NRSP10 Progress Report: October 2014 – May 2017

(1) Tripal Progress

- Core Tripal releases:
    - Tripal v2.0 released on June 1, 2016
    - Tripal v2.1 released on April 1, 2017
    - Tripal v3.0 beta 2 released February 10, 2017
    - Tripal v2.1-beta 3 released February 10, 2017
- Tripal Extension module releases:
    - Mainlab chado module to load sequence, map, marker, QTL, genotype, phenotype and germplasm released November 11, 2016
    - Mainlab chado module to search interfaces for sequence, map, marker, QTL, genotype, phenotype and germplasm released November 11, 2016
    - Mainlab Tripal module to display sequence, map, marker, QTL, genotype, phenotype and germplasm released November 11, 2016
    - Several extension modules developed and released by the larger Tripal community
    - New Tripal MapViewer v1.0 beta developed
    - New SNP marker search, SNP genotype search, and Reference transcriptome search modules developed
- Training and support for Tripal: monthly conference calls, 2 day hackathon and Tripal workshop held at Plant and Animal Genome Conference, more than 300 correspondences through the Tripal support mailing list, tripal.info website kept up-to-date (Figure 1)
    - Web service implementation underway that is backend independent so more users can use Tripal that don't need to have Chado as a database backend. This provides more flexibility for Tripal use.
    - Provided the carrot SCRI project with a copy of the CottonGen database (less cotton data) so they can get established quickly.
- Usage: Tripal now being used for more than 90 species/clade databases

Figure 1: Tripal Informational Site

(2) NRSP10 Databases

- Usage October 1, 2014 – May 1, 2017 (by Google Analytics)

Database	# Visitors	# Visits	# Pages Served	# Countries
Rosaceae	50,458	131,742	605,960	171
Cotton	25,262	60,331	349,731	163
Citrus	12,169	18,756	83,045	146
Legumes	8,027	13,218	60,606	144
Vaccinium	4,049	5,527	18,934	94
Total	99,965	229,574	1,118,276	

Table 1: NRSP10 Database Usage

- Training and outreach activities include presentations at several international, national and local meetings. These include holding user-taught training workshops and presentations at the 8th International Rosaceae Genomics Conference, the annual Plant and Animal Genome Conferences, the annual American Society of Horticultural Science Conferences, the Cotton Beltwide Conferences, a Crop Society of America annual conference, the 5th International Research Conference on Huanglongbing, etc., as well as webinars, brochures, posters. All presentations/webinars are available from database websites.

- New, more user-friendly website designed and implemented for all the databases using Tripal2 and Drupal7. The new designs include development of an NRSP customized logo for each of the five databases. So far, the CottonGen (Figure 2), CoolSeasonFoodLegume (CSFL, Figure 3), Genome Database for Rosaceae (GDR, Figure 4) and Citrus Genome Database (CGD, Figure 5) databases have gone live with the new design and the other three will go live later this year. Important new features in these designs include addition of major species overview pages to more easily view the data and tools available for the major species from a single page as well as quick links to the most used features from the home page. A significant volume of new data has been added to 4 out of the 5 databases, with major expansion planned for the remaining database (Vaccinium) later this year**.** The Citrus database has been expanded to include information specifically relevant to HLB research and Tripal extension modules are being developed relevant for disease pathogen and vector data and tools. The Genome Database for Vaccinium (GDV) has been redesigned in dev and will be made available publicly in June 2017. A significant volume of new data has been added to all of the 5 databases. For the 4 redesigned databases, it is estimated that between 90-95% of currently available public genomics, genetics and breeding data is available in these resources. This will be similar for the GDV when it is released.

Figure 2: CottonGen Database in Tripal2

Figure 3: Cool Season Food Legume Database in Tripal2

Figure 4: Genome Database for Rosaceae in Tripal 2/Drupal 7

Figure 5: Citrus Genome Database in Tripal2/Drupal 7

(3) Development of Breeding Tools

- Breeders Needs Assessment:
  Held NRSP breeders needs assessment workshops/meetings at NAPB 2015 in July 2015, the RosBREED participant project meeting in Michigan 2016 and the GDR workshop in San Diego in January 2017.

- Tools Development (Figure 6):
  - Major development ongoing of a comprehensive Breeding information Management System for Tripal databases (TripalBIMS) that will allow breeders to store, manage and analyze all their breeding data within a secure portal that connects up to all relevant public breeding data and genomic, genetic and breeding data in the crop database
    - Framework developed
    - Functionality developed to view and download breeding data by cross population
    - Functionality developed to generate the input files for Field Book App
    - Functionality developed to upload trait descriptor data has been updated to match the trait categories of the Field Book App.

Figure 6: Tripal BIMS – Todd Campbell Cotton Breeding Program

- Data in BIMS:
  - 11 years (2004 to 2014) of breeding data from Todd Campbell's cotton program (USDA-ARS South Carolina) has been converted from AgroBASE and loaded to BIMS in CottonGEN (Figure 7)
  - 3 years (2011-2013) of breeding data from Ksenija Gasic's peach breeding program (Clemson University) loaded to BIMS in GDR.

- 6 years (2011-2016) of breeding data from Kate Evans's apple breeding program (Washington State University) loaded to GDR
- 6 years (2012-2016) of breeding data from the USDA-ARS pea and lentil breeding program in process of being loaded to CSFL

- Field Book App for collection of phenotypic data
  - Development on Field Book has primarily focused on adding user-requested features and patching user-reported bugs. A new trait format, 'Location', was added to facilitate collection of location point data. A button was added to the main screen for missing values to help breeders distinguish between missing data and missing entries. Users can now load files directly from Dropbox, eliminating a file transfer step and streamlining the data collection process. Photos now also include the name of the trait to help researchers know better what they're looking at. A dedicated Android programmer was hired in the Poland lab in January who is learning how the apps work and working on rewriting parts of the apps to fix bugs, increase efficiency, and better-adhere to best programming practices.
  - Handheld Samsung tablets with Field Book have been provided to more than 40 NRSP10 associated breeders and researchers to test and use. Feedback critical in informing development through the Poland lab (provided with funds for a ½ time developer through NRSP10 since Sept 1, 2017)
  - Field Book and BIMS webinar held in November, 2016. FieldBook and BIMS presented at several conferences.

(4) Genome Sequence Annotation Server (GenSAS)

- GenSAS v3.0 released in December 2014 and GenSAS v4.0 released in January 2015Major improvements between them included the ability to use HPC to run jobs (enhanced speed) and addition of full functional annotation capability as well as additional tools for structural annotation, integration with JBrowse and WebApollo.
- GenSAS v5.0 released in January 2017 (Figure 7). Major improvements completed include; (1) ability to upload sequences before project creation so sequence subsets can be created from multiple-sequence fasta files. Sequence subsets can also be filtered by sequence name or minimum size (2) ability to upload RNA-seq reads use them to train gene model predictors (Augustus, during structural step) (3) addition of the tool Tophat to enable alignment specifically for the RNA-Seq reads. Alignment tools (blast, blat, PASA, tophat) are now included as a step before structural annotation (4) structural annotation (previously labelled as the "genes" step) now also has GeneMark for prokaryotic and eukaryotic gene prediction and (5) an official genes step (OGS) was added where users can either use EvidenceModeler to create a genes consensus and use it as the OGS, or can select another data track (from gene prediction program or alignment) as the OGS. Manual curation is merged with the OGS at the end of the GenSAS protocol.
- GenSAS was demonstrated/presented at several conferences since the grant started. Since GenSAS v3.0 was released it has been accessed by 2,129 visitors from 82 countries, with 5,701 sessions and 24,419 pages viewed
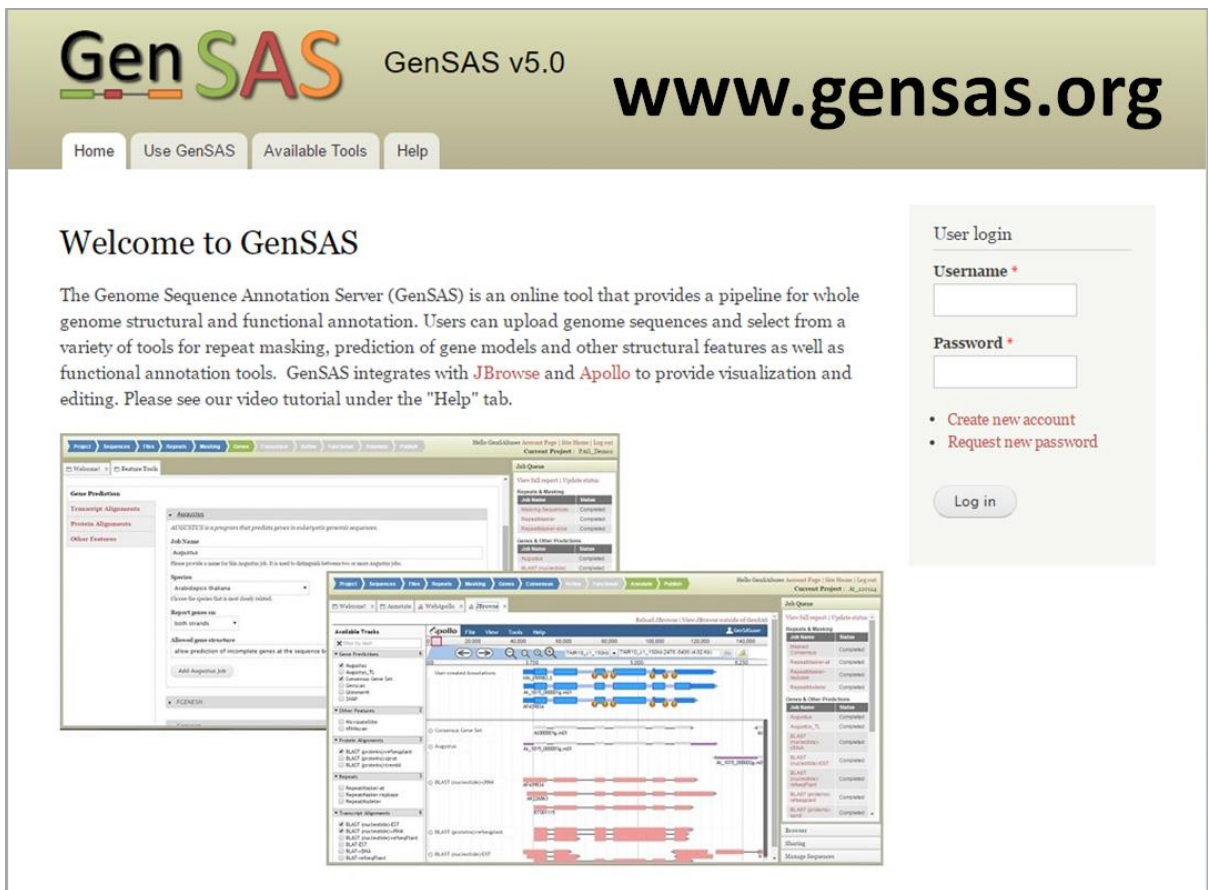
Figure 7: GenSAS Home Page

(5) Supporting Funding (~ $8M in large grants)

- "Standards and CyberInfrastructure that Enable Big-Data Driven Discovery for Tree Crop Research" (Role – PI) NSF Plant Genome Research Program, July 2016-July 2018, **$2,983,307**.
- "Further Development of CottonGEN: A Genomics, Genetics and Breeding Database for Cotton Improvement" (Role - PI). Cotton Incorporated and USDA-ARS, January 1, 2017 – December 31, 2020, **$741,889**.
- "CIF21 DIBBS: Tripal Gateway, a platform for next-generation data analysis and sharing" (PI Ficklin, CoPI Main) NSF CIF21 Program, Jan 2015 – Dec 31 2017, **$1,498,066**
- "Genome Database for Rosaceae: Empowering Specialty Crop Research through Big-Data Driven Discovery and Application in Breeding" (PI Main) USDA NIFA Specialty Crop Research Initiative, Sept 1 2014 – Aug 31 2019, **$2,700,000**.
- CottonGEN: A Genomics, Genetics and Breeding Database for Cotton Improvement" (Role - PI). Cotton Incorporated and USDA-ARS, October 2011 - December 2016, **$887,258**.
- US Land Grand Universities – participating scientist salaries
- Tree Fruit Research Commission, USA Dry Pea and lentil Commission and Northern Pulse Growers

(6) Other activities (Year 3)

- Provided complete copy of CottonGen database (less data) to the SCRI funded CarrotREP project for use as a project database

- AgBioData Workshop
  - Funded by NSF award to PI Main
  - Hosted and Co-Organized a 2 day AgBioData workshop (April 19-20, 2017) (Figure 8)
  - 43 scientists participated representing over 20 databases
  - Whitepaper on Agricultural Biological Databases by August
  - Opinion papers following the whitepaper

Figure 8: AgBioData website and workshop

- NRSP 10 Workshop
  - Holding NRSP10 Workshop at ASHS 2017 (Table 2) – focused on user presentations by users and covering application for specialty crops, with dedicated time for discussion on the future direction of NRSP.

(id=25937) What Is NRSP10 and What Can It Do for Specialty Crop Research . D. Main
(id=25935) NRSP10 Resources for Translational Tree Fruit Research . C. P. Peace
(id=25936) Using Tripal As a Project Database . M. Staton
(id=25900) NRSP10 Resources for Small Fruit Research . N. Bassil
(id=25901) Using Gensas for Specialty Crop Commmunity Genome Annotation . J. Humann
(id=25899) Using Fieldbook and Other Free Apps for Specialty Crop Field and Lab Data Collection . T. Rife
(id=25898) A Breeders Perspective on Using the Breeding Information Management System for Fruit Breeding . K. Gasic
(id=25939) Open Discussion on Direction of NRSP10 . M. Kahn

Table 2: NRSP10 Workshop Presentations scheduled for ASHS 2017