National Research Support Project Summary

Project Number: NRSP_TEMP321

Title: Database Resources for Crop Genomics, Genetics and Breeding Research

Duration: October 2014 to September 30, 2019

Administrative Advisor(s): [Susan K. Brown NE] [Steven Lommel S] [James W. Moyer (main) W] [Karen Plaut NC]

NIFA Reps:

Statement of Issues and Justification

Prerequisite Criteria

How is the NRSP consistent with the mission?

Recent advances in sequencing, genotyping, and phenotyping technologies have led to a paradigm shift in crop science research. Scientists now routinely sequence and genotype genomes from populations, families and individuals of interest, pursue large-scale gene expression studies, create highly saturated genetic maps, identify loci influencing traits of interest, and conduct large-scale standardized phenotyping. This is generating petabytes of data that must be organized, stored, analyzed, curated, and integrated with other genomic, genetic, and breeding data to promote access and enhance their utility to various research communities. Many community databases have been developed over the last 20 years to provide this critical, enabling role for many crops (MaizeGDB, Lawrence et al. 2008; GrainGenes, Carollo et al. 2005; Solanaceae Genomics Network, Bombarely et al. 2011; CottonDB, Yu et al. 2012; SoyBase, Grant et al. 2010; Genome Database for Rosaceae, Jung et al. 2013a). Although excellent resources, most of these databases were developed in isolation without access to a common database platform. Most function with a crop- or clade-specific focus and customized database schemas and therefore are complex to manage, have difficulty accommodating new data types, and are resource-intensive to implement for other crops or organisms.

The Genome Database for Rosaceae (GDR) development team led by Dr. Dorrie Main at Washington State University (WSU) faced the same legacy database issues when asked in 2008 by USDA-ARS and Mars, Inc. to create a database (www.cacaogenomedb.org) to house the genomic and genetic data generated from the Theobroma cacao Matina1-6 genome sequence project. Rather than implement a cacao version of the Rosaceae database with the inherent problems mentioned above, we decided to build using Chado (Mungall and Emmert 2007), an open-source, generic, and modular relational database schema that underlies many of the Generic Model Organism Database (GMOD) applications. Ontology driven, modular, and highly flexible, the Chado design enables the same schema to be used in projects with very different metadata. The modular design allows developers to use subsets of the schema needed for their specific data or request addition of new modules as new data types become available. Due to these advantages, funding agencies repeatedly asked us to adopt Chado when we

proposed renewal funding for GDR. Chado at the time lacked a module to store large-scale phenotypic and genotypic data, so we formed an international consortium of researchers to develop the ?natural diversity module?. The consortium ensured the module would meet all their data storage needs and also be compliant with Chado design principles and requirements. This module expanded Chado's capacity for storing data both from multiple experiments and from specimens collected, treated and evaluated in multiple locations, environments and time points. Subsequently, databases for widely different organisms have been built and made available to relevant research communities using the enhanced version of Chado (Jung et al. 2011).

In addition to using custom database schemas, the web interfaces used to query and display the biological data in community database have also been constructed using custom scripts developed using various computer languages which cannot be easily shared with other databases. The result is that there has been significant redundancy in database development, with too much time and effort spent on parallel infrastructure development and database management, time that could have been more valuably employed on curation, analysis and integration of data. To mediate this problem, the GDR team opted to continue the work initiated at Clemson University on Tripal, an open-source web front-end for the Chado database schema (Sanderson et al. 2013, Ficklin et al. 2011). Recruiting Tripal?s two lead developers (Drs. Stephen Ficklin and Margaret Staton) and collaborating with the Dr. Kirsten Betts group at the University of Saskatchewan, the GDR team subsequently led its transformation from a primarily transcriptome platform to its current status as a more comprehensive genomics and genetics database management solution (described below).

Tripal is an open-source, efficient, flexible and modular platform for building online genome databases. It is designed to provide data pages and search tools for genomic, genetic, stock, cultivar, library and species performance data stored in Chado as well as provide tools for parsing and uploading data from computational analyses such as BLAST (Altschul et al. 1990), Interpro (Hunter et al. 2012), KEGG (Kanehisa et al. 2012), and blast2GO (Conesa et al. 2005). Tripal was developed for use with the popular open-source content management system Drupal, which affords non-technical users the ability to add content and functionality easily and without the need for programming. By bridging Chado and Drupal, Tripal marries the power of a biological data storage schema (Chado) with a web development platform (Drupal) to decrease the cost and time associated with development of genomic, genetic and breeding websites for diverse biological research communities. The core Tripal package includes data loaders for common file formats and provides a web-enabled bulk loader that facilitates construction of custom data loaders for tab-delimited data sets. For sites that require more than default functionality, Tripal has a well-defined and documented Application Programming Interface (API) that provides access to data in Chado and other Tripal modules. Therefore, site developers can tailor and customize the look-and-feel of a site for their respective communities as well as display data in novel ways. Support is readily available via an active mailing list, online tutorials and documentation (http://tripal.info). Tripal is the only software categorized as ?website front end for Chado? in GMOD and we are not aware of other tools similar to Tripal that integrate a database schema for storage of biological data (i.e. Chado) with a Content Management System (e.g. Drupal) and provide an API to allow for extension by other developers.

Following development of the Cacao Genome Database (now managed by Dr. Steven Cannon, USDA-ARS Iowa State University), the GDR team used Tripal to construct several other community databases for crops of substantial economic importance to the U.S. (Table 1). Funded by a combination of federal, university and industry support, these databases provide centralized resources for data archiving, data mining, and analysis tools designed to assist scientists carrying out basic, translational, and applied

research (Jung et al. 2013; Yu et al. 2013; Main et al. 2013a; Main et al. 2013b). They are now considered the worldwide databases for the communities they serve and include the: Citrus Genome Database (Main et al. 2013a), CottonGen (Yu et al. 2013), Cool Season Food Legume Genome Database (Main et al. 2013b), and Genome Database for Vaccinium (www.vaccinium.org).

The GDR team has now converted two well established legacy-type databases to Tripal, demonstrating the feasibility of upgrading large and complex legacy databases to the more effective and functional Tripal platform. Firstly, CottonDB (Yu et al. 2012), a non-relational AceDB-type database established in 1995 was converted to Tripal in one year, followed by conversion of GDR.

Tripal applications developed for any one of the five WSU databases are made accessible to the other database as resources become available to collect the underlying data. This model of operation, targeting applications to meet demonstrated needs in a given research community, together with collaborations with the University of Saskatchewan and Clemson University, has allowed us to effectively leverage funding from a variety of sources and operate in a resource efficient way. To provide all the tools and functionalities of GDR in other databases, however, we need to continue with our curation effort and ensure the databases are current with publicly available data. More development of Tripal is also necessary to provide a complete toolkit for efficient construction of biological databases. The provision of this open source software for biological database creation will allow underrepresented institutions and/or crop communities to build comprehensive genomics, genetics and breeding databases, which would otherwise be cost and expertise prohibitive.

We propose to continue our current mission to (1) provide database resources for target crops and (2) further develop a standardized database platform for use by other communities, with the following specific objectives:

(1) Expand online community databases currently housing high quality genomics, genetics and breeding data for Rosaceae, citrus, cotton, cool season food legume and Vaccinium crops.

(2) Develop a tablet application to collect phenotypic data from field and laboratory studies.

(3) Develop a Tripal Application Programming Interface for building breeding databases.

(4) Convert GenSAS, the community genome annotation tool, to Tripal.

(5) Develop Web Services to promote database interoperability.

Providing tools for standardized database construction and continuing to develop efficient and widely used databases that provide a centralized repository for data archiving, resources for data mining, and analysis tools designed to assist scientists perform basic, translational and applied research is consistent with stated NRSP missions:

(1) Development of enabling technologies and/or support activities (such as to collect, assemble, store, and distribute materials, resources and information

(2) Sharing of facilities needed to accomplish high priority research.

How does this NRSP pertain as a national issue?

Recent advances in sequencing and genotyping technologies have led to an exponentially growing volume of data describing our target crops and have created grand challenges for data management, data mining, data querying, and data visualization. Community databases are the logical home for these data and enable post analysis integration with associated phenotypic data to maximize their utility to scientists and return of investment to funding organizations. Collectively, the 24 crops targeted in this proposal are grown commercially in all four SAES regions, in all 50 states, and had a value of production in 2012 exceeding $23.6 billion (Table 1). These databases are accessed routinely by researchers from all 50 states and territories (as recorded by Google Analytics), and the U.S. scientists who collect and provide the data, and use these databases are predominantly based at Land Grant Universities and USDA-ARS. This project would support the core infrastructure needed to establish and maintain these databases, creating a dynamic national resource that is broadly useful across crop agriculture and can be accessed by stakeholders with crop-specific interests. Community databases that are widely available and actively developed and curated will facilitate further applications in genomics-assisted breeding as well as advances in the genomics, genetics, and physiology knowledge base. This NRSP will result in the building of a national system for underserved crops which do not have access to federal flow-through funds individually but are regionally critical. Providing support to continue development and deployment of these fundamental, research-enabling databases, used ubiquitously in the U.S. for target crops and readily adoptable for other crops, classifies this NRSP as a national issue.

National issue: To increase data utilization across disciplines and facilitate research activities

This NRSP proposal will greatly accelerate integration of large scale breeding data with the genetic and genomic data of Rosaceae, citrus, cotton, cool season food legume and Vaccinium crops. Recent advances in genomic technologies have already enhanced research in crop physiology and genetic improvement in many commodity crops (e.g. maize, many cereal grains, soybean, tomato). Similar access to such molecular resources has enormous potential for the crops targeted in this proposal. A central goal of this proposed NRSP is to maximize the impact of these genomics, genetics and breeding resources by providing a web-based platform by which massive amounts of data can be routinely collected, curated, integrated and made available to scientists in formats that best meet their diverse needs.

National issue: To provide enabling technologies for efficient database construction

The use of Tripal as standard platform greatly simplifies the construction of biological databases that integrate large-scale genotypic and phenotypic data with genomic and genetic data. One critical next step to further expedite data integration involves development of novel methods to efficiently store and integrate the vast amounts of dense genotypic and resequencing data that are being generated from large populations. As an example, the USDA-ARS group at Geneva, NY, wishes to submit resequencing data for over 1000 Malus accessions to the GDR. Other groups within Rosaceae, cotton and citrus have indicated they will also be providing these type of data to our respective databases and we currently have no way to integrate them efficiently in Tripal. This is presented as an example of the type of demand that is being made across the cropping spectrum that must be addressed. Since major time and effort is being spent by both data providers and database curators to collect phenotypic data, change

formats when necessary and correct any errors introduced by manual recording, it is important to create efficient and broadly applicable ways of solving these problems. As one part of our effort in this, we will develop a tablet application or modify an existing one to collect phenotypic data in the field that allows direct uploading to the crop-specific Chado databases and will greatly reduce the time and effort needed for the data collection and management.

National issue: To promote interoperability and sharing of data among databases

Using ontology to describe traits and genes and providing programmatic access for data sharing between databases through web services will promote standardization of the data formats and interoperability of the associated analytical applications. It will improve the integration, preservation and utilization of data in and across various databases.

National issue: To promote community building

The application of molecular data in crop physiology and genetic improvement is greatly enhanced by a collaborative community of researchers, crop breeders, industry sector participants and extension professionals exchanging needs, ideas, and resources. In one recent example (RosBREED project, Iezzoni et al., 2010), the GDR served as both a scientific resource and a communication hub that propelled the development of an international, cohesive, well-organized and growing body of basic, translational and applied researchers. There are now elected steering committees at both the national (U.S. Rosaceae Genomics, Genetics and Breeding Committee) and international levels (Rosaceae International Genomics Initiative). The Rosaceae community has developed assets such as a priority-documenting White Paper; a stakeholder-driven technology roadmap; an annual Fruit and Nut Crops Workshop at the Plant and Animal Genome Conference; and a biennial International Rosaceae Genomics Conference. Rosaceae researchers across the world routinely partner on projects (RosBREED - Iezzoni et al. 2010, and FruitBreedOmics - www.fruitbreedomics.com) and share research data, often through the GDR. A similar role is emerging for the cotton database, CottonGen, which also houses communication resources as well as important data and tools. The community has a steering committee that meets every quarter. The database is presented at industry stakeholder conferences and the biennial cotton breeders tour to help ensure active participation from all types of users. This project would enhance community building across the 24 crops targeted immediately and to others that we anticipate will be joining in the combined effort.

Rationale

Priority Established by ESCOP/ESS

This NRSP proposal targets six of the seven grand challenge priorities:

Grand Challenge 1: Enhance the sustainability, competitiveness, and profitability of U.S. food and agricultural systems.

Superior cultivars can contribute directly to the profitability and sustainability of the agricultural sector. In the citrus industry, for example, the arrival and rapid spread of Huanglongbing (HLB), a bacterial disease also known as citrus greening and transmitted by psyllids, has caused thousands of acres of citrus in Florida to be abandoned and threatens other U.S. production areas. Variation exists for host

plant tolerance to the pathogen, but no sources of genetic resistance are known for most Citrus trees. The severity of this threat is underlined by the investments made by the Florida, Texas, and California citrus industries and federal funding agencies on research projects to better understand the disease and generate solutions. One foundational effort is to determine the sequence of Citrus genomes in order to probe genetic mechanisms underlying the disease process and to devise sustainable genetic solutions through the development of resistant rootstock and scion cultivars.

With the development of DNA-based screening technologies, more breeding programs can include genotyping in addition to phenotyping for performance evaluation. The integration of breeding data with other genomic and genetic data is required to develop and implement marker-assisted breeding tools, enhance genetic understanding of important crop traits and maximize access and utility by crop breeders and allied scientists. Breeders need the ability to search datasets in a targeted manner and retrieve and compare performance data from multiple selections, years and sites and then to output the data needed for variety release publications and patent applications. The Breeders Toolbox in GDR (Evans et al. 2013) enables exactly such activities. The GDR provides ready access to high-quality data of the genetic variation for traits of interest such as yield, fruit quality, abiotic and biotic stresses across the available germplasm pool, as well as specific tools for breeding decisions in parental and seedling selection. Using these data and tools, apple, cherry, peach and strawberry researchers identify optimal parental combinations to obtain the desired traits/trait levels in offspring with significantly enhanced efficiency and effectiveness. The project would develop the necessary Tripal Breeders Toolbox for target crops, provide similar enhancement for their breeding programs, and demonstrate the general utility of Tripal for other community databases.

Grand Challenge 2: Adapt to and mitigate the impacts of climate change on food, feed, fiber, and fuel systems in the United States.

The availability of comprehensive databases to provide access to germplasm evaluation and other breeding data by location and climate is crucial in developing new cultivars that are adapted to these environments. Integration of these breeding data with other genomic and genetic information such as expression and trait loci data will also help identify genes that are responsible for traits sensitive to climate change. The general threats to crop agriculture posed by climate change are well-known (Walthall et al. 2012). The specific impact on a crop or region can be devastating. For example, atypical low temperatures during bloom in 2012 destroyed up to 90% of the Michigan apple and tart cherry crop. On a regional basis, all the fruit bearing trees included in this proposal require a certain amount of winter chilling and even slight increases in winter temperatures on the West Coast would significantly imperil fruit and nut production. Integrated databases like the GDR can accelerate identification of genes affecting relevant plant traits and provide markers to facilitate the development of cultivars that require fewer chilling hours and/or are more tolerant of abiotic stresses. Similarly, researchers have been using the Cool Season Food Legume Genome Database to identify candidate genes or trait loci associated with drought or heat resistance in chickpea, pea and lentil and develop useful markers. This issue is of critical importance not only for U.S. producers, but for countries where these legumes provide an essential component of the human diet. Collaboration with scientists in these countries, facilitated by shared use of the Cool Season Food Legume Genome Database, has provided U.S. researchers with access to germplasm and associated data useful for introgressing traits from wild accessions into new cultivars of value for the U.S. Support from ICARDA led to allocation of $15,000 in 2013 to implement publicly available trait loci data for lentil. Further support is anticipated to house phenotypic and genotypic data from their lentil and chickpea breeding programs.

Grand Challenge 3: Support energy security and the development of the bioeconomy from renewable natural resources in the United States.

The provision of an easy to use, manageable, standardized platform for genome database development also enables databases to be developed with organismal data relevant to support energy security based on renewable natural resources. Many crops of potential use completely lack access to well-curated genome databases. For example, in collaboration with researchers from the University of Hawaii, the GDR team is developing a Pearl Millet NapierGrass Hybrids Breeding Database using Tripal. It will provide access to breeding-decision tools to enable comparison of parental and selection evaluation data that cannot be readily done in any free breeding data management software. Outcomes of this proposal include the essential infrastructure and expertise to have similar impact on crop breeding programs for high-yielding, low input bioenergy feedstocks.

Grand Challenge 4: Play a global leadership role to ensure a safe, secure, and abundant food supply for the United States and the world.

Every crop targeted in this proposal plays a role in food, feed, and fiber supply of the U.S. and the world. The community databases identified for further development will each enable domestic and international research programs to exploit the genomic, genetic and breeding resources made readily and widely available. The databases themselves are mostly crop-specific, but the software infrastructure and expertise to be developed in this proposal are explicitly intended to be exportable to other crops or organisms. Additionally, project participants will further solidify the significant place the current databases have gained as internationally recognized resources.

Grand Challenge 5: Improve human health, nutrition, and wellness of the U.S. population.

All of the specialty fruit and nut crops housed in the Rosaceae, Citrus and Vaccinium databases and the crops housed in the cool season food legume databases contribute significantly to a health-giving and nutritious human diet. Breeding programs in the U.S. will gain considerable advantages in efficiency and effectiveness if we can provide ready access to high quality integrated genomics, genetics and breeding databases. Further, converting and enhancing the existing breeding-decision support tools as Tripal applications will make them available to a much wider audience and encourage adoption of the platform by other crop or biological communities.

Grand Challenge 6: Heighten environmental stewardship through the development of sustainable management practices.

We will develop and curate the databases dynamically and in close collaboration with the individual research communities. This will give researchers steadily improving access to all the current data and literature on traits or markers associated with genetic tolerance or tolerance to critical biotic and abiotic stresses that currently require application of pesticides, other agricultural chemicals, or water resources. The databases will play a significant role in the provision of an integrated knowledge base for researchers to help understand the underlying biology of these traits, develop the genetic tests for marker-assisted breeding and guide multi-trait breeding strategies.

Relevance to stakeholders

The current stakeholders of the databases described in this NRSP include biologists, breeders, bioinformaticists, educators, and the industries based on the underlying crops. The databases will store and integrate data from various research projects, funded by government and industry, accelerating knowledge discovery from the integrated information, and maximizing the return on the genomics, genetics and breeding investment. As detailed in the management plan, the primary stakeholders from research institutions and industry will participate in project development and assessment as members of the steering committee for each database. Their participation will help ensure the development effort is prioritized to meet the needs of stakeholders. The ultimate stakeholders of this NRSP are consumers and U.S. taxpayers. Below is a detailed description of the benefits for each type of stakeholder.

Biologists:

The curated genomic and genetic data and analysis tools available in the databases developed in this NRSP will help basic biologists who are interested in the structure and evolution of genomes, gene function, genetic variability and the mechanisms underlying various traits. The integrated genomic and genetic data will help translational scientists who are interested in further QTL and marker discovery and genetic mapping studies. The integrated genomic, genotypic and phenotypic databases will also help applied scientists who are interested in developing methods for marker-assisted breeding. The Rosaceae, citrus and Vaccinium databases will have data from multiple species, which will enable transfer of knowledge among related species as well as studies on genome evolution. The molecular diversity data and germplasm data will help scientists to go beyond the well-known gene pools to explore other experiments to achieve their goals. The consistent interfaces will promote cross-utilization between communities by decreasing the time required to master each database.

Breeders:

The recent addition of an extensive breeding database in the GDR enables breeders to search integrated breeding-genetic-genomic datasets for apple, peach, cherry and strawberry in a fully targeted manner and to retrieve and compare performance data from multiple varieties and seedlings, years and sites. This facilitates the streamlining of selection decisions and output of data needed for variety release publications and patent applications. The GDR also provides breeding decision tools for parent selection, seedling selection and marker refinement. This NRSP will enable the same functionality in other databases. It will also result in availability of a tablet application either through modification of an existing application or development of a new one for data collection that will significantly help breeders and scientists reduce the time and cost of phenotypic evaluation. In this environment, breeders will have the option to keep their data in a private database while also linking to all relevant public data. We will implement the breeding decision tools currently available in the GDR for other crop databases developed under this NRSP. Users will be able to upload their private data to use the analysis tools without storing them in the database. The availability of the tablet application to collect data is expected to facilitate the standardization of phenotypic evaluation methods and development of collaborative research projects on cultivar evaluation and breeding. Programs using the commercial AgroBase breeding management software will be able to upload their data files directly to the online database to enable data mining capability. A prototype is currently being developed for the wheat and barley breeders at WSU.

Bioinformaticists:

The activities in this NRSP will further develop Tripal as a freely available genomic, genetic and breeding database construction platform. This NRSP will add new components to Tripal for compression and storage of large scale genotypic and re-sequencing data into Chado, additions to the Tripal API, and a tablet application to collect and upload large scale phenotypic data sets to the database. These improvements will be greatly beneficial for bioinformaticists as they build databases housing large-scale genomic, genetic and breeding data for their respective research communities. The web services will help other bioinformaticists who require programmatic access to large-scale data to integrate with other databases or perform further analyses.

Educators:

Comprehensive tutorials, screen casts and videos developed for uses of each database will be useful in formal or informal, classroom or distance learning contexts. Development of databases offers an ideal opportunity for graduate student education. Specific objectives include: 1) establishing cooperative Community-of-Practice-like interactions between the crop-specific curators and the core personnel that encourages the appropriate development and extension of resources; 2) small-group, face-to-face workshops in Pullman, WA to facilitate transfer of expertise between participants and common establishment of priorities; 3) focused teaching modules for use in classroom or web-based delivery platforms that familiarize users and potential users with the resources that are available and help train them in the possibilities available to them.

Consumers:

Consumers will be provided with higher quality, more nutritious fruits, vegetables and staple crops as a result of the use of the output of this NRSP.

Crop Production Industry:

Several industries have provided funding and are participating in building and populating our target databases, led by key commodity associations (rosaceous tree fruit, cotton, citrus, and dry pea and lentil). This NRSP will ultimately benefit each of these crop industries since it will significantly enhance current, highly-utilized databases and expedite the development of cultivars that are competitive, profitable, sustainable and climate-adaptable. Our core strategy will allow additional crops to be added to the list on demand as support for further crop-specific efforts is generated. Establishment of this NRSP should substantially lower the expenses needed for a commodity to establish a database and the associated analytical tools.

Implementation

Objectives

1. Expand online community databases for Rosaceae, citrus, cotton, cool season food legumes and Vaccinium crops:

   We will continue ongoing curation to integrate new genomic, genetic, genotypic, phenotypic and germplasm data in all community databases. For example, in the GDR, we will add all the data from the large European FruitBreedOmics project (www.fruitbreedomics.com). We will also

introduce software innovations first developed in the GDR to other community databases. New or updated genome sequence and annotation data will be added to the databases with additional computational analysis performed to identify predicted genes with homology to known genes in other databases (our standard analysis pipeline). Whole genome sequences will also be used to construct orthologous regions among closely related genomes. PlantCyc metabolic pathway databases will be displayed using GBrowse_Syn (McKay et al. 2010) and PathwayTools (Paley et al. 2012). Genetic data will continue to be integrated and the trait loci will be associated with the Trait Ontology (TO) terms (Jaiswal et al. 2002), with new terms added as necessary. Large scale phenotypic and genotypic data will continue to be integrated. Within the framework of the NRSP, these types of targets may be nominated by a specific crop but will be chosen for development mostly on the perceived need for comparable resources by the NRSP crop communities

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

   A tablet application will be developed for field and laboratory studies to record phenotypes, take pictures and submit to the appropriate database. Users will be able to upload site information, trait descriptors, dataset names, germplasm, new observation values and comments. The app will have functionality to associate pictures with each observation of germplasm and will allow users to temporarily store data in the tablet for subsequent upload to the appropriate database. We will also add an option to store data from the tablet in the cloud so that users can keep the data until they are ready to upload to the database. We will review open-source tablet software applications, such as the one developed for maize by the Dr. Ed Buckler Lab at USDA (http://www.maizegenetics.net/field-informatics), to see if we can modify to work with Chado

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

   GDR has interfaces to search breeding data and decision tools such as Marker Converter and Cross Assist to help breeders refine markers and support crossing decisions, respectively. We will convert and expand functionalities of these interfaces into Tripal modules for transfer to our other databases or other Tripal-based databases

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

   GenSAS (Main et al. 2013c) is a web-based Genome Sequence Annotation Server that provides a one-stop website with a single graphical interface for running multiple structural and functional annotation tools, enabling visualization and manual curation of genome sequences. The availability of a single web application where users can combine analyses and curation for their locus of interest will accelerate the refinement of whole genome sequence data by community experts. In this NRSP, we will develop a Chado exporter for GenSAS to allow the genome curator to export completed structural and functional annotations into Chado. GenSAS in Tripal provides a complete whole genome annotation and visualization platform for any research community

5. Develop Web Services to promote database interoperability:

Web services will be enabled for retrieval of sequence and annotation data by applying the most commonly used technologies for Web Services such as Representational State Transfer (REST) and the Simple Object Access Protocol (SOAP). We will follow developing standards and recommendations from BioHackathon, an organization who represent the major biological databases and cyber-infrastructure projects worldwide (Katayama et al. 2013)

Projected Outcomes

1. Database resources that facilitate utilization and exchange of data and information among researchers across disciplines for Rosaceae, citrus, cotton, cool season food legumes and Vaccinium:

   (a) A regularly updated GDR and significantly enhanced citrus, cotton, cool season food legume and Vaccinium databases containing integrated up-to-date genomic, genetic and breeding data and GDR data mining and analysis capabilities.
   (b) An open-source tablet app for phenotype data collection to accelerate data transfer from breeders to community databases as well as enhances efficiency and accuracy of data collection.
   (c) An advanced Marker Converter tool for researchers to optimize marker efficiency.
   (d) An advanced Cross Assist tool identifying most efficient parental crosses to increase the probability of achieving desired traits in offspring.
   (e) Web services enabled for retrieval of genomic sequence and annotation data so bioinformaticists can programmatically access the data for use in other software or integrate with other databases.

2. An integrated genomics, genetics and breeding open-source database construction platform for building other biological databases:

   (a) Open-source tool for database construction, Tripal, with enhanced functionality.
   (b) Open-source genome sequence annotation server, GenSAS, fully compatible with Chado and Tripal.
   (c) Suite of training resources facilitating database construction and database use, and understanding of the importance of genomics, genetics and breeding to U.S. agricultural production and food security.

3. Augmented Trait Ontology to describe trait loci and germplasm:

   (a) An augmented trait ontology list that extends current terms (which mainly apply to grasses) to include traits of rosaceous and Vaccinium crops, food legumes and cotton.
   (b) Association of data with Trait Ontology promoting data sharing and interoperability.

4. Community databases promote community building by acting as communication hubs:

   (a) Enhanced collaboration and coordination of specific research and extension programs facilitated by access to data and communication tools in the target crop databases.
   (b) More exchange of ideas, data and tools within the individual communities and among communities through the standardization of databases.

(c) More standardization of protocols, nomenclatures and methodology to enhance data transfer.

Management, Budget, and Business Plan

Management

Our team will comprise the project director (Dr. Dorrie Main), lead curator (Dr. Sook Jung), lead Tripal software engineer (Dr. Stephen Ficklin Tablet application development and Web Services Implementation), a database/system administrator (Chun-Huai Cheng), a programmer (Taein Lee - GenSAS Tripal Conversion and Breeding Database Tripal Conversion), a data analyst (Dr. Ping Zheng) and two or more curators (Dr. Jing Yu, Dr. Julia Piaskowski) depending on the support from industry and research industry stakeholders. Each curator will have assigned specific crop database(s). The lead curator will meet weekly with any newly hired or collaborating curators to train them on scientific curation procedures and check progress. Subsequently, curators and the project director will meet monthly to set up goals, check progress and discuss any topics related to curation. The data curation and analysis procedure is very similar regardless of the crop, and this management system will ensure the expertise on data curation is transferred efficiently. The meetings will also allow the specific knowledge gained dealing with a specific crop be transferred to other crop curators.

The project director will oversee design of new software components by the lead curator and lead programmer. Subsequently, programmers have responsibility for coding, following Drupal coding styles and existing Tripal API. When the beta version of software is finished, the appropriate curators and other programmers will test the software. As is current practice, the entire project team will meet weekly using online teleconferencing to discuss tasks, timelines and progress. For the tablet application development, we will form a specific committee composed of the project director, the lead curator/scientific software designer, lead programmer, main programmer and at least one breeder from each database. The committee will meet monthly without the breeders but meet quarterly with breeders from year 2. As stated in the objectives section, we will review any open-source tablet application for collecting phenotypic data, such as the one developed for maize by Dr. Ed Buckler s lab at USDA or the wheat communities Field Book App, to see if it is possible to modify to work with Chado. The lead curator will work with breeders to provide sample data and design the functionality of the application. The lead programmer will work with the main programmer to design the software and the actual coding will be done by the main programmer. The lead curator and the breeders will serve as alpha and beta testers when the application is developed. To enhance community participation and project accountability we will adopt the successful cotton model for all target databases. This structure has ensured the database team is apprised of new research data and can then adapt to community needs regarding functionality and content of the database.

Each database will have a steering committee comprising appropriate representatives nominated or elected by the research community and industry stakeholders. The steering committees will each meet at least quarterly to review progress and decide on tasks for the next quarter. Specific benchmarks will be decided in consultation with the steering committees for each database at the start of the project and reviewed quarterly. The steering committees will also facilitate the standardization of various data, such as phenotypic data collection method, trait descriptors, gene naming and SNP naming, for their community using the expertise gained from the likes of the RosBREED project. RosEXEC, serving the role of the steering committee for GDR, has two active subcommittees for standardizing gene naming and

SNP naming. When the community comes up with standardized trait descriptors and scales, we can store the breeding data that have used different scales in both standard and breeder-specific scales, if applicable. The schema allows us to store using different scales, and users will be able to view the original value as well as the standardized value if they want to compare with data from other projects.

Tripal has been developed and funded by various projects from different institutions, including: Clemson Univ., WSU and the Univ. of Saskatchewan. Each of these groups will continue contributing to both core and expanded Tripal modules. We also expect to add other groups, e.g. Dr. Steve Cannon at the USDA-ARS in Ames, Iowa (legumes), and one of the instigators of Tripal, Dr. Meg Staton at the Univ. of Tennessee, (hardwoods) and will invite participation by iPlant personnel. For the long-term maintenance and sustainability of the Tripal core modules, the current developers have agreed to form a non-profit entity called the Tripal Foundation. This entity will ensure consistency in development of Tripal core modules and allow any Tripal stakeholder to voice opinions on development regardless of the funding state of the stakeholder. Changes to the Tripal core modules will be approved by a committee from this entity, and membership in this committee will be open to interested stakeholders. All interested Tripal developers will meet in face-to-face meetings once a year and quarterly by GoToMeeting. This will be organized and facilitated through the Tripal Foundation.

Timeline

Year 1:

(1) Collect and curate genomic, genetic and breeding data for all databases(2) QTL and genetic map curation up-to-date for GDR (3) Implement all currently available Tripal pages in citrus, legume and Vaccinium databases (4) Develop Tripal API for breeding database and implement in GDR (5) Develop webinars for Tripal, GDR and CottonGen (6) Ensure all module development is to Tripal standard and make publicly available (7) Update all database tutorials.

Year 2: (1) Curate genomic, genetic and breeding data for all databases (2) QTL and genetic map curation up-to-date for cotton and citrus databases (3) Design the breeding data app interface and functionality (4) Implement breeding data interface in Tripal in databases for cotton and citrus (5) Convert GenSAS to Tripal; (6) Ensure all module development is to Tripal standard and make publicly available (7) Develop web services for data retrieval in the GDR (8) Update webinars for Tripal, GDR and CottonGen and develop webinars for the Cool Season Food Legume Genome, Citrus Genome and Vaccinium Genome Databases (9) update all database tutorials.

Year 3: (1) Curate genomic, genetic and breeding data for all databases (2) QTL and genetic map curation up-to-date for legume and Vaccinium databases (3) Test breeding data app before making it available (4) Implement breeding data interface in Tripal in databases for cool season food legume and Vaccinium (5) Implement GenSAS in Tripal in databases for cotton and citrus (6) Update all webinars and tutorials (7) Ensure all module development is to Tripal standard and make publicly available.

Year 4: (1) Curate genomic, genetic and breeding data for all databases (2) Implement breeding data app (3) Further develop the currently available tool for converting markers, Marker Converter, in GDR (4) Further develop the currently available tool for parent selection, Cross Assist, in GDR (5) Implement GenSAS in Tripal in cool season food legume database (6) Ensure all module development is to Tripal

standard and make publicly available (7) Implement web services in the remaining databases (8) Update all webinars and tutorials.

Year 5: (1) Curate genomic, genetic and breeding data for all databases (2) Further assess breeding data app and provide a newer version if necessary (3) Implement Cross Assist and Marker Converter in databases for cotton, citrus, cool season food legumes and Vaccinium (4) Ensure all module development is to Tripal standard and make publicly available (5) Update all webinars and tutorials.

Budget

A total of $1,991,190 (Table 2) is requested over five years from this NRSP to support the core database development activities described in this proposal. Additional funding of $2,166,942 from aligned objective support (Table 3) is projected over the course of this project from WSU ($1,319,275) and Industry/Regional/Federal grants ($847,667). This includes funds for cotton from industry, USDA-ARS and SAES stakeholders; the Washington Tree Fruit Research Commission for tree fruit data analysis and curation; a subaward to WSU from UC Davis for annotation editing software development; the USA Dry Pea and Lentil Council for pea and lentil data analysis and curation; and the WA SCRI Block program for development of a pea breeders toolbox. Combined with funds requested from NRSP, this totals $4,158,132 for the complete project. We anticipate additional aligned support as the project progresses but this budget only includes committed and currently available funding.

Business Plan

We seek to replace prior ad hoc funding of critical crop databases with a more sustainable two-stage model. This two-stage model will include (1) support for core database activities from this NRSP and (2) funding for data curation and analysis activities from industry stakeholders and regional/federal grant competitive sources.

Development and maintenance of the current Rosaceae, citrus, cotton, cool season food legume and Vaccinium databases targeted in this proposal have been funded to date (Table 3) via numerous sources, including: NSF Plant Genome Program (award # 0320544 - Rosaceae), USDA NIFA SCRI program (award #2009-51181-06036 - Rosaceae, and award # 2009-51181-05808 - Rosaceae and citrus), the Washington Tree Fruit Research Commission (tree fruit), Cotton Incorporated (cotton), USDA-ARS (cotton, cacao), the Southern Association of Agricultural Experiment Station Directors (cotton), Dow AgroSciences (cotton), Monsanto (cotton), Bayer Crop Science (cotton) the Citrus Research Board (citrus), the USA Dry Pea and Lentil Council (pea and lentil), and several universities for database development and/or specific data curation efforts. Universities include WSU, Univ. of Florida, Clemson Univ., Boyce Thompson Inst. of Plant Sciences, North Carolina State Univ., Michigan State Univ., Univ. of Minnesota, Cornell Univ., Univ. of Arkansas, Univ. of New Hampshire and Texas A&M Univ. This funding was effective to develop the target databases and Tripal to their current level of functionality.

This model leveraged numerous funding sources to permit creation and maintenance of these five databases and simultaneously advanced development of a common platform for genome database construction. However, we are convinced that this is not a sustainable model. In particular, federal funding agencies are unlikely to continue supporting community database development and maintenance once the pioneers have been developed and the concepts have been developed. Nowhere is this more evident than in the withdrawal of funds from the Arabidopsis thaliana database by the NSF

(Lamesch et al. 2011). This database has long been the international standard in plant functional annotation and has been utilized and referenced by many other plant databases, including our own.

We propose a new, split-function model to address community database development, maintenance and enhancement in a sustainable manner. We will separate the two major activity areas: 1) core database and functionality development; and 2) crop-specific database curation and enhancement. We expect to take the Tripal platform and our current databases to a level where all the major functionality that we believe our users will need is completed within the duration of this proposed project. Core development activities would be supported by funds from this NRSP and would include costs associated with database administration, development, data storage, IT server room space rental, server service contracts, data backup and support desk help for other Tripal adopters. It would also cover the cost to publish two peer-reviewed open access manuscripts per year and travel for one person to four scientific meetings a year. Funding to update computational database and web servers in years 1 and 3 are requested to ensure we can meet escalating demand for fast, efficient database access. As benchmarks, we plan to have non-NRSP funds equivalent to those needed to support 25% of the core database activities by the end of year 3 and to support 50% of the core database activities by the end of year 5. For our second major area of activity -- data curation and analysis support will be sought through industry and researcher stakeholders (Table 3: projected other support). We will explore and discuss other longer term funding options with our communities through our database steering committees. Options include their willingness to pay a fee to deposit their data in the appropriate community database (similar to the concept of paying to publish in open access journals) and/or charge an access fee to use the database, similar to that recently announced by The Arabidopsis Information Resource. An alternative is moving some or all of the databases to the USDA-ARS (as we did for the cacao genome database) on completion of this proposed five year project. Any such expansion of agency activity would, however, require significant resources and overall budget increases for the USDA-ARS.

The value of the two-stage funding system we are proposing is as follows: (1) Community database development builds on the core infrastructure funded by this NRSP and is very scalable. Commodity associations can determine a level of investment they think is appropriate and directly provide funds for full-time or part-time curators as needed; 2) Should additional crop groups seek development of new databases, an appropriate surcharge for core development would be necessary; (3) With the core infrastructure component funded for five years, crop-specific curator positions can be housed where the expertise resides for the crop of interest, a system that has worked very well to date. While project direction resides at WSU, Pullman WA, the cotton curator is housed in Texas, and the peach curator is housed in South Carolina; (4) Initially housing the core development team where the current expertise in Tripal database development exists (WSU, Pullman WA), enables more efficient tool development, data uploading and troubleshooting of problems during the critical five years it will take to fully develop Tripal into a comprehensive genomics, genetics and breeding platform; (5) For additional crop participants, easy access to working models and community expertise will substantially lower the risk and initial and recurring costs of participation.

Integration

This NRSP is highly integrated with academic and government research programs and is stakeholder-driven (as evidenced by the project participant list and supporting letters). Providing access to collated, curated, and integrated public genomics, genetics and breeding data will enable unanticipated scientific advances well beyond the research for which the data were originally collected. The enhanced databases will provide a catalytic environment where genomicists, geneticists, bioinformaticists,

breeders and growers can share data and ideas to elevate trans-disciplinary understanding of their crops, to suggest compelling directions for new methods and research, and to produce efficient and focused practical steps toward the common goal of crop improvement. Specifically, integration of the genetic, genomic and breeding data will enable members of the crop improvement community to develop new scientific hypotheses and theoretical models and test these using an appropriate database and tools. The results will improve understanding of the fundamental biology underlying the crops and their valuable traits. By integrating genetic data (such as quantitative trait loci, genetic markers, and pedigrees) used to make and populate genetic maps with genomic data (genome sequences, chromosomal physical arrangements, sequence variants from large populations, and gene expression measurements), genomics-genetics-breeding translational biologists will be able to develop better tools and knowledge for developing improved cultivars.

In addition to providing integrated databases for individual research programs worldwide, the database team has also been involved with several extension and academic programs in more direct ways. Both the GDR and the Citrus Genome Database were supported from 2009-2012 through an USDA NIFA-funded $3.99 million award. The GDR team was also integral to the 2009-2012 USDA NIFA SCRI-funded $14.2 million landmark project RosBREED: Enabling marker-assisted breeding in Rosaceae, providing that project with genomic and genetic data analysis and development of a full suite of online database and analysis tools to enable marker-assisted breeding in rosaceous crops (Chagne et al. 2012; Peace et al. 2012; Verde et al. 2012). The GDR team has also participated in other major research initiatives, providing both sequence analysis and database support to projects such as the International Peach Genome Initiative, the USDA-ARS/MARS-funded Cacao Genome Sequencing Initiative, the Murdoch Institute and NC State funded Blueberry Genome Sequence Initiative, the $15 million USDA-funded loblolly pine, sugar pine and Douglas fir reference genome projects, and the International Pea and Lentil Genome Sequence Consortiums. Three landmark scientific publications were recently co-authored by the GDR team (peach genome - IPGC. 2013; cacao genome Montamayor et al. 2013; Rosaceae ancestral genome Jung et al. 2012).

Through this NRSP, we hope to systematize the effort and use the experience acquired by the several crop research communities to maintain and improve existing databases and lower the barriers to entry for the construction of new ones. Industry stakeholders have been directly involved with the database development as funding agencies, users and members of advisory groups. The Washington Tree Fruit Research Commission (WTFRC) funded development of an apple and cherry cultivar performance database and toolbox built upon GDR resources originally funded by the NSF. This hybrid model has also worked well for establishing the new cotton, cacao and cool season food legume databases. Breeders and growers, as well as researchers, are integral to database development and use, and their participation and feedback is continually encouraged and actively solicited. In addition, a database representative is a permanent member of the elected steering committees at both the national (U.S. Rosaceae Genomics, Genetics and Breeding Committee, RosEXEC) and international level (Rosaceae Genomics Initiative, RosIGI). The GDR team is also leading a subcommittee of RosIGI, the Rosaceae Gene Naming Standardization Subcommittee, formed to develop a standard protocol for naming genes in Rosaceae and initiate a community-driven gene annotation effort through the GDR. A publication is in preparation to document this effort. A similar effort is also underway to develop a community-accepted SNP naming policy. Stakeholders from university, government and industry sectors from all the production and research regions of the U.S. are well represented in RosEXEC and are considered the de facto steering committee for the GDR. The GDR is home to registration and submission of abstracts for the 7th International Rosaceae Genomics Conference, organized by WSU and the WTFRC, to be held in Seattle in June 2014. For CottonGen, the steering committee, also composed of representatives from

universities, government and industry, meets quarterly to communicate the current and emerging database needs of the cotton research community and other stakeholders for development, implementation and dissemination of CottonGen.

Outreach, Communications and Assessment

Outreach

Our outreach effort will focus on both communities that we develop databases for and those who may be in need of such databases. With increasing data, a lot of underserved crop communities are in need of developing their databases and some communities with existing databases are searching for some alternative platforms to handle large-scale genotypic and phenotypic data. We will also approach iPlant to ascertain how they might become involved in this effort.

Training workshops will be held at the annual Plant and Animal Genome Conference (PAG) beginning in 2015our efforts will include continuing to apply yearly to the PAG computer demonstration session to hold training sessions for each of the databases and participating in the Plant Genome Database Booth. This training will show how to use the database and solicit feedback from the attendees. The workshops will be advertised via the mailing lists, newsletters, and Plant Workshop Sessions, to ensure maximum community participation. Participants for the workshops and computer demonstration sessions are composed of researchers from many other crop communities, these presentations at PAG will be a good opportunity to present the utility of the platform we develop. We will also present our databases at the annual American Society for Horticultural Science Conference, the Crop, Soil and Agronomy (ACS) annual meeting, and the annual Plant Biology meeting to reach researchers from additional communities. The databases will also be presented and training given where possible at the biennial conferences for each of the target database communities. Specifically this will include the North American Pulse Improvement Association in 2015, 2017; The International Cotton Genome Conference in 2014, 2016, 2018; The annual Cotton Beltwide Conference and biennial Cotton Breeders Tour; The International Rosaceae Genomics Conference in 2014, 2016, 2018; the annual North American Blueberry Research and Extension Workers Conference; the annual Citrus Genomics Conference. Responsibility for some of this will lie with the crop-specific curators, with core NRSP participation to keep the development aligned with stakeholder interests.

In addition to presentations at conferences, we will continue to update tutorials on how to use databases, develop webinars, maintain mailing lists and continue to publish when significant development has been made.

To reach database developers, Tripal will continue to be presented and training provided yearly at Generic Model Organism Database workshops at PAG and the GMOD Summer School. We will continue to provide tutorials and mailing list responses in collaboration with other Tripal developers. When this NRSP is funded, a dedicated help desk person for Tripal will help developers with questions. We will continue to participate in the Chado mailing list to provide examples of how we store data when questions arise. We have co-authored three publications on Chado and Tripal so far, and one publication on how we store data in Chado is in preparation. We will continue to provide tutorials and publish when we produce new modules and applications that run on top of Tripal and Chado.

Additionally, it is very important that breeders, stakeholders and the general public are trained about the usefulness of their community databases and functionalities the might be incorporated from other databases. Within the website a public component will be added, highlighting the utility of the databases in successful research stories that impact consumers, explaining the terminology at an appropriate level and including database training from the workshops. These extension and outreach activities will greatly enhance the research and extension programs of cotton, legumes and horticultural specialty crops. Utilizing the database in the process of breeding superior plant selections will distinguish the United States amongst others in the world market for those crops. In addition, in association with the continued growth of cotton, legumes and horticultural specialty crop production, it will lead to more jobs with higher incomes, which in turn will create economic development and prosperity, enhancing the quality of life in rural areas.

Communication

In addition to the workshops mentioned above other conduits to facilitate communication between the database developers and the users are needed. While we envision that implementation of each database will have significant individuality, we also will expect each database that associates with the NRSP to have a steering committee, composed of representatives from universities, government, consumer groups and industry for each crop, which will meet quarterly by online teleconferences to communicate the current and emerging database needs of their research community with other stakeholders and the NRSP staff to guide the development, implementation and dissemination of resources for the database. Any new major development such as the tablet app for phenotypic collection will be extensively discussed in these committee meetings. The meeting minutes will be posted on the NRSP and crop-specific websites as well as all quarterly reports and quarterly work plans. We will also have quarterly newsletter, twitter and LinkedIn accounts to notify the users of any new developments in the database and the crop research community.

Assessment

Various metrics will be used to assess the impact of the proposed project. For each database this will include usage statistics as measured by google analytics, feedback from the steering committee, annual online surveys of each community, number of publications, number of publications citing the databases and feedback via the online forms in the databases. Creation of new Tripal databases, number of projects adopting Tripal, and number of active developers with this project will all be indicators of the success of the use of Tripal. Very relevant measures of participation will be the number of curators and crops associated with the NRSP and the level of investment in NRSP by the users. Achieving our benchmark external funding for core database activities after 3 (25%) and 5 (50%) years with the goal of 100% funding (and the end of NRSP support) by year 10 will be further indicators of measurable success.

Projected Participation

Budget Requests Summary

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, Pujar A, Leto J, Gosselin J, Mueller LA. (2011) The sol genomics network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Res. 39:D1149-D1155.

Carollo V, Mathews DE, Lazo GR, Blake TK, Hummel DD, Lui N, Hane DL, Anderson OD (2005) GrainGenes 2.0. An improved resource for the small-grains community. Plant Phys. 139(2) 643-651

Chagné D, Crowhurst RN, Troggio M, Davey MW, Gilmore B, Lawley C, Vanderzande S, Hellens RP, Kumar S, Cestaro A, et al. (2012) Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. PLoS One 7, e31745.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 2, 3674-3676.

Evans K, Jung S, Lee T, Brutcher L, Cho I, Peace C, Main D. (2013) Addition of a breeding database in the Genome Database for Rosaceae. Database (Oxford) bar078.

Ficklin SP, Sanderson L, Cheng CH, Staton ME, Lee T, Cho IH, Jung S, Bett KE, Main D. (2011) Tripal: a construction toolkit for online genome databases. Database (Oxford) bar044.

Gene Ontology Consortium. (2013) Gene Ontology annotations and resources. Nucleic Acids Res. 2013, 41, D530-535.

Grant D, Nelson RT, Cannon SB, Shoemaker RC. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38, D843-D846.

Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res. 40, D306-12.

Iezzoni A, Weebadde C, Luby J, Yue C, van de Weg E, Fazio G, Main D, Peace CP, Bassil NV, McFerson J. (2010) RosBREED: Enabling marker-assisted breeding in Rosaceae. Acta Hortic., 859, 389394.

International Peach Genome Initiative. (2013) The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat. Genet. 45, 487-94.

Jaiswal P, Ware D, Ni J, Chang K, Zhao W, Schmidt S, Pan X, Clark K, Teytelman L, Cartinhour S, et al. (2002) Gramene:Development and integration of trait and gene ontologies for rice. Comp. Funct. Genom. 3: 132136.

Jung S, Jesudurai C, Staton M, Du ZD, Ficklin SP, Cho I-H, Abbott AG, Tomkins J, Main D. (2004) GDR (Genome Database for Rosaceae): integrated web resources for Rosaceae genomics and genetics research. BMC Bioinformatics 5:130.

Jung S, Staton ME, Lee T, Blenda A, Svancara R, Abbott AG, Main D. (2008) GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. Nucleic Acids Res. 36:D1034-1040.

Jung S, Menda N, Redmond S, Buels RM, Friesen, M, Bendana Y, Sanderson LA, Lapp H, Lee T, MacCallum B, et al. (2011) The Chado Natural Diversity module: a new generic database schema for large-scale phenotyping and genotyping data. Database (Oxford) bar051.

Jung S, Cestaro A, Troggio M, Main D, Zheng P, Cho I, Folta KM, Sosinski B, Abbott AG, Celton JM, et al. (2012) Whole genome comparisons of Fragaria, Prunus and Malus reveal different modes of evolution between Rosaceous subfamilies. BMC Genomics. 13, 129.

Jung S, Ficklin SP, Lee T, Cheng C-H, Blenda A, Zheng P, Yu J, Bombarely A, Cho I, Ru S, et al. (2013) The Genome Database for Rosaceae (GDR): year 10 update. Nucleic Acids Res. gkt1012.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. (2012) KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res. 40, D109-D114.

Katayama T, Wilkinson MD, Micklem G, Kawashima S, Yamaguchi A, Nakao M, Yamamoto Y, Okamoto S, Oouchida K, Chun HW, et al. (2013) The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies.J Biomed Semantics, 4:6.

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. (2011) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40, D1202-10.

Lawrence CJ, Harper LC, Schaeffer ML, Sen TZ, Seigfried TE, Campbell DA (2008) MaizeGDB: The Maize Model Organism Database for Basic, Translational, and Applied Research. Int J Plant Genomics. 496957.

Main D, Jung S, Ficklin SP, Zheng P, Cheng C-H, Olmstead M, Abbott AG, Blenda AV, LeeT, Chen C, et al. (2013a) Citrus Genome Database: Updates and New Functionality. In: Plant and Animal Genome Conference XXI: Jan 12-16, San Diego, CA; 2013.

Main D, Cheng C-H, Ficklin SP, Jung S, Zheng P, Lee T, Coyne C, McGee RJ, Mockaitis K. (2013b) The Cool Season Food Legume Database: An Integrated Resource for Basic, Translational and Applied Research In: Plant and Animal Genome Conference XXI: Jan 12-16, San Deigo, CA; 2013.

Main D, Lee T, Zheng P, Jung S, Ficklin SP, Humann J, Wegrzyn J, Neale DB. (2013c) GenSAS: A Genome Sequence Annotation Server, a Tool for Online Annotation and Curation. In: Plant and Animal Genome Conference XXI: Jan 12-16, San Deigo, CA; 2013.

McKay SJ, Vergara IA, Stajich JE. (2010) Using the Generic Synteny Browser (GBrowse_syn). Curr Protoc Bioinformatics Chapter 9:Unit 9.12

Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, Cornejo O, Findley SD, Zheng P, Utro F, Royaert, et al. (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biology 14:R53.

Mungall CJ and Emmert DB. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics, 23, i337–i346.

National Agricultural Statistics Service (2013) Annual Crop Summary Report.

Paley SM, Latendresse M, Karp PD. (2012) Regulatory network operations in the Pathway Tools software. BMC Bioinformatics. 13, 243.

Peace C, Bassil N, Main,D, Ficklin S, Rosyara UR, Stegmeir T, Sebolt A, Gilmore B, Lawley C, Mockler TC, et al. (2012) Development and evaluation of a genome-wide 6K SNP array for diploid sweet cherry and tetraploid sour cherry. PLoS One 7, e48305.

Sanderson LA, Ficklin SP, Cheng C-H, Jung S, Feltus FA, Bett KE, Main D. (2013) Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. Database (Oxford) bat075.

Stein, LD, Mungall C, Shu SQ, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. (2002) The Generic Genome Browser: A building block for a model organism system database. Genome Res. 12, 1599–1610.

The UniProt Consortium Update on activities at the Universal Protein Resource (UniProt) in 2013 (2013) Nucleic Acids Res. 41, D43-D47.

Verde I, Bassil N, Scalabrin S, Gilmore B, Lawley CT, Gasic K, Micheletti D, Rosyara UR, Cattonaro F, Vendramin E, et al. (2012) Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. PLoS One 7, e35668.

Walthall CL, Hatfield J, Backlund P, Lengnick L, Marshall E, Walsh M, Adkins S, Aillery M, Ainsworth EA, et al. (2012) Climate Change and Agriculture in the United States: Effects and Adaptation. USDA Technical Bulletin 1935. Washington, DC. 186 pages.

Youens-Clark K, Faga B, Yap IV, Stein LD, Ware D. (2009) CMap 1.01: A comparative mapping application for the Internet. Bioinformatics, 25, 3040-3042.

Yu J, Kohel R, Hinze L, Yu JZ, Frelichowski J, Ficklin SP, Main D, Percy RG. (2012) CottonDB. In: Plant and Animal Genome XX: Jan 14-18, 2013 2012; San Diego, CA USA; 2012.

Yu J, Jung S, Cheng CH, Ficklin SP, Lee T, Zheng P, Jones D, Percy R, Main D. (2013) CottonGen: A Genomics, Genetics and Breeding Database for Cotton Research. Nucleic Acids Res. gkt1064.

Land Grant Participating States/Institutions

CA-R, FL, GA, MI, MN, ND, OR, SC, WA, Washington Cooperative Extension

Non Land Grant Participating States/Institutions

Cotton Incorporated, PWA, Salve Regina University, SPA, USDA-ARS/Georgia, USDA-ARS/Iowa, USDA-ARS/South Carolina, USDA-ARS/WA, Washington State University, West Virginia, other:WA