CIF21 DIBBS: Tripal Gateway, a Platform for Next-Generation Data Analysis and Sharing

Stephen P. Ficklin (PI) [1], Alex Feltus (Co-PI) [2], Dorrie Main (Co-PI) [1], Meg Staton (Co-PI) [3], Jill Wegerzyn (Co-PI) [4], Sook Jung (SP) [1], Kuangching Wang (SP)[2]
Students, staff & postdocs w/ major contributions: Ming Chen[3], Nate Henry[3], Chun-Huai Cheng[1], Brian Soto[1], Connor Wytko[1], Mark Clytus[1], Nick Mills[2], Nick Watts[2], Emily Grau[4], Nic Herndon[4]

[1] Washington State University. Pullman, WA.
[2] Clemson University. Clemson, SC.
[3] University of Tennessee, Knoxville, TN.
[4] University of Connecticut, Storrs, CT.

Project Overview

Our work as proposed for the NSF DIBBs award #1443040 will provide extended functionality for the Tripal toolkit (http://tripal.info), which is an open-source, freely available software package that provides a framework to assist research groups publish genomic, genetic and related biological data in an online searchable format within Drupal, a popular content-management toolkit. This DIBBs award will allow us to provide greater cyberinfrastructure to these communities through development of three modules: 1) Data exchange module for a global federated network 2) scientific workflows module for large-scale data analysis using the Galaxy workflow tool and 3) a module for improved data transfer methods, including integration with national research networks (i.e. Internet2)

Accomplishments

Journal Articles

- Watts NW & Feltus FA. Big Data Smart Socket (BDSS): A System that Abstracts Data Transfer Habits from End Users. Bioinformatics 2016 (in press)
- Nicholas Mills, F. Alex Feltus, Walter B. Ligon III. "Maximizing the Performance of Scientific Data Transfer by Optimizing the Interface Between Parallel File Systems and Advanced Research Networks" Future Generation Computer Systems. in press, 2016.
- Connor Wytko, Brian Soto, Stephen P. Ficklin "blend4php: a PHP API for Galaxy". 2016 Oxford Database. In press.

Conference Papers & Presentations

- Stephen P. Ficklin, Lacey-Anne Sanderson, Chun-Huai Cheng, Connor Wytko, Brian Soto, Mark Clytus, Kirstin Bett, Dorrie Main. (2016) "The Future of Tripal: intuitive content creation, flexible data storage and web services". Plant and Animal Genome Conference XXIV. San Diego. Jan 2016.
- Staton M., Chen M., Henry N., Grau E., Wytko C., Soto B., Jung S., Wang KC., Watts N, Cheng CH., Sanderson L., Wegrzyn J., Main D., Feltus F., Ficklin S. (2016) Moving data from the warehouse to the workbench: a bridge to Galaxy from the Tripal community genome database software platform. Galaxy Community Conference, Bloomington, IN.
- Herndon, N., Grau, E. S., Batra, I., Demurjian Jr, S. A., Vasquez-Gross, H. A., Staton, M. E., and Wegrzyn, J. L. (2016). CartograTree: Enabling landscape genomics for forest trees. PeerJ Preprints, e2345v1.
- Demurjian Jr. S. A., Grau E., Vasquez-Gross H., Gessler D., Neale D. B., Wegrzyn J. L. (2016) TreeGenes and CartograTree: Community Resources for Forest Tree Genomics. Plant and Animal Genome XXII Conference, San Diego, CA USA.
- Grau E., Demurjian Jr. S. A., Vasquez-Gross H., Gessler D., Wegrzyn J. L. (2016) TreeGenes: Enabling Visualization and Analysis in Forest Tree Genomics. Plant and Animal Genome XXII Conference, San Diego, CA USA.

Software Releases

- Tripal v3 alpha (Jan 2016)
- blend4php (Jun 2016)
- BDSS v1.0.1b1 (August 2016)
- BDSS v1.0.1b2 (Nov 2016)

Other Resources

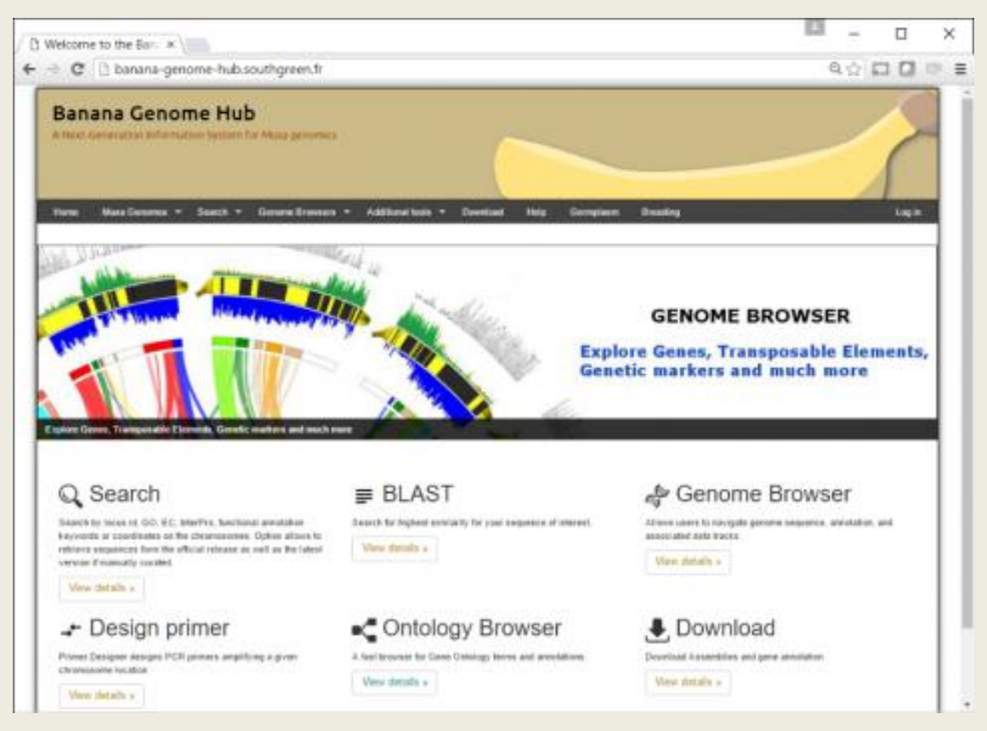
- Galaxy Instance for analytics: https://galaxy.bioinfo.wsu.edu/
- Tripal v3 demo site: http://demo.tripal.info/3.x/

GitHub Repositories

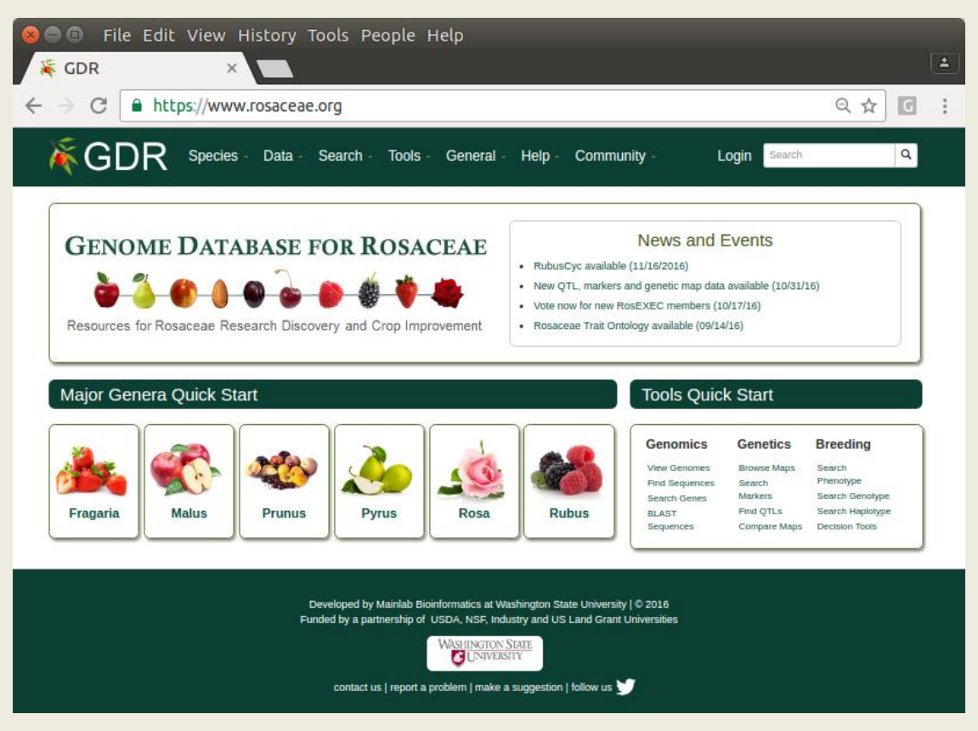
- Tripal v3 development: https://github.com/tripal/tripal/tree/7.x-3.x
- Tripal Galaxy module: https://github.com/tripal/tripal_galaxy
- blend4php: https://github.com/galaxyproject/blend4php
- BDSS: https://github.com/feltus/BDSS
- Galaxy Workflows: https://github.com/MingChen0919/docker-galaxy-dibbss/tree/master/my_workflows
- Galaxy Workflows in Docker: https://github.com/MingChen0919/docker-galaxy-dibbss/tree/master/my_workflows

Example Community Databases

The following sites are just a few example community databases that use Tripal. The technologies developed by our DIBBs work can have immediate impact for these sites and more…

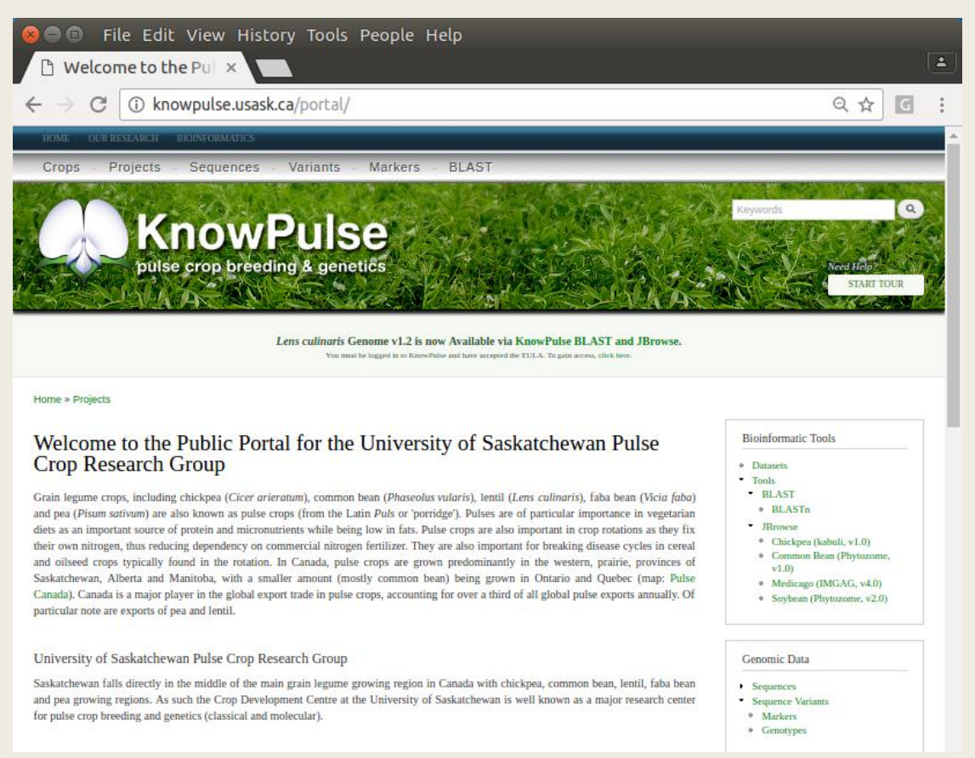


The Banana Genome Hub

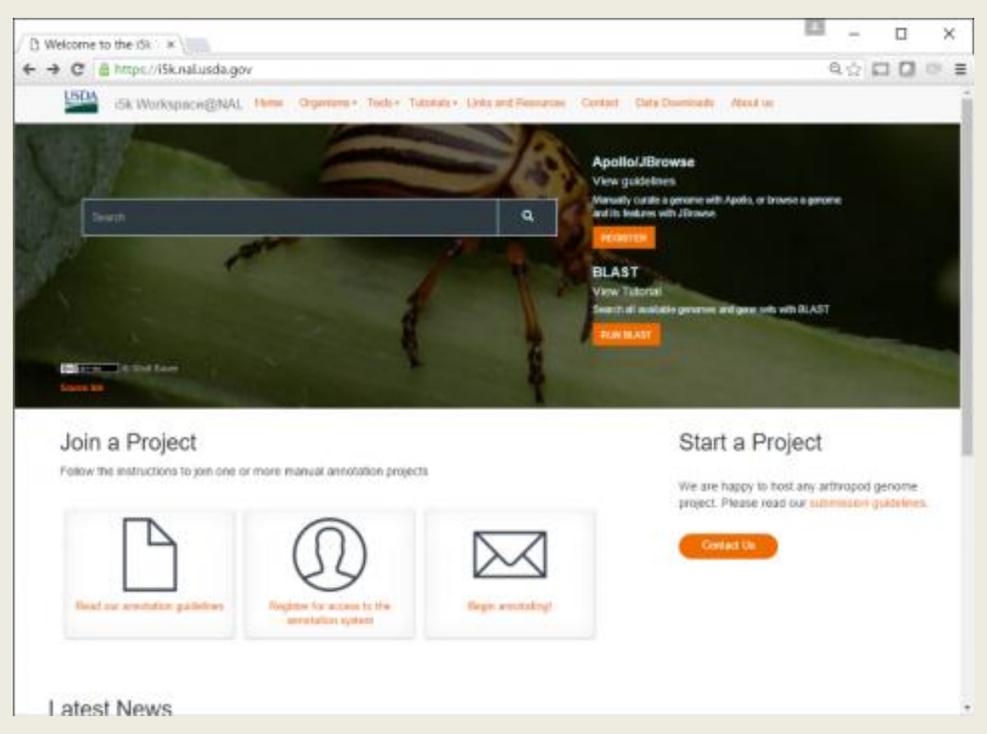


The Genome Database for Rosaceae

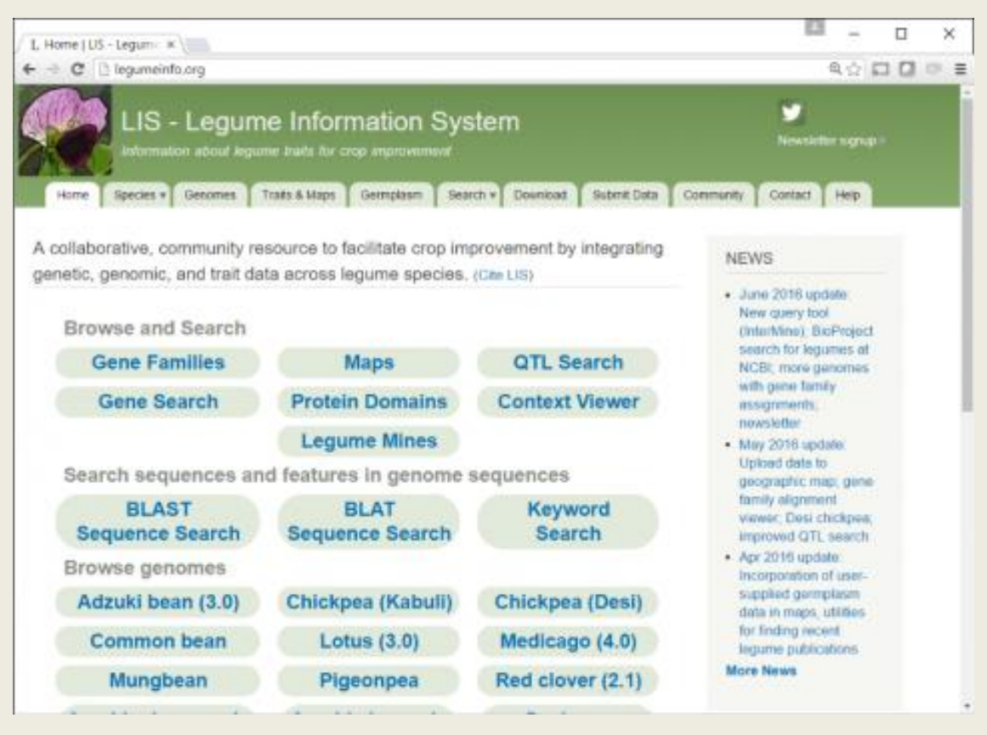
CottonGen



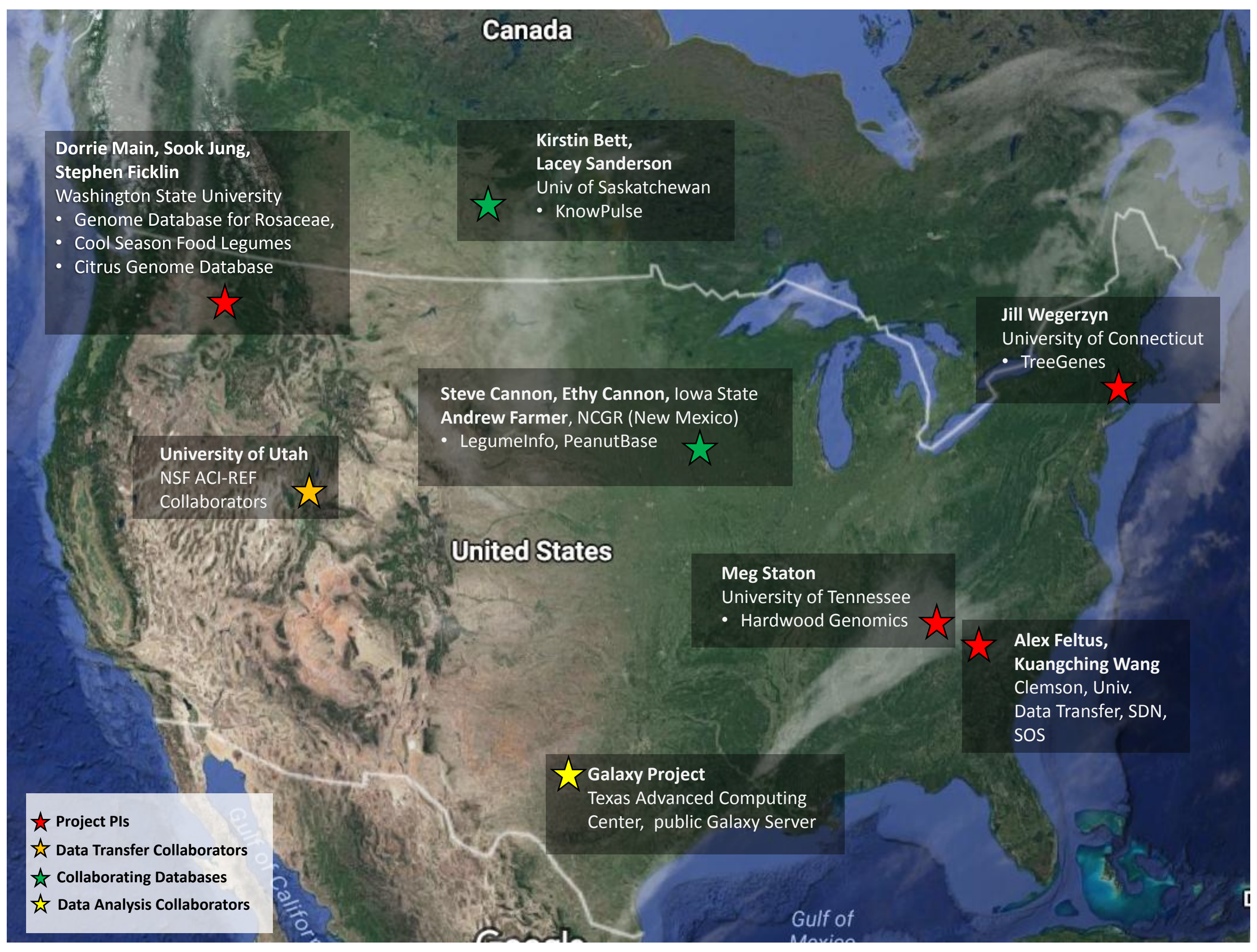
KnowPulse:



i5K



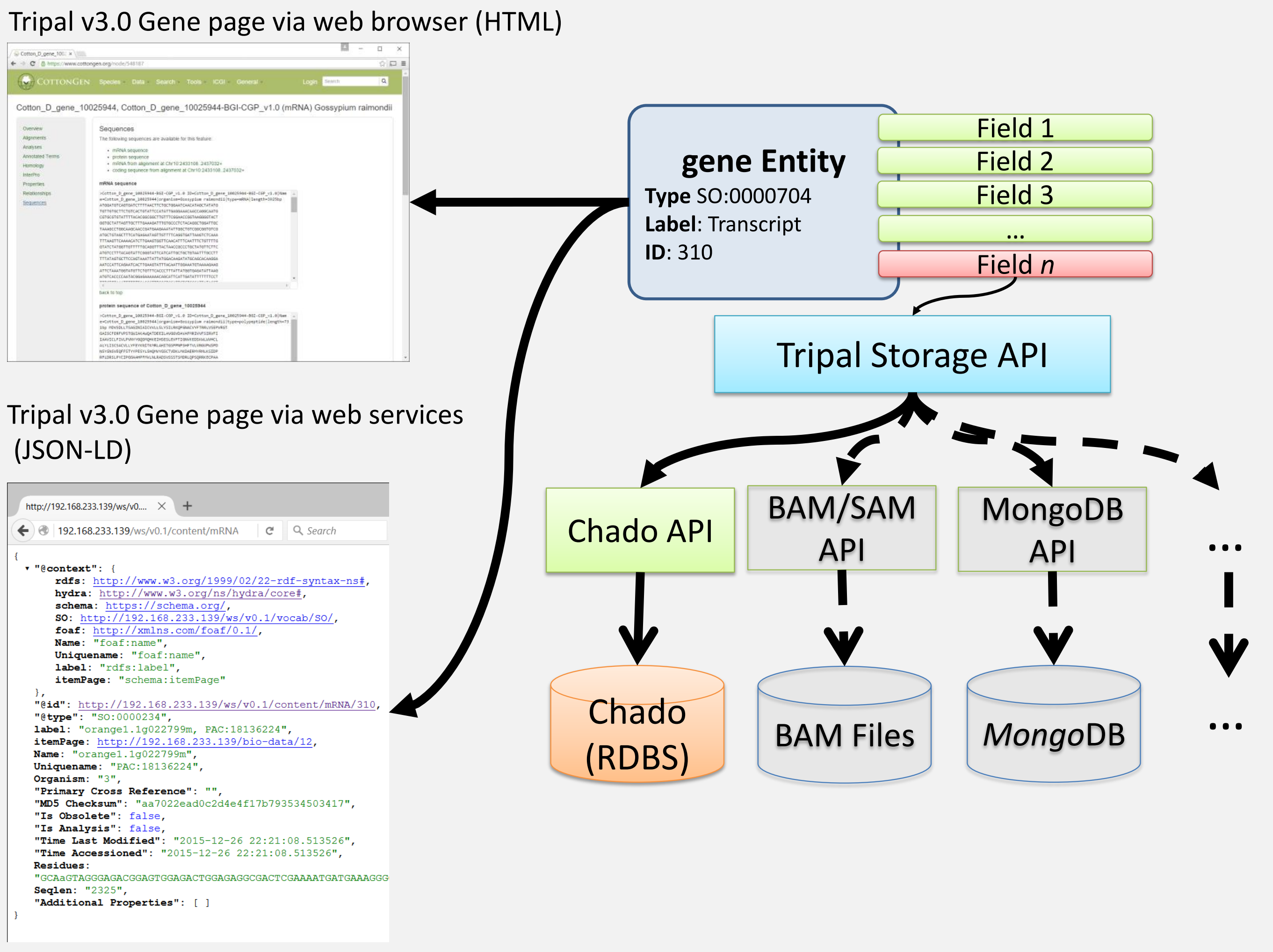
Legume Information System



Canada
Dorrie Main, Sook Jung, Stephen Ficklin
Washington State University
• Genome Database for Rosaceae,
• Cool Season Food Legumes
• Citrus Genome Database
Kirstin Bett, Lacey Sanderson
Univ of Saskatchewan
• KnowPulse
Jill Wegerzyn
University of Connecticut
• TreeGenes
Steve Cannon, Ethy Cannon, Iowa State
Andrew Farmer, NCGR (New Mexico)
• LegumeInfo, PeanutBase
University of Utah
NSF ACI-REF
Collaborators
United States
Meg Staton
University of Tennessee
• Hardwood Genomics
Alex Feltus, Kuangching Wang
Clemson, Univ.
Data Transfer, SDN, SOS
Galaxy Project
Texas Advanced Computing Center, public Galaxy Server
★ Project PIs
★ Data Transfer Collaborators
★ Collaborating Databases
★ Data Analysis Collaborators
Gulf of Mexico

Project PIs & Collaborators. Our DIBBs award includes researchers with both scientific and technical experience from around the US. Tree Community databases that will be integrated are hosted by Dorrie Main (WSU), Meg Staton (U. Tennessee) and Jill Wegerzyn (U. Connecticut).

Module #1 Data Exchange



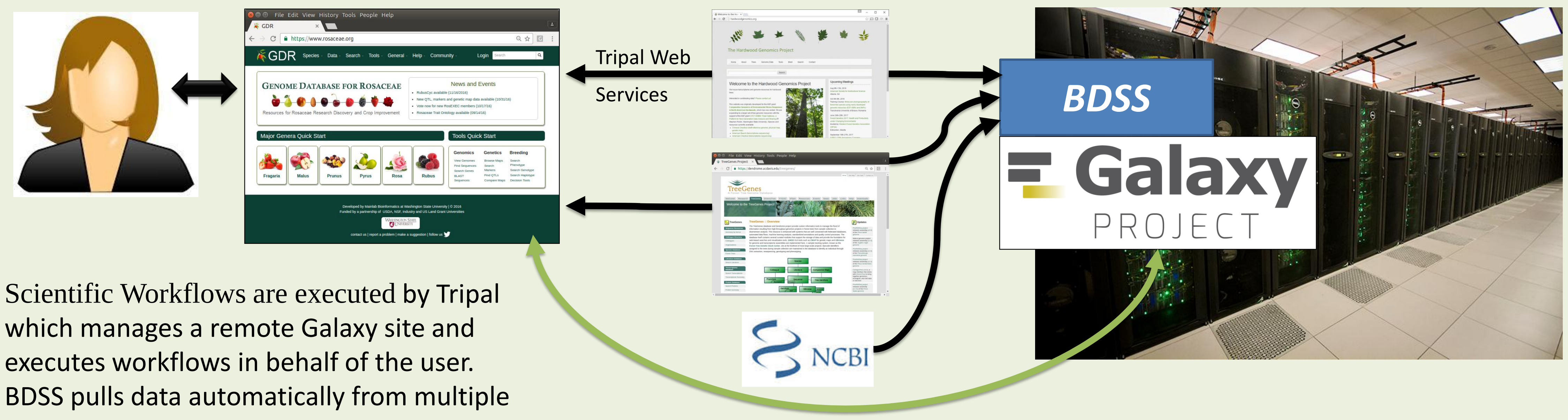
Tripal v3.0 Gene page via web browser (HTML)

Tripal v3.0 Gene page via web services (JSON-LD)

```
192.168.233.139/ws/v0.1/content/mRNA
{
 ▾ "@context": {
     rdfs: http://www.w3.org/1999/02/22-rdf-syntax-ns#,
     hydra: http://www.w3.org/ns/hydra/core#,
     schema: https://schema.org/,
     SO: http://192.168.233.139/ws/v0.1/vocab/SO/,
     foaf: http://xmlns.com/foaf/0.1/,
     Name: "foaf:name",
     Uniquename: "foaf:name",
     label: "rdfs:label",
     itemPage: "schema:itemPage"
   },
   "@id": http://192.168.233.139/ws/v0.1/content/mRNA/310,
   "@type": "SO:0000234",
   label: "orange1.1g022799m, PAC:18136224",
   itemPage: http://192.168.233.139/bio-data/12,
   Name: "orange1.1g022799m",
   Uniquename: "PAC:18136224",
   Organism: "3",
   "Primary Cross Reference": "",
   "MD5 Checksum": "aa7022ead0c2d4e4f17b793534503417",
   "Is Obsolete": false,
   "Is Analysis": false,
   "Time Last Modified": "2015-12-26 22:21:08.513526",
   "Time Accessioned": "2015-12-26 22:21:08.513526",
   Residues:
   "GCAaGTAGGGAGACGGAGTGGAGACTGGAGAGGCGACTCGAAAATGATGAAAGGG
   Seqlen: "2325",
   "Additional Properties": [ ]
}
```

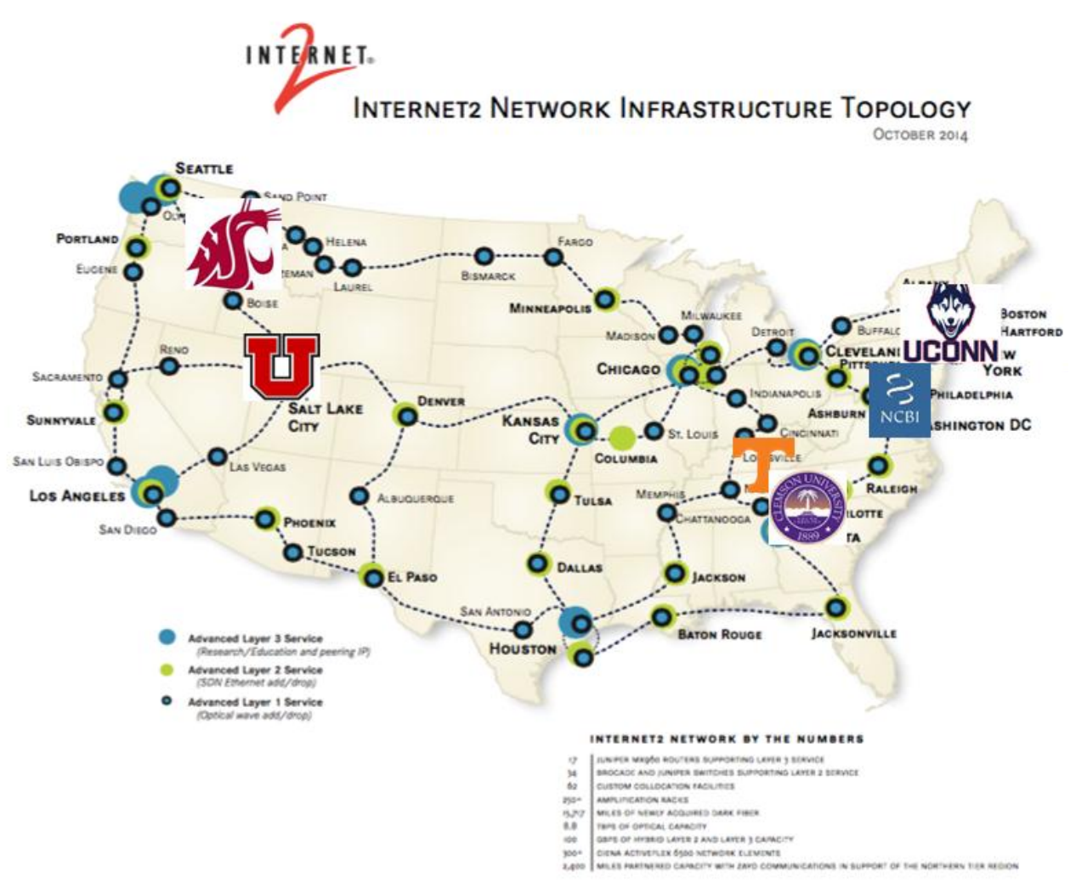
Module #1 of this project comprises creation of web services and infrastructure to support data exchange. The goal is to create a federated group of community databases that store their own data but share it for integration with other collaborative databases. The flowchart above represents the redesign of Tripal for support of the data exchange by representing data types as "entities" and "fields" that are all both mapped to controlled vocabularies. Because all data in Tripal is "typed", site admins can create custom web service APIs by adding data. Web services become discoverable via W3C Hydra protocol.

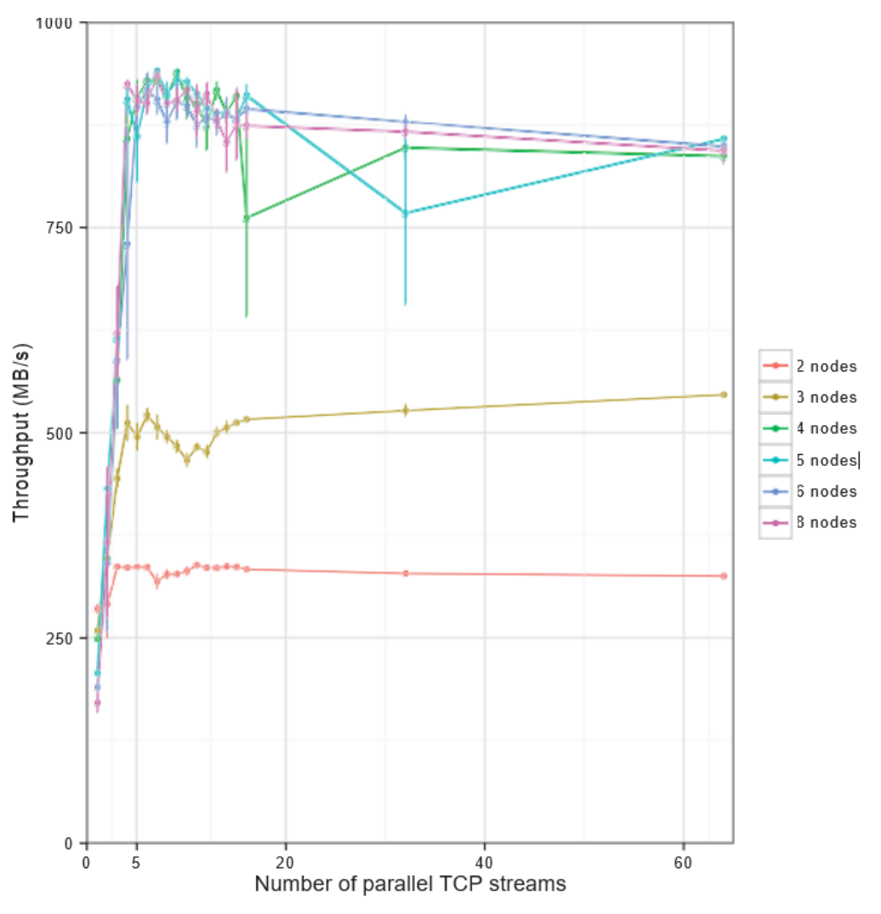
Module #2 Scientific Workflow Execution: Integration with Galaxy.



Scientific Workflows are executed by Tripal which manages a remote Galaxy site and executes workflows in behalf of the user. BDSS pulls data automatically from multiple locations specified by the user.

Module #3: Data Transfer





Our goal is to ensure best transfer speeds for data exchange (module #1) and movement of data for scientific workflow execution (module #2). Leveraging an NSF ACI-REF award (#1341935) with collaborators at Clemson we hope to develop protocols for Software Defined Networking (SDN) to support Tripal-site data exchange. We have developed Big Data Smart Socket (BDSS) software to use knowledge of existing network paths, remote repositories transfer protocols, and client software available on the local machine to make smart decisions about data transfer. The graph to the left demonstrates use of BDSS on a CloudLab-based DTN with a parallel file system transferring data between Clemson and University of Utah over Internet2 AL2S.