Tripal Overview & Future Developments

Stephen P. Ficklin, Lacey-Anne Sanderson, Chun-Huai Cheng, Sook Jung, Kirstin Bett, Dorrie Main**

** Presenting authors*

Tripal Database Network and Initiatives Workshop
PAG XXIII, San Diego CA. Jan 11, 2015.

WASHINGTON STATE UNIVERSITY

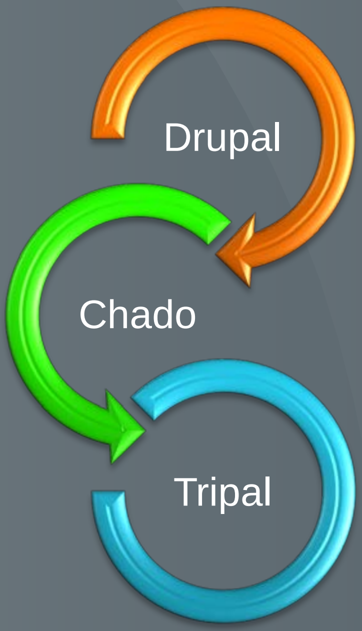
GMOD

UNIVERSITY OF SASKATCHEWAN

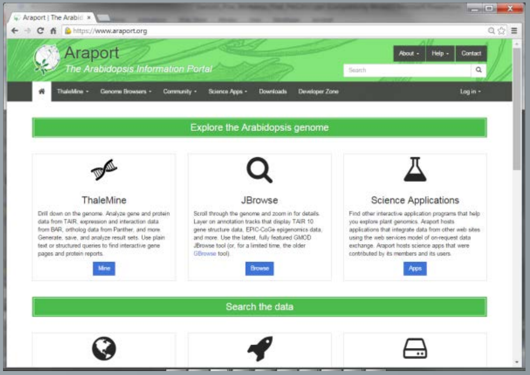
Tripal

- Toolkit for construction of online biological websites
  - A suite of Drupal modules
  - Uses the GMOD Chado schema for data storage
  - Goals
    - Decreases cost
    - Decrease time
    - Facilitate use of standards
    - Reduce “re-invention of the wheel”

http://tripal.info



Example Sites Using Tripal



Explore the Arabidopsis genome

ThaleMine

Drill down on the genome. Analyze gene and protein data from TAIR, expression and interaction data from BAR, ortholog data from Panther, and more. Generate, save, and analyze result sets. Use plain text or structured queries to find interactive gene pages and protein reports.

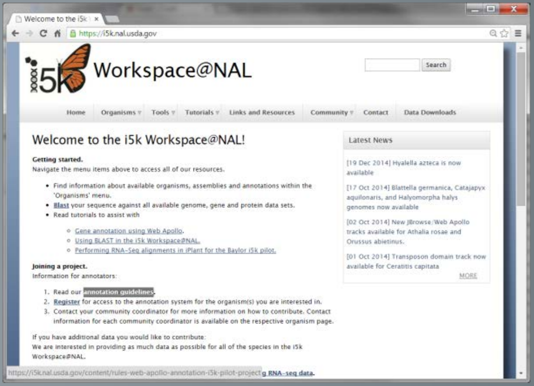
JBrowse

Scroll through the genome and zoom in for details. Layer on annotation tracks that display TAIR 10 gene structure data, EPIC-CoGe epigenomics data, and more. Use the latest, fully featured GMOD JBrowse tool (or, for a limited time, the older GBrowse tool).

Science Applications

Find other interactive application programs that help you explore plant genomics. Araport hosts applications that integrate data from other web sites using the web services model of on-request data exchange. Araport hosts science apps that were contributed by its members and its users.

Search the data



Welcome to the i5k Workspace@NAL!

Getting started.
Navigate the menu items above to access all of our resources.

- Find information about available organisms, assemblies and annotations within the 'Organisms' menu.
- Blast your sequence against all available genome, gene and protein data sets.
- Read tutorials to assist with
  - Gene annotation using Web Apollo.
  - Using BLAST in the i5k Workspace@NAL.
  - Performing RNA-Seq alignments in iPlant for the Baylor i5k pilot.

Joining a project.
Information for annotators:

1. Read our annotation guidelines.
2. Register for access to the annotation system for the organism(s) you are interested in.
3. Contact your community coordinator for more information on how to contribute. Contact information for each community coordinator is available on the respective organism page.

If you have additional data you would like to contribute:
We are interested in providing as much data as possible for all of the species in the i5k Workspace@NAL.

Latest News

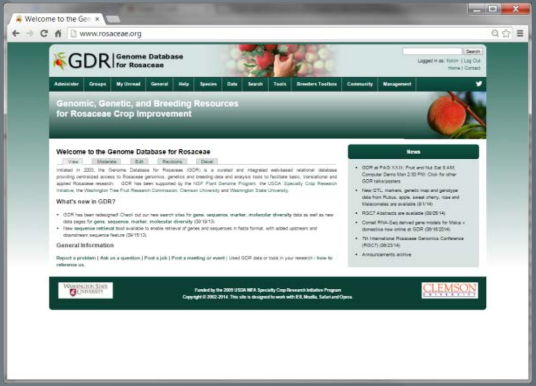
[19 Dec 2014] Hyalella azteca is now available

[17 Oct 2014] Blattella germanica, Catajapyx aquilonaris, and Halyomorpha halys genomes now available

[02 Oct 2014] New JBrowse/Web Apollo tracks available for Athalia rosae and Orussus abietinus.

[01 Oct 2014] Transposon domain track now available for Ceratitis capitata

MORE



Genomic, Genetic, and Breeding Resources for Rosaceae Crop Improvement

Welcome to the Genome Database for Rosaceae

Initiated in 2003, the Genome Database for Rosaceae (GDR) is a curated and integrated web-based relational database providing centralized access to Rosaceae genomics, genetics and breeding data and analysis tools to facilitate basic, translational and applied Rosaceae research. GDR has been supported by the NSF Plant Genome Program, the USDA Specialty Crop Research Initiative, the Washington Tree Fruit Research Commission, Clemson University and Washington State University.

What's new in GDR?

- GDR has been redesigned! Check out our new search sites for gene, sequence, marker, molecular diversity data as well as new data pages for gene, sequence, marker, molecular diversity (09/19/13)
- New sequence retrieval tool available to enable retrieval of genes and sequences in fasta format, with added upstream and downstream sequence feature (09/15/13)

General Information

Report a problem | Ask us a question | Post a job | Post a meeting or event | Used GDR data or tools in your research - how to reference us.

News

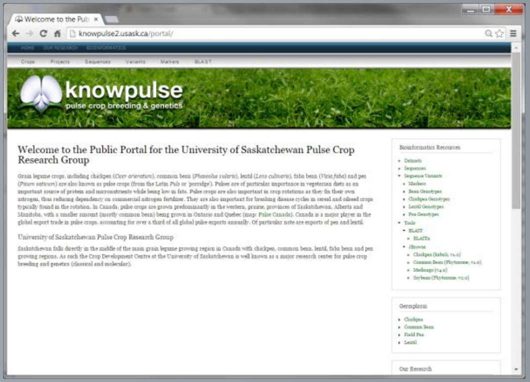
- GDR at PAG XXIII: Fruit and Nut Sat 8 AM; Computer Demo Mon 2:30 PM; Click for other GDR talks/posters
- New QTL, markers, genetic map and genotype data from Rubus, apple, sweet cherry, rose and Maloideae are available (9/1/14)
- RGC7 Abstracts are available (08/05/14)
- Cornell RNA-Seq derived gene models for Malus x domestica now online at GDR (08/18/2014)
- 7th International Rosaceae Genomics Conference (RGC7) (08/23/14)
- Announcements archive

Funded by the 2009 USDA NIFA Specialty Crop Research Initiative Program
Copyright © 2002-2014. This site is designed to work with IE8, Mozilla, Safari and Opera.

Arabidopsis Information Portal
https://www.araport.org//

i5K Workspace@NAL
http://i5k.nal.usda.gov/

Genome Database for Rosaceae
http://www.rosaceae.org



Welcome to the Public Portal for the University of Saskatchewan Pulse Crop Research Group

Grain legume crops, including chickpea (*Cicer arietinum*), common bean (*Phaseolus vulgaris*), lentil (*Lens culinaris*), faba bean (*Vicia faba*) and pea (*Pisum sativum*) are also known as pulse crops (from the Latin *Puls* or 'porridge'). Pulses are of particular importance in vegetarian diets as an important source of protein and micronutrients while being low in fats. Pulse crops are also important in crop rotations as they fix their own nitrogen, thus reducing dependency on commercial nitrogen fertilizer. They are also important for breaking disease cycles in cereal and oilseed crops typically found in the rotation. In Canada, pulse crops are grown predominantly in the western, prairie, provinces of Saskatchewan, Alberta and Manitoba, with a smaller amount (mostly common bean) being grown in Ontario and Quebec (map: Pulse Canada). Canada is a major player in the global export trade in pulse crops, accounting for over a third of all global pulse exports annually. Of particular note are exports of pea and lentil.

University of Saskatchewan Pulse Crop Research Group

Saskatchewan falls directly in the middle of the main grain legume growing region in Canada with chickpea, common bean, lentil, faba bean and pea growing regions. As such the Crop Development Centre at the University of Saskatchewan is well known as a major research center for pulse crop breeding and genetics (classical and molecular).

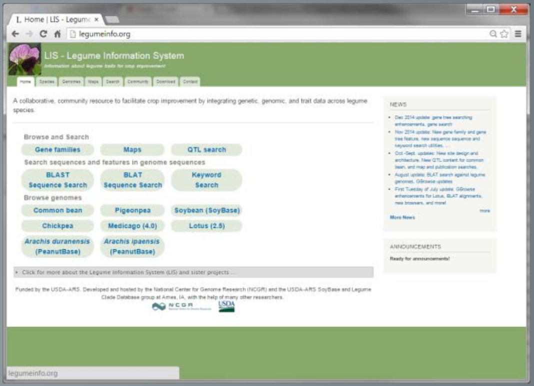
Bioinformatics Resources

- Datasets
- Sequences
- Sequence Variants
  - Markers
  - Bean Genotypes
  - Chickpea Genotypes
  - Lentil Genotypes
  - Pea Genotypes
- Tools
  - BLAST
    - BLASTn
  - JBrowse
    - Chickpea (kabuli, v1.0)
    - Common Bean (Phytozome, v1.0)
    - Medicago (v4.0)
    - Soybean (Phytozome, v2.0)

Germplasm

- Chickpea
- Common Bean
- Field Pea
- Lentil

Our Research



A collaborative, community resource to facilitate crop improvement by integrating genetic, genomic, and trait data across legume species.

Browse and Search

Search sequences and features in genome sequences

Browse genomes

Click for more about the Legume Information System (LIS) and sister projects ...

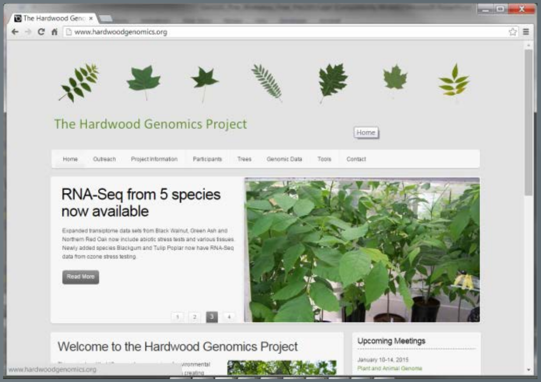
Funded by the USDA-ARS. Developed and hosted by the National Center for Genome Research (NCGR) and the USDA-ARS SoyBase and Legume Clade Database group at Ames, IA, with the help of many other researchers.

NEWS

More News

ANNOUNCEMENTS

Ready for announcements!



The Hardwood Genomics Project

RNA-Seq from 5 species now available

Expanded transiptome data sets from Black Walnut, Green Ash and Northern Red Oak now include abiotic stress tests and various tissues. Newly added species Blackgum and Tulip Poplar now have RNA-Seq data from ozone stress testing.

Welcome to the Hardwood Genomics Project

Upcoming Meetings

January 10-14, 2015
Plant and Animal Genome

Knowpulse: pulse crop genomics & breeding
http://knowpulse2.usask.ca/portal

Legume Information System
http://legumeinfo.org/

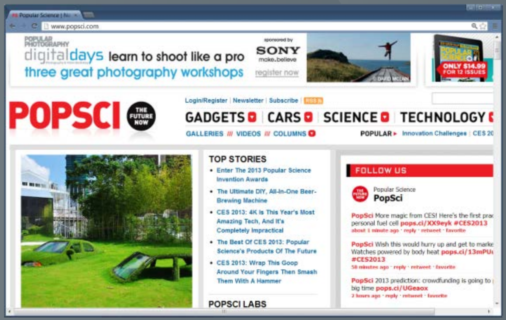
Hardwood Genomics
http://www.hardwoodgenomics.org/

Drupal

- Drupal is an open-source, free Content Management System (CMS)
- Used by millions of websites world-wide
- High profile sites include:

US White House
http://www.whitehouse.gov

British Medical Journal
http://www.bmj.com/



Popular Science
http://www.popsci.com/

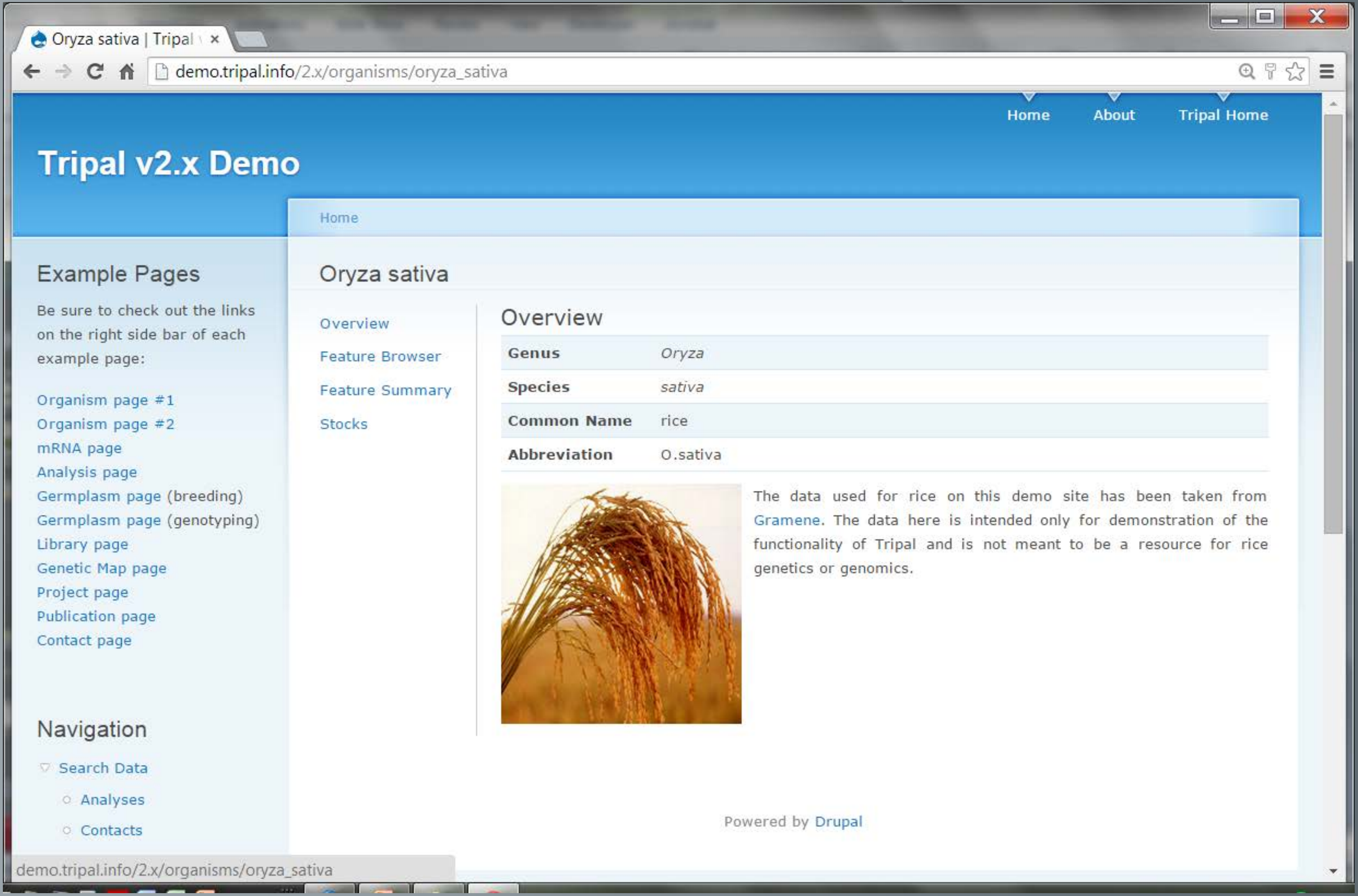
- Drupal provides tools for non-biological needs.
- Thousands of user-contributed functional plugins
- Tripal is a suite of Drupal modules.

Chado

Chado Module	Data Stored	Tripal Support
Cell Line	Cell lines with links to associated data	None
Companalysis	Computational analysis with links to associated data.	Full
Contact	People and links to associated data	Full
CV	Stores terms for controlled vocabularies and ontologies	Full
Expression	Gene expression Data	None
General	Catch-all for database cross-references and projects	Full
Genetic	Linking genotypes, environments and phenotypes	Partial
Group	Grouping data records (under construction)	None
Library	Genomic libraries and ancillary details	Full
Mage	Assay, their projects (intended for microarrays)	None
Map	Genetic or Physical maps.	Full
Natural Diversity	Associates phenotype, genotype, location from multiple accessions	Full
Organism	Species information	Full
Phylogeny	Represents phylogenetic trees	None *
Publication	Publication details with links to associated data	Full
Sequence	Genomic features	Full
Stock	Stocks (germplasm), stock collections, individual specimens and ancillary data	Full

* Support currently under construction as an extension module

Example: Organisms



Tripal v2.x Demo

Home

Example Pages

Be sure to check out the links on the right side bar of each example page:

- Organism page #1
- Organism page #2
- mRNA page
- Analysis page
- Germplasm page (breeding)
- Germplasm page (genotyping)
- Library page
- Genetic Map page
- Project page
- Publication page
- Contact page

Navigation

- Search Data
  - Analyses
  - Contacts

Oryza sativa

- Overview
- Feature Browser
- Feature Summary
- Stocks

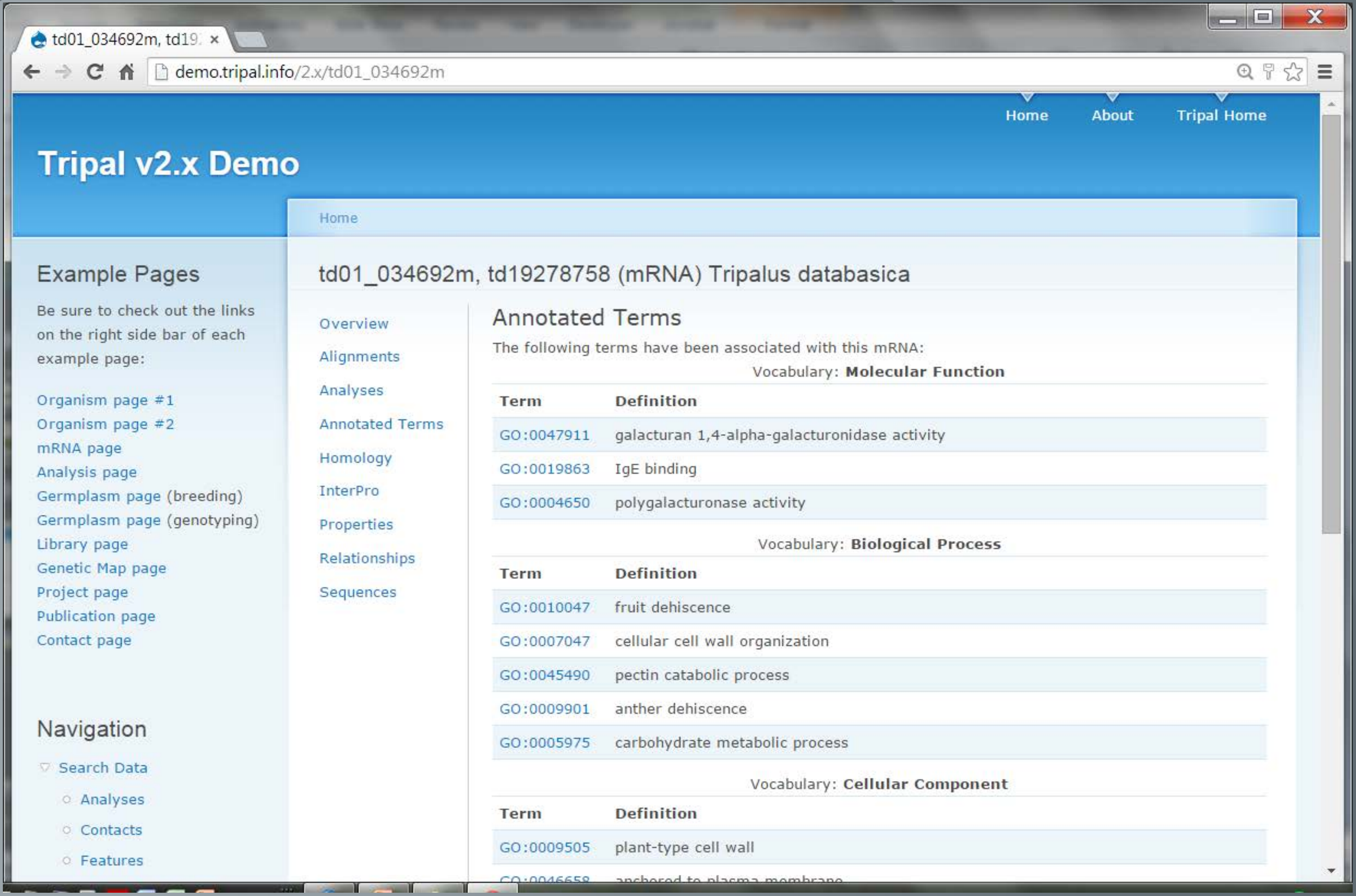
Overview

Genus	*Oryza*
Species	*sativa*
Common Name	rice
Abbreviation	O.sativa

The data used for rice on this demo site has been taken from Gramene. The data here is intended only for demonstration of the functionality of Tripal and is not meant to be a resource for rice genetics or genomics.

Powered by Drupal

Example: Gene (mRNA) Pages



Tripal v2.x Demo

Home

Example Pages

Be sure to check out the links on the right side bar of each example page:

- Organism page #1
- Organism page #2
- mRNA page
- Analysis page
- Germplasm page (breeding)
- Germplasm page (genotyping)
- Library page
- Genetic Map page
- Project page
- Publication page
- Contact page

Navigation

- Search Data
  - Analyses
  - Contacts
  - Features

td01_034692m, td19278758 (mRNA) Tripalus databasica

- Overview
- Alignments
- Analyses
- Annotated Terms
- Homology
- InterPro
- Properties
- Relationships
- Sequences

Annotated Terms

The following terms have been associated with this mRNA:

Vocabulary: **Molecular Function**

Term	Definition
GO:0047911	galacturan 1,4-alpha-galacturonidase activity
GO:0019863	IgE binding
GO:0004650	polygalacturonase activity

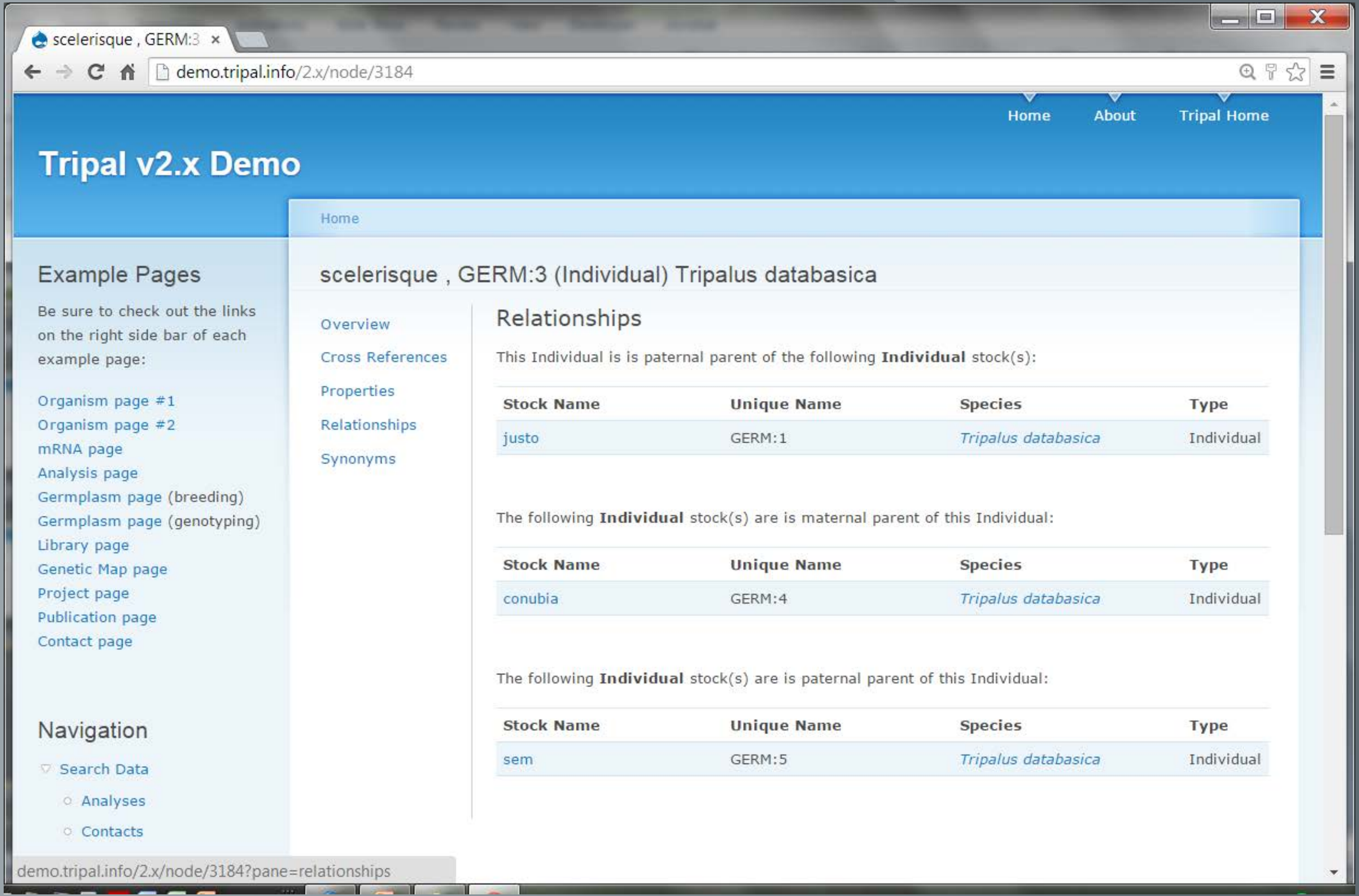
Vocabulary: **Biological Process**

Term	Definition
GO:0010047	fruit dehiscence
GO:0007047	cellular cell wall organization
GO:0045490	pectin catabolic process
GO:0009901	anther dehiscence
GO:0005975	carbohydrate metabolic process

Vocabulary: **Cellular Component**

Term	Definition
GO:0009505	plant-type cell wall
GO:0046658	anchored to plasma membrane

Example: Germplasm in Breeding



Tripal v2.x Demo

Example Pages

Be sure to check out the links on the right side bar of each example page:

- Organism page #1
- Organism page #2
- mRNA page
- Analysis page
- Germplasm page (breeding)
- Germplasm page (genotyping)
- Library page
- Genetic Map page
- Project page
- Publication page
- Contact page

Navigation

- Search Data
  - Analyses
  - Contacts

scelerisque , GERM:3 (Individual) Tripalus databasica

Relationships

This Individual is is paternal parent of the following **Individual** stock(s):

Stock Name	Unique Name	Species	Type
justo	GERM:1	*Tripalus databasica*	Individual

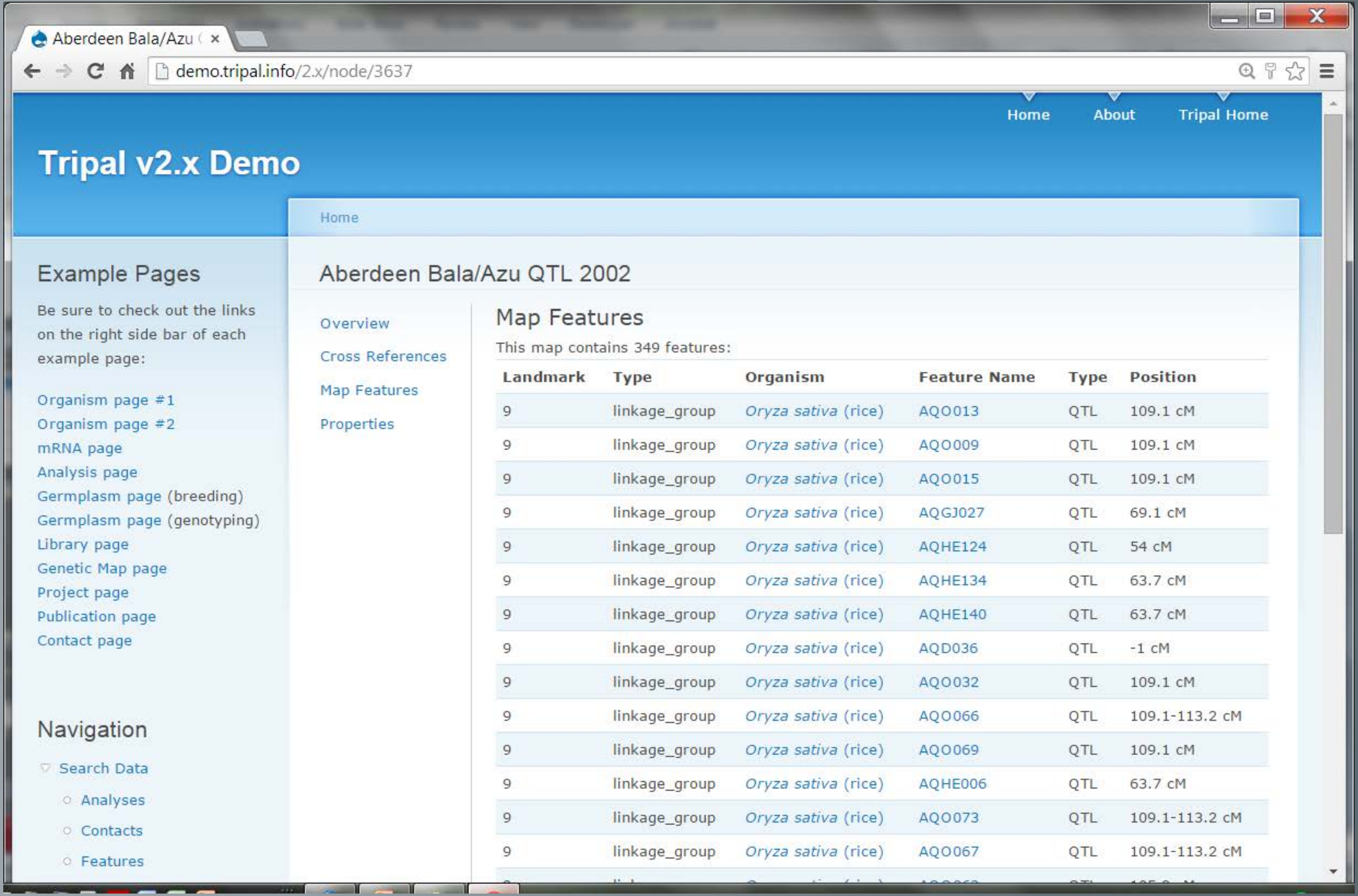
The following **Individual** stock(s) are is maternal parent of this Individual:

Stock Name	Unique Name	Species	Type
conubia	GERM:4	*Tripalus databasica*	Individual

The following **Individual** stock(s) are is paternal parent of this Individual:

Stock Name	Unique Name	Species	Type
sem	GERM:5	*Tripalus databasica*	Individual

Example: Genetic Map



Tripal v2.x Demo

Example Pages

Be sure to check out the links on the right side bar of each example page:

- Organism page #1
- Organism page #2
- mRNA page
- Analysis page
- Germplasm page (breeding)
- Germplasm page (genotyping)
- Library page
- Genetic Map page
- Project page
- Publication page
- Contact page

Navigation

- Search Data
  - Analyses
  - Contacts
  - Features

Aberdeen Bala/Azu QTL 2002

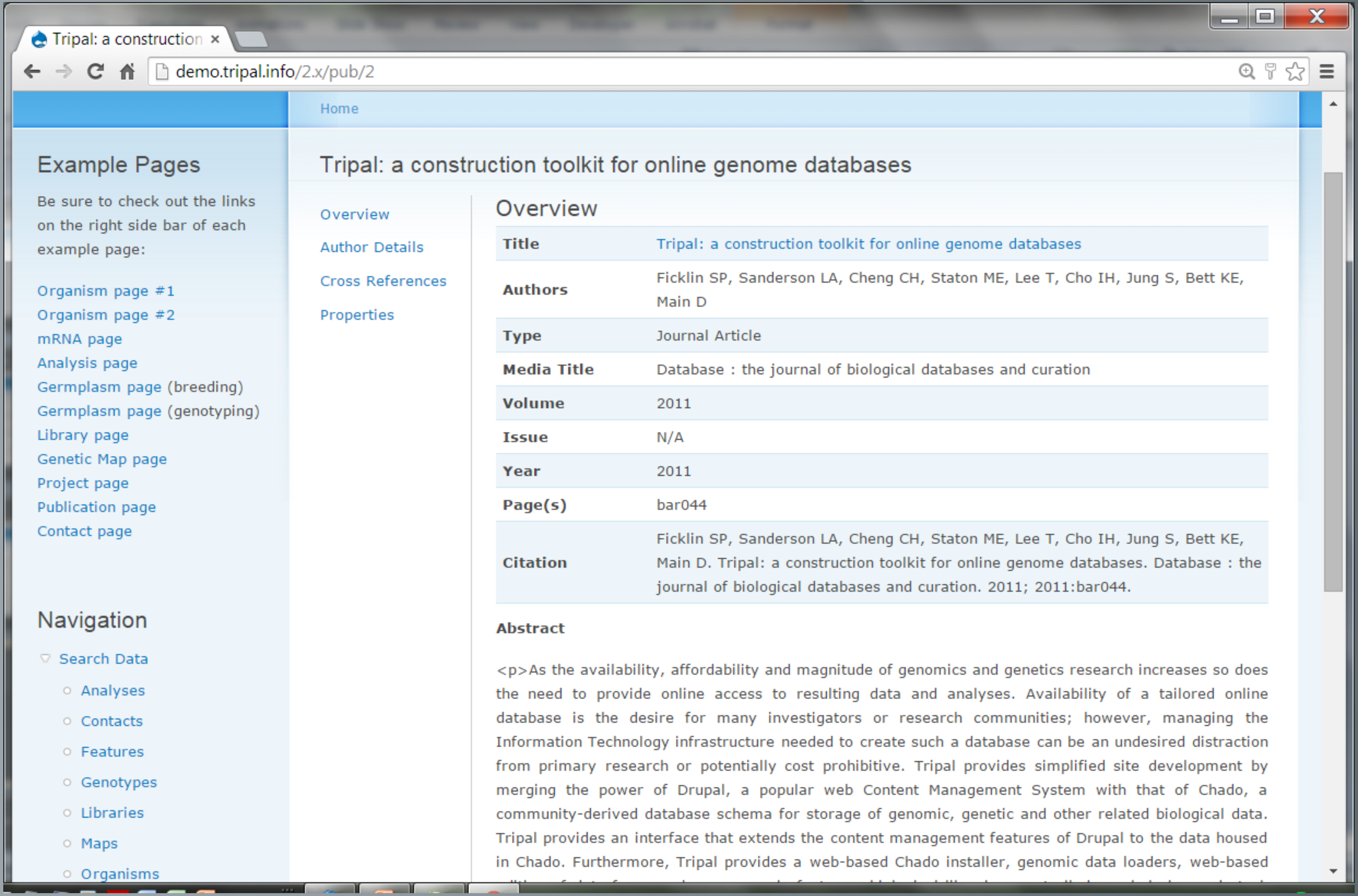
- Overview
- Cross References
- Map Features
- Properties

Map Features

This map contains 349 features:

Landmark	Type	Organism	Feature Name	Type	Position
9	linkage_group	*Oryza sativa* (rice)	AQO013	QTL	109.1 cM
9	linkage_group	*Oryza sativa* (rice)	AQO009	QTL	109.1 cM
9	linkage_group	*Oryza sativa* (rice)	AQO015	QTL	109.1 cM
9	linkage_group	*Oryza sativa* (rice)	AQGJ027	QTL	69.1 cM
9	linkage_group	*Oryza sativa* (rice)	AQHE124	QTL	54 cM
9	linkage_group	*Oryza sativa* (rice)	AQHE134	QTL	63.7 cM
9	linkage_group	*Oryza sativa* (rice)	AQHE140	QTL	63.7 cM
9	linkage_group	*Oryza sativa* (rice)	AQD036	QTL	-1 cM
9	linkage_group	*Oryza sativa* (rice)	AQO032	QTL	109.1 cM
9	linkage_group	*Oryza sativa* (rice)	AQO066	QTL	109.1-113.2 cM
9	linkage_group	*Oryza sativa* (rice)	AQO069	QTL	109.1 cM
9	linkage_group	*Oryza sativa* (rice)	AQHE006	QTL	63.7 cM
9	linkage_group	*Oryza sativa* (rice)	AQO073	QTL	109.1-113.2 cM
9	linkage_group	*Oryza sativa* (rice)	AQO067	QTL	109.1-113.2 cM

Example: Publication Page



Be sure to check out the links on the right side bar of each example page:

Tripal: a construction toolkit for online genome databases

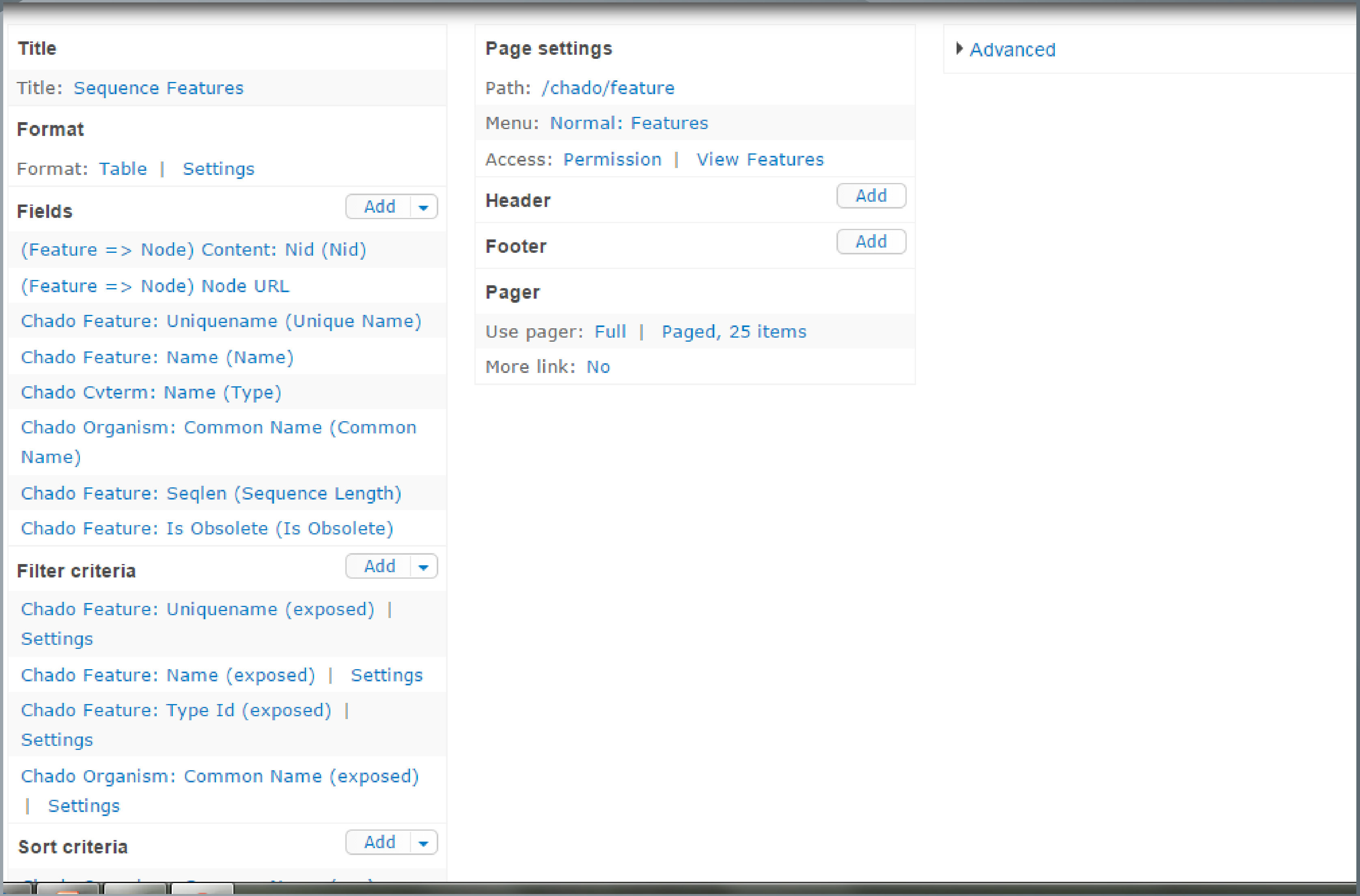
Overview

Title	Tripal: a construction toolkit for online genome databases
Authors	Ficklin SP, Sanderson LA, Cheng CH, Staton ME, Lee T, Cho IH, Jung S, Bett KE, Main D
Type	Journal Article
Media Title	Database : the journal of biological databases and curation
Volume	2011
Issue	N/A
Year	2011
Page(s)	bar044
Citation	Ficklin SP, Sanderson LA, Cheng CH, Staton ME, Lee T, Cho IH, Jung S, Bett KE, Main D. Tripal: a construction toolkit for online genome databases. Database : the journal of biological databases and curation. 2011; 2011:bar044.

Abstract

<p>As the availability, affordability and magnitude of genomics and genetics research increases so does the need to provide online access to resulting data and analyses. Availability of a tailored online database is the desire for many investigators or research communities; however, managing the Information Technology infrastructure needed to create such a database can be an undesired distraction from primary research or potentially cost prohibitive. Tripal provides simplified site development by merging the power of Drupal, a popular web Content Management System with that of Chado, a community-derived database schema for storage of genomic, genetic and other related biological data. Tripal provides an interface that extends the content management features of Drupal to the data housed in Chado. Furthermore, Tripal provides a web-based Chado installer, genomic data loaders, web-based

Drupal Views + Tripal



Title

Title: Sequence Features

Format

Format: Table | Settings

Fields

(Feature => Node) Content: Nid (Nid)
(Feature => Node) Node URL
Chado Feature: Uniquename (Unique Name)
Chado Feature: Name (Name)
Chado Cvterm: Name (Type)
Chado Organism: Common Name (Common Name)
Chado Feature: Seqlen (Sequence Length)
Chado Feature: Is Obsolete (Is Obsolete)

Filter criteria

Chado Feature: Uniquename (exposed) | Settings
Chado Feature: Name (exposed) | Settings
Chado Feature: Type Id (exposed) | Settings
Chado Organism: Common Name (exposed) | Settings

Sort criteria

Page settings

Path: /chado/feature
Menu: Normal: Features
Access: Permission | View Features

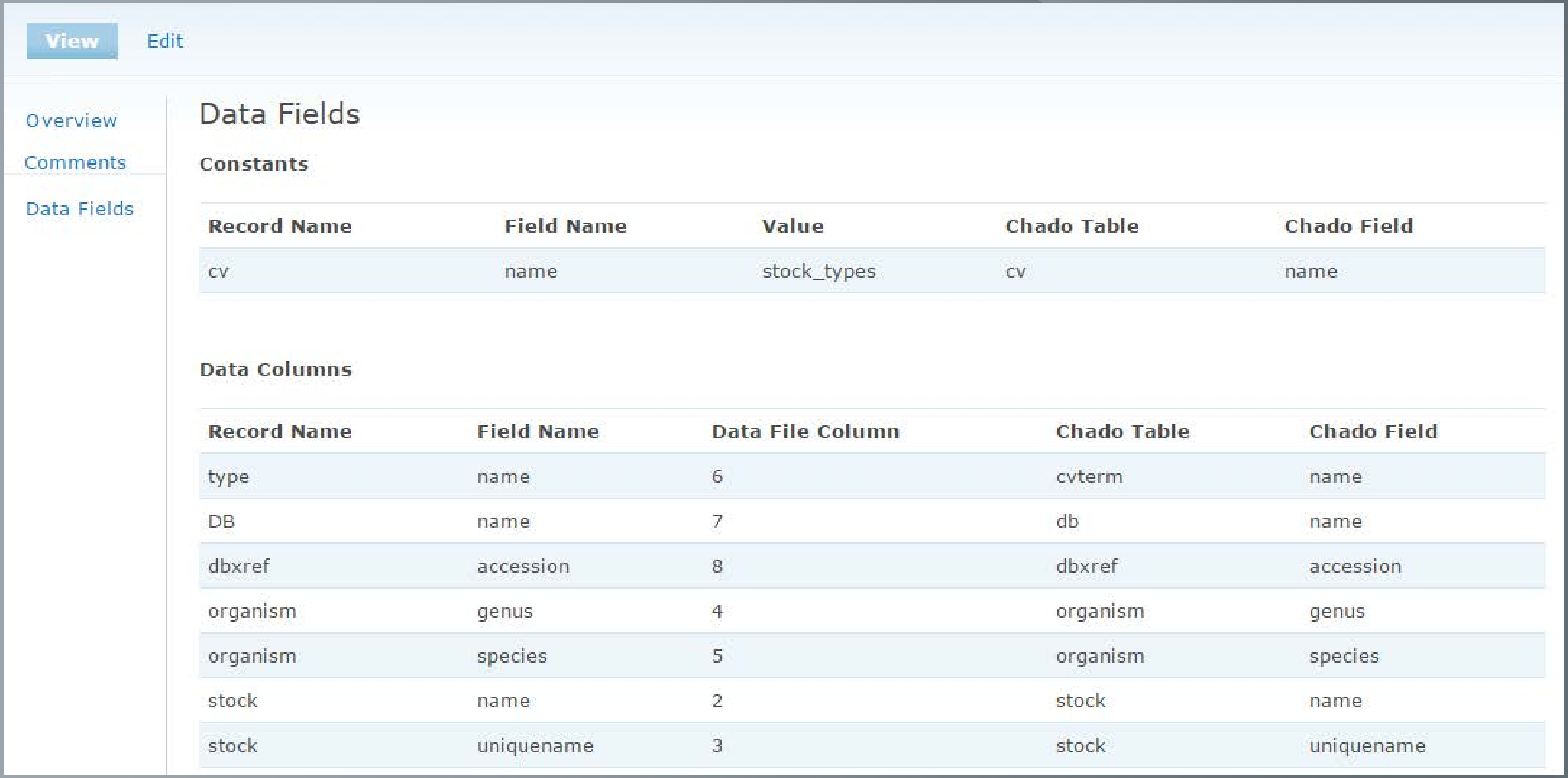
Header

Footer

Pager

Use pager: Full | Paged, 25 items
More link: No

Bulk Loader



Data Fields

Constants

Record Name	Field Name	Value	Chado Table	Chado Field
cv	name	stock_types	cv	name

Data Columns

Record Name	Field Name	Data File Column	Chado Table	Chado Field
type	name	6	cvterm	name
DB	name	7	db	name
dbxref	accession	8	dbxref	accession
organism	genus	4	organism	genus
organism	species	5	organism	species
stock	name	2	stock	name
stock	uniquename	3	stock	uniquename

Two Ways to Use Tripal

1. “Out of the box”.
   - Use an existing Drupal theme for look-and-feel of site.
   - Use provided data importers.
   - Use bulk loader for data in non standard file formats.
   - Use provided data viewers.
   - Use provided search tools.
   - Use Drupal & Tripal admin interface to manage data, create custom views and build new search tools.
   - No programming required

2. Customize
   - Use Application Programming Interface (API)
   - Customize how data is presented
   - Create new modules that extend Tripal functionality
   - Share customizations with other Tripal users.

Funding for Future Development

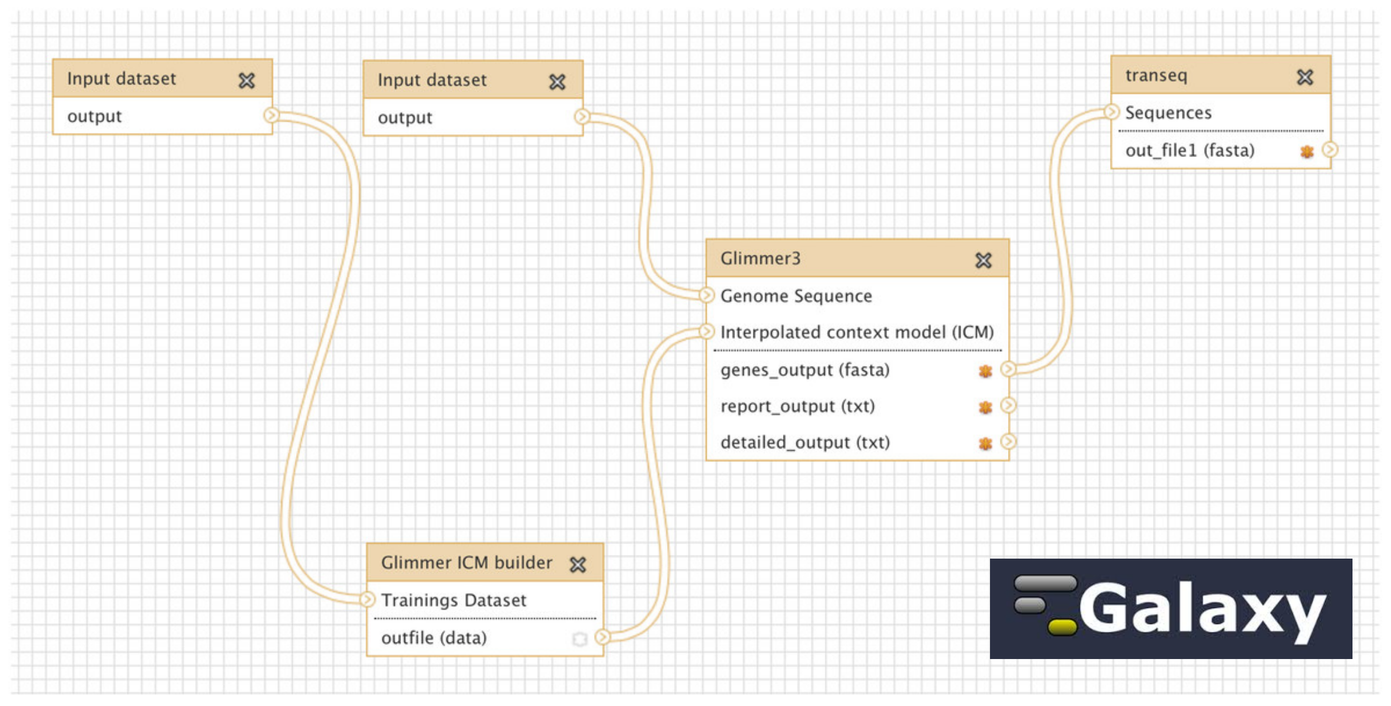
- Current funding at WSU
  - $1.5M NSF DIBBs Award #1443040 (2015-2018). PI Ficklin
  - $2M USDA NRSP10 Award (2014-2019). (PI Main)
  - $2.7M USDA SCRI Award # 2014-51181-22376 (2014-2019) (PI Main)
- Current funding at Univ. Saskatchewan
  - Lentil Genome Sequencing (LenGen) Establishing a Comprehensive Platform for Molecular Breeding: Saskatchewan Pulse Growers (SPG), July 2013 - Jun 2015 (PI Bett)
  - Sequencing the Pea Genome: Saskatchewan Pulse Growers (SPG), August 2013 - July 2015 (PI Bett)
- Additional funding for Tripal “Extensions”

NSF DIBBs Part 1

- Web Services
  - Goals:
    - Provide programmatic access to data
    - Allow Tripal sites to exchange and show data between sites.
  - Legume and tree communities are test cases
    - Legumes
      - KnowPulse Database
      - Cool Season Food Legume Database
      - LegumeInfo
    - Trees
      - Hardwood Genomics Database
      - Genome Database for Rosaceae
      - Citrus Genome Database
      - TreeGenes
  - Will be integrated into core of Tripal

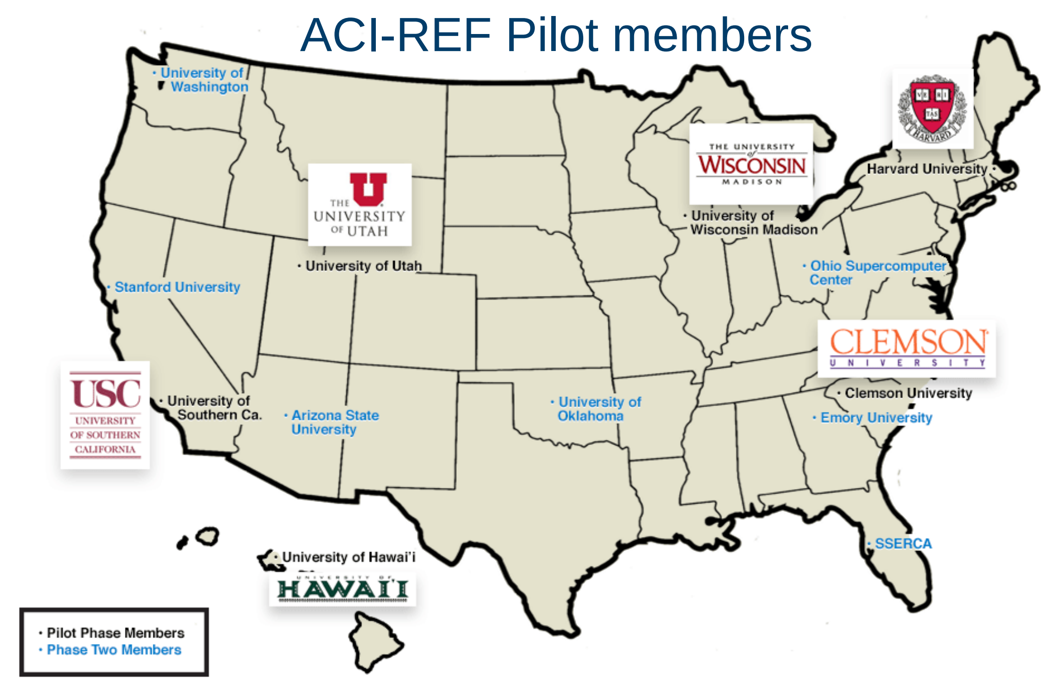
NSF DIBBs Part 2

- Galaxy Integration
  - Development of PHP bindings for Galaxy API
  - Analytical workflows are constructed in Galaxy
  - Data is available via web services for use in workflows
  - Workflows useful for legume and tree communities



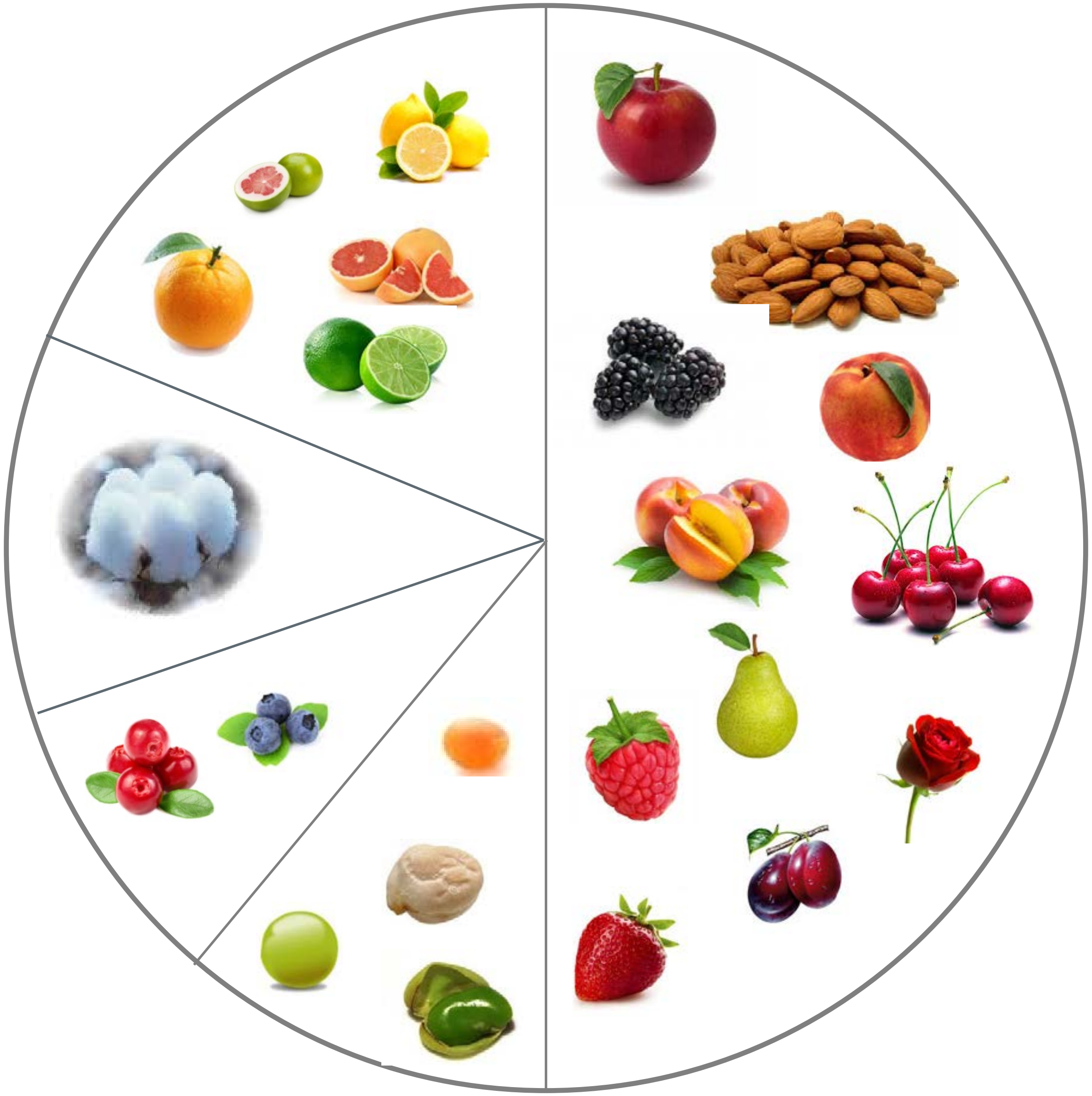
NSF DIBBs Part 3

- Software Defined Networking
  - Leverages SDN research part of NSF ACI-REF project.
  - Attempts to optimize path data passes over the network
  - Can significantly improve data transfer rates.
  - Explore SDN integration of Tripal-hosting facilities
  - Share experiences to help others do the same

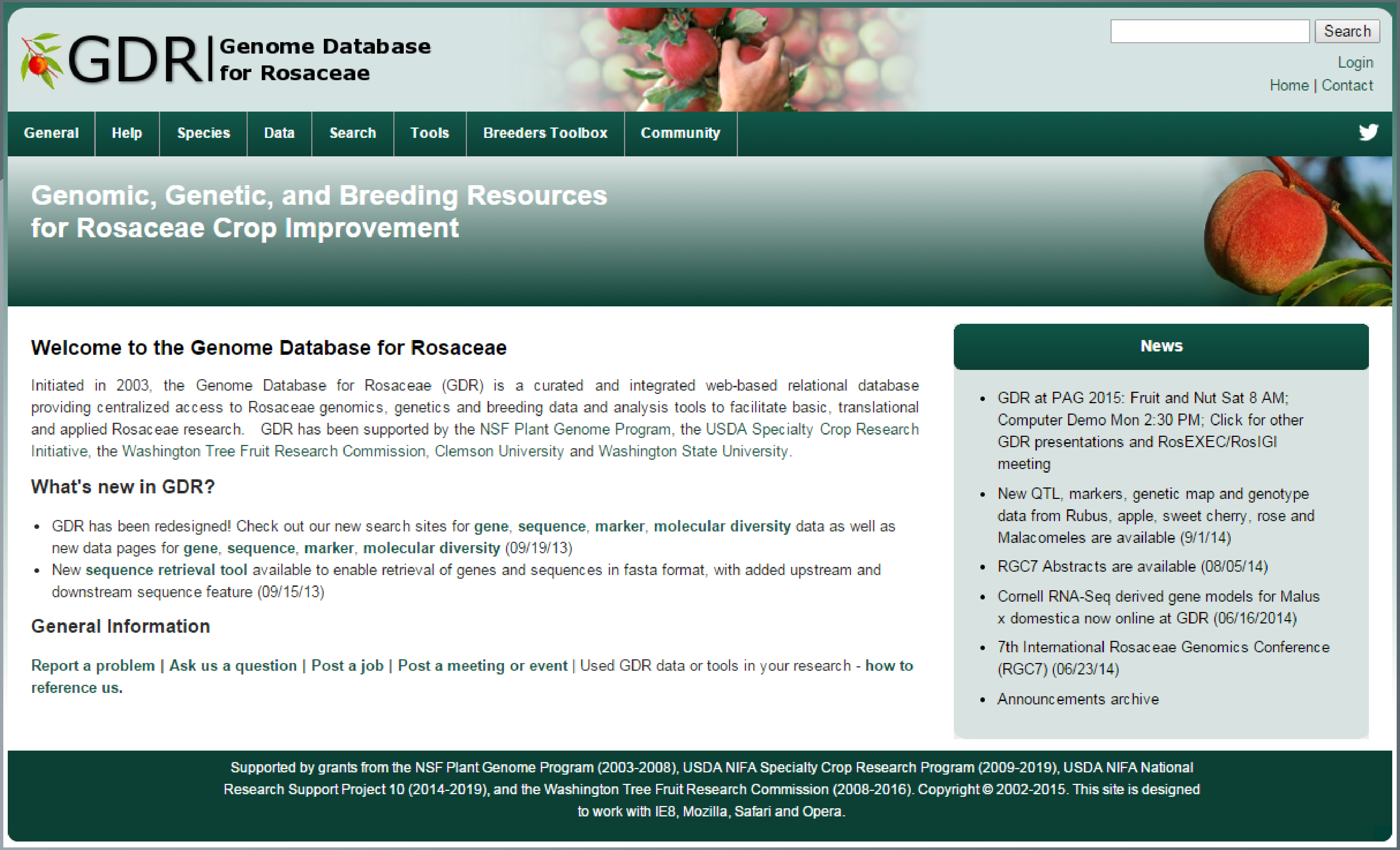

ACI-REF Pilot members

USDA NIFA NRSP10 and GDR

- Release custom modules developed for Main Lab databases
- Tripal API for software that we have developed
    - Breeders toolbox
    - GenSAS, the community genome annotation tool
- Tripal API for a tablet app to collect phenotypic data from the field
- Web Services to allow programmatic access to data to support remote data collection
- Data management: online data submission and editing
- Help Desk Support



- 24 crops (no access to federal flow through funds)
- Annual US value of production > $23 Billion
- Databases serve > 25,000 users from 153 countries



Genomic, Genetic, and Breeding Resources for Rosaceae Crop Improvement

Welcome to the Genome Database for Rosaceae

Initiated in 2003, the Genome Database for Rosaceae (GDR) is a curated and integrated web-based relational database providing centralized access to Rosaceae genomics, genetics and breeding data and analysis tools to facilitate basic, translational and applied Rosaceae research. GDR has been supported by the NSF Plant Genome Program, the USDA Specialty Crop Research Initiative, the Washington Tree Fruit Research Commission, Clemson University and Washington State University.

What's new in GDR?

- GDR has been redesigned! Check out our new search sites for **gene**, **sequence**, **marker**, **molecular diversity** data as well as new data pages for **gene**, **sequence**, **marker**, **molecular diversity** (09/19/13)
- New **sequence retrieval tool** available to enable retrieval of genes and sequences in fasta format, with added upstream and downstream sequence feature (09/15/13)

General Information

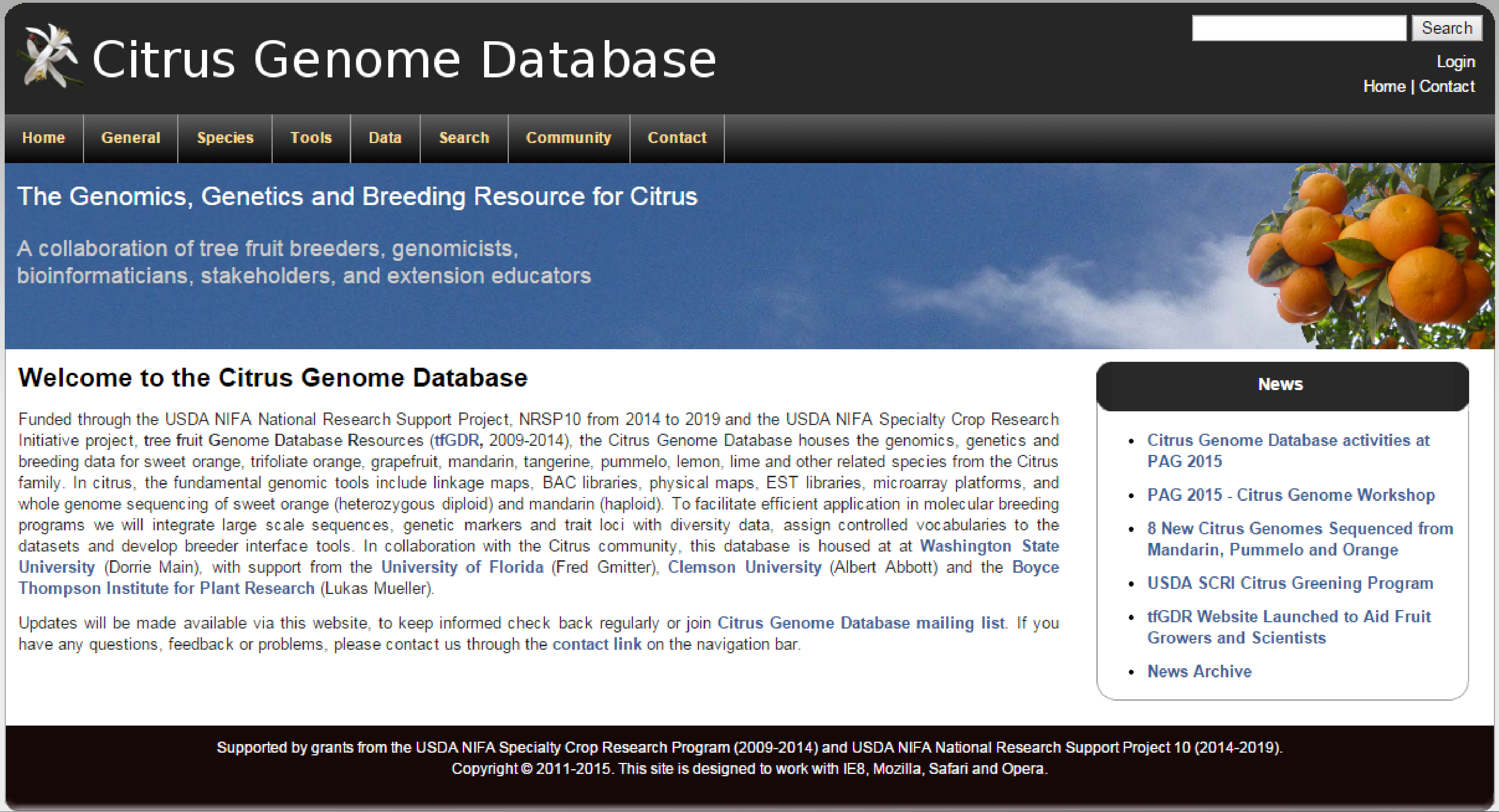
Report a problem | **Ask us a question** | **Post a job** | **Post a meeting or event** | Used GDR data or tools in your research - **how to reference us.**

News

- GDR at PAG 2015: Fruit and Nut Sat 8 AM; Computer Demo Mon 2:30 PM; Click for other GDR presentations and RosEXEC/RosIGI meeting
- New QTL, markers, genetic map and genotype data from Rubus, apple, sweet cherry, rose and Malacomeles are available (9/1/14)
- RGC7 Abstracts are available (08/05/14)
- Cornell RNA-Seq derived gene models for Malus x domestica now online at GDR (06/16/2014)
- 7th International Rosaceae Genomics Conference (RGC7) (06/23/14)
- Announcements archive

Supported by grants from the NSF Plant Genome Program (2003-2008), USDA NIFA Specialty Crop Research Program (2009-2019), USDA NIFA National Research Support Project 10 (2014-2019), and the Washington Tree Fruit Research Commission (2008-2016). Copyright © 2002-2015. This site is designed to work with IE8, Mozilla, Safari and Opera.

www.rosaceae.org



Citrus Genome Database

The Genomics, Genetics and Breeding Resource for Citrus

A collaboration of tree fruit breeders, genomicists, bioinformaticians, stakeholders, and extension educators

Welcome to the Citrus Genome Database

Funded through the USDA NIFA National Research Support Project, NRSP10 from 2014 to 2019 and the USDA NIFA Specialty Crop Research Initiative project, tree fruit Genome Database Resources (**tfGDR**, 2009-2014), the Citrus Genome Database houses the genomics, genetics and breeding data for sweet orange, trifoliate orange, grapefruit, mandarin, tangerine, pummelo, lemon, lime and other related species from the Citrus family. In citrus, the fundamental genomic tools include linkage maps, BAC libraries, physical maps, EST libraries, microarray platforms, and whole genome sequencing of sweet orange (heterozygous diploid) and mandarin (haploid). To facilitate efficient application in molecular breeding programs we will integrate large scale sequences, genetic markers and trait loci with diversity data, assign controlled vocabularies to the datasets and develop breeder interface tools. In collaboration with the Citrus community, this database is housed at at **Washington State University** (Dorrie Main), with support from the **University of Florida** (Fred Gmitter), **Clemson University** (Albert Abbott) and the **Boyce Thompson Institute for Plant Research** (Lukas Mueller).

Updates will be made available via this website, to keep informed check back regularly or join **Citrus Genome Database mailing list**. If you have any questions, feedback or problems, please contact us through the **contact link** on the navigation bar.

- Citrus Genome Database activities at PAG 2015
- PAG 2015 - Citrus Genome Workshop
- 8 New Citrus Genomes Sequenced from Mandarin, Pummelo and Orange
- USDA SCRI Citrus Greening Program
- tfGDR Website Launched to Aid Fruit Growers and Scientists
- News Archive

Supported by grants from the USDA NIFA Specialty Crop Research Program (2009-2014) and USDA NIFA National Research Support Project 10 (2014-2019).

www.citrusgenomedb.org

Cool Season Food Legume Genome Database

Building a Genomics, Genetics and Breeding Resource for Cool Season Food Legume Improvement

Welcome to the Cool Season Food Legume Genome Database

Pea (*Pisum sativum* L.), **lentil** (*Lens culinaris* Medik.), **chickpea** (*Cicer arietinum* L.) and faba bean (*Vicia faba* L.) are cool season food legume crops. Members of the Fabaceae family, these legumes are distinguished by providing delicious and nutritious grains to the human diet. These crops serve as an important part of sustainable agricultural systems by providing biologically fixed nitrogen, made possible through symbiosis with rhizobia bacteria, and serving as a key rotation crop. Genomic information from the sequenced legume crops (Chickpea, *Medicago*, Lotus and Soybean) along with genomic and genetic data from the cool season food legumes provide a framework and road map for crop improvement. This database is being developed to serve as a resource to facilitate cool season food legume plant breeding and research by providing centralized, integrated access to relevant genetic and genomic information.

What's New in the CSFL Genome Database

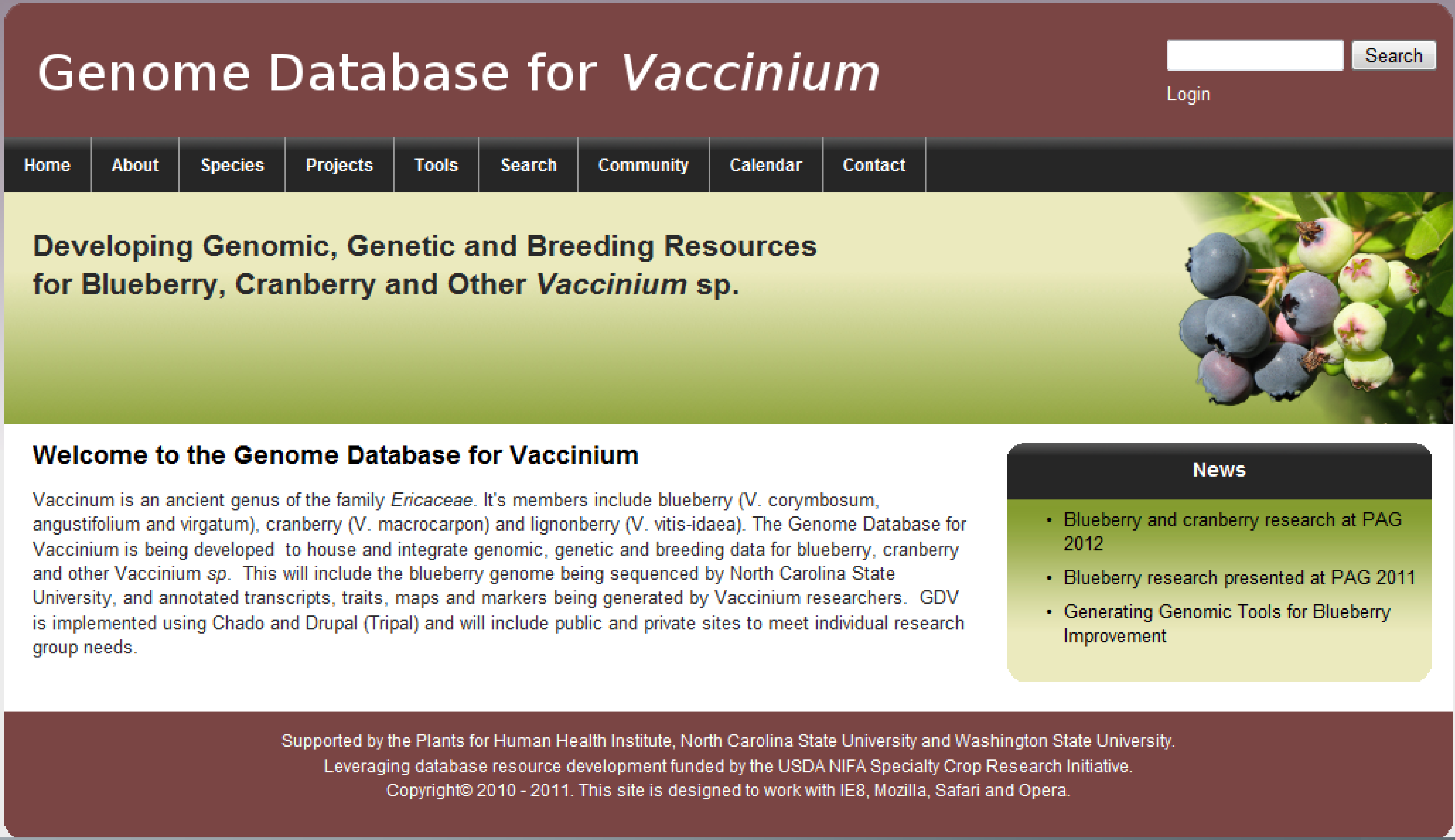
- Chickpea GBrowse | QTL search | Marker search | Sequence search | Franssen et al., 2012 Pea Transcriptome
- Data submission | Report a problem | Ask us a question | Post a job | Post a meeting or event | What's been added/fixed? | What are we working on? | Used CSFL Genome DB data or tools in your research - how to reference us.

News

- NAPIA meeting to be held at Niagra Falls, Ontario. 5 - 7 November 2015
- Cool Season Legume Workshop at PAG 2015
- Mendel's Legacy Conference in the Czech Republic, Sept. 7-10, 2015
- Mendel's legacy: International Pea Genome Sequencing Consortium formed!
- "Aragorn" and pea v2 unigenes now available.
- News archives

Supported by grants from the USDA NIFA WA Specialty Crop Block Research Program (2013-2016), USDA NIFA National Research Support Project 10 (2014-2019), the US Dry Pea and Lentil Council and the Northern Pulse Growers Association.

www.coolseasonfoodlegume.org



Genome Database for *Vaccinium*

Developing Genomic, Genetic and Breeding Resources for Blueberry, Cranberry and Other *Vaccinium* sp.

Welcome to the Genome Database for Vaccinium

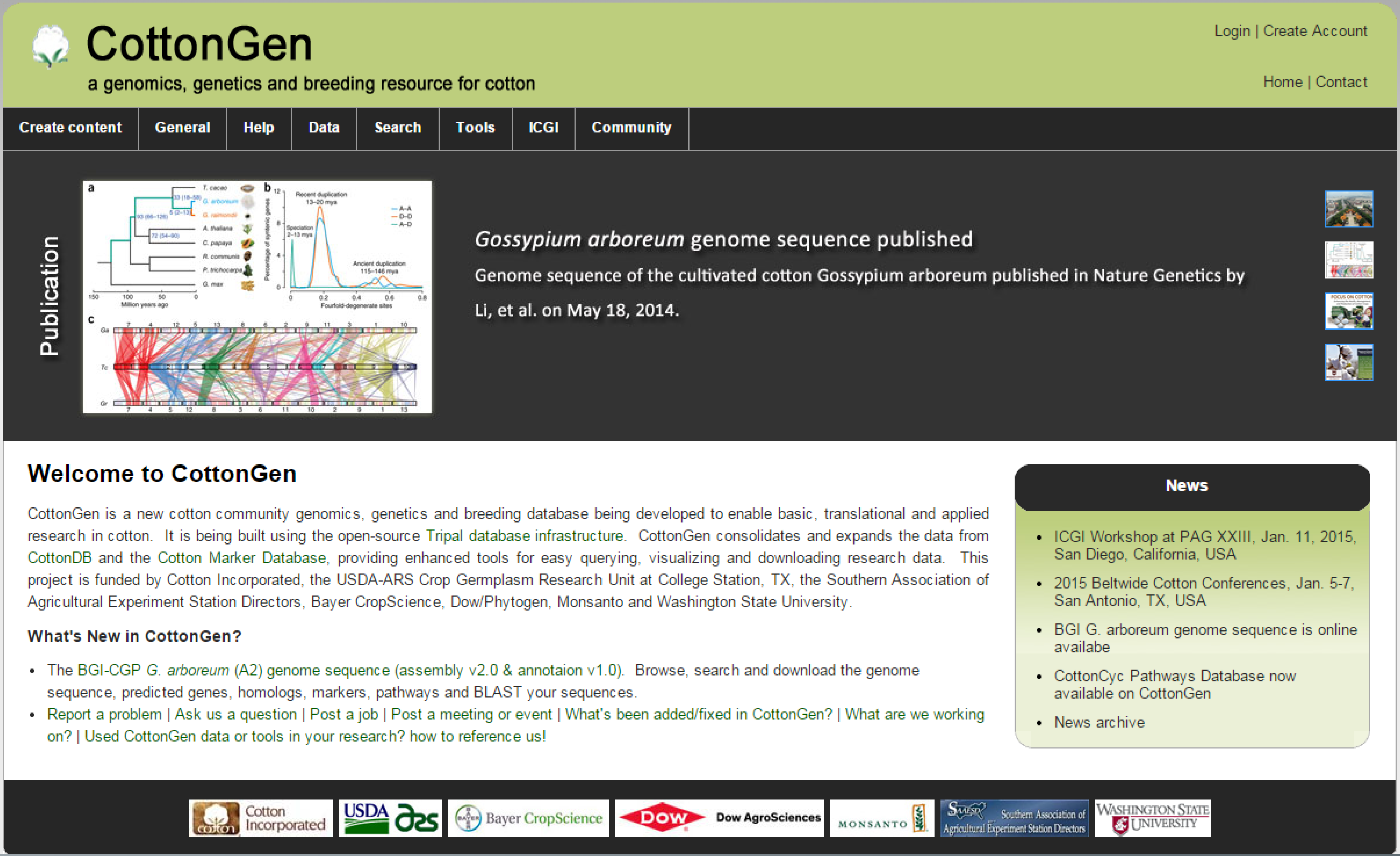
Vaccinum is an ancient genus of the family *Ericaceae*. It's members include blueberry (V. corymbosum, angustifolium and virgatum), cranberry (V. macrocarpon) and lignonberry (V. vitis-idaea). The Genome Database for Vaccinium is being developed to house and integrate genomic, genetic and breeding data for blueberry, cranberry and other Vaccinium *sp*. This will include the blueberry genome being sequenced by North Carolina State University, and annotated transcripts, traits, maps and markers being generated by Vaccinium researchers. GDV is implemented using Chado and Drupal (Tripal) and will include public and private sites to meet individual research group needs.

News

- Blueberry and cranberry research at PAG 2012
- Blueberry research presented at PAG 2011
- Generating Genomic Tools for Blueberry Improvement

Supported by the Plants for Human Health Institute, North Carolina State University and Washington State University.
Leveraging database resource development funded by the USDA NIFA Specialty Crop Research Initiative.

www.vaccinium.org



Gossypium arboreum genome sequence published

Genome sequence of the cultivated cotton Gossypium arboreum published in Nature Genetics by Li, et al. on May 18, 2014.

Welcome to CottonGen

CottonGen is a new cotton community genomics, genetics and breeding database being developed to enable basic, translational and applied research in cotton. It is being built using the open-source Tripal database infrastructure. CottonGen consolidates and expands the data from CottonDB and the Cotton Marker Database, providing enhanced tools for easy querying, visualizing and downloading research data. This project is funded by Cotton Incorporated, the USDA-ARS Crop Germplasm Research Unit at College Station, TX, the Southern Association of Agricultural Experiment Station Directors, Bayer CropScience, Dow/Phytogen, Monsanto and Washington State University.

What's New in CottonGen?

- The BGI-CGP *G. arboreum* (A2) genome sequence (assembly v2.0 & annotaion v1.0). Browse, search and download the genome sequence, predicted genes, homologs, markers, pathways and BLAST your sequences.
- Report a problem | Ask us a question | Post a job | Post a meeting or event | What's been added/fixed in CottonGen? | What are we working on? | Used CottonGen data or tools in your research? how to reference us!

News

- ICGI Workshop at PAG XXIII, Jan. 11, 2015, San Diego, California, USA
- 2015 Beltwide Cotton Conferences, Jan. 5-7, San Antonio, TX, USA
- BGI G. arboreum genome sequence is online availabe
- CottonCyc Pathways Database now available on CottonGen
- News archive

www.cottongen.org

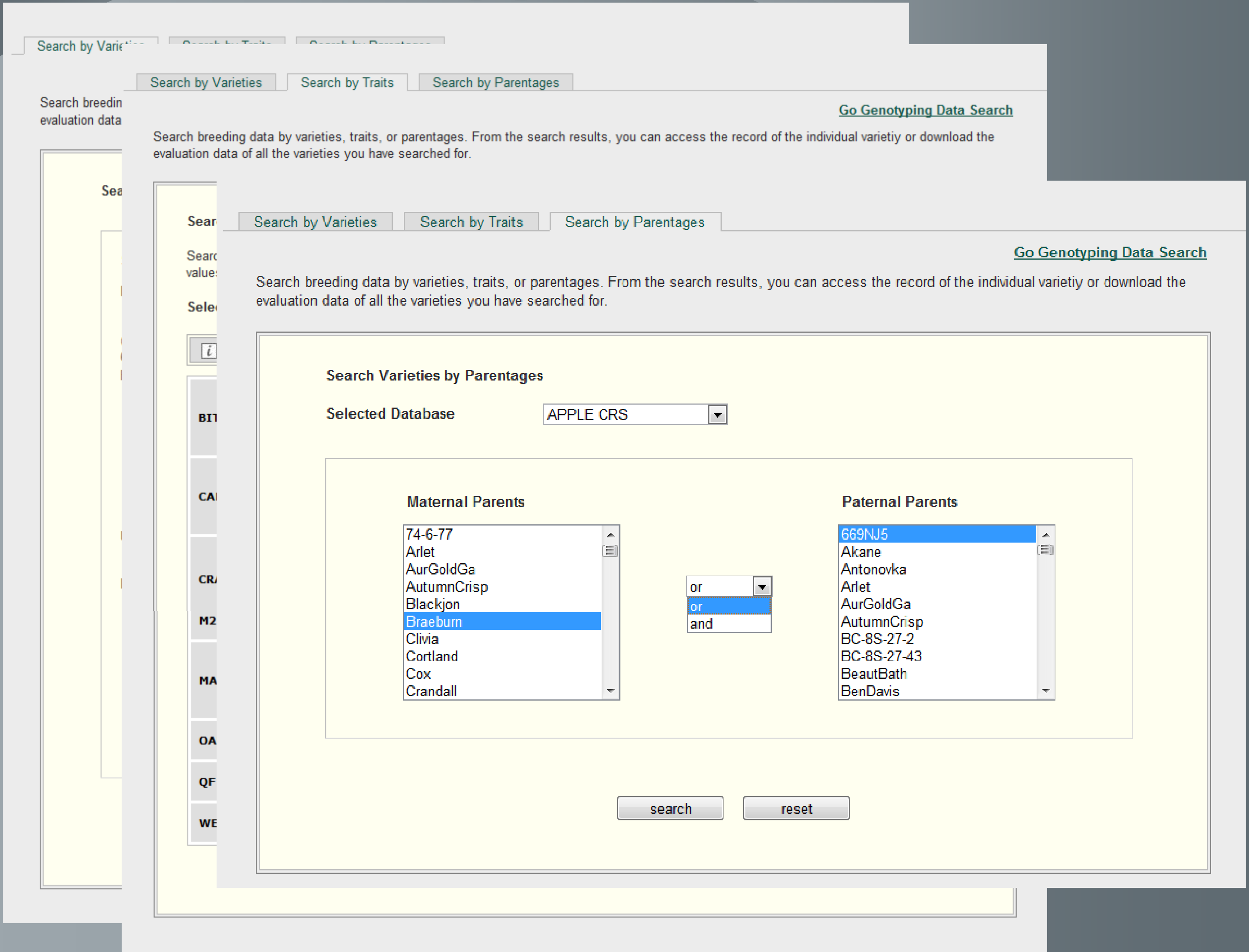
Future Development

- Release custom modules developed for Main Lab databases
- Tripal API for software that we have developed
  - Breeders toolbox
  - GenSAS, the community genome annotation tool
- Tripal API for a tablet app to collect phenotypic data from the field
- Web Services to allow programmatic access to data.
- Data management: online data submission and editing
- Help Desk Support

Breeders Toolbox Demo

- Data Management (Browse, Search and Download)
- Data Conversion (Generate Input files for Pedimap)
- Decision Support
  - Trait Locus Warehouse (QTL search page)
  - Marker Converter
  - Technology Portfolio
  - Cross Assist
  - Seedling Select

Phenotypic Data Search



Search breeding data by varieties, traits, or parentages. From the search results, you can access the record of the individual varietiy or download the evaluation data of all the varieties you have searched for.

Search breeding data by varieties, traits, or parentages. From the search results, you can access the record of the individual varietiy or download the evaluation data of all the varieties you have searched for.

Search Varieties by Parentages

Selected Database APPLE CRS

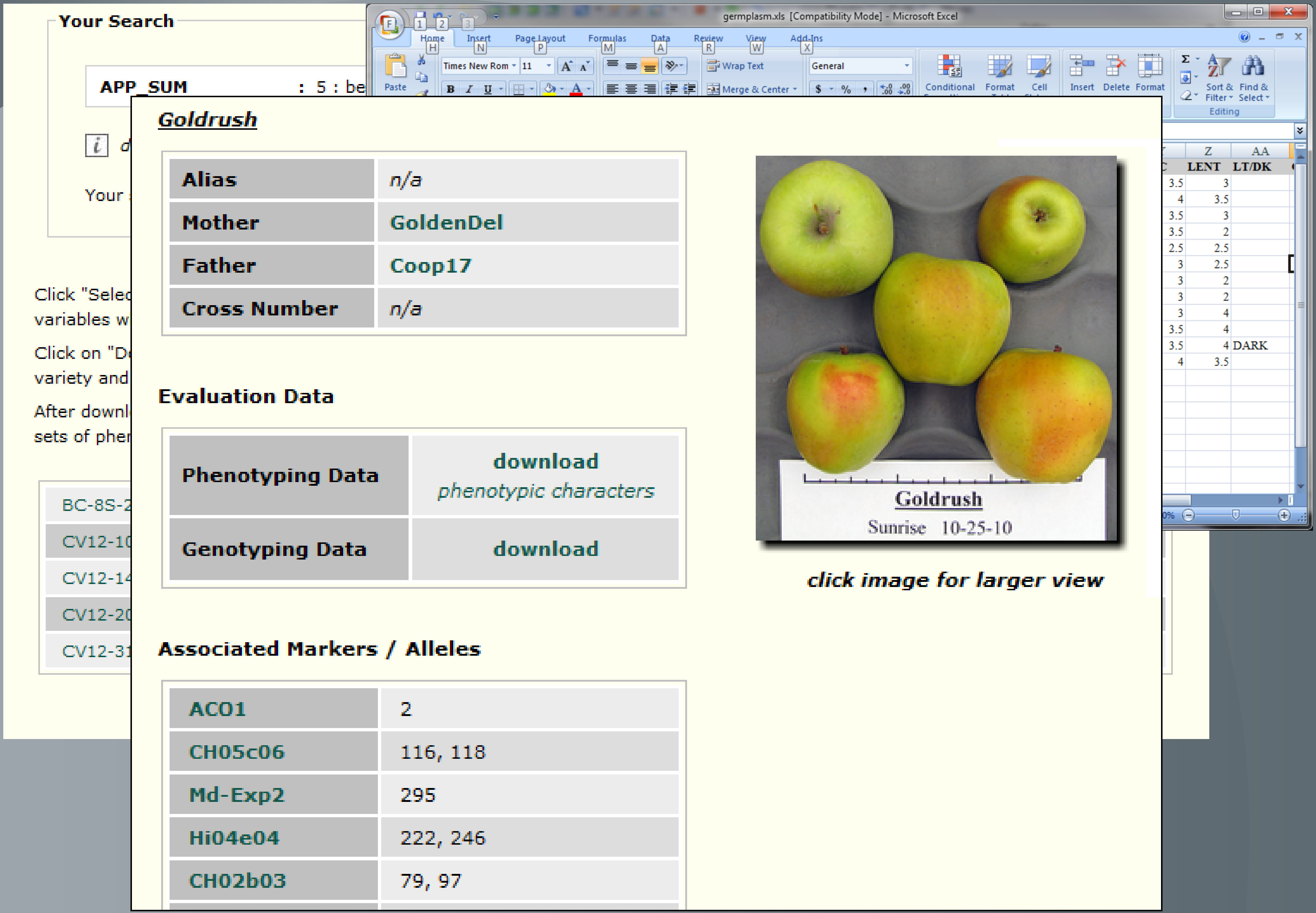
Maternal Parents

- 74-6-77
- Arlet
- AurGoldGa
- AutumnCrisp
- Blackjon
- Braeburn
- Clivia
- Cortland
- Cox
- Crandall

or
- or
- and

Paternal Parents

- 669NJ5
- Akane
- Antonovka
- Arlet
- AurGoldGa
- AutumnCrisp
- BC-8S-27-2
- BC-8S-27-43
- BeautBath
- BenDavis



Your Search

APP_SUM : 5 : be

Your

Click "Selec
variables w

Click on "D
variety and

After downl
sets of phe

BC-8S-2
CV12-10
CV12-14
CV12-20
CV12-31

germplasm.xls [Compatibility Mode] - Microsoft Excel

	Z LENT	AA LT/DK
3.5	3	
4	3.5	
3.5	3	
3.5	2	
2.5	2.5	
3	2.5	
3	2	
3	2	
3	4	
3.5	4	
3.5	4	DARK
4	3.5	

Goldrush

Alias	*n/a*
Mother	GoldenDel
Father	Coop17
Cross Number	*n/a*

Goldrush
Sunrise 10-25-10

click image for larger view

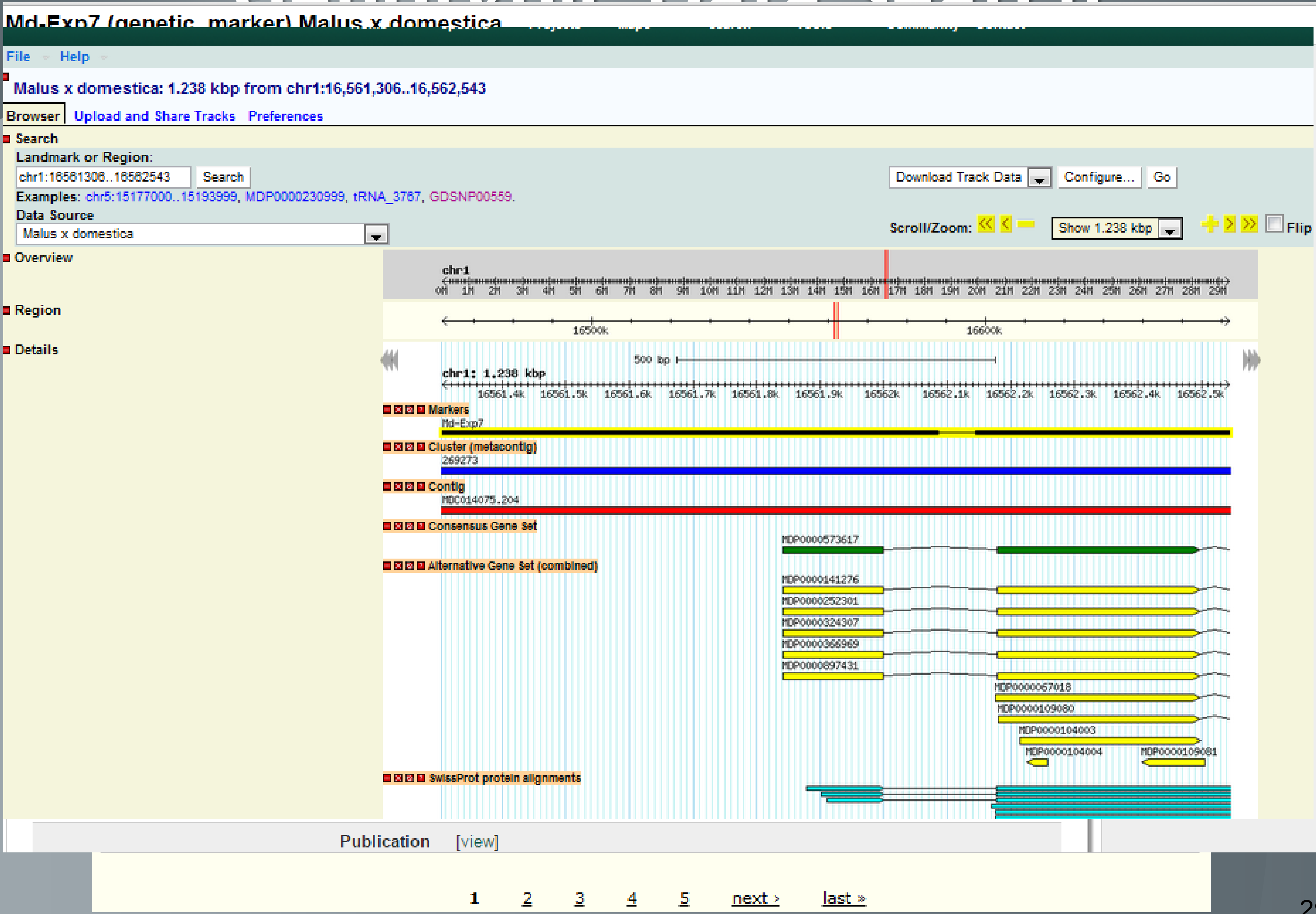
Evaluation Data

Phenotyping Data	download *phenotypic characters*
Genotyping Data	download

Associated Markers / Alleles

ACO1	2
CH05c06	116, 118
Md-Exp2	295
Hi04e04	222, 246
CH02b03	79, 97

Genotypic Data Search



Md-Exp7 (genetic marker) Malus x domestica

Malus x domestica: 1.238 kbp from chr1:16,561,306..16,562,543

Landmark or Region: chr1:16561306..16562543

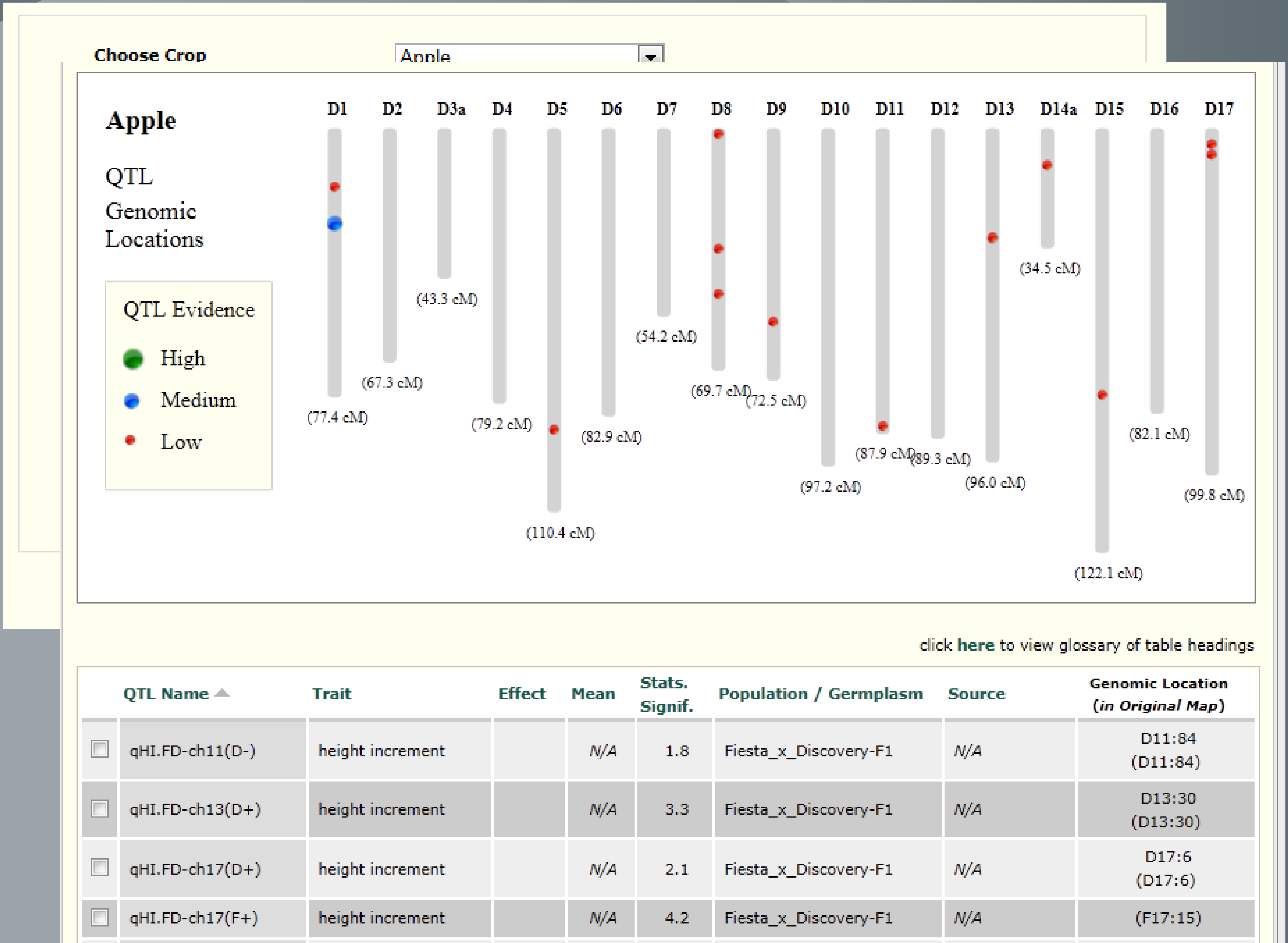
Examples: chr5:15177000..15193999, MDP0000230999, tRNA_3767, GDSNP00559.

Data Source: Malus x domestica

Scroll/Zoom: Show 1.238 kbp

Publication [view]

Trait Locus Warehouse



click **here** to view glossary of table headings

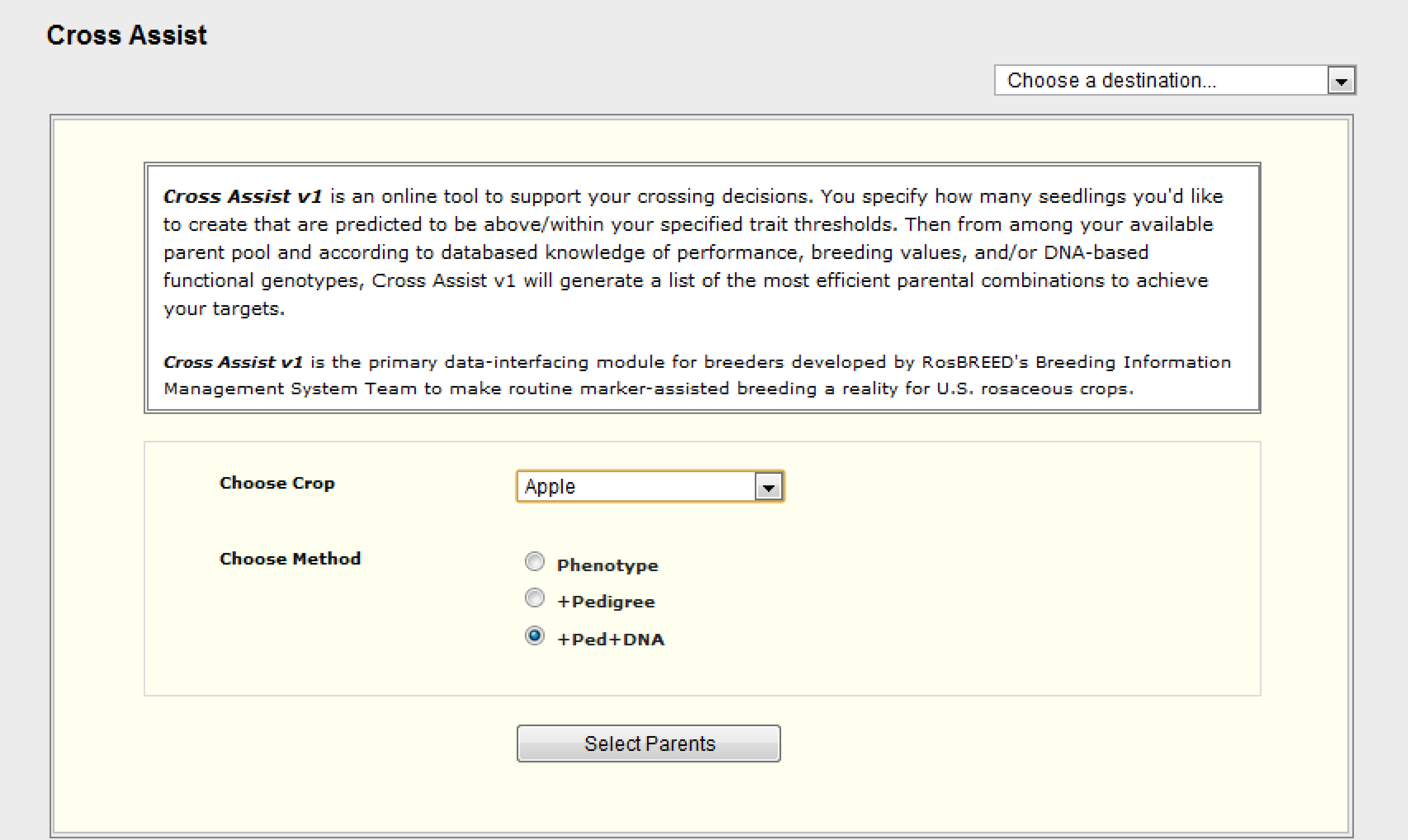
	QTL Name	Trait	Effect	Mean	Stats. Signif.	Population / Germplasm	Source	Genomic Location (*in Original Map*)
☐	qHI.FD-ch11(D-)	height increment		*N/A*	1.8	Fiesta_x_Discovery-F1	*N/A*	D11:84 (D11:84)
☐	qHI.FD-ch13(D+)	height increment		*N/A*	3.3	Fiesta_x_Discovery-F1	*N/A*	D13:30 (D13:30)
☐	qHI.FD-ch17(D+)	height increment		*N/A*	2.1	Fiesta_x_Discovery-F1	*N/A*	D17:6 (D17:6)
☐	qHI.FD-ch17(F+)	height increment		*N/A*	4.2	Fiesta_x_Discovery-F1	*N/A*	(F17:15)

Cross Assist

- A web interface to generate a list of parents and the number of seedlings to get the progeny with desired traits

- Methods
  - “*Phenotype*” (uses only phenotypic information of individuals in the dataset),
  - “*+Pedigree*” (uses both phenotypic and pedigree information)
  - “*+Ped+DNA*” (uses phenotypic, pedigree information and information provided by DNA-based functional genotypes).

Step 1: Select Method



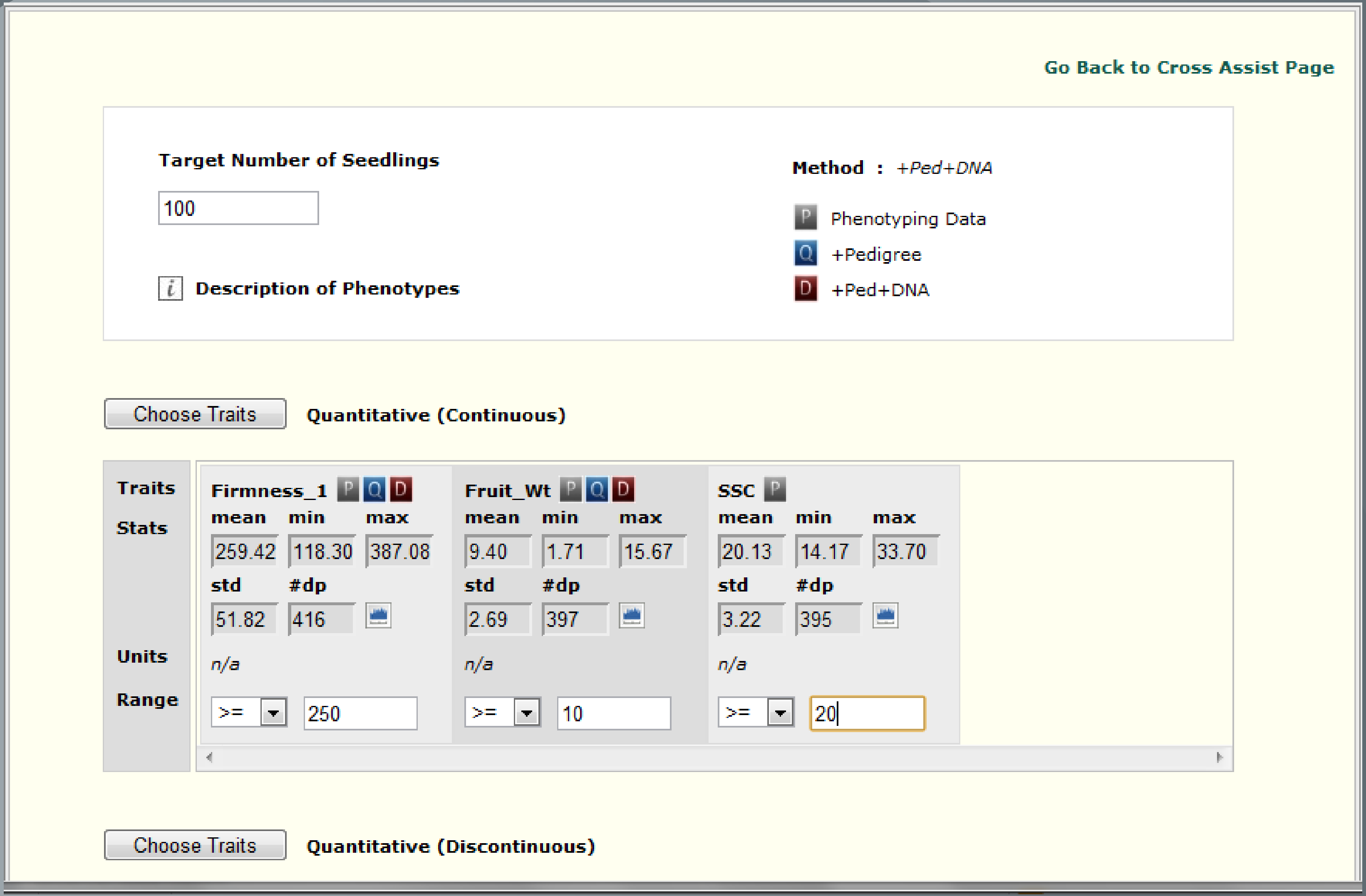
Cross Assist v1 is an online tool to support your crossing decisions. You specify how many seedlings you'd like to create that are predicted to be above/within your specified trait thresholds. Then from among your available parent pool and according to databased knowledge of performance, breeding values, and/or DNA-based functional genotypes, Cross Assist v1 will generate a list of the most efficient parental combinations to achieve your targets.

Cross Assist v1 is the primary data-interfacing module for breeders developed by RosBREED's Breeding Information Management System Team to make routine marker-assisted breeding a reality for U.S. rosaceous crops.

Choose Crop: Apple

Choose Method
- ○ **Phenotype**
- ○ **+Pedigree**
- ● **+Ped+DNA**

Step 2: Select target number and trait thresholds



Target Number of Seedlings

100

Description of Phenotypes

Method : *+Ped+DNA*

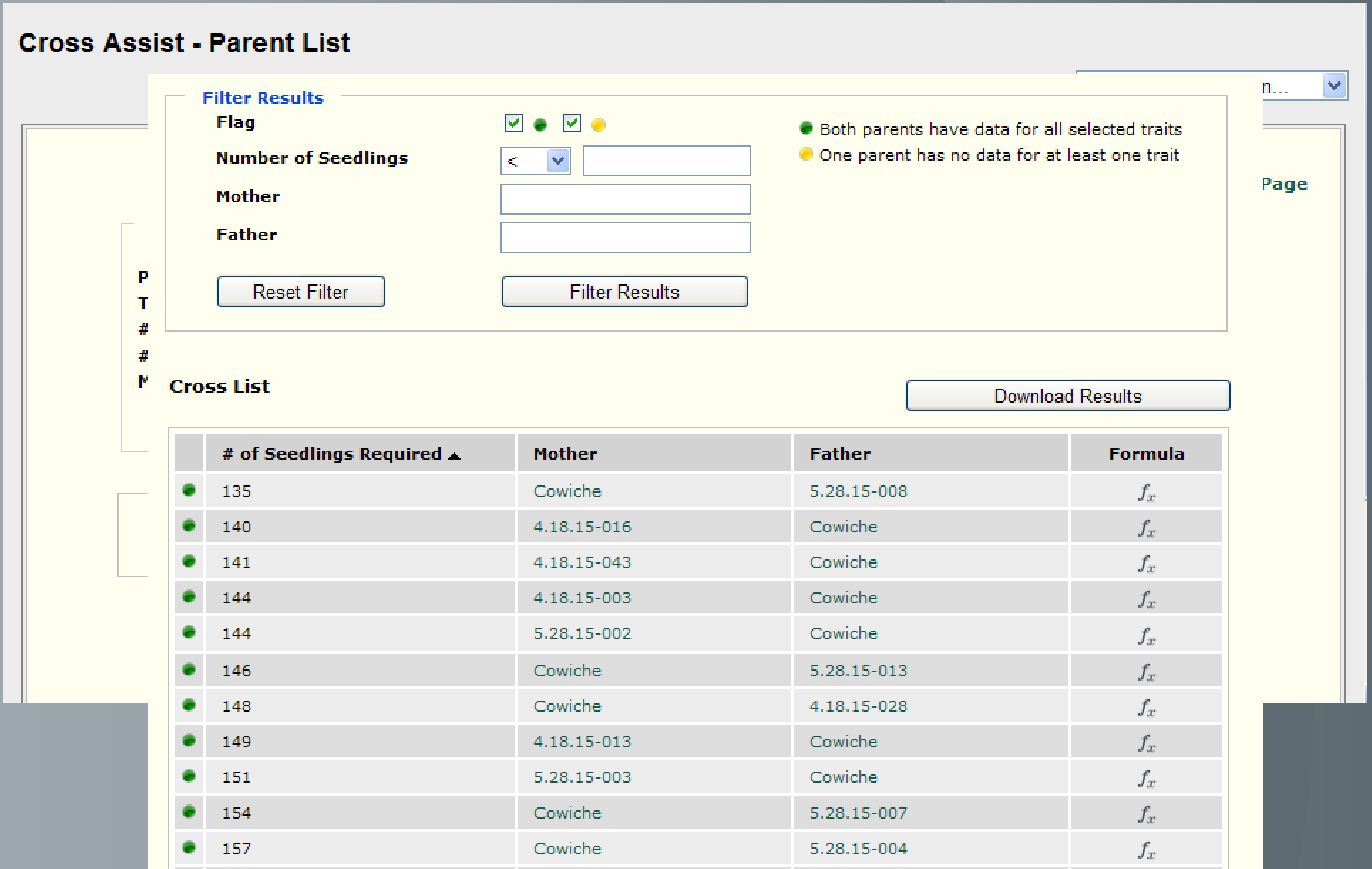
P Phenotyping Data
Q +Pedigree
D +Ped+DNA

Quantitative (Continuous)

Traits	**Firmness_1** P Q D			**Fruit_Wt** P Q D			**SSC** P		
Stats	**mean**	**min**	**max**	**mean**	**min**	**max**	**mean**	**min**	**max**
	259.42	118.30	387.08	9.40	1.71	15.67	20.13	14.17	33.70
	std	**#dp**		**std**	**#dp**		**std**	**#dp**	
	51.82	416		2.69	397		3.22	395	
Units	*n/a*			*n/a*			*n/a*		
Range	>=	250		>=	10		>=	20	

Quantitative (Discontinuous)

Step 3: Filter results by data completeness, required number of seedlings, and parentage



Cross Assist - Parent List

Filter Results

Flag ☑ 🟢 ☑ 🟡

Number of Seedlings < ☐

Mother ☐

Father ☐

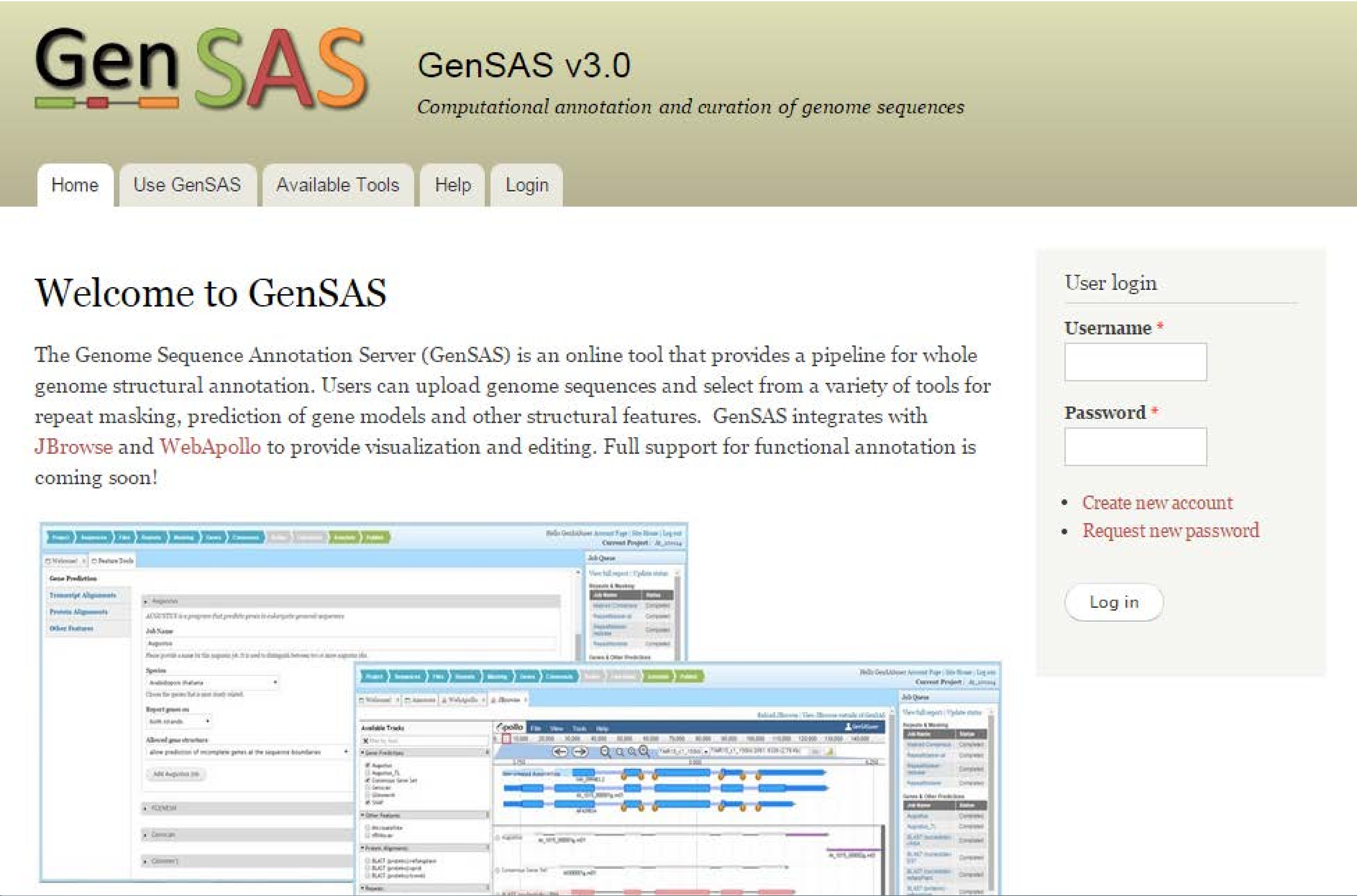
🟢 Both parents have data for all selected traits

🟡 One parent has no data for at least one trait

Cross List

	# of Seedlings Required ▲	Mother	Father	Formula
🟢	135	Cowiche	5.28.15-008	f_x
🟢	140	4.18.15-016	Cowiche	f_x
🟢	141	4.18.15-043	Cowiche	f_x
🟢	144	4.18.15-003	Cowiche	f_x
🟢	144	5.28.15-002	Cowiche	f_x
🟢	146	Cowiche	5.28.15-013	f_x
🟢	148	Cowiche	4.18.15-028	f_x
🟢	149	4.18.15-013	Cowiche	f_x
🟢	151	5.28.15-003	Cowiche	f_x
🟢	154	Cowiche	5.28.15-007	f_x
🟢	157	Cowiche	5.28.15-004	f_x

http://gensas.bioinfo.wsu.edu/



GenSAS v3.0

Computational annotation and curation of genome sequences

Welcome to GenSAS

The Genome Sequence Annotation Server (GenSAS) is an online tool that provides a pipeline for whole genome structural annotation. Users can upload genome sequences and select from a variety of tools for repeat masking, prediction of gene models and other structural features. GenSAS integrates with JBrowse and WebApollo to provide visualization and editing. Full support for functional annotation is coming soon!

User login

Username *

Password *

- Create new account
- Request new password

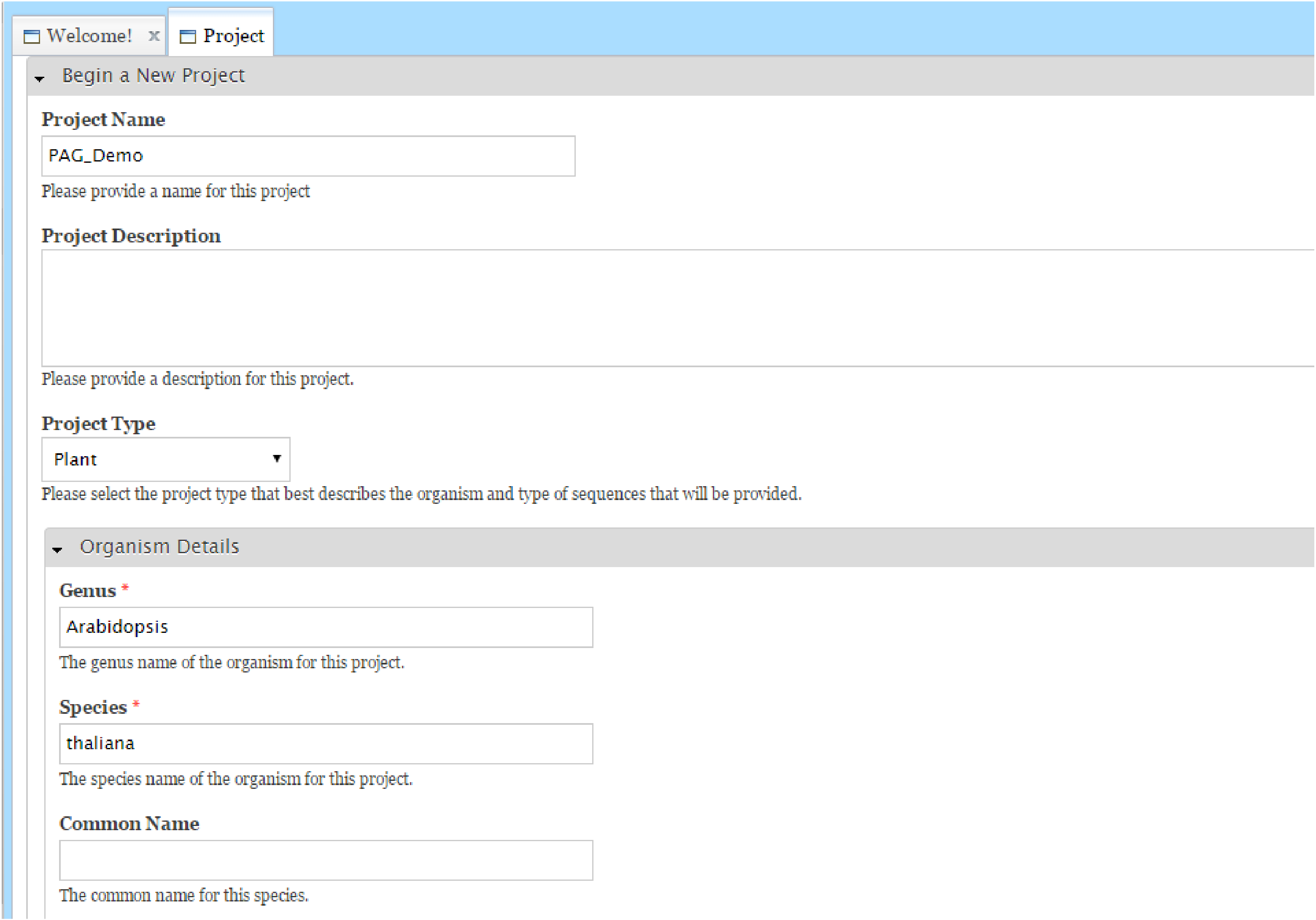
Functionality Overview

- Repeat finding & masking:
- Intrinsic and extrinsic gene Prediction
  - Augustus, FgeneSH (parsing only), Genscan, Glimmer3, GlimmerM, SNAP, tRNAScan, getorf
  - BLAT, BLAST with available and user uploaded transcript and protein databases

Functionality Overview

- Gene Consensus Prediction
    - EvidenceModeler
- Gene Visualization and Curation
    - WebApollo, JBrowse
- Publish
    - Generates GFF3 & FASTA files with properly versioned naming.

Step 1: Start a project



Project Name

PAG_Demo

Please provide a name for this project

Project Description

Please provide a description for this project.

Project Type

Plant

Please select the project type that best describes the organism and type of sequences that will be provided.

Genus *

Arabidopsis

The genus name of the organism for this project.

Species *

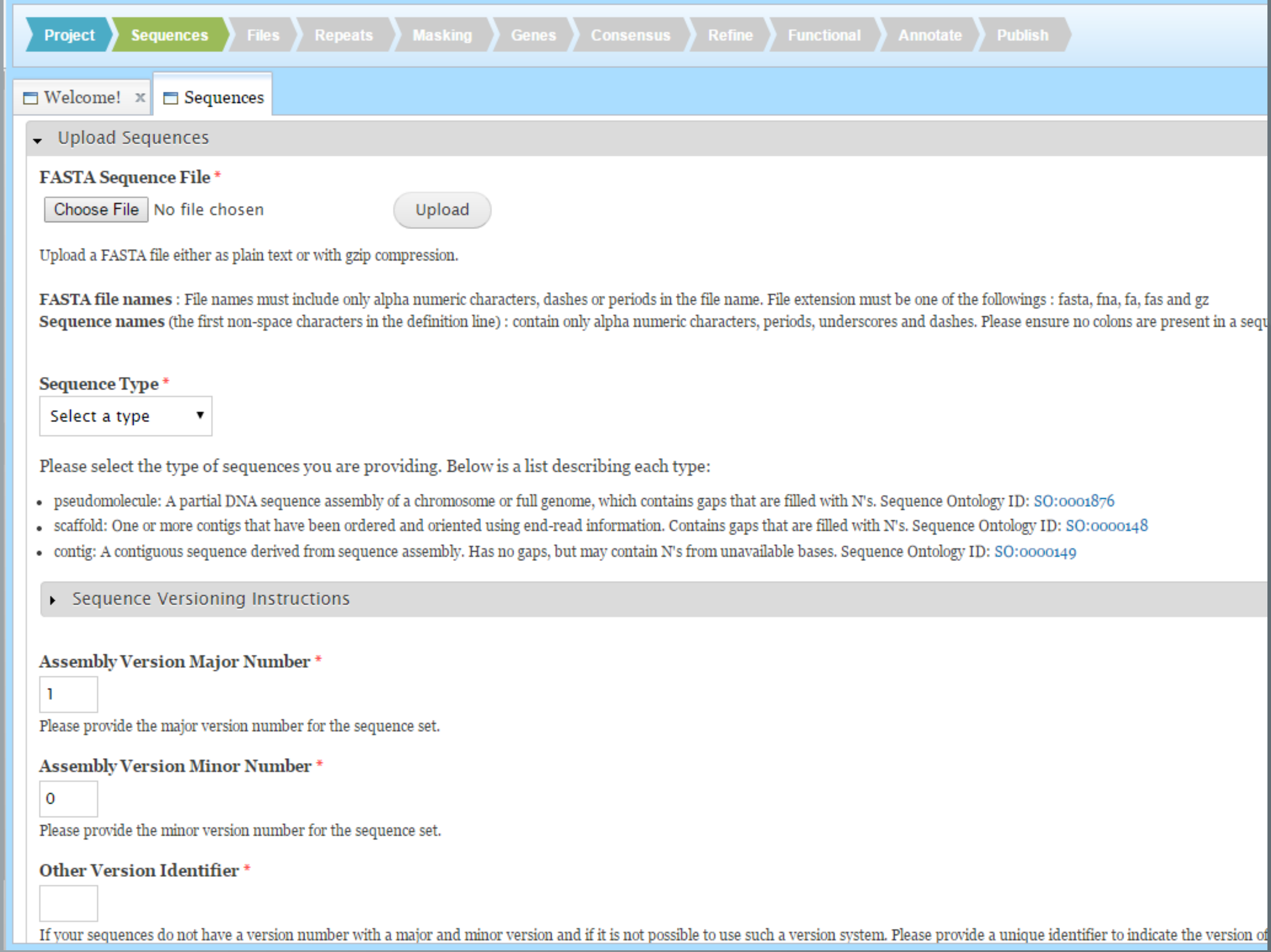
thaliana

The species name of the organism for this project.

Common Name

The common name for this species.

Step 2: Upload Sequences



Upload Sequences

FASTA Sequence File *

No file chosen

Upload a FASTA file either as plain text or with gzip compression.

FASTA file names : File names must include only alpha numeric characters, dashes or periods in the file name. File extension must be one of the followings : fasta, fna, fa, fas and gz
Sequence names (the first non-space characters in the definition line) : contain only alpha numeric characters, periods, underscores and dashes. Please ensure no colons are present in a sequ

Sequence Type *

Select a type ▾

Please select the type of sequences you are providing. Below is a list describing each type:

- pseudomolecule: A partial DNA sequence assembly of a chromosome or full genome, which contains gaps that are filled with N's. Sequence Ontology ID: SO:0001876
- scaffold: One or more contigs that have been ordered and oriented using end-read information. Contains gaps that are filled with N's. Sequence Ontology ID: SO:0000148
- contig: A contiguous sequence derived from sequence assembly. Has no gaps, but may contain N's from unavailable bases. Sequence Ontology ID: SO:0000149

▸ Sequence Versioning Instructions

Assembly Version Major Number *

1

Please provide the major version number for the sequence set.

Assembly Version Minor Number *

0

Please provide the minor version number for the sequence set.

Other Version Identifier *

If your sequences do not have a version number with a major and minor version and if it is not possible to use such a version system. Please provide a unique identifier to indicate the version of

Step 3: Upload Supporting Files

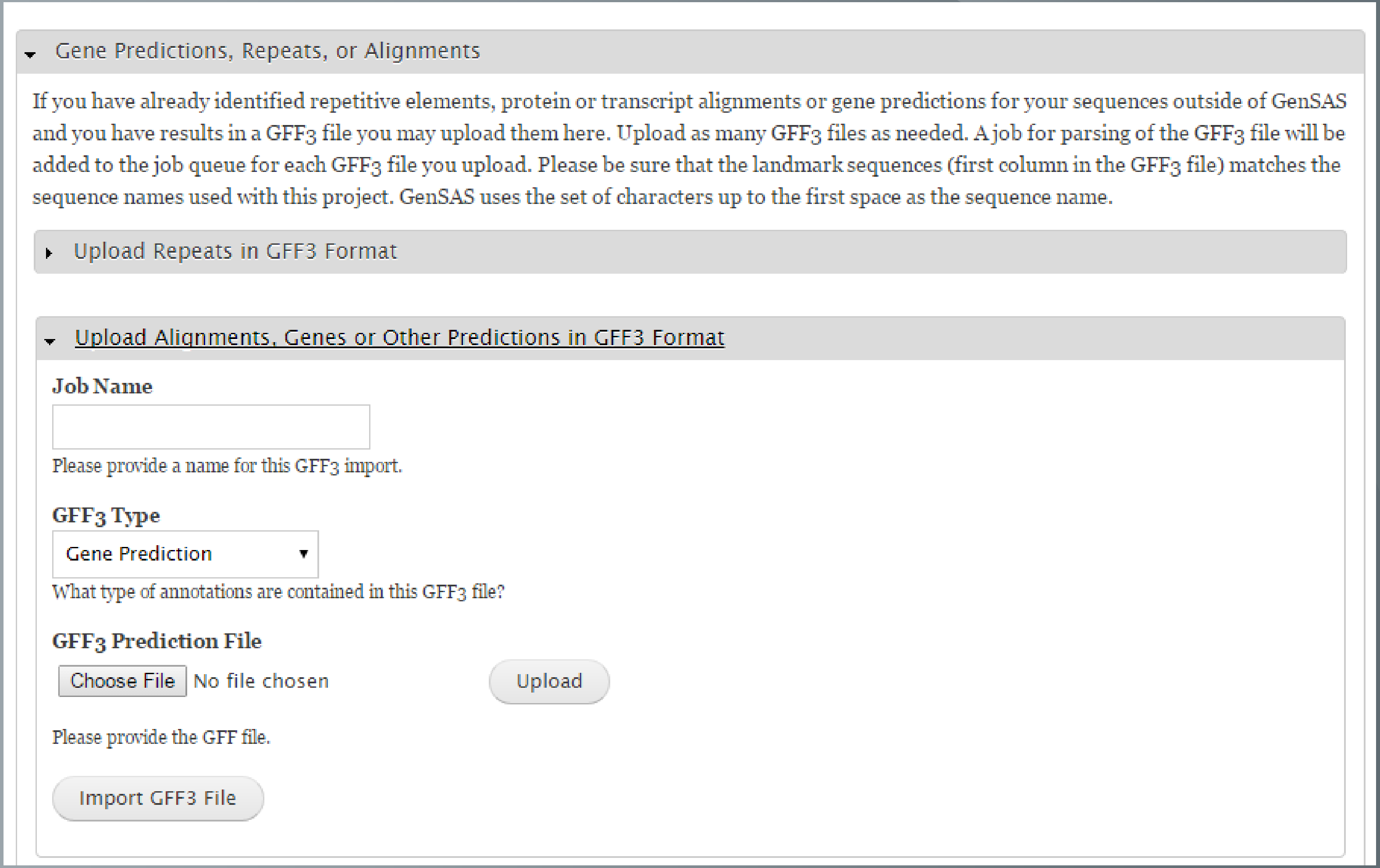
Evidence Files

One of the best ways to identify genes is to provide transcript (e.g. ESTs, full length cDNA, RNA-seq) and protein files from the species or from closely related species. These files should be provided in FASTA or FASTQ format, and when aligned to the genomic sequence serve as direct evidence of gene expression. Transcripts and proteins of closely related species can be provided when species-specific files are not available or are insufficient.
Supplying a species-specific (or closely-related species) file of known repeats, in FASTA format, can help with identification of repetitve elements more accurate for the species.

Uploaded Evidence Files

File	Type	Size	Action
There are currently no files uploaded.			

Step 3: Upload Supporting Files



If you have already identified repetitive elements, protein or transcript alignments or gene predictions for your sequences outside of GenSAS and you have results in a GFF3 file you may upload them here. Upload as many GFF3 files as needed. A job for parsing of the GFF3 file will be added to the job queue for each GFF3 file you upload. Please be sure that the landmark sequences (first column in the GFF3 file) matches the sequence names used with this project. GenSAS uses the set of characters up to the first space as the sequence name.

Job Name

Please provide a name for this GFF3 import.

GFF3 Type

Gene Prediction ▾

What type of annotations are contained in this GFF3 file?

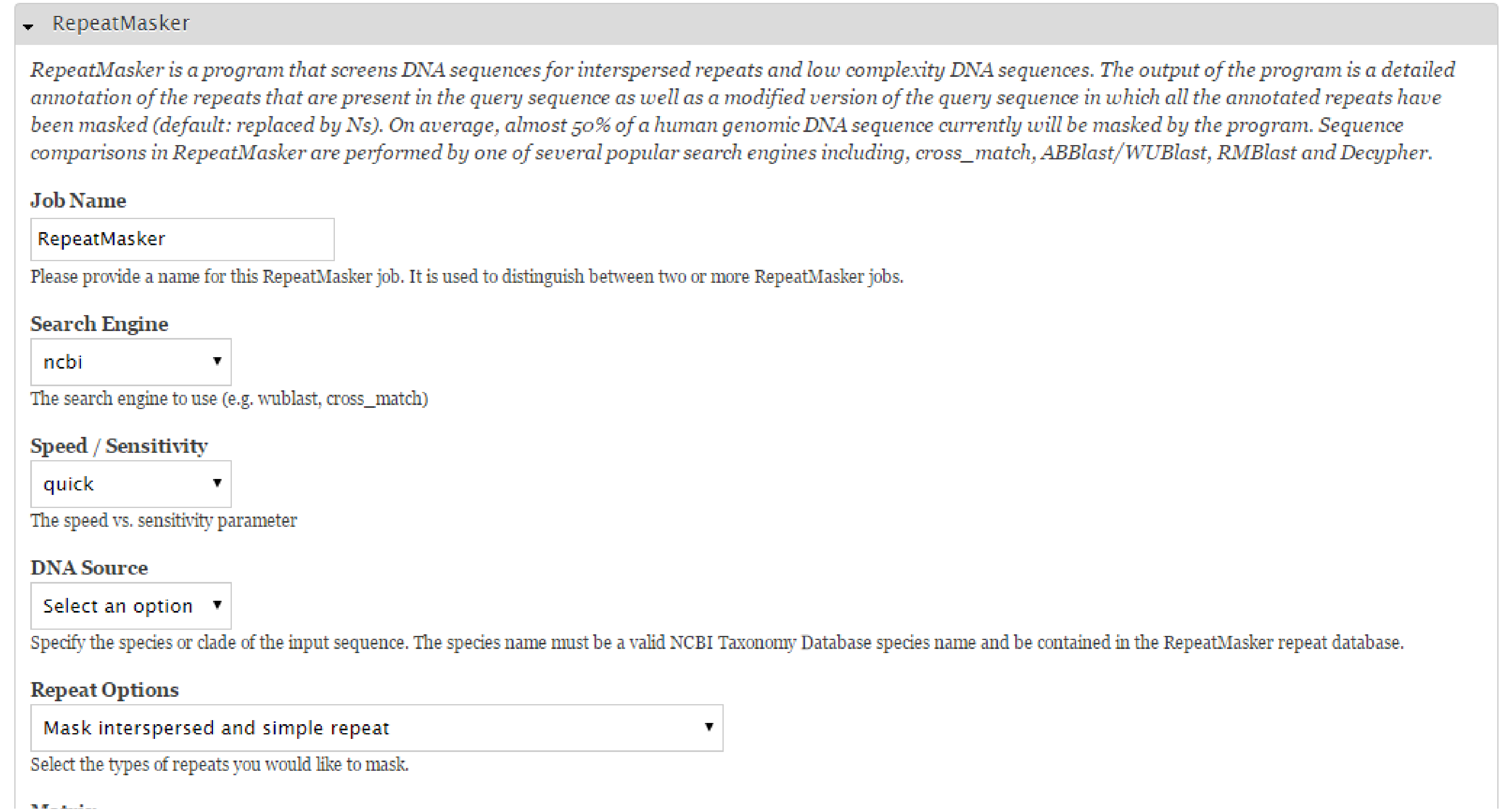
GFF3 Prediction File

No file chosen

Please provide the GFF file.

Current *Pinus taeda* Maker predictions loaded in this way

Step 4: Repeat Finding



▾ RepeatMasker

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (default: replaced by Ns). On average, almost 50% of a human genomic DNA sequence currently will be masked by the program. Sequence comparisons in RepeatMasker are performed by one of several popular search engines including, cross_match, ABBlast/WUBlast, RMBlast and Decypher.

Job Name

RepeatMasker

Please provide a name for this RepeatMasker job. It is used to distinguish between two or more RepeatMasker jobs.

Search Engine

ncbi ▾

The search engine to use (e.g. wublast, cross_match)

Speed / Sensitivity

quick ▾

The speed vs. sensitivity parameter

DNA Source

Select an option ▾

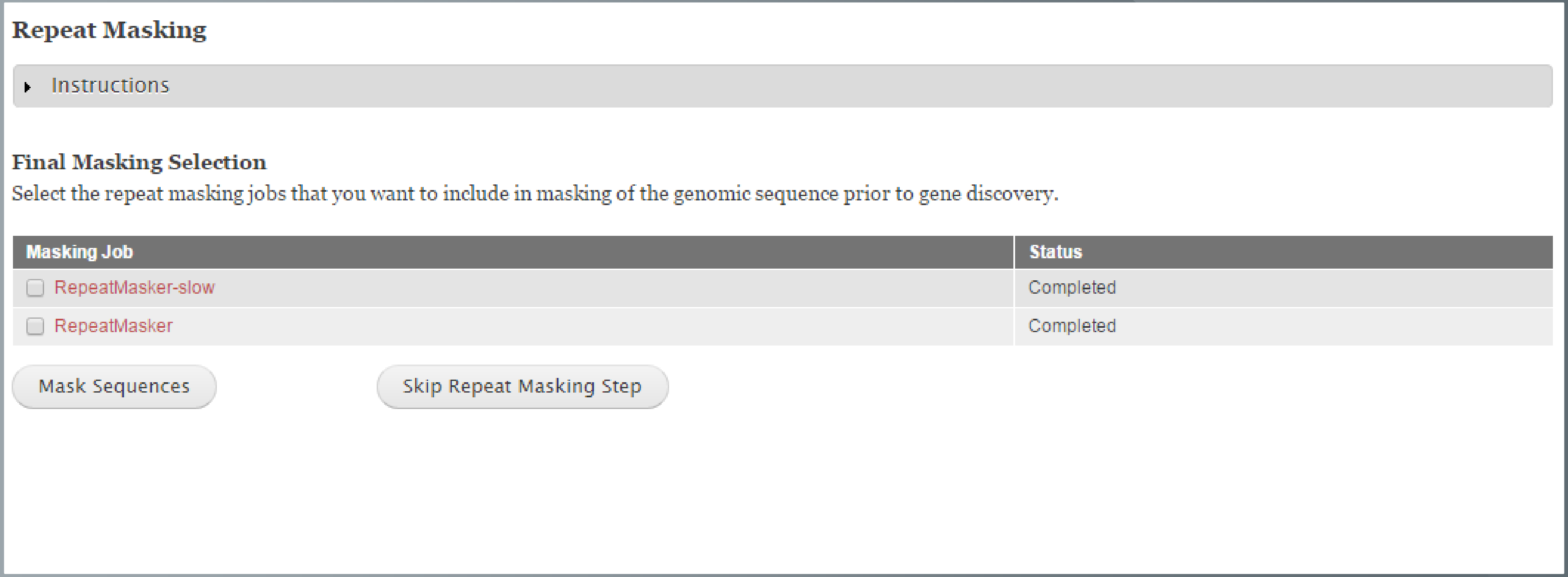
Specify the species or clade of the input sequence. The species name must be a valid NCBI Taxonomy Database species name and be contained in the RepeatMasker repeat database.

Repeat Options

Mask interspersed and simple repeat ▾

Select the types of repeats you would like to mask.

Step 5: Masking



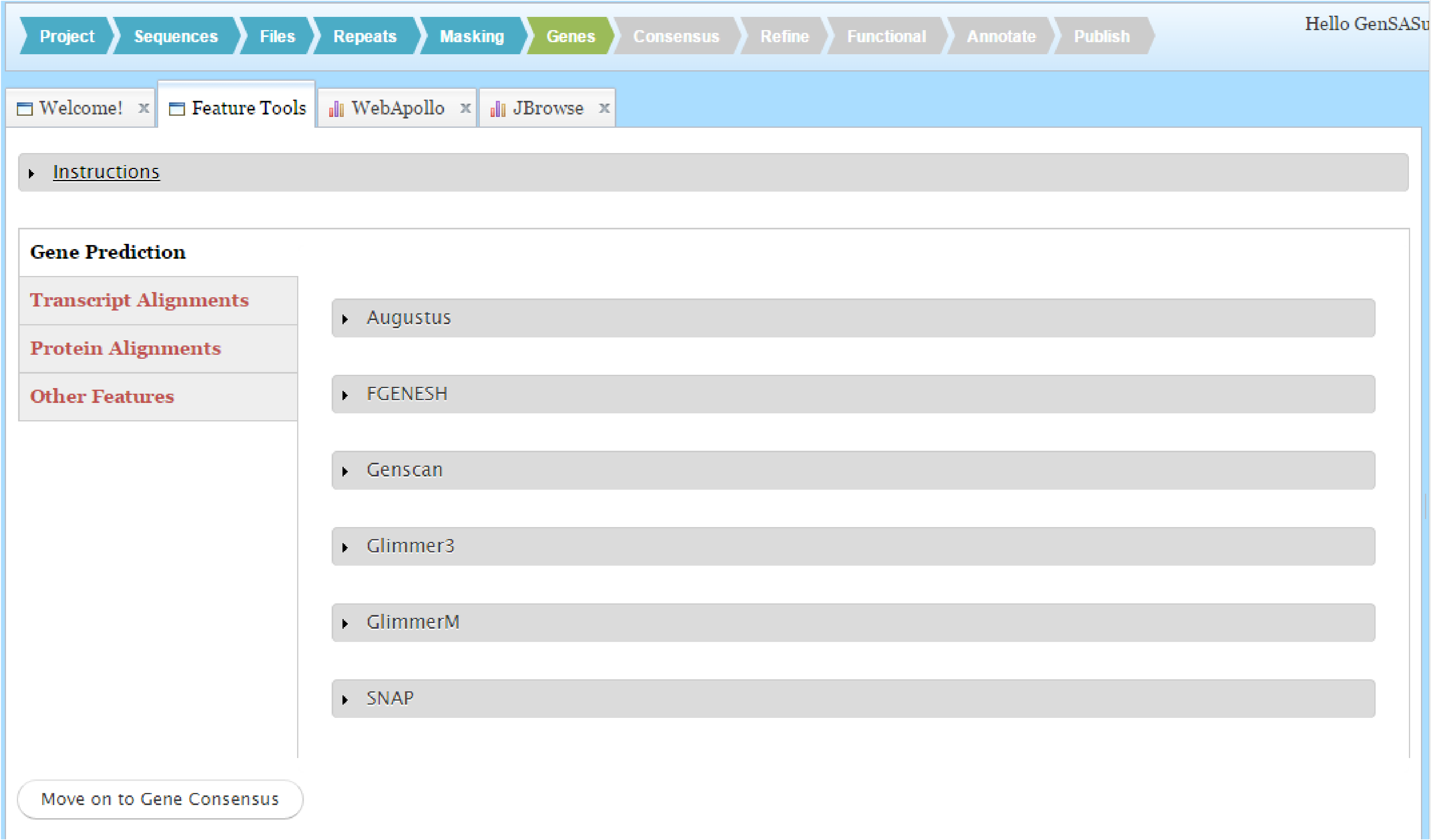
Repeat Masking

Final Masking Selection

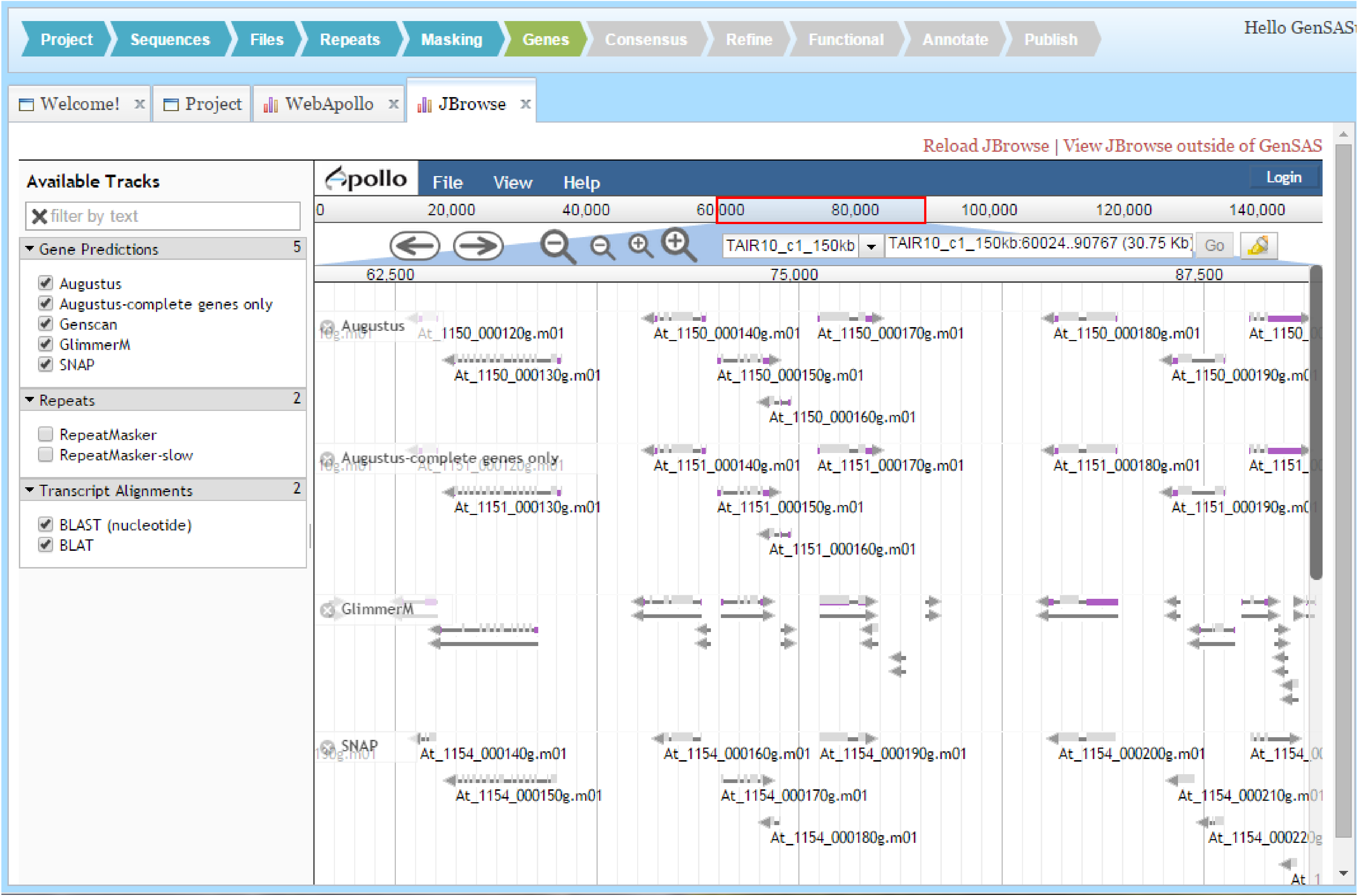
Select the repeat masking jobs that you want to include in masking of the genomic sequence prior to gene discovery.

Masking Job	Status
☐ RepeatMasker-slow	Completed
☐ RepeatMasker	Completed

Step 6: Gene Prediction



View Results



Available Tracks

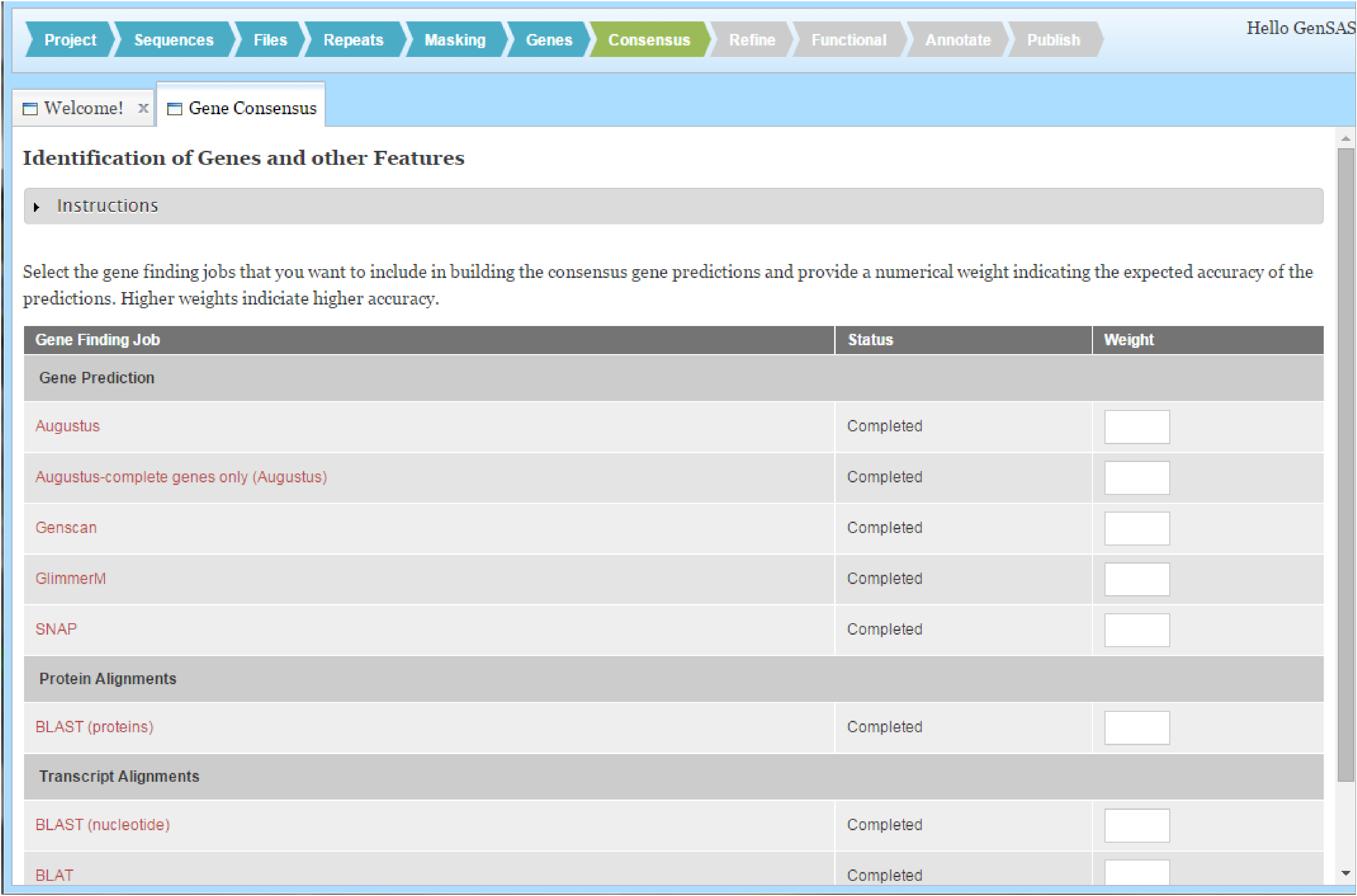
filter by text

- Gene Predictions 5
  - ☑ Augustus
  - ☑ Augustus-complete genes only
  - ☑ Genscan
  - ☑ GlimmerM
  - ☑ SNAP
- Repeats 2
  - ☐ RepeatMasker
  - ☐ RepeatMasker-slow
- Transcript Alignments 2
  - ☑ BLAST (nucleotide)
  - ☑ BLAT

TAIR10_c1_150kb

TAIR10_c1_150kb:60024..90767 (30.75 Kb)

Step 7: Consensus



Identification of Genes and other Features

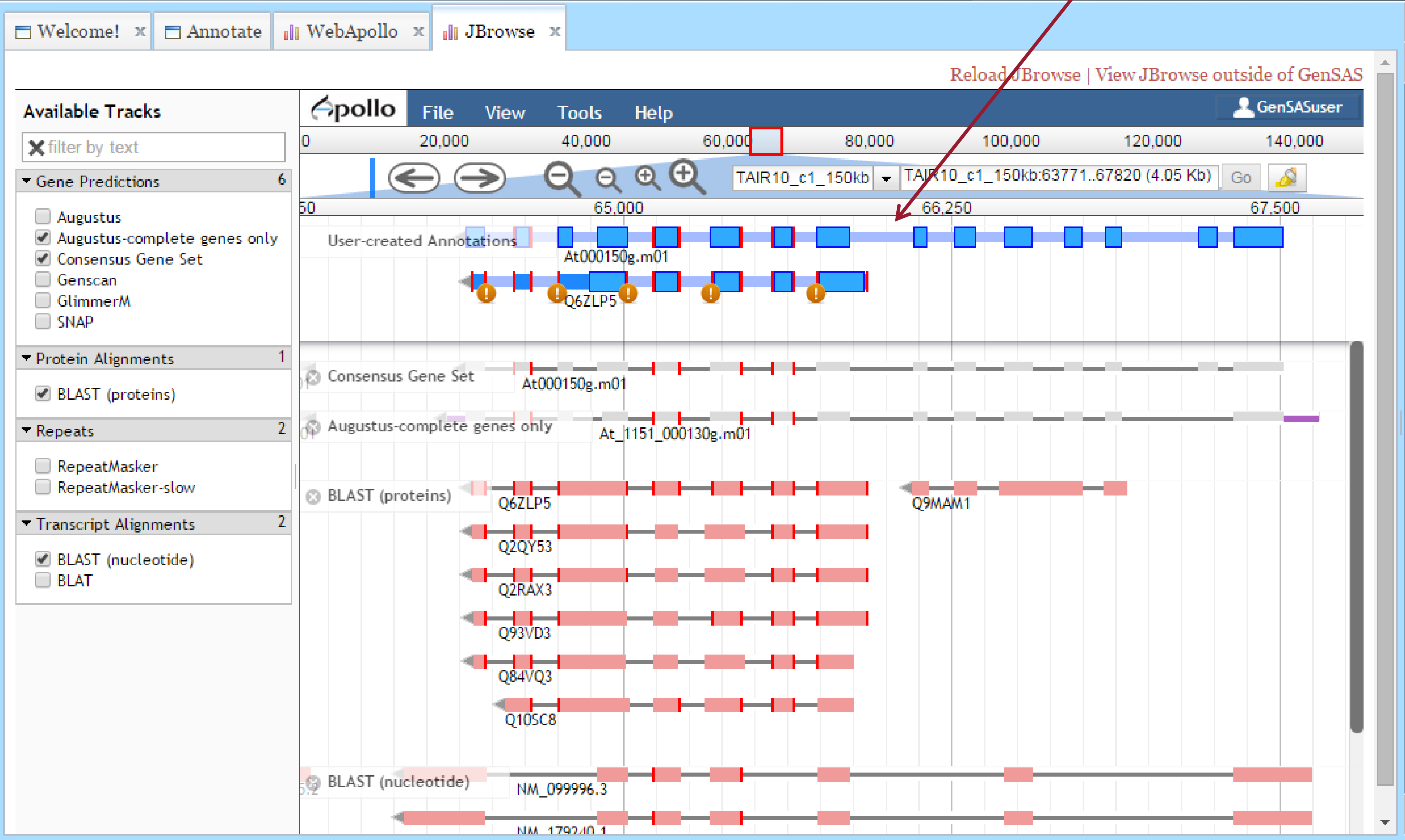
▸ Instructions

Select the gene finding jobs that you want to include in building the consensus gene predictions and provide a numerical weight indicating the expected accuracy of the predictions. Higher weights indiciate higher accuracy.

Gene Finding Job	Status	Weight
Gene Prediction		
Augustus	Completed	
Augustus-complete genes only (Augustus)	Completed	
Genscan	Completed	
GlimmerM	Completed	
SNAP	Completed	
Protein Alignments		
BLAST (proteins)	Completed	
Transcript Alignments		
BLAST (nucleotide)	Completed	
BLAT	Completed	

Step 8: Manual Curation

Curation Track



Available Tracks

filter by text

Gene Predictions 6

- ☐ Augustus
- ☑ Augustus-complete genes only
- ☑ Consensus Gene Set
- ☐ Genscan
- ☐ GlimmerM
- ☐ SNAP

Protein Alignments 1

- ☑ BLAST (proteins)

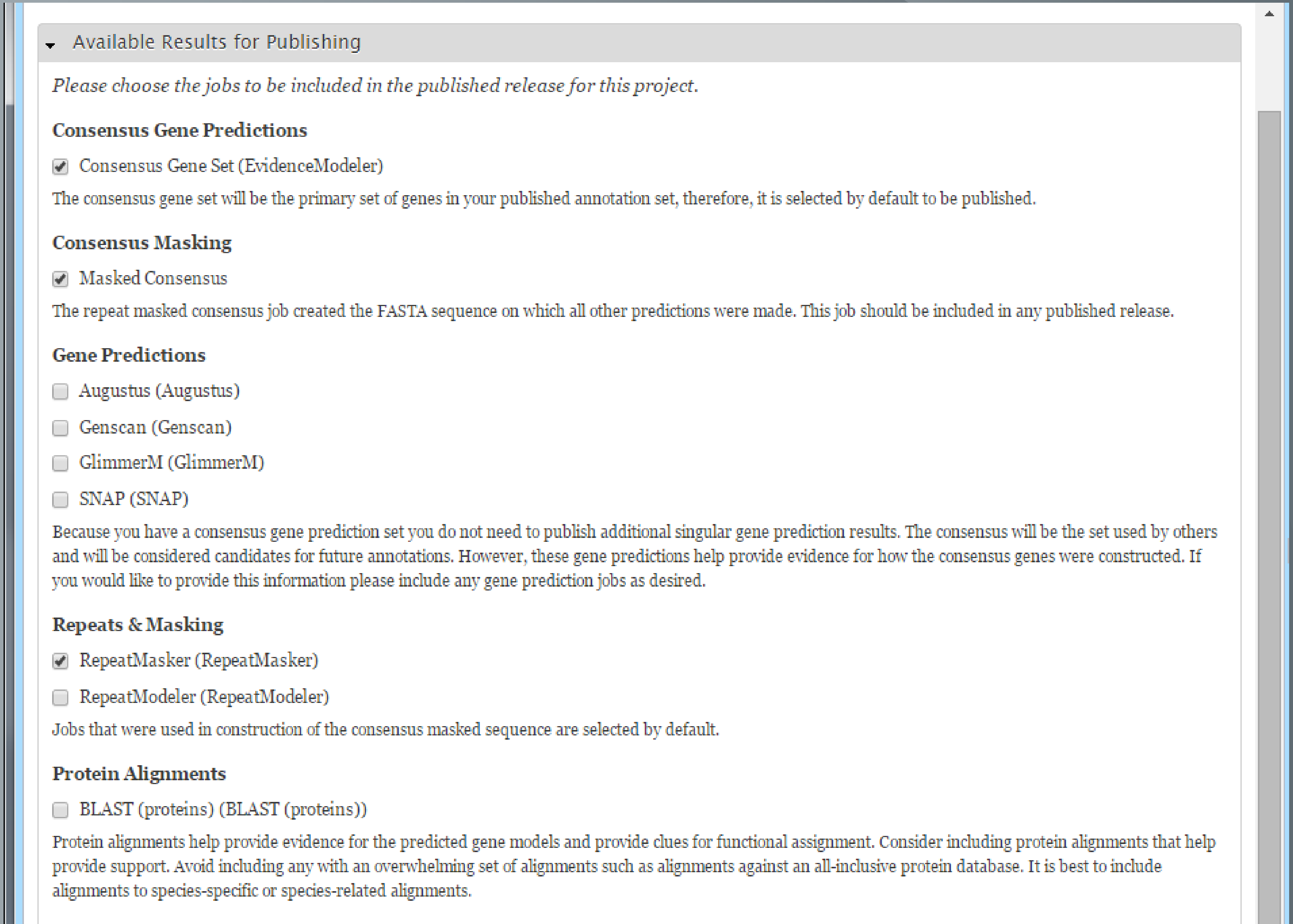
Repeats 2

- ☐ RepeatMasker
- ☐ RepeatMasker-slow

Transcript Alignments 2

- ☑ BLAST (nucleotide)
- ☐ BLAT

Step 9: Publish



Available Results for Publishing

Please choose the jobs to be included in the published release for this project.

Consensus Gene Predictions

☑ Consensus Gene Set (EvidenceModeler)

The consensus gene set will be the primary set of genes in your published annotation set, therefore, it is selected by default to be published.

Consensus Masking

☑ Masked Consensus

The repeat masked consensus job created the FASTA sequence on which all other predictions were made. This job should be included in any published release.

Gene Predictions

☐ Augustus (Augustus)

☐ Genscan (Genscan)

☐ GlimmerM (GlimmerM)

☐ SNAP (SNAP)

Because you have a consensus gene prediction set you do not need to publish additional singular gene prediction results. The consensus will be the set used by others and will be considered candidates for future annotations. However, these gene predictions help provide evidence for how the consensus genes were constructed. If you would like to provide this information please include any gene prediction jobs as desired.

Repeats & Masking

☑ RepeatMasker (RepeatMasker)

☐ RepeatModeler (RepeatModeler)

Jobs that were used in construction of the consensus masked sequence are selected by default.

Protein Alignments

☐ BLAST (proteins) (BLAST (proteins))

Protein alignments help provide evidence for the predicted gene models and provide clues for functional assignment. Consider including protein alignments that help provide support. Avoid including any with an overwhelming set of alignments such as alignments against an all-inclusive protein database. It is best to include alignments to species-specific or species-related alignments.

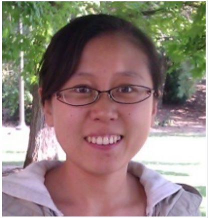
The Team

8

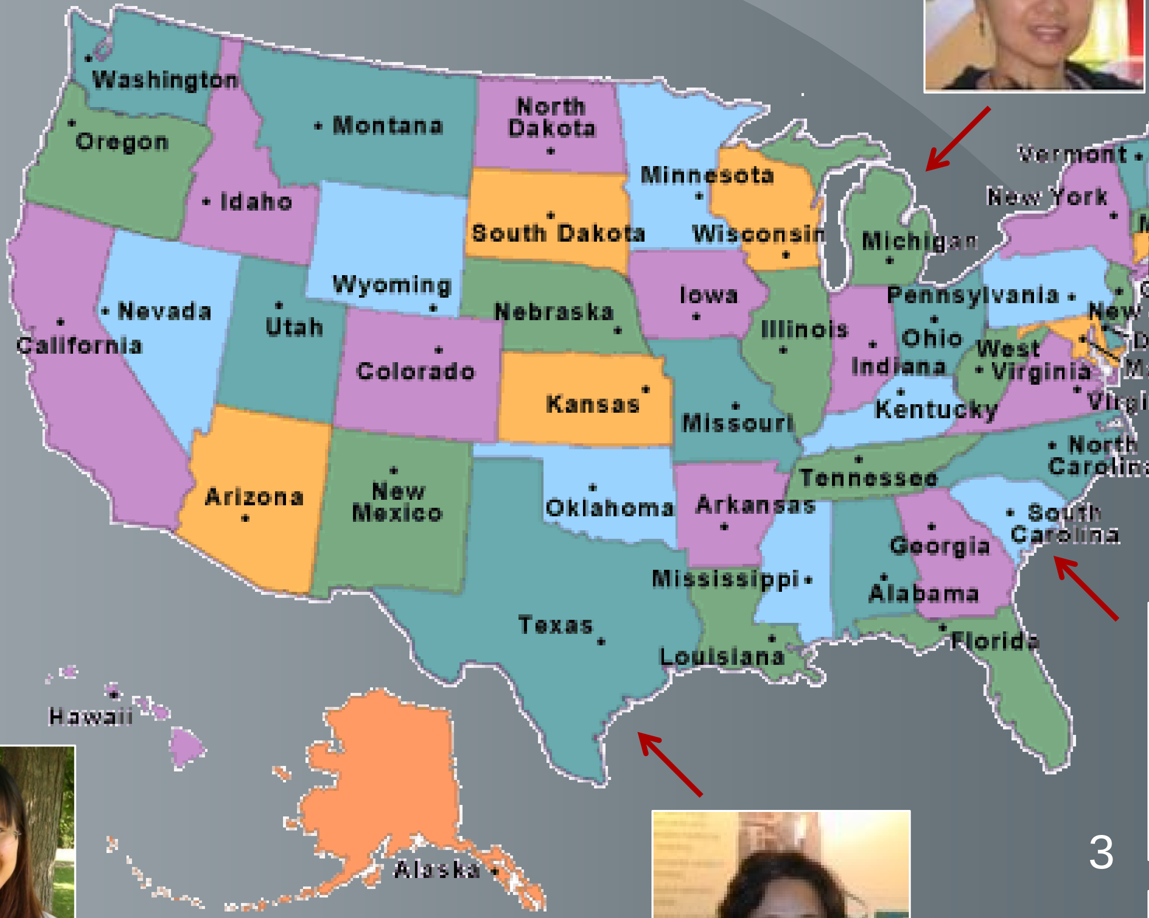
3

8

2



1



12

3

3

2

Acknowledgements

- Project Investigators
  - tfGDR (GDR and Citrus); Cacao Genome Database; Pine Genome Sequencing Project; Genome Database for Vaccinium; Cool Season Food Legume Database; CottonGen

- Rosaceae, Citrus, Cacao, Blueberry, Pea, Chickpea, Lentil, Cotton and Bioinformatics Community ☺

- Funding Sources: USDA NIFA SCRI, USDA DOE, NSF Plant Genome Program, USDA-ARS, Mars Inc, Washington Tree Fruit Research Commission, USA Dry Pea and Lentil Commission, Washington State University, Clemson University, University of Florida, Boyce Thompson Institute