

INTRODUCTION

NRSP10 is a National Research Support Project funded by the State Agricultural Experiment Stations (SAES) and administered by NIFA (Representative Dr. Ed. Kaleikau). Started on October 1, 2014, NRSP10 is due to complete on Sept 31, 2019. In this Appendix we report on productivity and accomplishment of objectives which we were unable to document in the objectives section of the new proposal due to a limitation of 4000 characters for both new objectives and progress of previous ones.

NRSP10 (<https://www.nrsp10.org>, Figure 1) provides stable funding for underserved fruit, nut, berry and cotton crop databases while further developing a standardized, open-source, resource-efficient genomics, genetics and breeding (GGB) database platform, Tripal. The project supports the **core infrastructure** needed to establish and maintain these databases, creating a dynamic national resource that is broadly useful across agriculture and can be accessed by stakeholders with crop-specific interests. Explicit in this proposal was the need to secure aligned support from industry and regional/federal funding competitive sources to fund data curation and analysis activities for the 25 target crops of the five NRSP10 databases; Genome Database for Rosaceae (GDR, www.rosaceae.org); CottonGen (www.cottongen.org); Citrus Genome Database (CGD, www.citrusgenomedb.org); Cool Season Food Legume Genome Database (CSFL, www.coolseasonfoodlegume.org); and the Genome Database for Vaccinium (GDV, www.vaccinium.org), and additional cyberinfrastructure as needed.

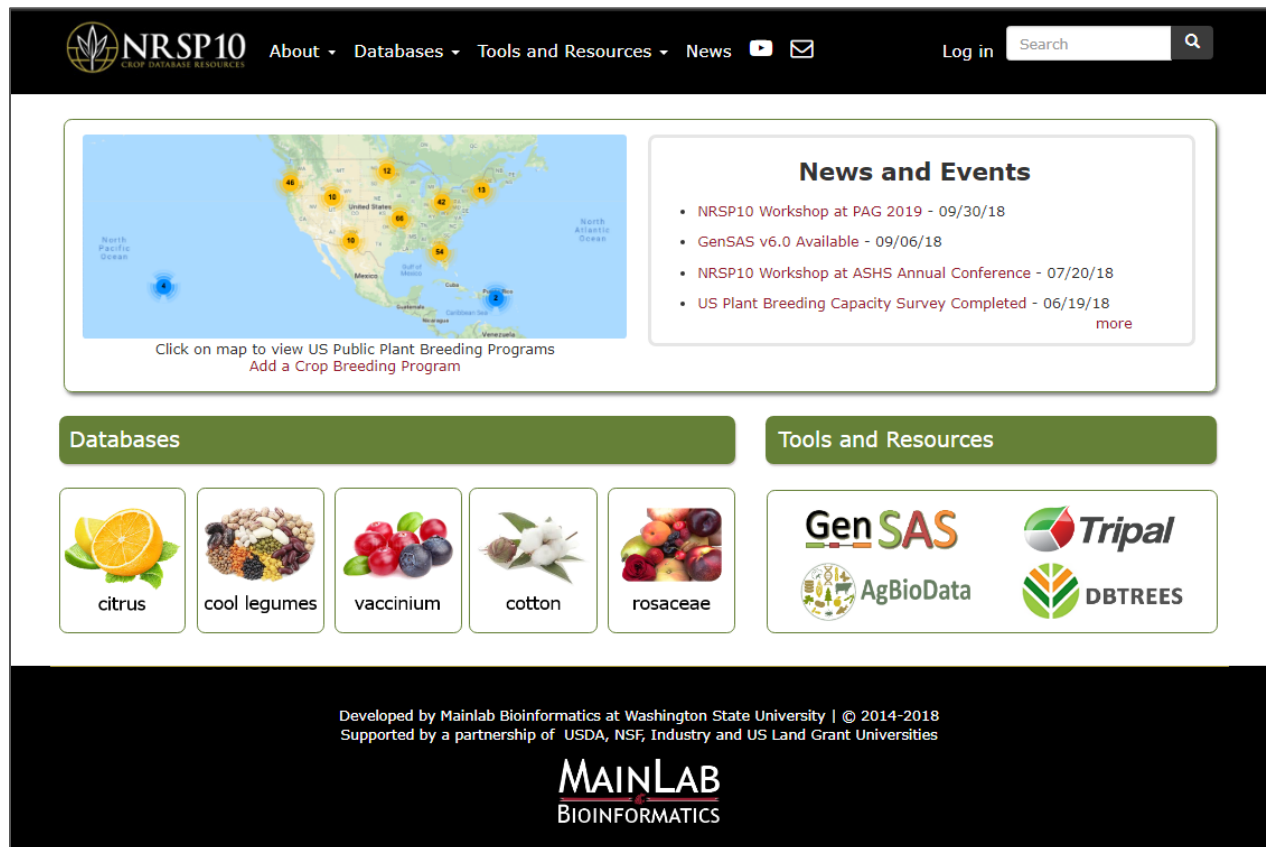


Figure 1: NRSP10 Project Website

The specific objectives were as follows:

Objective 1: Expand online community databases currently housing high quality genomics, genetics and breeding data for Rosaceae, citrus, cotton, cool season food legumes and *Vaccinium* crops.

Objective 2: Develop/implement a tablet application to collect phenotypic data from field and laboratory studies.

Objective 3: Develop a Tripal Application Programming Interface for building breeding databases.

Objective 4 – Convert GenSAS, the community genome annotation tool, to Tripal.

Objective 5 – Develop Web Services to promote database interoperability

ACCOMPLISHMENTS

(1) Tripal Progress

Major progress for Tripal in Years one to four included development and release of six Core Tripal versions (v2.0-rc1, v2.0, v2.1, v3.0.rc1, v3.0.rc2, and v3.0). Key improvements included (a) implementation of RESTful web services to support remote client queries, (b) interface to directly query the tripal.info site and find available extension modules, (c) improved performance by fixing memory leaks, (d) new and updated loaders, and (e) redesign of Tripal to make it back-end independent. This increased flexibility widens access to Tripal for databases that use a different database schema and helps resolve possible performance issues using Chado with very large datasets. Extension module releases include modules to load, search and display sequence, map, marker, QTL, genotype, phenotype and germplasm data, the comparative genetic mapping viewer TripalMap; and implementation of the popular search engine Elasticsearch for the chado database tables that enables cross-database site querying capability; Tripal Analysis Expression modules and Tripal Galaxy and associated workflows.

There are now 36 Tripal Extension Modules released or developed by the Tripal community (see <https://tripal.info/extensions#main-content>), 12 groups active in Tripal development in three countries, over 150 downloads of the Tripal platform and over 500 help desk support questions submitted and answered over the last 4 years. We estimate that over 4000 crop and wild relative species are now served through a Tripal database. Monthly Tripal meetings are regularly attended by 20 to 30 developers and the yearly codefest (formerly hackathon) attracts a similar number of developers. A Tripal workshop is held yearly at PAG with participant numbers exceeding 60 in the last 2 years following the inaugural workshop in 2015. The Tripal website (<https://tripal.info>, Figure 2) is kept current with tutorials and documentation, and regular webinars are provided for the community. Since implementing google analytics at the start of year 4 of NRSP, the site has been accessed by 6,760 users from 163 countries (1,332 U.S.) with 10,085 visits and 23,574 page views. All code is checked and approved by the Project Management Committee (PMC) before it is released to ensure standards are maintained. For newer projects, when required, we provide images of established databases, such as CottonGen to accelerate the transition to production, or provide programatic support to update to newer Tripal versions, e.g. PeanutBase. While all new versions of Tripal core are backward compatible, a lack of Drupal experience can make it a challenging for developers at times. Having access to an experienced support system led by initiators of Tripal, Drs. Stephen Ficklin and Meg Staton is very important and highly effective when accompanied by training for the developer.

[v2.x User's Guide](#)
[v3.x User's Guide](#)
[v3.x Developer's Handbook](#)
[Tripal @ gitHub.com](#)
[Tripal @ drupal.org](#)

Tripal is a toolkit for construction of online biological (genetics, genomics, breeding, etc), community database, and is a member of the [GMOD](#) family of tools. Tripal provides by default integration with the [GMOD Chado database schema](#) and [Drupal](#), a popular Content Management Systems (CMS). Tripal is licensed under the [GNU General Public License version 2](#).

Genomics, genetics, breeding and other biological data are increasingly complicated and time consuming to publish online for other researchers to search, browse and make discoveries. Tripal provides a framework to reduce the complexity of creating such a site, and provides access to a community of similar groups that share community-standards, and interact to address questions and learn best practices for sharing, storing and visualizing complex biological data.

Compatible With

Tripal directly supports or can be installed side-by-side with these popular tools:

News

[Tripal Help Desk Schedule](#)
02/22/2018 - 18:35
Times and meeting links for the Tripal Help Desk meetings for Friday, February 23, 2018.

[Online Help Desk Session](#)
02/20/2018 - 22:17
Tripal Help Desk, a series of virtual collaboration meetings, Kicks off this Friday, February 23, 2018.

[October Training](#)
10/25/2017 - 16:25
The Tripal Fall Training series will continue this Friday with an overview of the efforts to integrate Tripal and Galaxy.

[Online Training Event: Tripal v3](#)
09/28/2017 - 19:58
Join us Friday Sept 29th for an online training event to learn how to setup and manage a Tripal v3 site.
[More](#)

Sites Using Tripal

[Curcubit Genomics](#)

11 of 18
[\[View more\]](#)

Contributors

The following organizations contribute to development of the core Tripal package or to extension modules that are freely available. Funding agencies that contribute funds for core Tripal development are also shown.

Tripal Tweets

Tweets by [@TripalProject](#)

Figure 2: Tripal Informational Site

(2) NRSP10 Databases Progress:

A marked expansion of NRSP10 databases has been accomplished in terms of data (Table 2), tools and infrastructure. The utility of these databases has been demonstrated by a significant increase of users each year, over 2.5 million pages served and more than 1000 citations of the databases (see Table 3). For example, the number of data in GDR increased from 4 to 21 whole genome assemblies, 2.2 million to 3.3 million markers, 2,234 to 3,799 QTL, 783,639 to 878,476 phenotypes, and 37,795 to 10,806,569 genotypes. Similar increases in the number of data have been achieved in the other databases as well. Newly integrated data types in each database include whole genome assemblies and cultivar evaluation data in CGD and GDV, expression data in GDV, re-sequencing data in GDR and CottonGEN, synteny data in CottonGEN, CSFL and CGD, pathway data for all databases, and distillation of published RNASeq data for major species/genera into reference transcriptomes. 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. In addition, new interfaces and tools have been added to allow better access to these data. The new or enhanced tools include those developed by other databases such as Tripal synteny module (Fei Program, Boyce Thompson

Institute, <http://bioinfo.bti.cornell.edu/>) Tripal expression module (Staton Program, University of Tennessee, BLASTX (Bett Program, University of Saskatchewan and Ethy Cannon, Iowa State University/USDA-ARS), and multi-variant viewer plugin for JBrowse, as well as our own tools such as MapViewer, BIMS and search pages for SNP, genotype, gene/sequence data. The Tripal extension modules we developed have been released and published (Jung et al. 2017) or in preparation (BIMS and TripalMap). All databases have been upgraded either to Tripal v2.1 or v3.0 and the Citrus database has web services and ElasticSearch capability implemented, enabling users to search across the Hardwood Genomics and TreeGenes databases. The other 4 NRSP10 databases are being upgraded to Tripal v3.0 and will also enable seamless searching across Tripal v3.0 databases.

Table 2: NRSP10 Crop Database Data Summary (Oct 15, 2018)

Data Type	NRSP10 Database				
	Rosaceae	Cotton	Citrus	Legumes	Vaccinium
Gene	597226	258462	243066	0	0
Genome	22	7	19	3	1
Genotype	10,806,569	64775	0	?	0
Germplasm	14,411	17640	1426	2429	355
Map	316	109	70	167	10
Marker	3,286,882	573381	48154	137268	4253
mRNA	659,987	503516	448144	85359	1157
Pathway	1797	750	901	673	0
Phenotype	389,191	383431	23070	0	429
Polypeptide	155,001	154,213	67598	90,216	0
Publication	7,515	16852	4045	4988	477
QTL/MTL	3,902	4923	579	3009	26
Reference Transcript	259,213	214,180	33,111	191,154	39,461
RNASeq read	9.31 billion	5.85 billion	1.2 billion	5.11 billion	757 million
Species	1967	73	67	33	15

Table 3: Metrics of Usage and Citations for the NRSP10 Databases

Genome Database for Rosaceae					
Year	# Visitors	# Visits	# Pages Served	# US Visitors	# Citations
Year 1	18253	43958	197161	3649	169 (2015)
Year 2	19226	49529	231964	4079	181 (2016)
Year 3	21537	57217	312476	4148	199 (2017)
Year 4	24331	66591	641518	5947	175 (2018)
CottonGen					
Year	# Visitors	# Visits	# Pages Served	# US Visitors	# Citations
Year 1	9182	17,219	97047	2011	33 (2015)
Year 2	9464	23972	134123	2875	41 (2016)
Year 3	10946	29799	196013	2743	70 (2017)
Year 4	12098	32647	274665	3207	97 (2018)
Citrus Genome Database					
Year	# Visitors	# Visits	# Pages Served	# US Visitors	# Citations
Year 1	4350	6328	22659	1049	19 (2015)
Year 2	4525	6512	24064	1178	28 (2016)
Year 3	5165	9207	63475	1459	19 (2017)
Year 4	5931	10587	68683	1613	12 (2018)
Cool Season Food Legume Genome Database					
Year	# Visitors	# Visits	# Pages Served	# US Visitors	# Citations
Year 1	2842	4293	17064	547	12 (2015)
Year 2	2975	4765	22141	744	13 (2016)
Year 3	3375	6005	31965	941	10 (2017)
Year 4	3629	6228	47203	1039	8 (2018)
Genome Database for Vaccinium					
Year	# Visitors	# Visits	# Pages Served	# US Visitors	# Citations
Year 1	1302	1722	6685	577	5 (2015)
Year 2	1,339	1,811	6,106	628	3 (2016)
Year 3	2245	3312	17,380	990	10 (2017)
Year 4	2664	4361	30997	973	10 (2018)

New, more user-friendly websites has been designed and implemented for all the databases using Drupal7 and the code base upgraded to core Tripal v2.1/v3.0 from v1.1. The new designs include development of an NRSP customized logo for each of the five databases (Figures 3-7). 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. Stakeholder feedback through surveys and meetings has indicated the design has indeed made it easier for researchers to access data and tools in the databases relevant to their areas of interest.



Figure 3: CottonGen Database

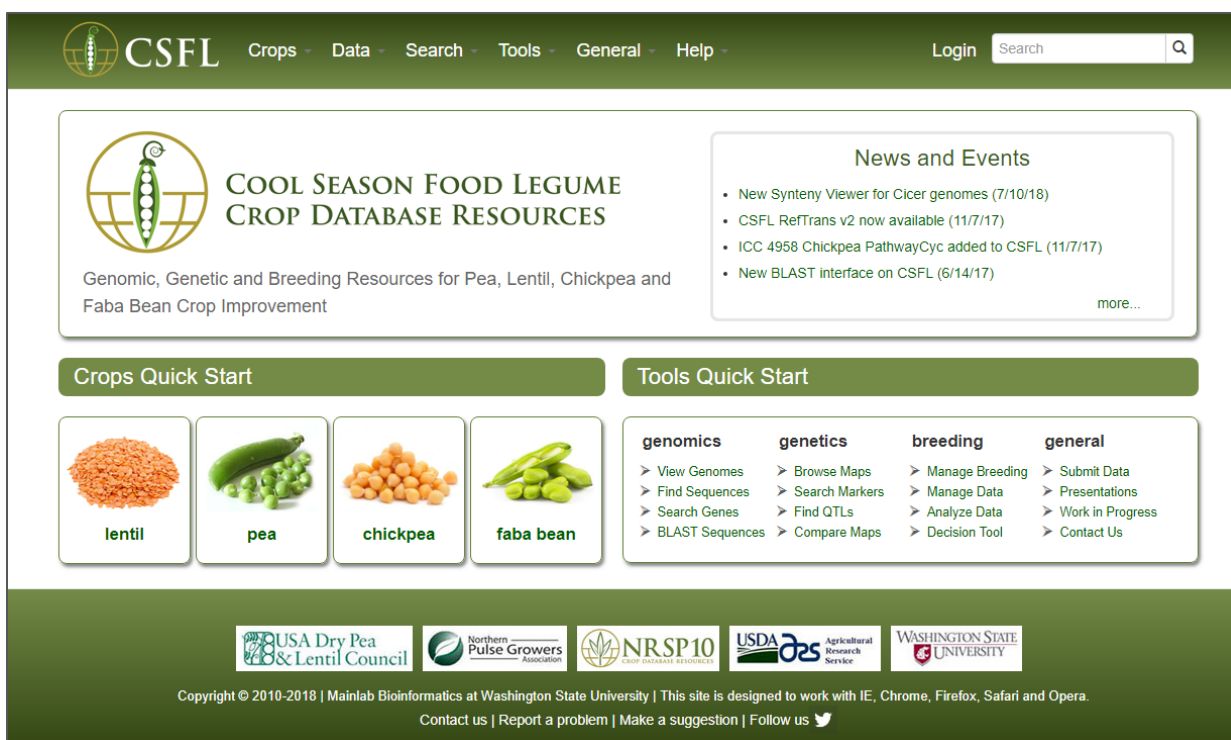


Figure 4: Cool Season Food Legume Database

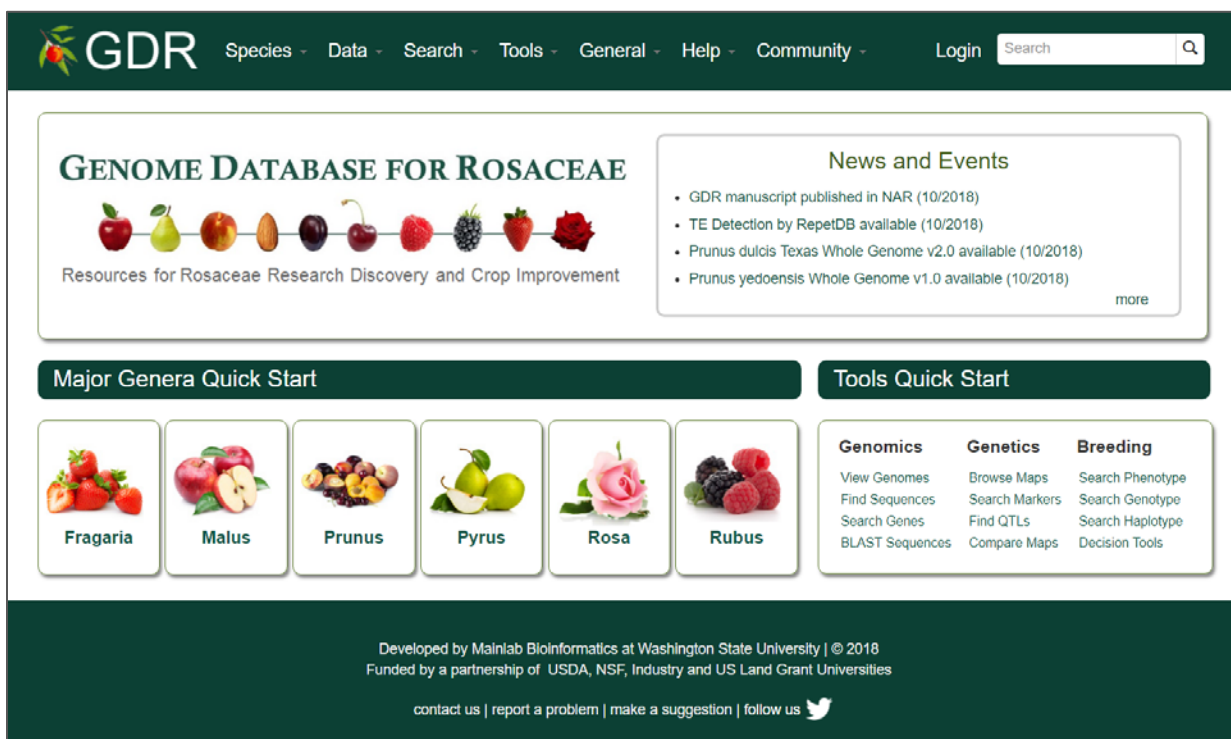


Figure 5: Genome Database for Rosaceae

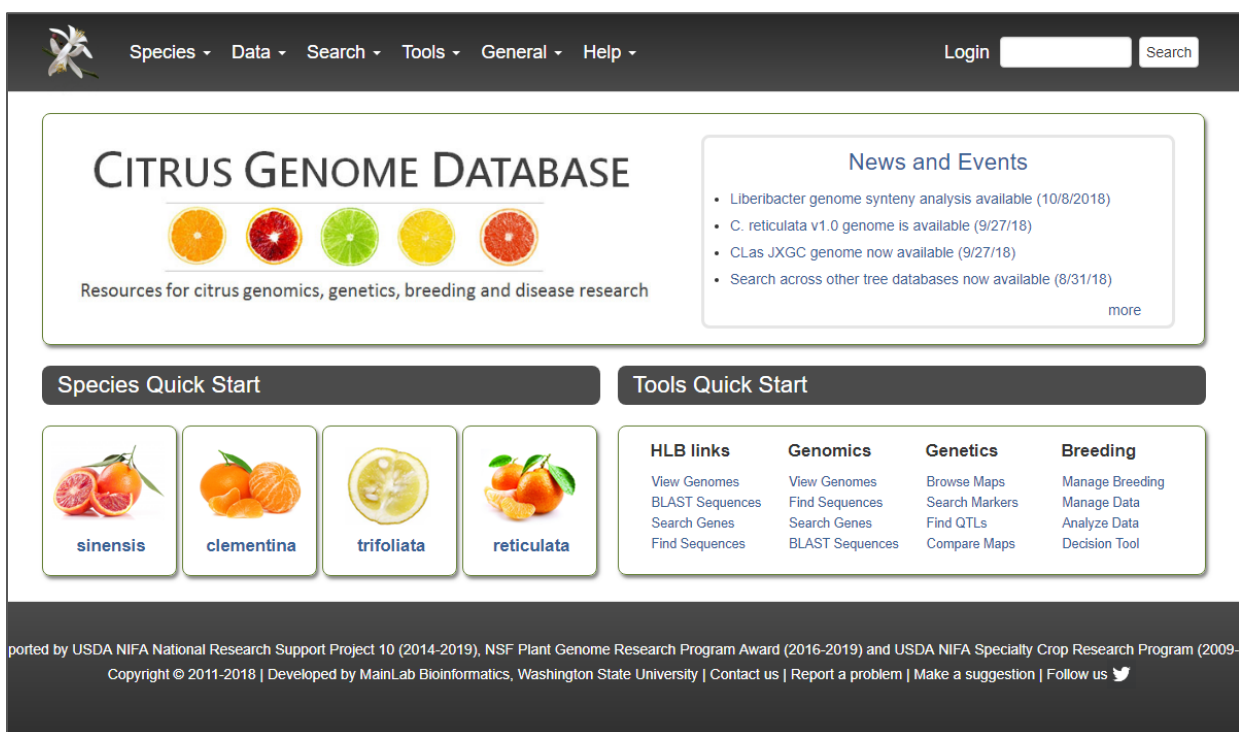


Figure 6: Citrus Genome Database

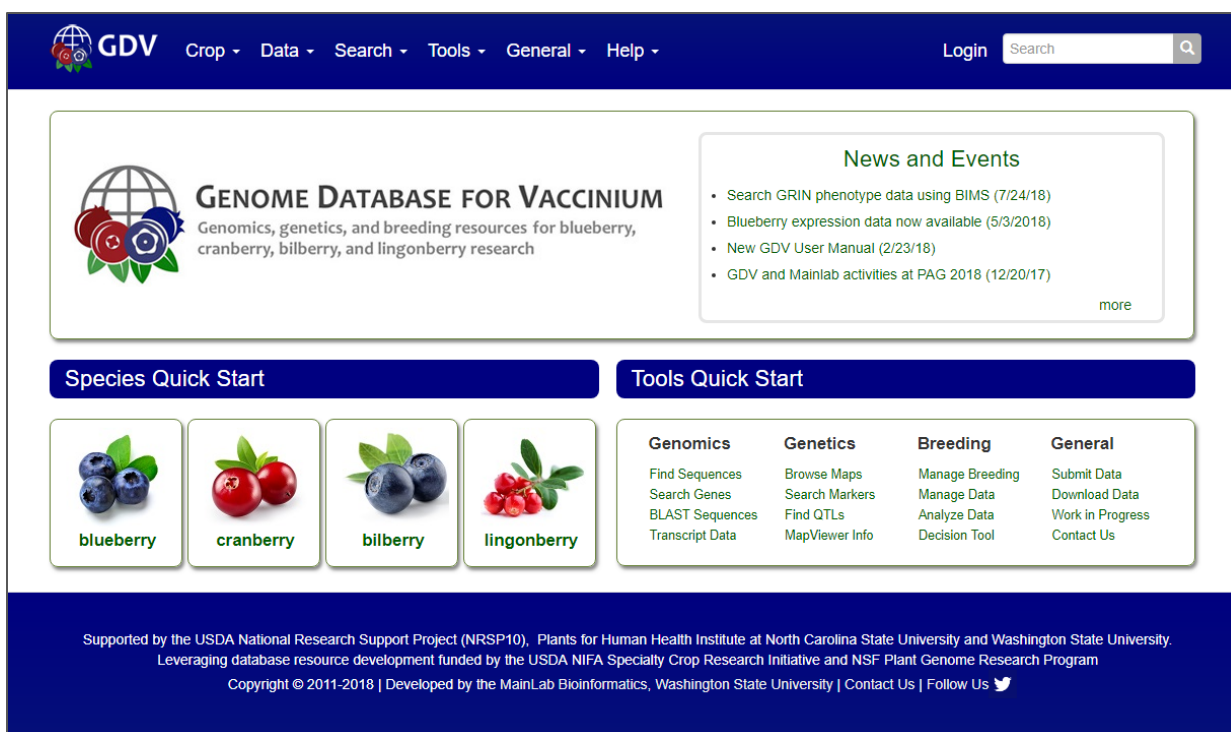


Figure 7: Genome Database for Vaccinium

(3) Development of Breeding Tools

We have developed and implemented the Tripal Breeding Information Management System (BIMS, Figure 8) in all five NRSP10 databases. BIMS allows individual breeders to integrate their phenotypic and genotypic data with public genomic and genetic data and at the same time have complete control of their own breeding data and access to tools such as data import/export, data analysis and a data archive. BIMS also allows site visitors to browse and analyze publicly available breeding data such as GRIN data or public RosBREED phenotype data. The current functionality includes manage breeding, data import, search and download and statistical analysis. We had multiple webinars and presentations at conferences, including a one day workshop for Rosaceae breeders with at least two more dedicated workshops planned in the final year of NRSP10. We have also recently hired an experienced translational genetics technician who will visit several breeding programs to provide in person training on using BIMS and converting and/or uploading the data into BIMS. Currently breeders of several crops including peach, apple and cotton are using BIMS for their private breeding programs. In the last year we will ability for breeders to also upload genotype data, and connect the private breeding data with all associated marker, trait, germplasm, genome, etc. data in the public side of the databases, without releasing any of the breeders private data. We will release BIMS as a major extension module in year 5 of the NRSP10. There has been considerable interest in this module from many communities and we suspect it might help further accelerate the adoption of Tripal databases with access to this breeding system. We propose continuing to add more analytic and high performance computing capability over the next five years of NRSP10 if funded. As the largest of the Tripal extension modules to be released, we expect less experienced Tripal developers will require support to implement fully.

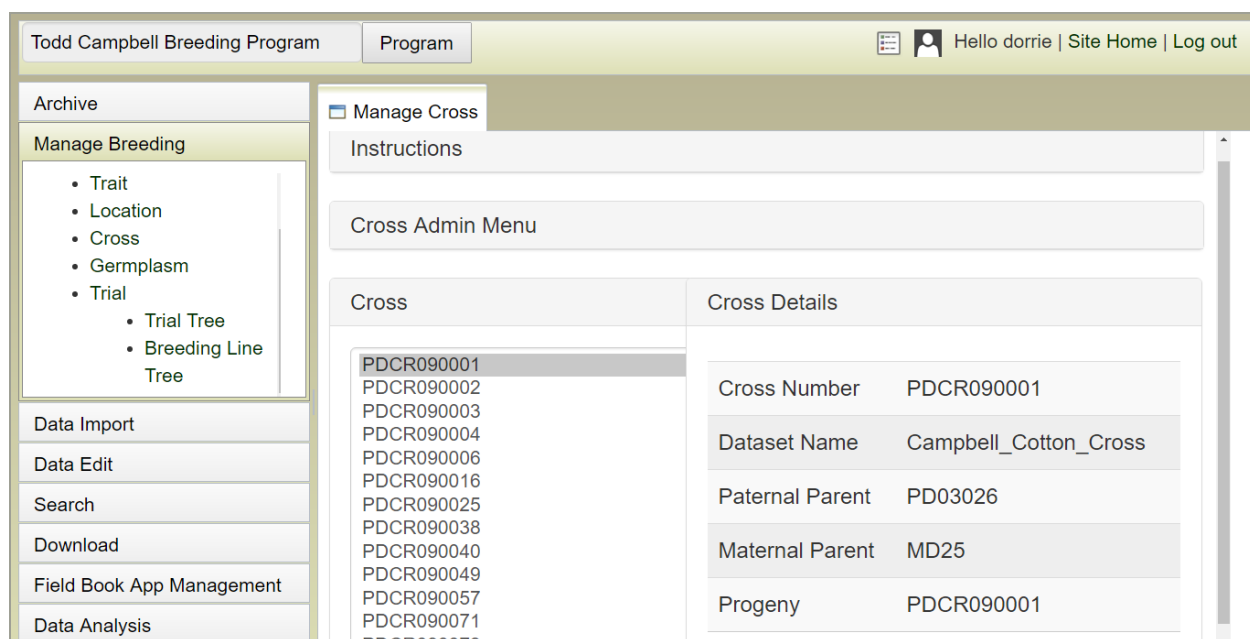


Figure 8: Tripal BIMS – Todd Campbell Cotton Breeding Program

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
- GRIN phenotype data for several crops
- RosBREED publicly available phenotype data (genotype data being added)

Field Book App for collection of phenotypic data

Following a review by the NRSP10 team and representative breeders at a NRSP10 breeders workshop in year 1, we recommended adoption of the Field Book App, a free Android App developed by Jesse Polands Program at Kansas State University. In years 2-5 we provided funds to support a 50% developer to adapt the App so it could also work better for specialty crop breeding as well as row crops. Development 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. We provided Handheld Samsung tablets with Field Book installed to over 40 NRSP10 associated breeders and researchers to test and use. Feedback has been critical in informing development. As the software will be fully functional for specialty crops by the end of the current NRSP10, we do not anticipate a need for continued NRSP10 funding for Field Book.

(4) Genome Sequence Annotation Server (GenSAS)

In the first four years of NRSP10, we have significantly redeveloped and improved the functionality of the Genome Sequence Annotation Server (GenSAS, www.gensas.org, Figure 9). GenSAS is a web-based annotation platform that combines several common annotation tools into one easy-to-use, integrated resource. The annotation process, which uses HPC, is carried out in user-friendly interfaces with embedded instructions and only requires a user account and internet access. GenSAS has annotation tools for eukaryotes and prokaryotes and supports model and non-model organisms. Users can upload Illumina RNA-Seq reads (or specify datasets from the NCBI SRA database), align the reads to the genome using HISAT2 or TopHat2, and use the aligned data to train the gene model prediction programs AUGUSTUS and BRAKER2, which allows for more accurate gene models for eukaryotic genomes, especially non-model organisms. JBrowse and Apollo are integrated into GenSAS allowing for structural annotation results to be easily viewed and manual curation performed. GenSAS projects can be shared with other users enabling collaborative manual curation. GenSAS has a functional annotation step to assign protein functions and identify putative functional domains for the official gene set. After annotation is complete, the required files for publication are generated and users can run BUSCO on predicted proteins to assess the completeness of the annotation. We have released v3.0, V4.0, v5.0, v5.1 and recently v6.0. In year 3, GenSAS v4.0 was accessed by 1,777 users from 76 countries with 6,525 sessions and 23,574 pages served. In year 4, GenSAS v5 was accessed by four times as many users (6,453) from 101 countries with 17,586 sessions and 65,536 pages served. Even greater use is expected with the release of V6.0 and the publication of a peer-reviewed manuscript.



Figure 9: GenSAS home page

In the last year of NRSP10 we will implement GenSAS Community as a Tripal module that focuses on manual curation. GenSAS Community will be a Tripal module that can be installed on model organism databases to facilitate community annotation of genomes. This would not be a tool to do computational annotation of genomes like GenSAS, but rather a tool for managing multi-user, manual curation of genomes. GenSAS Community will expand on the Apollo/JBrowse functionality that is already present in GenSAS v6.0 (Figure 9). While GenSAS would/could still be used to generate the initial annotation, GenSAS community will be a “lighter” version that focuses on manual curation. The proposed function of GenSAS Community will include:

1. Import of annotation projects directly from GenSAS (Final annotation and all supporting data tracks) OR import of any GFF3 file the user provides. If non-GenSAS files are used, an OGS file will have to be designated during import.
2. Designate Lead Curator, Assistant Curator, and Guest Curator roles linked to Drupal user accounts. Lead Curators will manage the community curator projects and have permission to review and approve manual curations. Assistant Curators will have permission to perform manual curations and submit them for review. Guest Curators will have view-only access and will not be able to perform manual curation. This will allow for some of the genomes to be private and not viewable by anonymous Drupal users.
3. Lead Curator control panel for project management. The Lead Curator(s) will have a dashboard for managing the project. Functions will include adding users/adjusting permissions, assigning genome regions to users/list of region each user is working on, viewing annotations that have been submitted for approval and the ability to accept annotations or return with comments to curator.
4. GenSAS community will have an option to add approved manual curations into the Chado database as the project progresses or when the project is complete. It will generate output files of the annotation over the course of the project.

(5) Supporting Funding (~ \$10M in large grants, commission and university support)

An explicit component of NRSP10 success metrics was the ability to obtain aligned support to conduct data curation and analysis activities, as well as additional cyberinfrastructure as needed. This was a successful endeavor with funding from USDA NIFA SCRI and NSF to support curation efforts and extend the functionality of the Tripal platform. Examples include:

- “Standards and CyberInfrastructure that Enable Big-Data Driven Discovery for Tree Crop Research” (PI Main) 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” (PI Main). 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" (PI Main). Cotton Incorporated and USDA-ARS, October 2011 - December 2016, **\$887,258**.
- Tree Fruit Research Commission, USA Dry Pea and Lentil Commission and Northern Pulse Growers
- US Land Grant Universities – participating scientist salaries

(5) Outreach and other activities

The first four years of NRSP10 has seen a significant amount of outreach activities. These have included:

- Over 80 presentations at workshops and meetings, (likely to exceed 100 by the end of the first NRSP10 award). These include PAG, ASHS, ASPB, Cotton Beltwide, International Rosaceae, Cotton, Rubes and Ribes, Strawberry, Peach and Citrus Greening conferences and symposiums, as well as Bioinformatics, GMOD, and Galaxy conferences.
- Two dedicated NRSP10 participant training workshops at ASHS conferences in years 3 and 4, with a third to be held at PAG in January 2019
- Tripal annual workshops at PAG since 2015
- BIMS workshops at the Cotton Beltwide annual conference (2016-2019), biennial Cotton Breeders Tour (2015, 2017) as well as one day Rosaceae Breeders training workshop at the 2018 and 2019 RosBREED participants meeting.
- Webinars for components of the databases/tools and Tripal
- Monthly Tripal developers meetings and monthly AgBioData meetings as well as steering committee membership
- Steering Committee Meetings
- Hosting of community elections and conference abstract submissions through the CottonGen and Rosaceae databases
- Brochures at PAG and ASPB annual meeting database booths
- Peer-reviewed publications for Tripal extension modules, GDR, and AgBioData, with plans to submit GenSAS, MapViewer, CGD, and Tripal v3.0 for publication by end of 2018. One manuscript a month will be submitted monthly February through May for the remaining four database, so all the databases, core Tripal, and extension modules will be published by the end of year 5.
- Provision of Tripal support has been quite extensive hence the request for more an additional position to support this in the next round of NRSP10 and help ensure the Tripal momentum continues to grow.
- Developed and host the AgBioData consortium website (<https://www.agbiodata.org>) and helped organize an AgBioData consortium workshop (Figure 10) with funding from NSF PGRP award to PI Main. The 2 day workshop was held in Salt Lake City in April 2017, with 43 scientists participating (including NRSP8 PI Reecy). A white paper was developed and published in 2018.

Other activities have included developing, hosting and analyzing a U.S. Plant Breeding Capacity Survey at the request of the Plant Breeding Coordinating Committee (Chair Kate Evans). This involved developing a module to capture and store breeders program information, and display on a dynamic google map with searchable tables. Over 260 breeders completed the rather comprehensive survey and another 100 have since added their breeding program to the map. This resource enables government, universities, and other decision makers with access to the types and numbers of breeding programs in the U.S. and allows them to assess what additional resources/new programs might be needed to meet the demand for new high yielding, climate resilient cultivars. We are also in the process of developing a searchable website for the paper-based fruit and nut register, so breeder and industry stakeholders can readily access this important information. The website will allow fruit and nut registry data curators to access the portal and add/ edit information directly, with a history of the changes kept. All sites in the NRSP10 domain are backed up monthly, both onsite and offsite, with nightly incremental backups. We anticipate in the renewal that we might need to add additional tools useful for genome editing.



Figure 10: AgBioData website and workshop

In conclusion, we believe we have demonstrated the success of the NRSP10 proposal, both in terms of the growing adoption of the Tripal platform and expansion and use of databases for underserved fruit, nut, legume and cotton crops of economic importance (\$26.6 B in 2016, NASS) to the U.S. We have exceeded projections for aligned support and will make progress in collecting data needed to start recommending solutions for longer term sustainability to be tested and assessed in the proposed renewal project. We are embedded in the communities these crop databases serve and the wider agricultural GGB database community, with widespread support for the activities of this NRSP10.