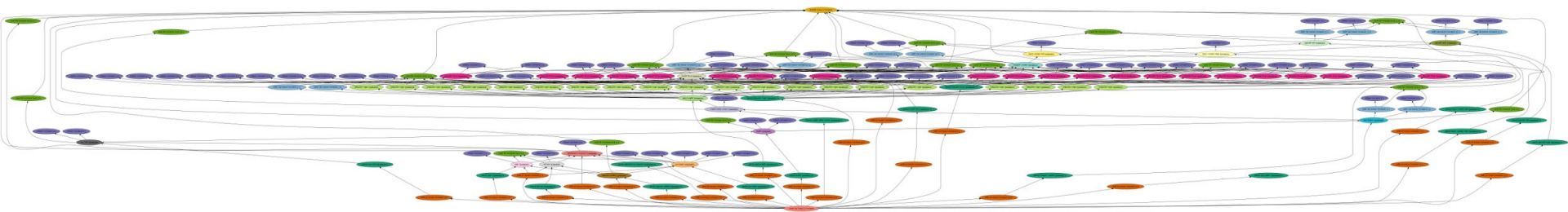
Flash Storage

Switched to glideins (pilots)

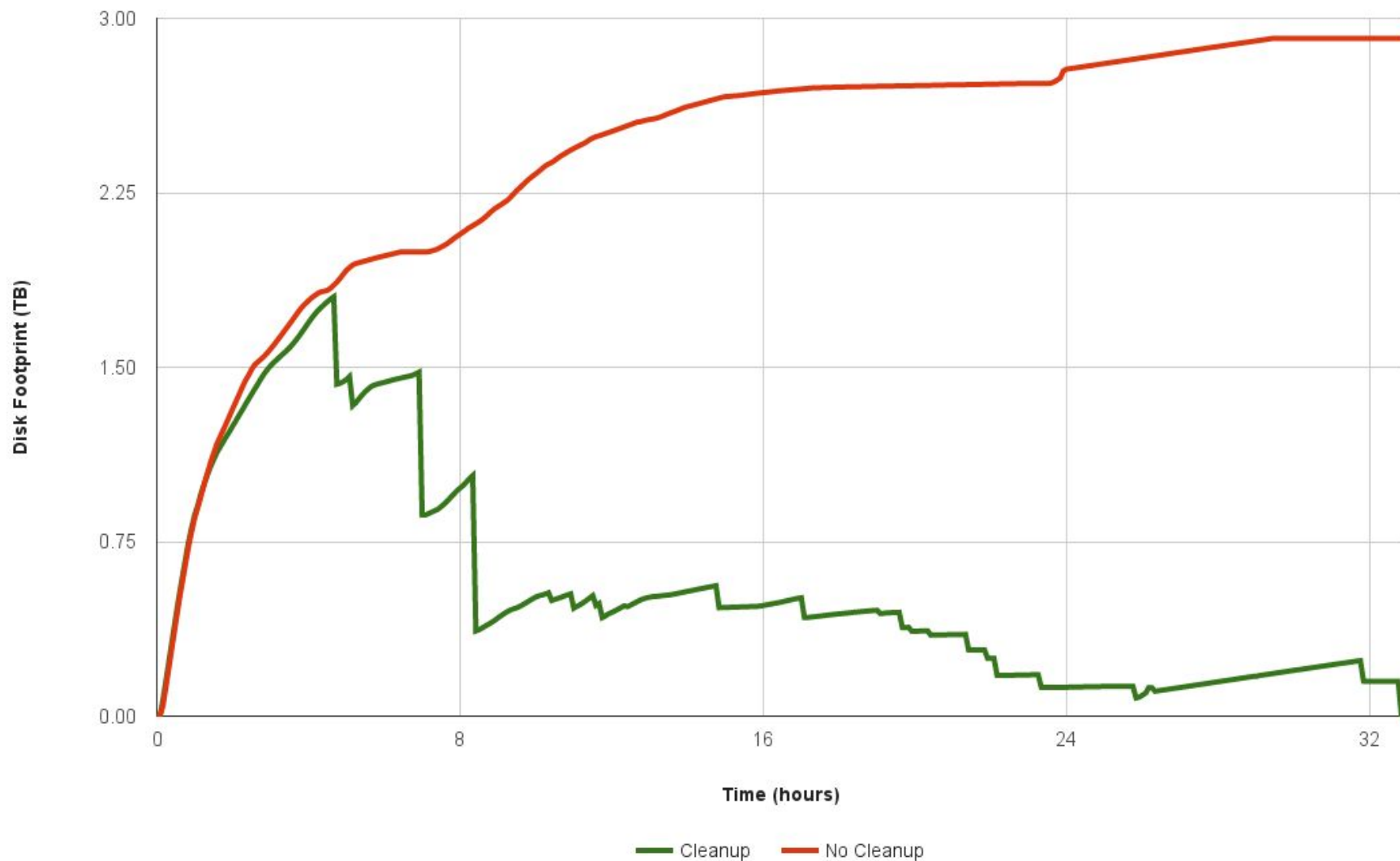
Brings in remote compute nodes and joins them to the HTCondor pool on in the submit host

Workflow runs at a finer granularity

Works well on Wrangler due to more cores and memory per node (48 cores, 128 GB RAM)



SoyKB - 20 Input Genomes - Disk Footprint



Conclusion

The SoyKB can efficiently map their computations to XSEDE resources, and use those resources in combination with iPlant.

Group has developed their own workflows

