



# MaizeGDB STATUS REPORT

---

## *UPDATES, ACTIVITIES, AND NEW INITIATIVES*

USDA-ARS

Project No. 3625-21000-061 (Ames, IA)

and

Project No. 3622-21000-038 (Columbia, MO)

Prepared by: The MaizeGDB Team

Carson M. Andorf, Darwin A. Campbell,  
Ethalinda (Ethy) Cannon, Jack Gardiner, Lisa C. Harper, John Portwood, Mary L. Schaeffer,  
Taner Z. Sen, Kokulapalan Wimalanathan,  
and Carolyn J. Lawrence

**AUGUST 2013**

**Contact: C. Lawrence**

USDA-ARS

1034 Crop Genome Informatics Laboratory

Iowa State University

Ames, IA 50011

Email: [Carolyn.Lawrence@ars.usda.gov](mailto:Carolyn.Lawrence@ars.usda.gov)

URL: <http://www.maizegdb.org>

515-294-8280 (fax)

515-294-4294

## TABLE OF CONTENTS

<b>1 – Meeting agenda and the Working Group’s role</b>	3
<b>2 – Executive summary</b>	4
<b>3 – Personnel list and project resources (both ARS and outside funds)</b>	5
<b>4 – Past year’s activities and accomplishments</b>	7
Overview (general)	
Interface Redesign	
<b>5 – MaizeGDB 5-year Project Plan for USDA-ARS</b>	
ARS process	10
Project Plan	
Ames	11
Columbia	91
<b>6 – Presentations/outreach</b>	94
<b>7 – Peer-reviewed publications since March 2012</b>	95
<b>Appendix: Known Issues</b>	96

## 1 – Meeting agenda

**Tuesday, August 6, 2013** *Times shown as Eastern/Central/Mountain/Pacific*

11am/12pm/1pm/2pm – **Connect to the Web** (available 30 min earlier to enable early connection for anyone who wants to try it out)

Online instructions: [http://alpha.maizegdb.org/about/working\\_group/2013](http://alpha.maizegdb.org/about/working_group/2013)

**11:00 am (Eastern) Presentation: MaizeGDB’s activities, accomplishments, and plans for the future**

- **Introduction (15 min)**
- **Past Year**
  - **Overview (15 min)**
  - **Interface Redesign (15 min)**
- Break (10 min)*–
- **5-Year Project Plan**
  - **Genome Stewardship and Diversity (15 min)**
  - **Metabolic and Breeding Networks (15 min)**
  - **Flexible Access to the Database (5 min)**
  - **Community Service (5 min)**
  - **Opportunities and Questions (5 min)**

–*Break (15 min)*–

**1:00 pm (Eastern) Working Group Executive Session**

**2:00 pm (Eastern) Working Group Summarizes for the MaizeGDB Team**

**Meeting adjourns**

### **Working Group’s Role**

The Working Group is tasked with evaluating MaizeGDB current status and recommending a course of action that will insure that the MaizeGDB project tracks the trajectory of maize research as closely as possible. The ultimate goal of MaizeGDB is to provide a robust and timely source of data and analysis tools that will help researchers to investigate the biology of maize, both as a research model and as a crop. **Note well: The role of the Working Group is to help the MaizeGDB Team with strategic planning.** Feedback on other topics including, but not limited to, site functionality issues and data access are desired and needed, but should be provided on an *ad hoc* basis and as ideas and issues emerge (either via the website directly or by communicating with project personnel directly) rather than via Working Group meetings and/or guidance documents. For a (non-exhaustive) list of known issues and plans for their resolution, see *Appendix*.

### **Current Working Group Membership**

Alice Barkan, Qunfeng Dong, David Jackson, Thomas Lübberstedt, Eric Lyons, Adam Phillippy (chair), Marty Sachs, Mark Settles, and Nathan Springer.

2013 new members: Qunfeng Dong, Adam Phillippy, and Mark Settles

Rotating off: Anne-Francoise Lamblin, Karen McGinnis, Lukas Mueller, and Mihai Pop

## 2 – Executive Summary

In the past year, personnel at MaizeGDB have supported community activities and curated information into the database including both genomic information as well as functional data to enable the use of genomic information toward understanding and improving corn. One major focus was the wholesale interface redesign (scheduled for full deployment toward the end of 2013). A second major focus was the development of the USDA-ARS project plan that outlines work to be accomplished over the next five years. Objectives for the next five years are to:

- 1) Support stewardship of maize genome sequences and forthcoming diverse maize sequences
- 2) Create tools to enhance access to expanded datasets that reveal gene function and datasets for genetic and breeding analyses
- 3) Deploy tools to increase user-specified flexible queries
- 4) Provide community support services, training and documentation, meeting coordination, and support for community elections and surveys

Beginning to work toward Objective 1, we have strengthened ties with NCBI toward genome stewardship going forward and are currently making changes to BAC and AGP files in GenBank to enable sustainable reference genome stewardship. For Objective 2, we have developed a visual interface to extract annotation from Cyc databases (collaboration with Jesse Walsh and Julie Dickerson at Iowa State University) and are working to create a proof-of-concept view of maize pedigrees from 9 states that displays family relationships among inbred lines. Objective 3 is in the planning stages and Objective 4 is an ongoing endeavor.

### 3 – Personnel list and project resources

#### Personnel

*Federal* (\* demarcates term appointments)

**Carolyn Lawrence**, research geneticist and lead scientist, USDA-ARS in Ames, IA

Responsible for project management, coordination with outside groups, financial planning, and defining program direction in consultation with the MaizeGDB Team. A member of the Maize Genetics Executive Committee as well as the Maize Nomenclature Committee.

**Carson Andorf**, IT specialist (bioinformatics engineer), USDA-ARS in Ames, IA

Lead programmer responsible for maintaining the MaizeGDB interface and the creation of data analysis tools unique to MaizeGDB. Also assists with the implementation of available software packages (like the MaizeGDB implementation of GBrowse for the Genome Browser and MaizeCyc for a Pathway View tool) and with server configuration and maintenance.

**Darwin Campbell**, IT specialist (database administrator), USDA-ARS in Ames, IA

Responsible for the relational databases, which are the backbone of the MaizeGDB project. Coordinates purchases (hardware, software, supplies/services) for the group, and is responsible for all associated accounting.

**Lisa Harper\***, geneticist (curator & outreach coordinator), half time, USDA-ARS in Albany, CA

Responsible for community outreach and education, video tutorials, Phenotypic Controlled Vocabulary, and literature curation.

**John Portwood\***, computer clerk (student), USDA-ARS in Ames, IA

Responsible for development work on interface redesign as well as administering and updating the MaizeGDB Genome Browser.

**Mary Schaeffer**, geneticist (curator), USDA-ARS in Columbia, MO

Responsible for curation involving maps, loci, literature, QTL, etc. A member of the Maize Nomenclature Committee and a co-editor of the Maize Newsletter. Mary is an excellent source of historical information on why certain data are represented in a particular way.

**Taner Sen**, computational biologist, USDA-ARS in Ames, IA

Responsible for MaizeCyc, CornCyc, and other systems biology and network analysis projects.

*State* (funded by NSF, NCGA, and ISU)

**Ethalinda (Ethy) Cannon**, Solution/application architect, ISU in Ames, IA

Responsible for design and development of the original POPcorn website, its integration with MaizeGDB, some aspects of the MaizeGDB Website Redesign, genome stewardship project implementation, and communication with groups providing data access.

**Jack Gardiner**, curator, ISU located on the University of Arizona campus, Tucson, AZ

Responsible for identifying and recruiting gene expression, physical and genetic mapping, and epigenetic datasets with a special emphasis on large data sets. Manages the MaizeGDB Editorial Board.

**Kokulapalan Wimalanathan**, Bioinformatics and Computational Biology doctoral student, ISU in Ames, IA

Focused on genome-wide functional annotation of maize genes. Co-mentored by Carolyn Lawrence and Erik Vollbrecht, collaborating with Carson Andorf.

**Funds: ~\$800,000 per year in Ames**

***Current:***

*ARS (annual)*

Ames: \$523,285 permanent  
+ \$100,000 temporary transfer = \$623,285 in Ames

Covers salaries, hardware, travel, etc. for:

Carolyn Lawrence, Ph.D.

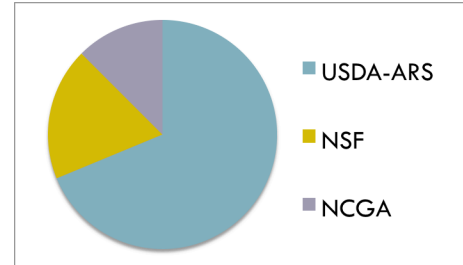
Taner Sen, Ph.D.

Lisa Harper, Ph.D.

Carson Andorf,

Darwin Campbell

Columbia: covers salary, hardware, travel, etc. for Mary Schaeffer, Ph.D.



*NSF Plant Genome Research Program*

Ames (Iowa State University)

Functional Structural Diversity Among Maize Haplotypes

Pat Schnable, PI, Carolyn Lawrence and Brent Buckner, coPIs

\$3,000,000 total, ~\$150,000 to coPI Lawrence.

Supports Ethy Cannon

Expires September 2013

Ames (Iowa State University)

Genetic Networks Regulating Structure and Function of the Shoot Apical Meristem

Mike Scanlon (Cornell), PI. Subcontract to Carolyn Lawrence

\$1,937,366 total, \$153,597 to coPI Lawrence

Expires January 2018

*National Corn Growers Association*

Ames (Iowa State University)

Functional Genomics Software Tools for MaizeGDB

~\$100,000 per year

Supports Jack Gardiner, Ph.D.

Current agreement provides funds through the end of February 2014

***Pending:***

*NSF Plant Genome Research Program*

Detection of Structural Variation and Epistasis via NGS-Enabled BAC-by-BAC Genome

Sequencing of Maize

Pat Schnable, PI, Carolyn Lawrence and Brent Buckner, coPIs

\$4,044,713 total over 3 years, ~\$275,000 to coPI Lawrence.

*United Sorghum Checkoff Program*

Ames (Iowa State University)

Coordination to Support Sorghum Genomics: Developing Community Resources with Emphasis on Informatics Systems

Carolyn Lawrence, PI, Maria Salas-Fernandez, coPI

\$552,938 over 3 years

#### 4 – Past year’s activities and accomplishments

In addition to the project accomplishments listed below, it should be noted that Andorf, Campbell, Harper, Schaeffer, and Lawrence contributed heavily to community support including training researchers in the use of MaizeGDB, administering community elections and surveys, organizing the Maize Meeting abstract book, and managing the IT needs of the Maize Meeting website and conference.

**Site Redesign is on schedule.** The redesign for the MaizeGDB web interface (first announced March 2011) was released in preliminary form March 2012, and was demonstrated in March 2013 at the Maize Genetics Conference, Chicago IL. This effort will result in improved operation, function and overall cosmetic appeal. A number of security features have been added. New tools provide access to the international standards such as the gene function vocabulary (Gene Ontology or GO) for cellular components, biological process and molecular function. The interface was beta-tested (Nov 2012-Mar 2013) by community volunteers (see <http://maizegdb.org/freeze.php>).

#### Data updates.

*Sequence-indexed mutations.* Some 15,611 new UniformMu insertions were added this year, from data supplied by the UniformMu project. These now number 42,785, representing 14,157 gene models in the filtered gene set and with insertions within 100bp of the start and end positions of genes. The 6130 stocks with the mapped insertions are available at the Maize Genetics Cooperation – Stock Center. To find those for your sequence of interest, use the MaizeGDB BLAST search tools, or look at the UniformMu track on the genome browser track. More details about these mutants and the project are posted at MaizeGDB (<http://www.maizegdb.org/documentation/uniformmu/>)

*Gene Expression.* Representation now includes embedded glyphs from the eFP browser at BAR (The Bio-Array Resource for Plant Biology, [http://bar.utoronto.ca/efp\\_maize/cgi-bin/efpWeb.cgi?dataSource=Sekhon\\_et\\_al](http://bar.utoronto.ca/efp_maize/cgi-bin/efpWeb.cgi?dataSource=Sekhon_et_al); Winter et al 2007 PloS One 2:e718.) and MaizeGDB histograms for the Maize Gene Expression Atlas (Sekhon et al 2011 Plant J 66:453-463).

*Metabolism.* Metabolic network representations have been computed by two groups, Gramene and the Plant Metabolic Network, both in collaboration with MaizeGDB, and both using the Pathway-Tools of the MetaCyc project (Caspi et al 2011 Nucleic Acids Res D742-753). Addition of CornCyc enhanced existing prediction of protein function and pathway assignments by adding the information extracted from non-enzyme proteins into the algorithmic pipeline. CornCyc was computationally created by Plant Metabolic Network. The project at MaizeGDB was coordinated by Taner Sen. MaizeGDB curators manually curated auxin, brassinosteroid, and gibberellin pathways into CornCyc based on experimental data available in the literature. Analysis of a set 197 UniProt proteins with experimental evidence show that two metabolic resources available at MaizeGDB, MaizeCyc and CornCyc, are complementary: while MaizeCyc has a higher coverage in terms of assigned enzymatic reactions and pathways, CornCyc, which includes spliced variants, is a higher stringency resource in assigning functions to enzymes.

*Protein classifications, gene homologs and syntelogs.* These annotations are listed on each gene model page with links to offsite resources. These links were made possible through collaboration with Gramene, Phytozome, and efforts from James Schnable and Mike Freeling. Linked protein classifications resources include Panther (<http://www.pantherdb.org>), PFAM (<http://pfam.sanger.ac.uk/>), and COG. (<http://www.ncbi.nlm.nih.gov/COG>). Homolog links are included from Gramene (<http://www.gramene.org>), TAIR (<http://www.tair.org>; Arabidopsis), and the MSU Rice Genome Annotation Project (<http://rice.plantbiology.msu.edu>). Syntenic orthologs (syntelogs) are provided for *Sorghum bicolor*, *Setaria italica*, *Oryza sativa japonica*, and *Brachypodium distachyon*, with links to those gene models at Phytozome. Links also are provided to tools (qTeller, CoGE) that were used to compute and analyze expression for these syntelogs

*Diversity.* Collaborated with Panzea to provide a genotype search tool at MaizeGDB that searches the following diversity datasets: HapMapv1, HapMapv2, raw SNP genotypes for 1144 SNP markers genotyped on the NAM lines,

and genotypic data for 282 inbred association panel using the 55K SNP array. Also developed TYPSimSelector, a tool that sorts lines from the Ames Diversity Panel based on IBS (Identity by State) score (see <http://alpha.maizegdb.org/TYPSimSelector>).

*Phenotypes.* Worked with the NSF-funded phenotype RCN to enable phenotypic queries across multiple plant datasets and data repositories by converting phenotypic descriptions using hierarchically related terms. Collaborator Ramona Walls from iPlant has been instrumental in preparing the maize datasets. Also transferred to MaizeGDB high-resolution images of maize phenotypes provided by Dr. Gerry Neuffer.

*Literature curation.* Prepared maize gene function training data sets to support development of tools for both automated extraction of gene functions from the literature, with the goal to support facile integration into genome databases such as MaizeGDB. This project is affiliated with the international BioCreative consortium, and is a collaboration of database curators from several model plant and animal species: maize (MaizeGDB), Arabidopsis, Drosophila, mouse and rat.

**New Genome Browser tracks** (more detail is found by clicking on the ‘?’ at each track).

*Assembly/Genome Features.*

*Knobs*, Both BLAST identified and genetically mapped. Ghaffari et al. 2012 Maize Meeting Abstract.

*Foreign contamination*. A subtrack for the pseudomolecule indicating regions found by GenBank to be sequence that is mitochondrial or from other species.

*Mitochondrial and chloroplast genome* sequences have been added for convenience, although they are not part of the B73 RefGen\_v2 assembly. They are annotated with their long established genes, rather than computed gene models.

*Nucleosome occupancy* predictions from H Bass and J Dennis. See also Gupta et al 2008 PLoS Comput Biol 4:e1000134.

*Reconstructed chromosomes from maize tetraploidy*. Provided by J Schnable and M Freeling; see also J Schnable et al PNAS 108:4069-4074; Schnable J and Freeling M 2011 PLoS One 6:e17855.

*B73/Mo17 Methylation* (Nathan Springer).

*Diversity*

*HapMap1 SNP* from Gore et al 2009 Science 326:1115-1117 and <http://www.panzea.org>.

*ISU SNPs on 291 IBM RILs* from Liu et al 2010 Genetics 184:19-26.

*Illumina SNP50 tracks* both genetically mapped and non-mapped.

*Expression and transcripts.* (see also above)

*IBM SAM eQTL*. RNA-SEQ data for SAM from 105 IBM RILs; from Muehlbauer and the Shoot Apical Meristem (SAM) Project (M Scanlon PI).

*5' Methylcytosine methylation* in B73 and Mo17 (Eichten et al 2011 PLoS One Genet 7:e1002372.)

*miRNA* mirBase data from Zhang et al (2009) PLoS Genet. 5:e1000716.

*KNOTTED1 binding regions* from Bolduc et al 2012 Genes and Dev 26:1685-1690.

*Agilent microarray annotations; anther stages and mutants*. From Ginny Walbot and Dave Berger. Ma et al 2006 Genome Biol 7:R22; Nan et al 2011 BMC Plant Biol 11:120; Wang et al 2010 Plant J 63:939-951.

*Gene models*

*Sorghum syntenic orthologs* for 24,000 maize genes identified using SynMap (<http://genomeevolution.org>) in collaboration with James Schnable. For details of methods see also Lyons et al 2008 Tropical Plant Biology 1:181-190; Tang et al 2011 BMC Bioinformatics 12:102, Schnable et al 2011 PLoS One 6:e17855.

*Split Genes*. These are putative gene annotation artifacts and determined by Gramene. There are 2 general categories. When two apparently paralogous genes lie on different strands in the assembly, with no overlap between the gene fragments and (2) when gene fragments are in close proximity on the same strand, but



have no overlapping sequence. Details about the methodology are at [http://useast.ensembl.org/info/docs/compara/homology\\_method.html](http://useast.ensembl.org/info/docs/compara/homology_method.html).

#### *Repetitive Elements*

*Sirevirus LTR retrotransposns* from Bousios A et al 2011 The Plant Journal 69:475-488.

*Ac/Ds and UniformMu* tracks were updated with links to Stocks at the Maize Genetics Cooperation – Stock Center (there are now 42,785 [Dec 2012] insertions, representing 14.157 gene models in the filtered gene set, with insertions within 1000 bp of th start/end positions of genes.

**Sorghum community interest.** The Sorghum research community recently met to discuss genomics and community needs. Two top needs articulated at that meeting were community support and information systems. There is an interest from the research community as well as from members of the Sorghum Checkoff to consider the services provided by MaizeGDB as a model for their development going forward and to learn from the maize community on how best to address these needs.

**A Chinese instance of MaizeGDB.** In collaboration with Jinsheng Lai at China Agricultural University and Pat Schnable at Iowa State University, a plan is in place to deploy Chinese versions of the Maize Genetics Conference website (within the next month) and the MaizeGDB website (in March of 2014 at the Maize Meeting in Beijing). Translations are currently underway with a plan to host the Chinese site in Ames. This ensures concerted development of functionality for both website instances and enables the continuation of a single resource for maize data. This project is simplified by the fact that the newly redesigned interface infrastructure separates web content from interface functionality. This project is especially exciting to us because ~25% of MaizeGDB users currently access the site from machines configured to use Chinese as their primary language.

## 5 – Five Year Project Plan

### ARS Process:

#### *National Program 301 (Plant Genetic Resources, Genomics and Genetic Improvement)*

- ✓ **Accomplishment report, 2006-2011:**
  - [http://www.ars.usda.gov/research/programs/programs.htm?np\\_code=301&docid=22191](http://www.ars.usda.gov/research/programs/programs.htm?np_code=301&docid=22191)
- ✓ **Assessment, (last 5 years):**
  - <http://www.ars.usda.gov/SP2UserFiles/Program/301/NP301%20Exec%20Summary%202011%20Assessment%20Report.pdf>
- ✓ **Stakeholders meeting:** November 15, 2011
- ✓ **Draft copy of the new Action Plan:** March 1, 2012
- ✓ **Scientists meeting:** March 16, 2012
- ✓ **MaizeGDB Working Group input:** March 30, 2012
- ✓ **Concept paper:** May 17, 2012
- ✓ **Project Plan due to the ARS Office of Scientific Quality Review:** January 3, 2013
- ✓ **Overall score (Ames):** 6 (of 8 possible). Minor revision.
- ✓ **Project start date:** April 10, 2013

## **Project Plan**

**NP 301 – Plant Genetic Resources, Genomics and Genetic Improvement**  
**Panel Review: January - April 2013**

**Old ARS Research Project Number**  
3625-21000-051-00D

**Research Management Unit**  
Corn Insects and Crop Genetics Research Unit

**Location**  
Ames, Iowa

**Project Title**  
MaizeGDB: Enabling Access to Basic, Translational, and Applied Research Information

**Investigator(s)**  
Carolyn J. Lawrence (Lead Scientist) 1.00  
Taner Z. Sen 1.00

**Scientific Staff Years**  
2.00

**Planned Duration**  
60 months

## Post-Peer Review Signature Page

**Carolyn Lawrence**, 3625-21000-051-00D

MaizeGDB: Enabling Access to Basic, Translational, and Applied Research Information

This project plan was revised, as appropriate, according to the peer review recommendations and/or other insights developed while considering the peer review recommendations. A response to each peer review recommendation is attached. If recommendations were not adopted, a rationale is provided.

Craig Abel

05-09-13

---

Research Leader

---

Date

The attached plan for the project identified above was created by a team of credible researchers and internally reviewed and recognized by the team's management and National Program Leader to establish the project's relevance and dedication to the Agricultural Research Service's mission and Congressional mandates. It reflects the best efforts of the research team to consider the recommendations provided by peer reviewers. The responses to the peer review recommendations are satisfactory. The project plan has completed a scientific merit peer review in accordance with the Research Title of the 1998 Farm Bill (PL105-185) and was deemed feasible for implementation. Reasonable consideration was given to each recommendation for improvement provided by the peer reviewers.

JL Willett

5-10-13

---

**Associate Area Director (original signature required)**

---

Date

**PrePlan Signature Page for ONP Validation(s)**

**Pre-Peer Review**

**Lawrence, 3625-21000-051-00D, NP301**

**MaizeGDB: Enabling Access to Basic, Translational, and Applied Research Information**

Signature Page Completed for Research Leader through Area Director

X The objectives in this PrePlan are those provided in the PDRAM or subsequently approved by the Office of National Programs and the approaches are suitable for achieving the objectives.

/s/ Jack Okamuro, NPL, Plant Biology

12.18.2012

\_\_\_\_\_  
National Program Leader

\_\_\_\_\_  
Date

Comments:



## PROJECT SUMMARY

For more than a decade maize has been the number one production crop in the world. Its success is largely due to high productivity and commercial versatility: not only is maize an excellent source of food, feed, and fuel, its byproducts are used to produce paint, soap, rubber, plastics, and various other commodities. Maize's unparalleled success in agriculture is partially a result of basic and applied research, the outcomes of which drive breeding and product development.

For applied researchers to benefit from basic investigations, generated data must be made freely and easily accessible. MaizeGDB, the Maize Genetics and Genomics Database (<http://www.maizegdb.org>), is the maize research community's central repository for genetics and genomics information. The overall aim of our work is to create and maintain unified public resources that facilitate access to the outcomes of maize research. Over the next five years, we will address four objectives: (1) to support reference genome stewardship within the context of extensive genomic diversity, (2) to deploy datasets and tools that reveal gene function and support genetic and breeding analyses, (3) to enable researchers to access data in a customized and flexible manner, and (4) to support the education, outreach, and organizational needs of the maize research community.

## OBJECTIVES

The long-term objectives of this project are to devise, synthesize, display, and provide access to maize genetics and genomics tools and data to enable the research community to investigate basic plant biology, accelerate the pace of genetic enhancement and breeding, and translate those findings into products that increase crop quality and production. The major challenge to be met over the next five years will be storing and providing useful access to a diversity of large-scale data types including, but not limited to, multiple whole-genome sequences and associated gene expression datasets. To address these challenges, we will:

**Objective 1:** Support stewardship of maize genome sequences and forthcoming diverse maize sequences.

**Objective 2:** Create tools to enhance access to expanded datasets that reveal gene function and datasets for genetic and breeding analyses.

**Objective 3:** Deploy tools to increase user-specified flexible queries.

**Objective 4:** Provide community support services, training and documentation, meeting coordination, and support for community elections and surveys.

The central goal of this project is to enable the maize genetics community and relevant informatics experts to identify and meet the intellectual, technological, and bioinformatics needs that currently limit the use of maize genetics and genomics information, both for crop improvement and as a model system. This will require not only stewardship and continued improvement of the assembly and annotation of the B73 reference genome, but also management of a variety of information becoming available as a result of the sequencing revolution. Data types that must be accommodated include genetic diversity data obtained via comparative sequencing of numerous maize haplotypes as well as myriad gene expression datasets, the availability of which will enable researchers to formulate and test functional hypotheses on a scale previously intractable. Flexible tools that enable easier, standardized means to conduct genetic and breeding analyses are required to interact with and make use of these datasets. To accomplish these tasks will require interactions with, and help from, researchers within and outside of the current maize research community.

## NEED FOR RESEARCH

Maize (referred to commonly as corn or by its botanical name *Zea mays* L. ssp. *mays*) is an important crop. Not only is it one of the most abundant sources of food and feed for people and livestock worldwide, it also is an important component of diverse consumer products where its content is less apparent. For example, maize is used to manufacture glue, paint, insecticides, toothpaste, rubber tires, rayon, and molded plastics. Maize is also the nation's major source of ethanol, a major biofuel that is more environmentally friendly than gasoline and that may be a more long-term economical fuel alternative.

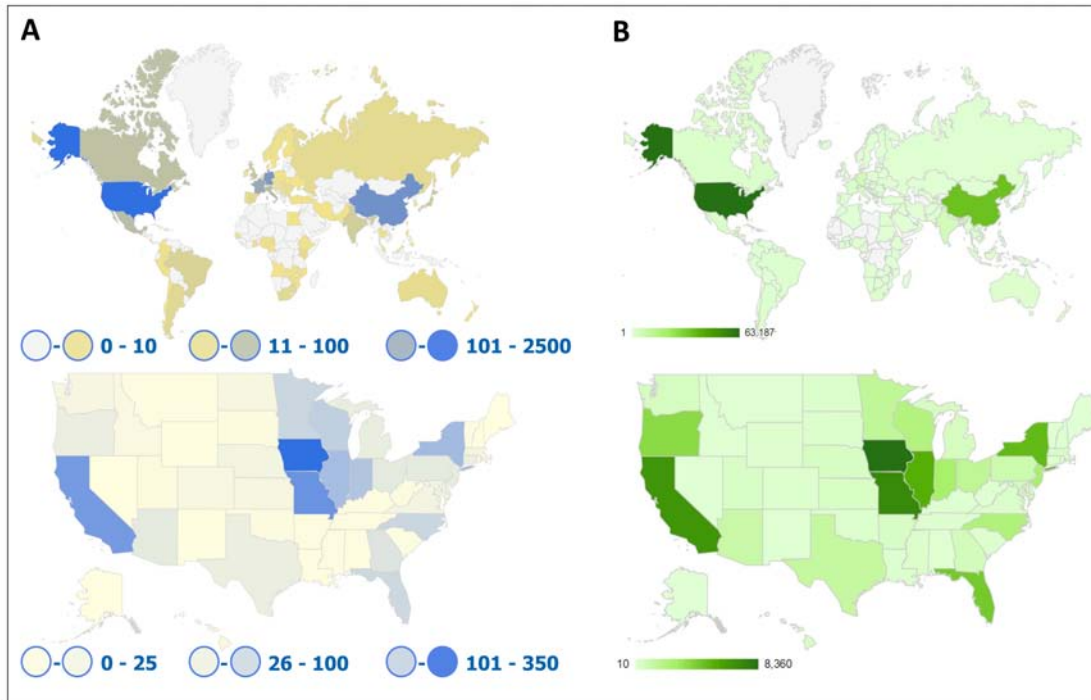
In addition to its value as a commodity, maize is an organism of historical importance to all biologists. Maize researchers including Emerson, Stadler, McClintock, and Rhoades made seminal genetic discoveries that hold true not only for maize but also for living organisms in general, thus setting the stage for maize to become one of the first model organisms for genetics research. More recently, maize has become a leading model for the development of sustainable feedstock grasses for biofuel production (Bosch et al. 2011; Carpita and McCann 2008; Lawrence and Walbot 2007; Penning et al. 2009), and the availability of genetics, genomics, and



sequence datasets has poised maize to become the premier species for plant biological research (Schnable et al. 2009). At the same time, maize must be improved per se as a crop given the impacts of global climate change that are anticipated and the need to feed an ever-growing world population (Grassini and Cassman 2012).

In order to be able to address these needs, data that describe the biology of maize must be catalogued, curated, and made accessible to researchers worldwide. MaizeGDB, the Maize Genetics and Genomics Database (<http://www.maizegdb.org>), makes accessible through a web-based portal genetic and genomic data and data analysis tools that are used by researchers to investigate basic biological concepts and translate findings into technology that is deployed in farmers' fields. The MaizeGDB project is a component of National Program (NP) 301, Plant Genetic Resources, Genomics, and Genetic Improvement, which supports research that expands, maintains, enhances, and protects the United States genetic resource and information base, and increases knowledge of the structure and function of plant genes, genomes, and biological and molecular processes. Through innovative research tools and approaches, this NP manages, integrates, and delivers vast amounts of genetic, molecular, biological, and phenotypic information to a diverse global clientele. The ultimate goals of these efforts are to improve production efficiency, yield, sustainability, resilience, health, product quality, and value of crops.

MaizeGDB Project team members work closely with stakeholders to meet their needs. Members of the Maize Genetics Cooperation (Cooperators) constitute the primary stakeholders of the work conducted by MaizeGDB team members. Cooperators are defined as researchers who have attended the Annual Maize Genetics Conference, those publishing frequently on maize, and any persons who specifically request to be listed as a maize Cooperator. Figure 1 (next page) compares the distribution of maize Cooperators worldwide to MaizeGDB accesses, demonstrating that defining primary customers of the research as members of the Cooperation is reasonable. Other customers include additional geneticists, plant breeders, corn producers, the Iowa Corn Growers Association, the National Corn Growers Association, organic food producers, consumers, and students. New customers, primarily plant biologists working on species other than maize, have emerged as the maize genome has become more fully sequenced and as comparative genomics resources at MaizeGDB offer new methods of access to maize data. Input from customers/stakeholders is gathered via email, telephone, and in-person meetings as well as documented via feedback mechanisms embedded in the MaizeGDB resource itself (i.e., <http://www.maizegdb.org>). Specific, project-level input is formalized by the MaizeGDB Working Group (WG), a group of ten maize geneticists and bioinformatics researchers who have interest in and expertise that is of use of MaizeGDB (for a recent example, see Appendix C). In addition, the Maize Genetics Executive Committee (MGEC) and the Maize Genetics Conference Steering Committee (MGCSC) play a major role in guiding the evolution of MaizeGDB. It is the mission of the MGEC to identify both the needs and the opportunities for maize genetics, and to communicate this information to the broadest possible life science community. This includes scientists, funding agencies, and the end users for the accomplishments of maize genetics, from farmers to consumers (Bennetzen 2001). The MGCSC plans the Annual Maize Genetics Conference. MaizeGDB plays an integral role in coordinating activities of the MGEC and MGCSC in addition to making available specific biological data collections. These services enable the MaizeGDB staff to bring useful statistics and data about the maize research community into MaizeGDB in addition to emphasizing the fact that MaizeGDB is the central resource for maize data as well as maize community information.



**Figure 1.** *Cooperator location closely matches MaizeGDB website usage.* A: Worldwide, areas with large numbers of Cooperators are shaded dark blue and include the US, China, and Europe (upper panel). Within the US, Cooperators are located primarily in the Midwest with additional concentrations in California, New York, Florida, Georgia, and North Carolina (lower panel). B: Visits at the MaizeGDB website (per year) are shaded green with areas of high usage shown in darkest green. Worldwide accesses come primarily from the US and China (upper panel). Within the US, visits are concentrated in the Midwest, California, New York, Oregon, Florida, and North Carolina. (MaizeGDB’s log-based usage statistics are available online at <http://usage.maizegdb.org> and Google Analytics-based usage statistics available by request.)

In March of 2012, a survey of Cooperators was conducted by the MGEC. Survey results (available online at [http://maizemeeting.maizegdb.org/mgec-survey12/analyze\\_final\\_sort.php](http://maizemeeting.maizegdb.org/mgec-survey12/analyze_final_sort.php)) indicate that priorities for information management and access should include: improving assemblies and annotations of the B73 reference genome sequence as well as the genome sequences of diverse inbred lines; advancing functional studies of maize genes, networks, and high-quality phenotypic descriptions; improving interoperability among the online information resources that serve the maize research community; and training at all levels in how best to make use of available genetics, genomics, and bioinformatics tools and resources. Based upon these data, areas of concentration for MaizeGDB development during the 5-year span covered by this Project Plan were devised by project personnel and evaluated by the MaizeGDB WG. Guidance from that group (2012 WG Report) is included in Appendix C. Objectives developed by the USDA-ARS Office of National Programs and outlined here closely match both survey results and WG recommendations.

The research described in this project plan will address problems outlined in the NP 301 Action Plan Component 2: Crop Genetic and Genomic Resources and Information Management Problem Statement 2A and 2B.

*Problem Statement 2A: Crop genomic information resources and bioinformatics.* This problem statement calls for transformation of the ARS crop genetic and genomic databases into portals that deliver up-to-date bioinformatics tools, cyber-infrastructure, and knowledge to address genetic and genomic problems. This will be accomplished by facilitating coordination and cooperation among researchers and data management experts to create software environments that enable researchers to interact with large, heterogeneous datasets. Such work requires management of large genomic and phenotypic datasets and development and implementation of tools that allow researchers to apply analytical methods, tools, and workflows to novel problems. Anticipated products of this research include actively curated, long-term and interconnected information resources and tools that manage high-throughput phenotypic and genotypic data.

The MaizeGDB team will make accessible high-quality, actively curated and reliable genetic, genomic, and phenotypic description data sets. At the root of a high-quality genome annotation lies a well-supported assembly and annotation. For this reason, we focus our efforts on benefitting researchers by developing a system to ensure long-term stewardship of the B73 reference genome sequence assembly and associated structural and functional annotations to allow researchers to plan their work relative to a well-defined release schedule. Building on the reference genome as a coordinate system to understand diversity and organize genes, we will create and deploy tools that allow access to the large-scale diversity and genome-wide expression profile datasets as well as functional annotations for maize genes coupled with genotype similarity and pedigree tools that will reveal gene function and enable highly-predictive breeding decisions to be made. To allow researchers real-time access to data sets, we will deploy tools that allow direct query of the database to produce customized, large-scale data downloads.

## **SCIENTIFIC BACKGROUND**

In 2005 the NSF, USDA, and DOE announced that the ~2.3 Gb genome of inbred line B73, a major contributor to much of the germplasm used for US grain production, would be sequenced using a BAC-by-BAC approach. The plan was to sequence BACs from a minimal tiling path (MTP) to ~6X coverage, and to further improve only the unique “genic” regions. These sequences would be labeled “Phase 1 HTGS\_IMPROVED” at GenBank (the sequence repository component of NCBI, the National Center for Biotechnology Information), and the GenBank record for each BAC was to include information on the improved regions as well as order and orientation, where available, as comments. The Maize Genome Sequencing Consortium (MGSC) planned to release all data via MaizeSequence.org, a project database, with a plan to transition all data into MaizeGDB and Gramene, a comparative resource for plant genomics (Youens-Clark et al. 2011), at project close.

Not only did the MGSC produce these sequences, they created reference assemblies for each chromosome (the first assembly was named “B73 RefGen\_v1”) as well as structural and functional annotations to genes (Schnable et al. 2009). The published B73 reference genome (RefGen\_v1) available from GenBank consists of 2,048 Mb in 125,325 sequence contigs (N50 of 40 kb), forming 61,161 scaffolds (N50 of 76 kb) anchored to a high-resolution genetic map (Wei et al. 2009). After predicting transposable elements (TEs), a combination of evidence-based and

*ab initio* approaches and stringent TE filtering resulted in a set of 32,540 high-confidence, predicted protein-encoding genes (the “Filtered Gene Set”). Due to incomplete sampling of the genome, the B73 reference genome is estimated to be missing ~5-10% of genes that are physically present in the B73 genome.

With timelines as shown in Table 1, following the release of the first draft, B73 RefGen\_v2 improved v1 by the addition of fosmid reads as well as by integrating genetic and optical map information. The assembly and various annotations are represented at MaizeSequence.org, MaizeGDB, and GenBank. For B73 RefGen\_v2, ~80% of the maize genome is ordered and oriented, and optical map and genetic map comparisons suggest that only 2-2.5% of the sequences are likely to be misplaced in the assembly (Fusheng Wei, Jeff Glaubitz, and Mike McMullen, personal communication). The set of gene predictions for RefGen\_v2 includes 110,028 models in the “Working Gene Set” with a subset of 39,656 high-confidence structures identified as the “Filtered Gene Set”.

In the last year of the project 454 WGS reads have been made available to improve the coverage of the gene space not included in the BAC MTP (and thereby identifying some of the estimated 10% of genes that were missed). Planned improvements include refinements to contig placement supported by recent improvements to the IBM genetic map and inclusion of 454 gene space contigs. It is anticipated that the improved genetic resolution of forthcoming genetic maps will provide support for future improvements of the B73 reference assembly.

**Table 1.** *Timeline for B73 reference genome sequence assemblies and structural annotation endeavors.*

Assembly and Comments	Release Date
<i>B73 RefGen_v1</i>	3/2009 (Community databases) 11/2009 (GenBank)
<i>B73 RefGen_v2</i>	4/2010 (Community databases)
GenBank records include functional annotation (GO) of the “filtered gene set”	12/2012 (GenBank)
<i>B73_RefGen_v3</i> Currently under consideration by GenBank	1/2013 (est. Community databases and GenBank)

The MGSC’s final product (B73 RefGen\_v3; currently computed but not yet publically available) is high quality sequence in genic regions, but BAC sequences remain in many pieces with some, but not all, order and orientation information annotated. The MGSC’s funding period has ended and the v3 assembly and annotation product, along with associated gene annotations, is anticipated to become available in early 2013 pending GenBank approval and release.

*Stewardship of the maize B73 reference genome is required for the maize community to efficiently exploit public-sector investments in maize genome sequencing.* Indeed the absence of a well-curated and up-to-date version of the B73 genome sequence is likely to have a negative impact on research that uses the maize genome sequence. For example, individual labs currently maintain internal lists of known sequencing errors for maize and use those data for their own research, but the reference genome has not been updated to reflect these known errors. The B73 reference genome is currently used to create assemblies of other inbred lines and as a substrate for short read mapping, and errors in the reference sequence propagate throughout downstream research. In addition, long delays between updates, and perhaps even more importantly uncertainties about when updates will occur, can lead to alternate assemblies being generated within the community as has happened for several other model organisms, notably rice (R. Buell, personal communication) and cow (Church and Hillier 2009). Failure to make available anticipated updates and known errors to an assembly in real-time (that is, as the errors are

reported) can also negatively impact researchers' interest to contribute their knowledge for improving an assembly and annotation product. Although it is acknowledged that independent assemblies have value, the lack of a single standard often leads to confusion in the community and difficulties in relating results among different studies. To this point, no group has taken on long-term stewardship of the B73 genome. This has resulted in lack of public awareness of plans for genome assembly and annotation release. Coupled with delays of anticipated release dates when dates have been articulated, the situation affects researchers' ability to plan their work and has resulted in some frustration among researchers waiting for improved assemblies with which to conduct their research.

*Accurate functional annotation of genes is required to predict phenotypic consequences, and the putative functions of genes identified in B73 are not well-documented or consistently supported.* In 2012, Cooperators identified "Advance functional studies of maize genes, gene families, and networks" as their top priority for research directions (2012 Survey of Cooperators; [http://maizemeeting.maizegdb.org/mgec-survey12/analyze\\_final\\_sort.php](http://maizemeeting.maizegdb.org/mgec-survey12/analyze_final_sort.php)). This outcome is perhaps not surprising given that most Cooperators share gene function analysis as their focus of research. In addition, from the perspective of data handling there are additional reasons this priority likely was ranked highly: maize functional annotations are entirely missing from the current set of pseudomolecules represented at GenBank (GK000031 - GK000034 and CM000777 - CM000786) causing some data repositories (e.g., PlantGDB, Gramene, and Phytozome) to create their own independent functional assignments that are not in agreement across repositories, and the most widely used computational pipelines to generate functional annotations (including those that have been used for maize) are known to assign inaccurate and incomplete functional annotations as described below. At a practical level, the absence of an accurate and authoritative centralized set of functional annotations causes researchers to be unable to improve functional annotations given that it is not clear which annotations should be improved. At a more applied level, the significant knowledge gap between our understanding of genes and how they function hampers development of predictive capability in research.

Missing functional annotations are easy to identify and correct. Inaccurate annotations are more problematic because researchers can waste months of research time and energy on erroneous hypotheses based on errant annotations. A recent cross-species study demonstrated that pathways in the KEGG, the Kyoto Encyclopedia of Genes and Genomes (Tanabe and Kanehisa 2012) pathway database, contain 5% to 63% misannotations across six protein superfamilies (Schnoes et al. 2009). It has been estimated that annotation methods such as BLAST (Altschul et al. 1997) that are based on sequence-similarity can have a misannotation rate of up to 49% (Jones et al. 2007). Although estimates of inaccuracy are not reported for maize, anecdotal problems with functional annotations are frequently communicated.

Improving the accuracy of predicted gene function for the high-quality "Filtered Gene Set" will aid in the development of metabolic pathways, as part of larger biological networks including interaction, genetic, and regulatory networks. Enhanced functional annotations at these levels are required to enable researchers to develop more accurate hypotheses, generate hypotheses with higher predictive accuracy, and save months of labor and resource expenditure. Metabolic pathways play a major role in the cellular environment to create biological products to control cellular responses. However, the enzymes and metabolites that are part of the metabolic pathways do not function independently of other types of physical and genetic interactions, which can have significant influence on the phenotype of the maize plants. Visualizing and analyzing interaction networks are activities critical to researchers who seek to understand the

cellular mechanisms that underlie phenotypes. Two maize metabolic pathway resources, MaizeCyc (<http://maizecyc.maizegdb.org/>) and CornCyc (<http://pmn.plantcyc.org/organism-summary?object=CORN>; (Chae et al. 2012), were created jointly between MaizeGDB personnel and the teams working at Gramene and the Plant Metabolic Network project, respectively. Both were developed and are displayed via the Pathway Tools (Caspi et al. 2012; Karp et al. 2010) software suite. Pathway Tools provides a powerful visualization of the metabolic pathways, but falls short of displaying customized views of physical and genetic interactions. The metabolic “connections” (i.e., relationships between enzymes and metabolites) that form the basis of Pathway Tools views are predetermined, and do not allow for addition of new pathways or connections by researchers who interact with the tools (such changes must be inserted manually by professional curators). In addition, the Pathway Tools software presents views of a single metabolic network, but it does not have the ability to show protein-protein or genetic interactions or genomic and metabolic similarities among maize lines. To enable maize researchers to develop their own interaction views based on their own public or private data, new visualization solutions are needed.

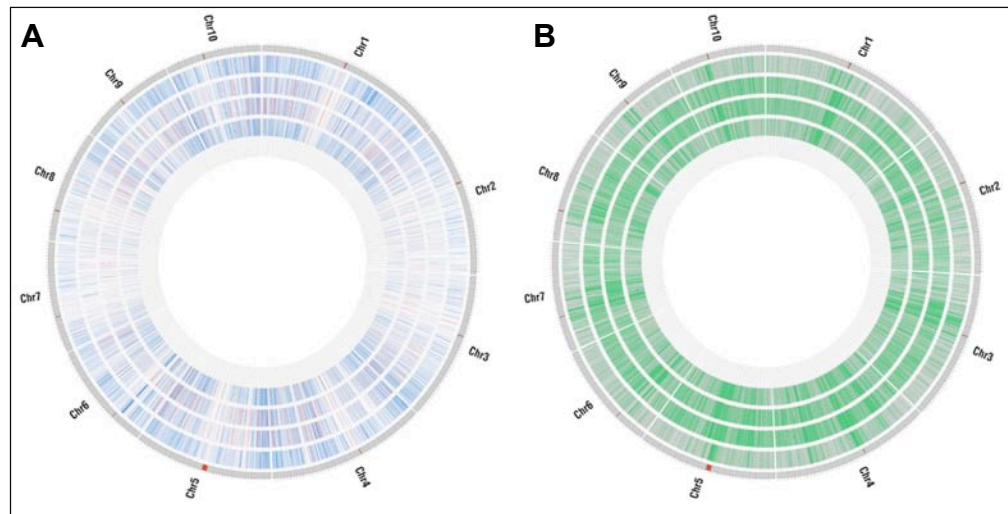
*Raw materials to study and document maize genomic diversity are maturing.* At the same time that needs for B73 assembly and annotation stewardship are becoming evident and urgent, the documentation of genomic diversity in maize has increased dramatically. This development is especially important given that diversity represents the materials available for crop improvement. John Doebley and colleagues (Wright et al. 2005) created a number of inbred teosintes (teosinte is the wild progenitor of maize), and the Maize Diversity Project (Dr. Ed Buckler, PI) has developed the Nested Association Mapping (NAM) population (McMullen et al. 2009). To create the NAM population, 25 lines that capture the diversity of maize were crossed to B73 and then 200 RILs were created from each of the 25 families, resulting in a total population of 5,000 recombinant inbred lines (RILs), with each RIL derived from a unique F<sub>2</sub> plant. NAM is a fairly new approach to map the genes underlying complex traits, in which the statistical power of QTL mapping is combined with the high (potentially gene-level) chromosomal resolution of association mapping. The NAM concept was explained and demonstrated via computer simulation (Yu et al. 2008) and has been applied to mapping phenotypes including flowering time (Buckler et al. 2009).

The genomic diversity available to address basic research questions as well as to improve maize as a crop has prompted the production of whole genome skim sequencing of six inbred lines (Lai et al. 2010) as well as 105 lines by another US-Chinese collaboration (Chia et al. 2012). Others are making use of a “Genotype by Sequencing” (GBS) approach that involves sequencing restriction enzyme fragments for reduced genome representations combined with high multiplexing. Two phases of discovery require data analysis support: (1) confirmation of previously identified variation and (2) identification of rare variants using modified pipelines such as those developed in the Ware and Buckler laboratories as part of the Maize HapMap project (Gore et al. 2009a; Gore et al. 2009b). GBS analysis pipelines serve as the foundation for downstream population analyses and marker assisted breeding pipelines, which is why CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo; International Maize and Wheat Improvement Center) in Mexico is using these techniques to evaluate their entire collection (more than 20,000 lines) over the course of the next few years. In the US, GBS is being used to characterize over 20,000 accessions including ~2,500 accessions of *Zea* from the NCRPIS (North Central Regional Plant Introduction Station, in collaboration with C. Gardner, NCRPIS Director). Mechanisms to store the resulting data in a way that allows fast query as well as

methods to visualize the data in meaningful ways are lacking, and standard methods for dealing with the data are needed across the genus.

In addition to the high levels of sequence diversity in maize, structural variations (SV's), a class of variation that includes copy number and presence/absence variations, (CNV and PAV, respectively), are rampant among maize haplotypes (see Figure 2; Swanson-Wagner et al. 2010). There is a growing appreciation in mammals for the roles of SV in explaining phenotypic variation (Beckmann et al. 2007; Cooper et al. 2007; Feuk et al. 2006; Hurles et al. 2008; Scherer et al. 2007; Sebat 2007; Sharp et al. 2006). But far less SV is observed in humans than is present in maize. For example, any two humans exhibit <100 CNVs and PAVs, while as detailed below maize lines exhibit hundreds of CNVs and thousands of PAVs. Recent work demonstrates that structural variation occurs in many lines is strongly correlated with agronomically important traits (Chia et al. 2012).

**Figure 2.**  
Results of CGH comparisons between B73 and an inbred teosinte (outer ring), Tx303 and Hp301 (NAM parents, middle rings) and Mo17 (inner ring). The positions of centromeres are shown in red. 2A: CGH "segments"



(chromosomal regions) that exhibit losses and gains in copy number relative to B73 are illustrated in blue and red, respectively. The fact that losses outnumber gains is likely an artifact that the CGH array was designed based on the B73 reference genome sequence. 2B: CGH "segments" that exhibit highly similar hybridization patterns between B73 and the other indicated haplotype are shown in green. Note that many regions are conserved among all four maize haplotypes and some regions are also conserved with the tested teosinte haplotype. Image kindly provided by collaborator P. Schnable at Iowa State University, Ames, IA.

*In order to make use of this level of diversity to improve maize as a crop, breeders need access to tools that allow them to use sequence and marker information to infer similarity among lines and compare inferred relationships to known pedigrees.* Graphical representation of the outcomes of such analyses is largely lacking. Pedigree representations showing the genetic relationships between generations are used extensively by breeders and patent examiners. One such pedigree graph is shown in Figure 3 (next page) for lines with the Stiff Stalk heterotic pattern (Mikel 2006). In the classical sense, pedigree visualizations are based on historical data describing breeding schemes. However, pedigrees do not specifically show which genes were associated with trait selections. With advances in genomic technologies, it is now possible to genotype individuals using technologies including, e.g., GBS for thousands of individuals at a relatively low cost. This makes it now possible to compare different lines based upon their genotype on a large scale and to infer similarity among lines and compare to the historically documented pedigrees.





**Objective 1:** Support stewardship of maize genome sequences and forthcoming diverse maize sequences. (CJL; non-hypothesis driven)

**Goal 1.a:** Enlist the community of maize researchers in the genome assembly and annotation process to enable their contributions to and use of improved reference genome sequences in real-time.

**Background and Related Research:** Researchers currently funded to improve the maize genome assembly and annotations include Drs. Doreen Ware (ARS in Cold Spring Harbor, NY), Mark Yandell (University of Utah, Salt Lake City, UT), and Volker Brendel (Indiana University, Bloomington, IN). Ware's Gramene group is funded by the NSF's Plant Genome Research Program to create a single assembly and annotation of B73 beyond RefGen\_v3 during the first year of the project period, which begins in late 2012. For annotation of additional assemblies, the Volker Brendel and Mark Yandell research groups are both currently funded by the NSF's Plant Genome Research Program to develop and improve plant genome structural annotation tools, but are not funded to annotate maize, *per se*. Brendel's group has three aims: (1) to implement and deploy pipelines that would automate the xGDB (eXtensible Genome Data Broker) pipeline (Duvick et al. 2008) for plant genome annotation and visualization such that any species could have an xGDB instance created fairly easily, (2) to enable gene structure prediction using machine learning approaches, and (3) to enable researchers to structurally annotate an entire gene family across many plant species (termed "vertical" annotation). The Yandell group plans to optimize MAKER (Cantarel et al. 2008), a structural annotation tool suite, to plant genomes. Once optimized for rice and Arabidopsis, Yandell's group will reannotate some plant genomes, with maize listed as one genome that will be reannotated with the improved and optimized software. Yandell's group is currently working in collaboration with Ware to release a set of high-quality gene models for the B73 RefGen\_v3 assembly in late 2012. Beyond this work, no one is currently funded to improve the maize B73 reference genome assembly and annotations and neither Yandell's nor Ware's group is focused on enabling manual curation by experts or end users to improve annotations.

**Approach:** Regularly scheduled statements describing the status of genome assembly and annotation along with anticipated dates of availability for public access to datasets in the offing are missing from the current maize genome assembly and annotation efforts. In addition, tools that enable independent researchers to document misassembled and/or misannotated regions of the genome are missing or are not a component of a recognized mechanism for genome assembly and annotation improvement. Our efforts are focused on improving this situation for the B73 reference genome assembly.

To address the need to keep researchers informed of the status of maize genome assembly and annotation efforts, a webpage that documents which groups are funded to provide assembly and annotation products along with estimated release dates for key datasets will be delivered via MaizeGDB. To enable research groups to maintain the content directly, MaizeGDB will permit them to edit content directly (data contributed by cooperating project personnel will go live pending daily approval by MaizeGDB personnel).

To enable researchers to document needed changes to the reference genome assembly and annotations, a set of tools that record suggested changes along with experimental

evidence that supports the change will be made available for maize researchers' use. MaizeGDB staff members will coordinate updates to the genome assembly representation at GenBank by brokering (i.e., arranging, negotiating, and organizing) collected community-reported annotations and changes and communicating timelines for when those community-reported changes will become available. Not only will researchers' documentation of errors be passed on to any group responsible for generating assembly and annotation data, they will be made available via MaizeGDB prior to incorporation into forthcoming assemblies to enable researchers to leverage the community's shared knowledge in real-time.

For example, to repair an error in the assembly and follow NCBI guidelines, the toolset must allow:

- documentation of the region(s) involved (in genomic coordinates from the B73 reference genome)
- specification of GenBank Accession and Version numbers of sequences from the B73 inbred that could be used to correct the error
- storage of that information and creation of documentation that could be provided to the group responsible for maintaining the B73 genome assembly, and ultimately to GenBank

To make available researchers' information in real-time, the MaizeGDB interface must be adapted to show:

- regions of the genome that are likely assembled incorrectly via a genome browser context
- alternate, community-contributed annotations via gene model pages where a gene model already exists
- additional community gene model pages for novel genes identified by community members but absent from the official genome annotation
- visual indications of alternate and additional gene models from a genome-browsing perspective including indicators where experimental evidence argues that an annotated gene model should be deleted

Existing tools will be applied wherever possible. For example, an excellent tool that currently enables this sort of assembly error documentation is available from the Genome Reference Consortium's (GRC) website for human, mouse, and zebrafish genomes (see <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/ReportAnIssue.shtml>), and could be used directly (if GRC assembly tools and software are adopted by MaizeGDB) or as a model for our development of in-house tools to be deployed via MaizeGDB (see Contingencies below). Note that it is not required that maize be voted into the GRC as a full member to use the GRC toolset: their data stores and interfaces are already being used by the MaizeGDB team as we develop a proof of concept. For structural annotation there are a few options to choose from. yrGATE (Wilkerson et al. 2006; <http://www.plantgdb.org>), which can be used to create or repair gene structure annotation, works well for collecting community annotations though it is not known to be user-friendly. The yrGATE tool currently is in use at PlantGDB for improving maize assembly gene structures wherein improvements are served by the Distributed Annotation System (DAS; Dowell et al. 2001) and represented within the MaizeGDB Genome Browser (Sen et al. 2009), but currently lacks an agreement with and mechanism for providing the new/corrected gene models to the Ware group. Another possible tool that should be evaluated for the same purpose is

WebApollo, a distributed version of Apollo (Lee et al. 2009). WebApollo is based on the genome browser software JBrowse (Skinner et al. 2009) that could be used pervasively for genome browsing and annotation throughout MaizeGDB and because the WebApollo tool is under consideration for implementation as a stop within the easy-to-use DNASubway genome annotation system currently under development as a component of the iPlant Consortium. In addition to yrGATE and WebApollo, other existing tools that will be evaluated include the original Apollo resource, JGI's Phytozome resource track editor tool (accessible via [http://genome.jgi-psf.org/help/track\\_editor.jsf](http://genome.jgi-psf.org/help/track_editor.jsf)), and other emerging software tools including, but not limited to, Artemis, Manatee, and Otterlace/ZMap. Lewis's Apollo development group has focused on deployment of WebApollo as a replacement to Apollo, so planning to work with Apollo, while stable, likely would not be a good longitudinal solution. The JGI solution would likely be compatible with MaizeGDB given that, as is true for MaizeGDB's Genome Browser, the JGI solution is GBrowse-based. However, the JGI solution requires a login via JGI and hence may not be practical for deployment to the MaizeGDB user community directly. Similar limitations are anticipated for the Artemis, Manatee, and Otterlace/ZMap tools given that they are not widely used, but all will be evaluated and carefully considered before a specific solution is selected.

The specifications for the set of tools to be adopted and/or developed will be documented in year one with tool release via MaizeGDB planned as part of the project's second year products (see Milestones Table). These tools will be maintained and modified according to researchers' input during the second year and long-term maintenance will be supported by MaizeGDB. Development will involve 5 fundamental tasks that will be carried out with input requested from the Cooperators, MaizeGDB WG, and *ad hoc* committees throughout. As documented in the Milestones Table, the development process and timing for deliverables are as follows:

- *Requirement documentation:* Known requirements include experimental evidence in many instances. In other situations where no sequence data exist to the contrary, re-assembly and reorder of contigs can occur in the absence of creating new sequence. Other requirements will be gathered by consulting with the Genome Reference Consortium, GenBank, and researchers currently funded to assemble and annotate the maize genome.
- *Functional design:* Navigation, functionality and user-interface design will be worked out in collaboration with a test group of maize Cooperators who volunteer or are requested to aid in this task.
- *Technical design:* A review of possible technologies and existing tools will occur simultaneously with gathering information on functional requirements. The design of the code and database architecture will follow, based on requirements, selected tools and technologies and functional design.
- *Implementation:* Actual construction of the documentation tool suite and visualization tools to be carried out during year 2 with release by the end of year 2 (see Milestones Table).
- *Outputs:* All data provided by the community for assembly and annotation improvements will be made available to all via MaizeGDB, and we will work with those assembling and annotating the maize genome to define file formats that simplify inclusion in their automated pipelines.
- *Revision:* After the suite is in public use, modifications are inevitable. Year 2 will be devoted to revising the tool according to community needs (as articulated by the

MaizeGDB WG and Cooperators). The final suite will be made available at the end of year 4.

- *Long-term maintenance:* MaizeGDB will assume maintenance of the tool suite.

For assembly stewardship, it is likely that a combination of tools would be used including third party tools (tools and pipelines available from the GRC and elsewhere as they emerge) as well as some to be developed in-house where third party tools do not meet needs articulated by the community of maize researchers.

We will be, and are already, developing close relationships with the GRC as well as the developers of third party tools such as WebApollo that show promise. (Note that the WebApollo tool can be used not only for annotation updates, but for updates to the assembly itself. Through WebApollo the assembly can be edited transiently, e.g., to specifically create a gene model that otherwise cannot be assembled, or permanently.)

For assembly update activities, researchers will receive “gold stars” and see accepted update counts for their efforts displayed on their person record at MaizeGDB. Though this sort of “award” for efforts may seem trite, we have found that researchers are delighted to work for gold stars. We have used this incentive/acknowledgement successfully already for service to the community through serving on the MaizeGDB Editorial Board (e.g., see James Schnable’s person record at <http://maizegdb.org/cgi-bin/displaypersonrecord.cgi?id=981154>) as well as for contributing expert, in-depth analyses of well-characterized loci in the maize genome (e.g., Alice Barkan has contributed 17 gene reviews as reported on her person record at <http://www.maizegdb.org/cgi-bin/displaypersonrecord.cgi?id=15938>).

For annotation updates, attribution to the community annotator (by name with linkage to the “Person Record” at MaizeGDB) will be stored and displayed for both structural and functional annotations via GBrowse (mouse over for structural attributions) as well as via gene model pages (structural and functional). As a set of annotations is the current version in use, all annotations and updates contributed will be made available. When new assemblies and annotations are released, the official or default structural annotation version for a given gene will be displayed along with linkages to the evidence that supported the updated annotation (expert annotations of the previously released structure) available. This process recapitulates MAKER-based annotation methods in that structural annotations contributed by researchers directly can be weighted differentially in updated genome annotations, thus enabling a consistent display of how contributions feed into new annotation releases.

It will be important to measure how successful the project to engage community annotation is. The number of gene models improved as an outcome of this plan can be reported annually, but it is not clear that numbers of gene models improved would be a reasonable measure of success or user satisfaction. To get at user satisfaction, questions relating to community perception of the project’s success will be included in community surveys (generally conducted by the Maize Genetics Executive Committee) as well as via a questionnaire about database usage that is accessible to all researchers who submit abstracts for the Annual Maize Genetics Conference. Information collected via these mechanisms will be evaluated and system changes to improve community engagement will be devised and implemented in keeping with collected feedback from these venues.

**Goal 1.b:** Deliver sequence-based representations of maize diversity, both with

respect to the B73 reference genome and in the absence of homologous reference sequence.

**Background and Related Research:** With the diversity of maize so well-documented, the lack of an accepted data model that enables alternate assemblies and/or coordinate systems would be a valuable tool for investigating diversity. In the past, there was an assumption that a higher-order genome (e.g., the genomes of human and maize) could be represented by end-to-end sequence with few gaps and separated into chromosomes, and that variations would be primarily single nucleotide variants. This proved wrong on both counts. In addition, it has become apparent that genome assemblies are never perfect and many contain contigs that have not been ordered and oriented on a chromosome or that have not been assigned to a particular chromosome because SVs (structural variations) often impact a large region of an assembly and can take the form of CNV (copy number variations), PAV (presence/absence variations), inversions, translocations, complex re-arrangements, etc. A single reference genome consequently does not provide a linear coordinate system onto which all observed variants can be represented across the genome. For example, if a haplotype does not exist in B73, it is not readily apparent how to represent the alternate genome. In the case of maize as a species, some sequences don't appear in the B73 reference assembly at all because the sequence doesn't exist in the B73 genome.

Community-driven standards have the best chance of success if developed within the auspices of international working groups. The Genomic Standards Consortium (GSC) was formed in September 2005 as an international working body, following a workshop held at the National Centre for Environmental E-Science in Cambridge, United Kingdom. The GSC has introduced the "Minimum Information about a Genome Sequence" (MIGS) guideline and its genomic contextual data markup language, GCDML, for MIGS and for "Minimal Information about a Metagenomic Sequence" (MIMS). Participants in the GSC include biologists, computer scientists, those building genomic databases and conducting large-scale comparative genomic analyses, and those with experience of building community-based standards. The mission of the GSC is to work with the wider community towards: the implementation of new genomic standards; methods of capturing and exchanging metadata; and harmonization of metadata collection and analysis efforts across the wider genomics community by encouraging vast, open membership.

Initial models for representing genome assemblies used a single preferred tiling path to produce a single consensus representation of the genome. Subsequent analysis has shown that for most mammalian genomes a single tiling path is insufficient to represent a genome in regions with complex allelic diversity. Indeed, various online resources have emerged for the representation and analysis of SV, mainly for human (Sneddon and Church 2012). The Genome Reference Consortium (GRC; mentioned above) has been at the forefront of developing mechanisms to enable alternate genomic assembly representation. The GRC recently has published a data model that enables the representation of alternate assemblies and diversity within a reference genome assembly and that provides more robust substrates for genome analysis (Church et al. 2011). In their model, a primary tiling path is assigned and alternates are stored in regions of known diversity and/or assembly alternates. In developing this data model, the GRC aimed to enable the continual improvement of mature reference genome assemblies with a mechanism that allows community feedback into the process of assembly curation. Their solution addresses both these needs and allows

incremental improvements (minor releases) so that researchers have early access to data, but without changing the coordinate system (major releases). The data model accommodates complex regions that have alternative tiling paths and tracks versions of *collections* of sequences rather than individual sequences. Currently only vertebrate genomes are managed using the GRC data model (i.e., human, mouse, and zebra fish), but the consortium is willing to consider partnering with other mature eukaryotic genome assemblies for which there is a dedicated commitment to continued improvement and associated funding. The assembly database is available online at <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/index.shtml> and mechanisms for community members to report issues on these assemblies are also available (<http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/ReportAnIssue.shtml>).

**Approach:** MaizeGDB will apply the alternate assembly/diversity data representation model as outlined by the GRC and develop alternate and additional community standards in collaboration with the GSC and GRC by:

- Adopting and developing mechanisms to enable diversity data visualization at MaizeGDB, building on the Human Hapmap project at NCBI where freely available GBrowse modules currently include mechanisms to view haplotypes, SNP prevalence, and LD across genomic segments including, but not limited to, segments represented in the reference genome assembly.
- Initially storing diversity data at MaizeGDB using the GRC-developed data model (for currently available diversity data including HapMap\_v1 and HapMap\_v2 (E. Buckler, PI) as well as data forthcoming from the Maize PAV/CNV project funded by the NSF (P. Schnable, PI) to accommodate immediate needs and familiarize the MaizeGDB team with these types of datasets.
- Implementing mechanisms to automatically pull and/or access diversity data from NCBI's dbVar and/or iPlant infrastructure for diversity representation at MaizeGDB.
- Instructing researchers how to deposit their diversity data directly to NCBI's dbVar.

Because the volume of data to be stored and analyzed will require large-scale storage and high compute power, multiple avenues of data storage and analysis will be pursued simultaneously to insure that a workable solution is developed. This will involve not only storing and analyzing data at MaizeGDB for modeling and developing query and visualization tools, but also experimenting with storing and accessing data within the iPlant Cyberinfrastructure via the MaizeGDB interface. For some data access needs, precomputed results datasets for large-scale queries, analyses, and visualization requests will be required. Specific examples of such usage cases will be gathered from Cooperators to ensure that common queries are given priority for development. Details regarding timeline, milestones, and deliverables are in the Milestones Table toward the end of this Project Plan.

**Contingencies:** Other, independent projects to assemble and annotate the B73 reference genome might not publicize their intentions or could miss established deadlines. In such an instance, personnel at MaizeGDB will directly maintain information describing those projects' timelines and deliverables as closely as possible. Additional projects and individual researchers might be funded to work on maize genome assembly and annotation. In such an instance, project personnel would be contacted by MaizeGDB team members and made

aware of the process in place to gather information toward future improvements to the genome. Because the GRC has tools and data structures to document known issues with assemblies and annotations these GRC- developed resources might be applied to maize directly. If this occurs, effort would be redirected to working with GRC to analyze and improve the maize assembly and annotation. Potential problems that can emerge as a direct result of incorporating third-party tools (like the GRC-developed resources) include difficulties in their adoption and rapid development, which can lead to software quality and accuracy issues. However, their use is planned for this project because 1) we lack sufficient in-house programming staff to develop all the tools needed, 2) the tools we are considering were intended for general use, 3) the developers of these tools are generally very eager to help make them work properly, and 4) to help justify investment in general tool development. We also are working and will continue to work to ease some anticipated ‘adoption pains’ by direct involvement in the development process as testers and/or code contributors. Though not anticipated to be a likely risk at this point, if collaboration with the GRC should happen to fail, porting GRC tools to MaizeGDB directly will not be an option given current resources and the resource investment that likely would be required to deploy those tools. However: parts of the GRC workflow have already been implemented at MaizeGDB. In particular, the free-form assembly error reporting mechanism already is in place for gene model record pages at the redesigned MaizeGDB Alpha website available at <http://alpha.maizegdb.org/> (currently accessible and planned to become the default MaizeGDB entry page in November of 2013). If tools developed to accommodate animal systems are not adequate to analyze and/or display the levels and types of sequence and structural variation present within maize, scaling back to representing the diversity within lines most frequently accessed by researchers may be required. In that case, Cooperators will be queried to determine which lines are most important across the research community and efforts will focus on supporting the diversity represented by those lines.

**Collaborations:** Drs. D. Ware, ARS, Cold Spring Harbor, NY and M. Yandell University of Utah, Salt Lake City, UT to document currently funded project products and schedules as well as routing community annotations for improvements to RefGen\_v4. Drs. V. Brendel, Indiana University, Bloomington, IN; S. Lewis, Lawrence Berkeley National Laboratory in Berkeley, CA; Drs. D. Micklos, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY; and V. Schneider, National Center of Biotechnology Information to develop and deploy tools to document community contributed improvements for genome assembly and annotation. Drs. E. Buckler, ARS, Ithaca, NY and P. Schnable, Iowa State University, Ames, IA to implement diversity models and data representations that accommodate diversity data their groups are generating as well as Drs. V. Schneider, National Center of Biotechnology Information (key member, GRC), J. Wooley, University of California, San Diego and (PI, GSC), and S. Goff (University of Arizona and iPlant Collaborative) to implement, develop, and deploy diversity storage and display mechanisms.

**Objective 2:** Create tools to enhance access to expanded datasets that reveal gene function and datasets for genetic and breeding analyses. (TZS; non-hypothesis driven)

**Goal 2.a:** Enable researchers to access high-quality functional descriptions for maize gene products by documenting their potential involvement in particular biochemical and metabolic pathways.

**Background and Related Research:** The MGSC assembled the B73 pseudomolecule sequences as well as gene models. However, knowledge of how these gene products interact in order to generate a desired phenotype is limited. Building a representation of metabolic pathways, i.e., a map of interactions among gene products (e.g., RNA, protein), carbohydrates, lipids, metabolites, etc., is a proven method to uncover the reaction routes of a gene product or a group of gene products to create a given phenotype. Documenting characterized reaction routes and biochemical networks helps researchers to focus their efforts to create discrete, testable hypotheses. However, generating metabolic networks is inherently challenging because pathways do not always evolve similarly in different species and the function, structure, and activity of even orthologous gene products are strongly influenced by their genomic context and evolutionary trajectory. For example, C<sub>3</sub> plants like rice are not anticipated to carry out photosynthesis using exactly the same pathways and genes as C<sub>4</sub> plants including maize and sorghum. Therefore, building species-specific metabolic pathways is an absolute necessity for hypothesizing the complex biological processes that underlie metabolic and phenotypic differences between species and among genotypes within a particular species.

The BioCyc databases (<http://www.biocyc.org>) aim to overcome this challenge by building different metabolic pathways for each species and for individual lines within a species (currently 1,962 pathway/genome databases are available). In the case of maize, two BioCyc-based metabolic networks are available with varying levels of accuracy and coverage: MaizeCyc and CornCyc. In 2011 and 2012, MaizeGDB personnel collaborated with Gramene to generate MaizeCyc, which is currently available at both MaizeGDB and Gramene (accessible at <http://maizecyc.maizegdb.org/> and at <http://www.gramene.org/pathway/maizecyc.html>). In 2012, another maize metabolic network, called CornCyc, was independently developed by Plant Metabolic Network (PMN) personnel working under the guidance of Dr. Sue Rhee at Stanford University. Different methodologies were used by each project to assign gene functions, but both used the BioCyc Pathway Tools suite to create metabolic network representation based on these functional assignments. In the case of MaizeCyc, the method is based on (1) sequence similarity score to UniProt peptides obtained via the Ensembl XRef pipeline (Slater and Birney 2005), (2) MaizeGDB curated pathways, and (3) orthologous gene function import based upon RiceCyc, AraCyc, and PlantCyc supported by phylogeny-based clustering methods (Monaco et al, unpublished results). CornCyc, in contrast, uses three different functional assignment methods. BLAST (Altschul et al. 1997), CatFam (Yu et al. 2009), and Priam (Claudel-Renard et al. 2003) are used to determine sequence similarity and protein families in tandem with an Ensembl pipeline.

CornCyc is a high-confidence, low-coverage set, whereas MaizeCyc is a lower-confidence, high-coverage set. The performances of these two methodologies are strikingly different. We compared enzymatic assignments of MaizeCyc and CornCyc based on UniProt



annotations for 198 maize proteins in collaboration with PMN (unpublished). For MaizeCyc, the F-measure (harmonic mean of precision and recall) for functional assignment is 0.365 (precision = 0.711, recall = 0.245), while for CornCyc, the F-measure is 0.893 (precision = 0.873, recall = 0.913). These measures indicate that CornCyc contains a higher-confidence set. In terms of coverage, a different picture emerges: while CornCyc has enzymatic assignments for 10,511 polypeptides, MaizeCyc provides assignments for a much larger set: 39,656 polypeptides (including gene products from transcripts encoded by the same gene, e.g., via alternative splicing events).

**Approach:** Because CornCyc is a high-confidence dataset, curation of literature-based functional annotations from MaizeGDB is slated for inclusion in the CornCyc resource. Collaborator Dr. Mary Schaeffer, a curator for the MaizeGDB project, will curate functional annotations of maize genes from the literature into MaizeGDB including select biochemical pathways including abscisic acid signaling and auxin biosynthesis. Other specific pathways to be curated will be selected based on available data and priorities identified by the MaizeGDB WG, with the bulk of activity focused on carbohydrate metabolism, photosynthesis, pest defense, flowering, and seed quality. Functional descriptors including Gene Ontology (GO) terms (GO Consortium 2012) will be entered into MaizeGDB database through customized curation tools that are available for use both by professional curators and by community curators via the MaizeGDB interface. These tools were recently developed and are slated to become available in March of 2013. Curators also can add more complex curations of gene function through the Pathway Tools interface to enable functional additions to the CornCyc resource. As shown in the Milestones Table, MaizeGDB is committed to annual releases of the CornCyc resource.

MaizeGDB will support collaborative curatorial activities by: coordinating the communication and curatorial activities between MaizeGDB and PMN personnel, ensuring timely delivery of data in different formats to the maize community and annotating groups, and mirroring CornCyc at MaizeGDB. Functional annotations made by MaizeGDB curators as well as the community at large will be available through MaizeGDB by way of gene model pages, by creating mechanisms for bulk downloads, and via a CornCyc or other visualization tool interface. GO term annotations also will be converted into Bioconductor's annotation package format, similar to the ones available for Arabidopsis and other species, which will enable researchers to analyze these annotations using R/Bioconductor. We will also make the pathways available in the BioPAX format, to which Pathway Tools can natively export. These BioPAX-formatted pathway files can then be used as input to Cytoscape, as well as to Bioconductor, for further visualization and analysis. MaizeGDB website users and Cooperators will be made aware of new releases of metabolic networks and serve these pathways through news items, updates, and tutorials made available at MaizeGDB and by on-site tutorials of research groups as well as online video tutorials (see Objective 4).

**Goal 2.b:** Develop and deploy network-based data access and analysis tools that support predictive biological investigations routinely pursued by basic biologists and breeders.

**Background and Related Research:** Physical, genetic, and expression network visualization solutions are inherently challenging as so many data points can create complex graphs with limited mechanisms available to extract useful information. A similar challenge exists in applied breeding, where the deluge of SNP data cannot be easily represented to show relationships among inbred lines and other breeding material. Some software packages specifically address, for example, haplotype and genotype visualization for germplasm, e.g., FlapJack software (Milne et al. 2010; <http://bioinf.scri.ac.uk/flapjack/>), which provides views of alleles colored according to their frequency or similarity. Other software suites aim to provide general visual solutions to represent relationships among diverse entities (e.g., proteins, genes, maize lines). CurlyWhirly (<http://bioinf.scri.ac.uk/curlywhirly>), for example, is a simple 3D Java viewer to visualize clusters of data points. Cytoscape (<http://www.cytoscape.org/>) and Cytoscape Web (Smoot et al. 2011) enable mapping numerical data such as gene expression levels on the network.

In addition to the need to enable researchers to visualize relationships among entities across data types (e.g., networks, similarities between lines), analysis tools are required to infer the relationships among entities. This sort of data analysis software is urgently needed for breeding analyses. Cytoscape, for example, has a wide range of applications (“plug-ins”) for analysis, visualization, and data transfer, but these tools have not been adapted to solve the needs of breeders who must select lines to include in germplasm enhancement pipelines. Tools specific to inferring similarities among breeding germplasm are also being developed such as computational pipelines that analyze GBS data to infer relationships among lines. GAPIT calculates kinship and performs genome-wide association studies (GWAS) analyses from the command line (Lipka et al. 2012). Once the similarity among lines is solved, the findings should be compared to documented pedigrees. TASSEL, a stand-alone software system can store pedigree information and genotypic data side-by-side (Bradbury et al. 2007) The Genetic Records Family Tree (GRFT) Web Applet enables researchers to visualize records of genetic crosses (Pimentel et al. 2011). For marker visualization, GGT, the Graphical Genotypes software package (van Berloo 2007) can be used to filter unwanted alleles of a genomic region.

Breeders possess in-depth pedigree knowledge of maize lines, and haplotype information obtained by GWAS and QTL studies can be then traced back to parents based on pedigrees. The addition of visualization and analyses tools to the MaizeGDB resource will assist researchers to associate underlying genotypes with desirable phenotypes and aid in genomic selection analyses.

**Approach:** Interaction network, breeding similarity network, and pedigree visualization software development will be pursued. A list of software options available for representing protein-protein interactions as well as genetic, metabolic, and breeding networks and pedigree representation will be compiled by project personnel in year 1 (see Milestones Table). To decide which software from the list would be most useful to address researchers’ stated needs, the MaizeGDB team will work with the MGEC to survey Cooperators to gather use cases and functionality requirements, and scientists (including breeders) will be invited to participate in discussions of the topic at the Annual Maize Genetics Conference in year 2. In addition to queries aimed at gathering information describing data analysis and visualization needs, the survey will include questions that aim to identify discrete data sets that should be brought into MaizeGDB to support such analyses. Once data analysis and software

visualization suites have been selected, fully functional versions of these tools will be implemented within the MaizeGDB environment over the course of years 2 and 3 with all three tools (i.e., interaction network, breeding similarity network, and pedigree visualization software) available by the end of year 3. Implemented tools will be maintained and modified according to researchers' input. Development will be carried out with input requested from Cooperators (via surveys and individual interactions), the MaizeGDB WG, and *ad hoc* committees throughout. The development process and timing for deliverables are documented in the Milestones Table.

**Contingencies:** Dr. Sue Rhee's CornCyc project is funded by the NSF until August 2015. Her group is committed to continuing to assist MaizeGDB curators by providing curatorial and technical assistance and updating CornCyc with literature-based data with until the end of their project's funded period. Early in 2015 MaizeGDB team members will consult with Dr. Rhee to determine how best to proceed with literature-based metabolic pathway data curation. Possible courses of action include continuing the collaboration in a similar fashion, adopting the CornCyc database and automated development pipelines at MaizeGDB, or choosing to enable access to metabolic pathway data using a different system. At the same time, it should be noted that other metabolic pathway databases may be developed that are superior to the currently available BioCyc-based CornCyc database. If such systems appear to better meet maize researchers' needs, alternate data storage and visualization relative to metabolic pathways will be evaluated and implemented and/or the annotations stored at MaizeGDB will be transitioned for representation into alternate venues. For network analyses and pedigree visualizations, we assume that third-party tools already available for implementation will meet the needs to be articulated by researchers via stakeholder feedback mechanisms. If existing third-party solutions do not meet these needs, we will adapt existing or create custom tools and interfaces to provide data access and analysis. Because third-party tools have been implemented successfully as reported in the literature, this contingency is not likely. If new software is developed that better address data access and analysis needs, plans for implementing alternate solutions will be developed and pursued.

**Collaborations:** Dr. M. Schaeffer, ARS, Plant Genetics Research, Columbia, MO for performing and providing curation; Dr. S. Rhee and PMN personnel at the Carnegie Institution for Science at Stanford University for training MaizeGDB curators, providing curatorial and technical assistance, and sharing the CornCyc database; Dr. D. Ware, ARS, Cold Spring Harbor, NY, for network data exchange between MaizeGDB and Gramene; Dr. Jack Gardiner, Iowa State University, Ames, IA, for expertise exporting, transforming, and loading expression data for use with data analysis and visualization tools including, but not limited to, R/Bioconductor and Cytoscape. Dr. G. Bader, Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada for providing technical assistance on visualizing interaction networks; Dr. E. Buckler, ARS, Plant, Soil and Nutrition Research at Cornell University, Ithaca, NY for providing TASSEL and GAPIT source code as well as maize genotyping; Drs. T. Lubberstedt and K. Lamkey, Iowa State University, Ames, IA as well as Dr. C. Gardner, ARS, Plant Introduction Research Unit, Ames, IA for providing guidance on the needs of breeders, pedigree information, and survey preparation.

**Objective 3:** Deploy tools to increase user-specified flexible queries. (CJL; non-hypothesis driven)

**Goal 3:** Enable researchers to quickly retrieve customized datasets from the MaizeGDB database.

**Background and Related Research:** Currently, the MaizeGDB database interface allows researchers to access data record-by-record. For example, query tools return lists of individual locus pages organized by name when the locus datacenter is queried rather than tables of data that include only the information relevant to a particular research need. Mechanisms and tools that allow researchers access to larger sets of data do exist, including web-based and command-line access though these are rarely used. Some tools deployed at MaizeGDB (e.g., the MaizeGDB Genome Browser) allow users to visualize large amounts of data. Other tools (e.g., the Bin Viewer) allow access to large sets of data that share common characteristics (for this example, results sets belong to the same ~10 cM chromosomal segment called a ‘Bin’). MaizeGDB also encourages users to request bulk-data reports, which are developed and then delivered via the website. The problem is that the outputs are usually static, predetermined, and inflexible.

Two of the most widely implemented software solutions that address this problem are BioMart (Zhang et al. 2011) and InterMine (Chen et al. 2011). Both systems’ goals are to provide increased performance on read-only datasets. BioMart and InterMine were designed specifically for biological databases and the complex relationships that are common among biological datasets. Both tools provide customizable web interfaces that allow researchers to interact with databases directly. BioMart is currently implemented in over 40 databases including plant databases Gramene, Ensembl Plants, Rice-Map, Potato database, Phytozome, KazusaMart, SalmonDB, Cildb, WormBase, Ensembl Plants; and other model organism databases including: WormBase, Mouse Genome Informatics, and SalmonDB. InterMine is currently being used at 6 databases including: Rat Genome Database's RatMine, Saccharomyces Genome Database's YeastMine, and FlyMine.

**Approach:** We will determine how best to implement software tools designed to address the need to access custom data sets (contrasting with currently available data point query tools as well as bulk tools that allow predetermined queries). Irrespective of which software is selected, implementation will involve mapping MaizeGDB-specific internal databases tables to the third-party tool’s schema. Regular updates of data and mappings would be required. Currently, there are several software solutions available. For some third-party software the user-interface and report formats are standardized. Researchers might have already interacted with the same tools at other databases. This creates for those implementing the software a pre-existing peer-group with whom to interact on deployment and allows a more gradual learning curve for researchers. Stakeholders will be surveyed to determine which software solution would best meet data access needs. MaizeGDB personnel will implement the software, work with stakeholders to identify issues, and improve the deployed tools to address identified issues as they are identified.

Additional pre-configured reports will be created based on customizable ‘views’ of the database (where the term view is defined to be a stored query against a database). The results of the query can be saved in a structured flat-file (e.g., tab-delimited format). A few potentially useful data views will be designed and constructed to provide a mechanism for

researchers to request additional views. These reports would be available for download and updated regularly (monthly or quarterly to enable researchers to access to all or large amounts of the data stored at MaizeGDB. These reports can subsequently be parsed by researchers directly to identify information relevant to particular research problems or incorporated into their data pipelines.

Potential tools to address these needs will be evaluated based on ease of implementation and breadth of query and report capabilities. Selected tools will be developed and released via MaizeGDB, and implemented tools will be maintained and modified according to researchers' input and long-term maintenance will be supported. Development will be carried out with input requested from Cooperators, the MaizeGDB WG, and *ad hoc* committees throughout. The development process and timing for deliverables are documented in the Milestones Table.

**Contingencies:** Potential problems that can emerge as a direct result of incorporating third-party tools include difficulties in their adoption and rapid development, which can lead to software quality and accuracy issues. However, their use is planned for this project because 1) we lack sufficient in-house programming staff to develop all the tools needed, 2) the tools we are considering were intended for general use, 3) the developers of these tools are generally very eager to help make them work properly, and 4) to help justify investment in general tool development. We also are working and will continue to work to ease some anticipated 'adoption pains' by direct involvement in the development process as testers and/or code contributors. If existing third-party solutions do not meet researchers' needs for real-time, flexible, user-customizable access to data at MaizeGDB, we will create internal tools and interfaces to provide access to the database. This would include enhanced interfaces and forms to gather requests from a user. Because third-party tools have been successfully implemented at a substantial number of other databases, this contingency is not likely. However, if users fail to utilize these tools, further outreach will be provided (examples include video tutorials, step-by-step documentation, and example use-case scenarios). Additionally, if new software is developed that better address our data-access needs or specific alternate software is requested by community members (e.g., the Table Browser tool, which is a component of the UCSC Genome Bioinformatics website), new, additional, or alternate software will be adopted and/or made available via the MaizeGDB website.

**Collaborations:** Dr. D. Ware, ARS, Cold Spring Harbor, NY for guidance and consultation based on their current implementation of GrameneMart (an instance of BioMart). Dr. R. Shoemaker, ARS, Ames, IA for guidance, consultation, and possible knowledge sharing as his project has similar goals to improve data access.

**Objective 4:** Provide community support services, training and documentation, meeting coordination, and support for community elections and surveys. (CJL; non-hypothesis driven)

**Goal 4:** Facilitate communication among maize researchers to support the needs of the research community and to create and leverage synergistic activities.

**Background and Related Research:** The maize research community's success with collaborating as a focused group can be largely attributed to the coordinated activities of the

MGEC, the MGCSC, and the cooperative spirit of individual maize researchers. MaizeGDB also is partially responsible for coordinating maize research (see Appendix C).

The first three Objectives of this proposal will provide requested and valuable new tools to the maize research and breeding community. The purpose of Objective 4 is threefold: to keep the maize community informed of new data as they are made available via MaizeGDB, educated in the use of new tools for accessing these new data, and to facilitate open communication with and among the community. Individual scientists may not be familiar with every type of data or tool we offer and must continually work to transfer information about the technologies and data made available via MaizeGDB. We consider it of the utmost importance to keep our stakeholders abreast of ongoing activities at MaizeGDB, and provide outreach and education whenever the need arises.

**Approach:** To publicize and refine standards for data inclusion, the MaizeGDB team will continue to work with researchers directly to further develop and refine the set of guidelines ensuring that their data may be made available through MaizeGDB, and to provide direct support if needed to incorporate their data into correct formats for meeting MaizeGDB quality standards. These guidelines will include descriptions of the types of data that can be incorporated directly, links to standards for data deposition and defined file formats for automated inputs, and schedules for incorporating data into the database that will also include a list of focus data types for curation activities by month. In addition, personnel at MaizeGDB will continue provide direct support, as needed, for researchers to transform their data into correct formats and to meet defined quality standards.

Second, researchers will be instructed in the use of the MaizeGDB website including, but not limited to, types of data that are available and how those data can be accessed and used. This will be ever more important as we deploy increasingly complicated tools, such as genome browsers that allow queries of hundreds of different maize sequences (Objective 1) and tools for analyzing biological networks (Objective 2). Live training sessions, web-based video tutorials, and online manuals as well as email communications and phone calls have been successfully employed to provide guidance and instruction in the past (Harper et al. 2011; Schaeffer et al. 2011). These activities will be continued and expanded to reach a broader audience.

Third, MaizeGDB will continue to provide the infrastructure for researchers to elect new members to the MGEC. This entails inviting researchers to nominate others for the ballot, confirming nominees' willingness to serve if elected, and sending out unique keys to researchers that will allow them to vote only once per person. We also will conduct surveys of the maize community as directed by the MGEC. Other activities that we will pursue include making available custom software that enables the submission of abstracts for the Annual Maize Genetics Conference, maintaining the conference website, managing reports that enable the MGCSC to choose speakers from submissions, and creating the conference program.

Finally, MaizeGDB will organize meetings of the MaizeGDB WG that are scheduled biannually (one will be a general meeting and the other will be an informal meeting focused on some particular data type or need; see letter from WG Chair Dr. Mihai Pop). The MaizeGDB team will create documents for the formal meeting that outline recent progress as well as current activities and future plans. Any reports or guidance from the WG will be considered based on the availability of resources, and personnel at MaizeGDB will draft and

post responses to their guidance online. As the availability of data and new tools increases exponentially, guidance from our WG will continue to be essential in helping MaizeGDB deploy tools and data that are of top priority to our stakeholders.

**Contingencies:** If the MGCSC or the MGEC decide to manage information technology-related tasks directly, the need to rely upon the MaizeGDB team would be obviated. In that case, efforts would be redirected to expand efforts in support of Objectives 1 and 2.

**Collaborations:** The MGEC (see <http://www.maizegdb.org/mgec.php> for current membership and affiliations) to keep abreast of current needs communicated by the community, the MGCSC (see the most recent conference site listed at [http://www.maizegdb.org/maize\\_meeting/](http://www.maizegdb.org/maize_meeting/) for current membership and affiliations), and the MaizeGDB WG (see [http://www.maizegdb.org/working\\_group.php](http://www.maizegdb.org/working_group.php) for current membership and affiliations).

## PHYSICAL AND HUMAN RESOURCES

### Personnel:

*Dr. Carolyn J. Lawrence*, Lead Scientist (Research Geneticist; under the supervision of Research Leader Dr. C.A. Abel).

*Dr. Taner Z. Sen*, SY (Cat 4 Computational Biologist; under the supervision of Research Leader (RL) Dr. C.A. Abel).

*Carson M. Andorf*, Bioinformatics Engineer (Information Technology Specialist; see CV under Appendix A; under the supervision of Dr. Lawrence).

*Darwin A. Campbell*, Database Administrator (Information Technology Specialist; under the supervision of Dr. Lawrence).

*Dr. Elisabeth C. Harper* (Cat 3 Curator/Geneticist; located at the Plant Gene Expression Center, Albany, CA; under the supervision of Dr. Lawrence) is 0.5 FTE on this project.

### Offices:

#### Ames, IA

The MaizeGDB group in Ames (i.e., Lawrence, Sen, Andorf, and Campbell) as well as the SoyBase (Dr. R. Shoemaker; ARS) and PLEXdb (Dr. R. Wise; ARS) groups, are located in the Crop Genome Informatics Laboratory, an office building on the Iowa State University campus. Also located in the same building are RL Abel and unit administrative support. These groups share ca. 6,000 ft<sup>2</sup> of space including offices and a state-of-the-art server room.

#### Albany, CA

Curator Harper (administratively associated with Ames but physically located in Albany) is on location at the Plant Gene Expression Center and occupies an office (ca. 150 ft<sup>2</sup>) set aside by Center Director Dr. S. Hake for MaizeGDB personnel (see attached letter).

### Server Room:

Ca. 330 ft<sup>2</sup> climate controlled room in Crop Genome Informatics Laboratory, shared with Drs. Shoemaker and Wise, houses production MaizeGDB servers.

### Project-level Machines:

1 x HP 2312FC DC Modular Smart Array – 1.8TB, redundant power and network fiber channel connectivity.

1 x HP DL785-G6 -- 8 processor/6 core = 48 CPUs @ 2.79 GHz, 262137 MB Ram, 1.6 TB on board data store.

3 x Dell PowerEdge 2850 2 processor x 2 core @ 2.8GHz, 16MB Ram, 500GB on board storage.

2 x Dell Power Edge 2950 2 processor x 4 core @ 2.493 GHz, 32 MB Ram, 1.09TB on board storage.

1 x Dell PowerEdge 2950 2 processor x 4 core @ 1.6GHz, 4MB Ram, 257Gb on board storage.

1 x HP x1600 Network Storage System 1 process 4 core @ 2.26GHz, 6GB Ram, 9TB on board storage.

Databases are backed up off-site weekly to Columbia, MO. Mission critical servers are replicated locally.



## **PROJECT MANAGEMENT AND EVALUATION**

Lawrence will be the overall project manager and the main contact for the MaizeGDB team to interact with the maize community. Lawrence will coordinate with Andorf, Harper, Campbell, and collaborators to accomplish Objectives 1, 3, and 4. For Objective 2, Sen will lead the effort and coordinate with Harper and collaborator Schaeffer to address data curation needs. Lawrence also will be responsible for coordination among aspects of the Objectives, implementing the Objectives in a timely manner, and providing guidance during the lifetime of this Project Plan.

The MaizeGDB team will accomplish these Objectives in coordination with the maize community, the MaizeGDB WG, the MGEC, and the MGCSC. Letters of collaboration are attached to substantiate the community support for and willingness to help MaizeGDB personnel to meet the goals described.

The MaizeGDB team arranges weekly phone conferences among Ames, IA, Columbia, MO, and Albany, CA to assess progress with collaborators involved regularly in those meetings. Several times a year, Schaeffer and Harper visit Ames to facilitate communication. An internal wiki also has been created and is used to document work, schedule meetings, and share information. The MaizeGDB team takes the initiative to establish and maintain contacts with the maize community. The team meets with the WG at least once a year formally, and communicates with the WG and other maize researchers informally numerous other times at conferences and through *ad hoc* phone calls and e-mail.

**MILESTONES AND EXPECTED OUTCOMES**

<b>Project Title</b>	MaizeGDB: Enabling Access to Basic, Translational, and Applied Research Information			<b>Project No.</b>	3625-21000-051-00D
<b>National Program</b>	<b>301 – Plant Genetic Resources, Genomics and Genetic Improvement</b>				
<b>Objective</b>	1: Support stewardship of maize genome sequences and forthcoming diverse maize sequences. (non-hypothesis driven)				
<b>NP Action Plan Component</b>	2: Crop Genetic and Genomic Resources and Information Management				
<b>NP Action Plan Problem Statement</b>	2A: Crop genomic information resources and bioinformatics.				
<b>Goal</b>	<b>SY Team</b>	<b>Months</b>	<b>Milestones</b>	<b>Progress/Changes</b>	<b>Products</b>
1.a Enlist the community of maize researchers in the genome assembly and annotation process to enable their contributions to and use of improved reference genome sequences in real-time.	CJL	<b>12</b>	Determine in consultation with GRC personnel whether maize assembly and annotation can be represented using GRC tools. If not, query Cooperators and document known requirements and user preferences for tools to be developed for assembly error documentation in-house. For structural annotation, consult with Cooperators to determine whether yrGATE, WebApollo, or other tools are preferable with regards to ease of use.		Requirements for collaboration with GRC (Genome Reference Consortium) documented or functional and technical design in place for in-house or third-party tools to be deployed for assembly and annotation error reporting and improvement via MaizeGDB.  Assembly/annotation status page released at MaizeGDB.
	CJL	<b>24</b>			Error documentation tool suites deployed (via GRC or at MaizeGDB directly).  Updated assembly/annotation status page at MaizeGDB.
	CJL	<b>36</b>	Survey Cooperators to determine whether assembly and annotation tool suites meets their needs. Determine whether and how to improve tools.  Revise tools based upon Cooperator feedback.		Updated assembly/annotation status page at MaizeGDB.  Report on the status of maize reference genome assembly and annotation via a peer reviewed journal.
	CJL	<b>48</b>			Updated assembly/annotation status page at MaizeGDB.  Release updated tool suite.
	CJL	<b>60</b>			Updated assembly/annotation status page at MaizeGDB.

Goal	SY Team	Months	Milestones	Progress/ Changes	Products
1.b: Deliver sequence-based representations of maize diversity, both with respect to the B73 reference genome and in the absence of homologous reference sequence.	CJL	12	Implement GRC-developed data model and populate with Reference Assembly, HapMap, and PAV/CNV data.  Adopt and develop mechanisms to enable diversity data visualization based on existing software developed for human.		Diversity interfaces available for deployment at MaizeGDB.
	CJL	24			Diversity data available via the MaizeGDB Genome Browser.
	CJL	36	Implement mechanisms to automatically pull and/or access diversity data from NCBI's dbVAR based on previous experience with accessing NCBI data directly.		Instructions for how to deposit diversity data at dbVAR in place at MaizeGDB.
	CJL	60	Assess recent developments on how best to store and access diversity data with respect to a reference genome sequence, implement necessary updates and changes.		Diversity data deposition instructions updated at MaizeGDB.

<b>Objective</b>	2: Create tools to enhance access to expanded datasets that reveal gene function and datasets for genetic and breeding analyses. (non-hypothesis driven)				
<b>NP Action Plan Component</b>	2: Crop Genetic and Genomic Resources and Information Management				
<b>NP Action Plan Problem Statement</b>	2A: Crop genomic information resources and bioinformatics.				
Goal	SY Team	Months	Milestones	Progress/ Changes	Products
2.a: Enable researchers to access high-quality functional descriptions for maize gene products by documenting their potential involvement in particular biochemical and metabolic pathways.	TZS	12			Literature-based curation of ABA, auxin, and select signaling pathways documented and available at MaizeGDB.
	TZS	24	Identify additional pathways for data curation.		Literature-based curation of carbohydrate and photosynthetic pathways documented and available at MaizeGDB.
	TZS	36	Determine whether CornCyc collaborations should be continued, an alternate tool should be deployed, or the system should be internalized for MaizeGDB administration.		Literature-based curation of select pest defense pathways documented and available at MaizeGDB.
	TZS	48			CornCyc or other metabolic

					pathway representations available at MaizeGDB.  Literature-based curation of select flowering pathways documented and available at MaizeGDB.
	TZS	60			CornCyc or other metabolic pathway representations available at MaizeGDB.  Literature-based curation of select seed quality pathways documented and available at MaizeGDB.
Goal	SY Team	Months	Milestones	Progress/ Changes	Products
2.b: Develop and deploy network-based data access and analysis tools that support predictive biological investigations routinely pursued by basic biologists and breeders.	TZS	12	Determine which software tools and network and pedigree data are available for implementation and deployment.		
	TZS	24	Survey maize researchers including basic scientists and breeders to determine their needs for network software data analysis and visualization.		Survey results available at MaizeGDB.
	TZS	36			Deploy visualization tools for displaying interaction and pedigree datasets.
	TZS	48	Survey Cooperators to determine new and additional datasets to be added to the MaizeGDB resource.		Identified datasets available via MaizeGDB
	TZS	60	Interact with stakeholders to determine necessary changes to data and tools.		Revised data and tools released at MaizeGDB

<b>Objective</b>	3: Deploy tools to increase user-specified flexible queries. (non-hypothesis driven)				
<b>NP Action Plan Component</b>	2: Crop Genetic and Genomic Resources and Information Management				
<b>NP Action Plan Problem Statement</b>	2A: Crop genomic information resources and bioinformatics.				
Goal	SY Team	Months	Milestones	Progress/ Changes	Products
3: Enable researchers to quickly retrieve customized datasets from the MaizeGDB database.	CJL (CMA)	12	Select software for improved data-access.		
	CJL (CMA)	24	Release first version of data-access tool.		Data-access tool is publically accessible from MaizeGDB.
	CJL (CMA)	36	Survey Cooperators to determine necessary changes to existing software.		
	CJL (CMA)	48	Revise tools based upon		Publication documenting

			feedback. Submit journal article outlining functionality of the data-access tool.		functionality of the data-access tool.
--	--	--	--	--	--

<b>Objective</b>	4: Provide community support services, training and documentation, meeting coordination, and support for community elections and surveys. (non-hypothesis driven)
<b>NP Action Plan Component</b>	2: Crop Genetic and Genomic Resources and Information Management
<b>NP Action Plan Problem Statement</b>	2A: Crop genomic information resources and bioinformatics.

Goal	SY Team	Months	Milestones	Progress/ Changes	Products
4: Facilitate communication among maize researchers to support the needs of the research community and to create and leverage synergistic activities.	CJL	12	Conduct annual MGEC Elections, collect abstracts for Annual Maize Genetics Conference, conduct two WG meetings, conduct two outreach visits to stakeholder locations.  Submit journal article outlining MaizeGDB's recent updates and activities to a peer-reviewed journal.		Book of abstracts for Annual Maize Genetics Conference, report document to WG describing recent activities, at least two new outreach video tutorials available online.  Updated data submission guidelines at MaizeGDB.  Publication documenting MaizeGDB's updates and activities.
	CJL	24	Conduct annual MGEC Elections, collect abstracts for Annual Maize Genetics Conference, conduct two WG meetings, conduct two outreach visits to stakeholder locations.		Book of abstracts for Annual Maize Genetics Conference, report to WG on recent activities, at least two new outreach video tutorials available online.
	CJL	36	Conduct annual MGEC Elections, collect abstracts for Annual Maize Genetics Conference, conduct two WG meetings, conduct two outreach visits to stakeholder locations.  Submit journal article outlining MaizeGDB's recent updates and activities to a peer-reviewed journal.		Book of abstracts for Annual Maize Genetics Conference, report to WG on recent activities, at least two new outreach video tutorials available online.  Publication documenting MaizeGDB's updates and activities.
	CJL	48	Conduct annual MGEC Elections, collect abstracts for Annual Maize Genetics Conference, conduct two WG meetings, conduct two outreach visits to stakeholder locations.		Book of abstracts for Annual Maize Genetics Conference, report to WG on recent activities, at least two new outreach video tutorials available online.
	CJL	60	Conduct annual MGEC Elections, collect abstracts for Annual Maize Genetics Conference, conduct two WG meetings, conduct two outreach visits to		Book of abstracts for Annual Maize Genetics Conference, report to WG on recent activities, at least two new outreach video tutorials available online.

			stakeholder locations. Submit journal article outlining MaizeGDB's recent updates and activities to a peer- reviewed journal.		Updated data submission guidelines at MaizeGDB.  Publication documenting MaizeGDB's updates and activities.
--	--	--	--	--	--

## ACCOMPLISHMENTS FROM PRIOR PROJECT PERIOD

**Terminating ARS Research Project Number:** 3625-21000-051-00D

**Title:** The Maize Genetics and Genomics Database

**Project Period:** April 10, 2008 to April 9, 2013

**Investigators and FTE:**On this CRIS:

Carolyn J. Lawrence – 1.00 FTE Cat 1 Research Geneticist (Lead Scientist)

Taner Z. Sen – 1.00 FTE Cat 4 Computational Biologist

Elisabeth C. Harper – 0.50 FTE Cat 3 Geneticist

Darwin A. Campbell – 1.00 FTE Cat 6 Information Technology Specialist

Carson M. Andorf – 1.00 FTE Cat 6 Information Technology Specialist

Craig A. Abel – 0.03 FTE Cat 1 Research Leader

Collaborating in Columbia, Missouri:

Mary L. Schaeffer – 1.00 FTE Cat 4 Geneticist

**Project Accomplishments and Impact:**

**Accomplishment 1:** Made the genome sequences of B73 and Palomero Toluqueño accessible to scientists worldwide. The maize B73 genome reference sequence comprises both an assembly of each chromosome and associated, annotated gene structures. To enable researchers to interact with the maize genome sequence, MaizeGDB team members developed and deployed of a genome browser at MaizeGDB. To allow visualization of gene structures and to improve poorly supported models, the team worked with Dr. V. Brendel to make the PlantGDB yrGATE tool (described in Objective 1) accessible via the MaizeGDB genome browser. Team members also developed the Locus Lookup tool, which allows researchers to specify a locus name and arrive at its genomic region even if the locus has not been placed on the genome assembly. At the same time that the B73 reference genome was being sequenced in the US, Mexican scientists at CINVESTAV (in English, the Center for Research and Advanced Studies of the National Polytechnic Institute) were sequencing Palomero Toluqueño, a Mexican popcorn landrace. Unlike B73, Palomero's genome sequence was not deposited at GenBank, so MaizeGDB personnel to made the GenBank submission directly. **Impact:** The genomes of B73 and Palomero Toluqueño are accessed over 2,500 times daily via the MaizeGDB genome browser. Scientists designing experiments to test the function of maize genes are able to locate regions of the genome likely to contain their gene or region of interest, evaluate the quality of gene structures before beginning experiments, and improve structural annotations of the maize genome directly. (Andorf et al. 2010; Sen et al. 2009; Sen et al. 2010)

**Accomplishment 2:** Created novel venues for technology transfer, interaction, and outreach that enable diverse groups to make use of bioinformatics tools and benefit from the availability of biological information and opportunities. Informatics resources are routinely criticized for failure to document details of how information is accumulated and

analyzed as well as for the sometimes significant effort required on the part of researchers to learn how best to interact with available data. To address these needs, community bulletin boards have been made available for researchers at MaizeGDB, an outreach video tutorial site was developed and deployed, outreach workshops are regularly conducted at the Maize Genetics Conference, and curators have been physically located at institutions where many maize researchers work to increase opportunities for direct interaction (i.e., the University of Missouri; the University of Arizona; and the USDA-ARS Plant Gene Expression Center [PGEC] in Albany, CA. **Impact:** MaizeGDB team members consider outreach to be a primary service associated with the project. The MaizeGDB project is cited in publications as well as by funding agencies as a model for how best to interact with stakeholders including scientists, growers, and the general public. (Harper et al. 2011; Lawrence 2011; Schaeffer et al. 2011; Sen et al. 2010)

**Accomplishment 3:** Created and deployed novel, state of the art methods to access distributed data and heterogenous datasets. Because the volume of biological data available has increased and myriad data repositories are now available, it has become difficult to address complex biological questions using only one data source. To address this problem, MaizeGDB team members worked collaboratively with teams of scientists and IT specialists to develop the Bioextract Server and POPcorn resource to enable access to multiple repositories' data from a single location. BioExtract offers access to distributed data, allows the incorporation of user-specified tools, and can document and reproduce defined workflows. POPcorn's developed functionalities address four specific challenges: researchers' inability to locate all websites serving particular datatypes, the repetitive nature of performing the same sequence search at multiple websites, the need to discover all types of data related to a particular sequence, and the problems associated with long-term data storage once small-scale research projects are completed. A third system called VPhenoDBS takes as input an image of a mutant phenotype and searches for similar images, thus taking a researcher's subjective interpretation of the phenotype away and retrieving images based upon visual similarity. VPhenoDBS also serves a text-based search tool that makes use of controlled vocabularies (ontologies) to classify mutant phenotypes and weight and sort returned results by relevance. **Impact:** Researchers now access information that otherwise would not be brought to bear on their scientific investigations. For example, scientists visit the POPcorn site more than 2,000 times per month with accesses primarily from the US (43%) and China (28%). All three projects are the first of their kind and are the foundation for revolutionary tool development going forward. (Cannon et al. 2011; Green et al. 2011; Lushbough et al. 2010; Lushbough et al. 2008)

#### **How Past Objectives and Accomplishments Relate to the Current Plan:**

The MaizeGDB team has created a sequence-centric genetics and genomics database and website (Accomplishment 1) that serves as the basis for developing Objectives 1 and 2 of the current Project Plan. Building toward the activities outlined in Objectives 2 and 3, the MaizeGDB team is currently in the process of deploying a new website that represents a significant upgrade to state-of-the-art web programming and technologies (see <http://alpha.maizegdb.org>; slated to go live in March of 2013) and has demonstrated extensive experience with implementing both third-party (e.g., GBrowse and



BioCyc/Pathway Tools) and custom (e.g., Locus Lookup and POPcorn) database and Web Services-based query tools. The MaizeGDB team's ability to interact with and support stakeholder needs is well documented by Accomplishment 3, demonstrating that Objective 4 for the current Project Plan is very likely to succeed.

**LITERATURE CITED**

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402
- Andorf CM, Lawrence CJ, Harper LC, Schaeffer ML, Campbell DA, Sen TZ (2010) The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. *Bioinformatics* 26(3):434-436
- Beckmann JS, Estivill X, Antonarakis SE (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet* 8(8):639-646
- Bennetzen J (2001) The Maize Genetics Executive Committee (MGEC). *Maize Genetics Cooperation Newsletter* 75:v-vi
- Bosch M, Mayer CD, Cookson A, Donnison IS (2011) Identification of genes involved in cell wall biogenesis in grasses by differential gene expression profiling of elongating and non-elongating maize internodes. *J Exp Bot* 62(10):3545-3561
- Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23(19):2633-2635
- Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Sanchez Villeda H, da Silva HS, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. *Science* 325(5941):714-718
- Cannon EK, Birkett SM, Braun BL, Kodavali S, Jennewein DM, Yilmaz A, Antonescu V, Antonescu C, Harper LC, Gardiner JM, Schaeffer ML, Campbell DA, Andorf CM, Andorf D, Lisch D, Koch KE, McCarty DR, Quackenbush J, Grotewold E, Lushbough CM, Sen TZ, Lawrence CJ (2011) POPcorn: An Online Resource Providing Access to Distributed and Diverse Maize Project Data. *Int J Plant Genomics* 2011:923035
- Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18(1):188-196
- Carpita NC, McCann MC (2008) Maize and sorghum: genetic resources for bioenergy grasses. *Trends Plant Sci* 13(8):415-420
- Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 40(Database issue):D742-753
- Chae L, Lee I, Shin J, Rhee SY (2012) Towards understanding how molecular networks evolve in plants. *Curr Opin Plant Biol* 15(2):177-184
- Chen YA, Tripathi LP, Mizuguchi K (2011) TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One* 6(3):e17844
- Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, Elshire RJ, Gaut B, Geller L, Glaubitz JC, Gore M, Guill KE, Holland J, Hufford MB, Lai J, Li M, Liu X, Lu Y, McCombie R, Nelson R, Poland J, Prasanna BM, Pyhajarvi T, Rong T, Sekhon RS, Sun

- Q, Tenaillon MI, Tian F, Wang J, Xu X, Zhang Z, Kaeppler SM, Ross-Ibarra J, McMullen MD, Buckler ES, Zhang G, Xu Y, Ware D (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* 44(7):803-807
- Church DM, Hillier LW (2009) Back to Bermuda: how is science best served? *Genome Biol* 10(4):105
- Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR, Albracht D, Kremitzki M, Rock S, Kotkiewicz H, Kremitzki C, Wollam A, Trani L, Fulton L, Fulton R, Matthews L, Whitehead S, Chow W, Torrance J, Dunn M, Harden G, Threadgold G, Wood J, Collins J, Heath P, Griffiths G, Pelan S, Grafham D, Eichler EE, Weinstock G, Mardis ER, Wilson RK, Howe K, Flicek P, Hubbard T (2011) Modernizing reference genome assemblies. *PLoS Biol* 9(7):e1001091
- Claudiel-Renard C, Chevalet C, Faraut T, Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* 31(22):6633-6639
- Cooper GM, Nickerson DA, Eichler EE (2007) Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* 39(7 Suppl):S22-29
- Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L (2001) The distributed annotation system. *BMC Bioinformatics* 2:7
- Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res* 36(Database issue):D959-965
- Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. *Nat Rev Genet* 7(2):85-97
- GO Consortium (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res* 40(Database issue):D559-564
- Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES (2009a) A first-generation haplotype map of maize. *Science* 326(5956):1115-1117
- Gore MA, Wright MH, Ersoz ES, Bouffard P, Szekeres ES, Jarvie TP, Hurwitz BL, Narechania A, Harkins TT, Grills GS, Ware DH, Buckler ES (2009b) Large-Scale Discovery of Gene-Enriched SNPs. *The Plant Genome* 2(2):121-133
- Grassini P, Cassman KG (2012) High-yield maize with large net energy yield and small global warming intensity. *Proc Natl Acad Sci U S A* 109(4):1074-1079
- Green JM, Harnsomburana J, Schaeffer ML, Lawrence CJ, Shyu CR (2011) Multi-source and ontology-based retrieval engine for maize mutant phenotypes. *Database (Oxford)* 2011:bar012
- Harper LC, Schaeffer ML, Thistle J, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Braun BL, Birkett SM, Lawrence CJ, Sen TZ (2011) The MaizeGDB Genome Browser tutorial: one example of database outreach to biologists via video. *Database (Oxford)* 2011:bar016
- Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. *Trends Genet* 24(5):238-245
- Jones CE, Brown AL, Baumann U (2007) Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics* 8:170
- Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, Lee TJ, Kaipa P, Gilham F, Spaulding A, Popescu L, Altman T, Paulsen I, Keseler IM, Caspi R (2010) Pathway

- Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief Bioinform* 11(1):40-79
- Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Zhang J, Li D, Guo X, Ye K, Jian M, Wang B, Zheng H, Liang H, Zhang X, Wang S, Chen S, Li J, Fu Y, Springer NM, Yang H, Wang J, Dai J, Schnable PS (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 42(11):1027-1030
- Lawrence CJ (2011) MaizeGDB – past, present, and future. *Maydica* 56(1-2):3-6
- Lawrence CJ, Walbot V (2007) Translational genomics for bioenergy production from fuelstock grasses: maize as the model species. *Plant Cell* 19(7):2091-2094
- Lee E, Harris N, Gibson M, Chetty R, Lewis S (2009) Apollo: a community resource for genome annotation editing. *Bioinformatics* 25(14):1836-1837
- Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28(18):2397-2399
- Lushbough C, Bergman MK, Lawrence CJ, Jennewein D, Brendel V (2010) BioExtract server--an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data. *IEEE/ACM Trans Comput Biol Bioinform* 7(1):12-24
- Lushbough CM, Bergman MK, Lawrence CJ, Jennewein D, Brendel V (2008) Implementing bioinformatic workflows within the bioextract server. *Int J Comput Biol Drug Des* 1(3):302-312
- McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Oropeza Rosas M, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES (2009) Genetic properties of the maize nested association mapping population. *Science* 325(5941):737-740
- Mikel MA (2006) Availability and Analysis of Proprietary Dent Corn Inbred Lines with Expired U.S. Plant Variety Protection. *Crop Science* 46:2555-2560
- Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WT, Flavell AJ, Marshall D (2010) Flapjack--graphical genotype visualization. *Bioinformatics* 26(24):3133-3134
- Penning BW, Hunter CT, 3rd, Tayengwa R, Eveland AL, Dugard CK, Olek AT, Vermerris W, Koch KE, McCarty DR, Davis MF, Thomas SR, McCann MC, Carpita NC (2009) Genetic resources for maize cell wall biology. *Plant Physiol* 151(4):1703-1728
- Pimentel S, Walbot V, Fernandes J (2011) GRFT - Genetic Records Family Tree Web Applet. *Front Genet* 2:14
- Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Sen TZ, Lawrence CJ (2011) MaizeGDB: curation and outreach go hand-in-hand. *Database (Oxford)* 2011:bar022
- Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, Feuk L (2007) Challenges and standards in integrating surveys of structural variation. *Nat Genet* 39(7 Suppl):S7-15
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L,

- Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddelloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956):1112-1115
- Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 5(12):e1000605
- Sebat J (2007) Major changes in our DNA lead to major changes in our thinking. *Nat Genet* 39(7 Suppl):S3-5
- Sen T, Andorf C, Schaeffer M, Harper L, Sparks M, Duvick J, Brendel V, Cannon E, Campbell D, Lawrence C (2009) MaizeGDB becomes 'sequence-centric'. *Database* 2009:bap020
- Sen T, Harper L, Schaeffer M, Andorf C, Seigfried T, Campbell D, Lawrence C (2010) Choosing a genome browser for a Model Organism Database: surveying the maize community. *Database* 2010:baq007
- Sen TZ, Harper LC, Schaeffer ML, Andorf CM, Seigfried TE, Campbell DA, Lawrence CJ (2010) Choosing a genome browser for a Model Organism Database: surveying the maize community. *Database (Oxford)* 2010:baq007
- Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, Eichler EE (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet* 38(9):1038-1042
- Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH (2009) JBrowse: a next-generation genome browser. *Genome Res* 19(9):1630-1638
- Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31
- Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3):431-432
- Sneddon TP, Church DM (2012) Online resources for genomic structural variation. *Methods Mol Biol* 838:273-289

- Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res* 20(12):1689-1699
- Tanabe M, Kanehisa M (2012) Using the KEGG database resource. *Curr Protoc Bioinformatics* Chapter 1:Unit1 12
- van Berloo R (2007) GGT 2.0: Versatile Software for Visualization and Analysis of Genetic Data. *Journal of Heredity* 99(2):232-236
- Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K, Kudrna D, Faga BP, Wissotski M, Golser W, Rock SM, Graves TA, Fulton RS, Coe E, Schnable PS, Schwartz DC, Ware D, Clifton SW, Wilson RK, Wing RA (2009) The physical and genetic framework of the maize B73 genome. *PLoS Genet* 5(11):e1000715
- Wilkerson MD, Schlueter SD, Brendel V (2006) yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes. *Genome Biol* 7(7):R58
- Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The effects of artificial selection on the maize genome. *Science* 308(5726):1310-1314
- Youens-Clark K, Buckler E, Casstevens T, Chen C, Declerck G, Derwent P, Dharmawardhana P, Jaiswal P, Kersey P, Karthikeyan AS, Lu J, McCouch SR, Ren L, Spooner W, Stein JC, Thomason J, Wei S, Ware D (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res* 39(Database issue):D1085-1094
- Yu C, Zavaljevski N, Desai V, Reifman J (2009) Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases. *Proteins* 74(2):449-460
- Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178(1):539-551
- Zhang J, Haider S, Baran J, Cros A, Guberman JM, Hsu J, Liang Y, Yao L, Kasprzyk A (2011) BioMart: a data federation framework for large collaborative projects. *Database (Oxford)* 2011:bar038

## PAST ACCOMPLISHMENTS OF INVESTIGATORS – Carolyn J. Lawrence

### *EDUCATION*

<b>B.A.</b>	Biology	1992-1996	Hendrix College, Conway, AR
<b>M.S.</b>	Biology	1996-1997	Texas Tech University, Lubbock, TX
<b>Ph.D.</b>	Botany	1997-2003	University of Georgia, Athens, GA

### *WORK EXPERIENCE*

2003-2005	Postdoctoral Research (Bioinformatics)	Iowa State University, Ames, IA
2005-2008	GS-12 Cat 1 Research Geneticist	USDA-ARS CICGR, Ames, IA
2008-2012	GS-13 Cat 1 Research Geneticist	USDA-ARS CICGR, Ames, IA
2012-present	GS-14 Cat 1 Research Geneticist	USDA-ARS CICGR, Ames, IA
2005-present	Collaborator/Assistant Professor	Genetics, Development and Cell Biol., Iowa State University, Ames, IA

### *ACCOMPLISHMENTS*

Early work focused on improving data access and analysis tools for flowering plants. This involved: (1) developing comparative genomics resources that enable plant biologists to make use of large-scale genome and transcriptome sequencing datasets (Dong et al. 2005; Duvick et al. 2008), (2) developing stand-alone bioinformatics tools that solve common, well-defined data analysis problems (Lawrence et al. 2006; Lawrence et al. 2004; Yi et al. 2009), and (3) developing and deploying MaizeGDB in 2003 (Lawrence et al. 2004; Lawrence et al. 2007; Lawrence et al. 2005). Recent work has involved (4) demonstrating the use of maize and, more specifically, MaizeGDB to solve real-world problems (Lawrence et al. 2008; Lawrence and Walbot 2007), (5) making the genome sequences of B73 and Palomero toluqueño accessible to scientists worldwide (Andorf et al. 2010; Sen et al. 2009; Sen et al. 2010), (6) creating novel venues for technology transfer and outreach that enable diverse groups to make use of biological findings and opportunities (Gray et al. 2009; Lawrence 2011; Schaeffer et al. 2011), and (7) creating and deploying novel, state of the art methods to access distributed data and heterogeneous datasets (Cannon et al. 2011; Green et al. 2011; Lushbough et al. 2010; Lushbough et al. 2008).

### *PEER REVIEWED PUBLICATIONS (underline indicates corresponding author)*

- Andorf CM, **Lawrence CJ**, Harper LC, Schaeffer ML, Campbell DA, Sen TZ (2010) The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. *Bioinformatics* 26(3):434-436
- Cannon EK, Birkett SM, Braun BL, Kodavali S, Jennewein DM, Yilmaz A, Antonescu V, Antonescu C, Harper LC, Gardiner JM, Schaeffer ML, Campbell DA, Andorf CM, Andorf D, Lisch D, Koch KE, McCarty DR, Quackenbush J, Grotewold E, Lushbough CM, Sen TZ, **Lawrence CJ** (2011) POPcorn: An Online Resource Providing Access to Distributed and Diverse Maize Project Data. *Int J Plant Genomics* 2011:923035
- Dong Q, **Lawrence CJ**, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C, Brendel V (2005) Comparative plant genomics resources at PlantGDB. *Plant Physiol* 139(2):610-618

- Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, **Lawrence CJ**, Lushbough C, Brendel V (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res* 36(Database issue):D959-965
- Gray J, Bevan M, Brutnell T, Buell CR, Cone K, Hake S, Jackson D, Kellogg E, **Lawrence C**, McCouch S, Mockler T, Moose S, Paterson A, Peterson T, Rokshar D, Souza GM, Springer N, Stein N, Timmermans M, Wang GL, Grotewold E (2009) A recommendation for naming transcription factor proteins in the grasses. *Plant Physiol* 149(1):4-6
- Green JM, Harnsomburana J, Schaeffer ML, **Lawrence CJ**, Shyu CR (2011) Multi-source and ontology-based retrieval engine for maize mutant phenotypes. *Database (Oxford)* 2011:bar012
- Lawrence CJ** (2011) MaizeGDB - Past, Present, and Future. *Maydica* 56(1):3-5
- Lawrence CJ**, Dong Q, Polacco ML, Seigfried TE, Brendel V (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res* 32(Database issue):D393-397
- 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* 2008:496957
- Lawrence CJ**, Schaeffer ML, Seigfried TE, Campbell DA, Harper LC (2007) MaizeGDB's new data types, resources and activities. *Nucleic Acids Res* 35(Database issue):D895-900
- Lawrence CJ**, Seigfried TE, Bass HW, Anderson LK (2006) Predicting chromosomal locations of genetically mapped loci in maize using the Morgan2McClintock Translator. *Genetics* 172(3):2007-2009
- Lawrence CJ**, Seigfried TE, Brendel V (2005) The maize genetics and genomics database. The community resource for access to diverse maize data. *Plant Physiol* 138(1):55-58
- Lawrence CJ**, Walbot V (2007) Translational genomics for bioenergy production from fuelstock grasses: maize as the model species. *Plant Cell* 19(7):2091-2094
- Lawrence CJ**, Zmasek CM, Dawe RK, Malmberg RL (2004) LumberJack: a heuristic tool for sequence alignment exploration and phylogenetic inference. *Bioinformatics* 20(12):1977-1979
- Lushbough C, Bergman MK, **Lawrence CJ**, Jennewein D, Brendel V (2010) BioExtract server--an integrated workflow-enabling system to access and analyze heterogeneous, distributed biomolecular data. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* 7(1):12-24
- Lushbough CM, Bergman MK, **Lawrence CJ**, Jennewein D, Brendel V (2008) Implementing bioinformatic workflows within the bioextract server. *International journal of computational biology and drug design* 1(3):302-312
- Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Sen TZ, **Lawrence CJ** (2011) MaizeGDB: curation and outreach go hand-in-hand. *Database (Oxford)* 2011:bar022
- Sen T, Andorf C, Schaeffer M, Harper L, Sparks M, Duvick J, Brendel V, Cannon E, Campbell D, **Lawrence CJ** (2009) MaizeGDB becomes 'sequence-centric'. *Database* 2009:bap020
- Sen TZ, Harper LC, Schaeffer ML, Andorf CM, Seigfried TE, Campbell DA, Lawrence CJ (2010) Choosing a genome browser for a Model Organism Database: surveying the maize community. *Database (Oxford)* 2010:baq007
- Yi G, Luth D, Goodman TD, **Lawrence CJ**, Becraft PW (2009) High-throughput linkage analysis of Mutator insertion sites in maize. *Plant J* 58(5):883-892



## PAST ACCOMPLISHMENTS OF INVESTIGATORS – TANER Z. SEN

### *EDUCATION*

<b>B.S.</b>	Chemical Engineering	1992-1996	Bogazici University, Istanbul, Turkey
<b>M.S.</b>	Chemical Engineering	1996-1998	Bogazici University, Istanbul, Turkey
<b>Ph.D.</b>	Polymer Engineering	1998-2003	University of Akron, Akron, OH

### *WORK EXPERIENCE*

2003-2007	Postdoctoral Research	Iowa State University, Ames, IA
2007-present	GS-12 Cat 4 Computational Biologist Collaborator/Assistant Professor	USDA-ARS CICGR, Ames, IA Genetics, Development and Cell Biol., Iowa State University, Ames, IA

### *ACCOMPLISHMENTS*

Surveyed stakeholders (i.e., the maize research community) to determine genome browser needs, and implemented the MaizeGDB Genome Browser. Applied eigenvector analysis to the yeast protein interaction network to reveal significant sub-networks. Showed that the protein interaction network is wired in such a way that proteins with similar degrees (i.e., number of interaction partners) tend to interact with each other to infer other interactions not previously reported. Developed a rule-based consensus approach for protein-protein interaction site predictions based on four methods: Conservatism-of-Conservatism, threading, support vector machines, and phylogenetic trees. Improved binding site accuracy in hydrolase-inhibitor complexes. Improved secondary structure prediction accuracy by data mining protein structure fragments. Developed consensus secondary structure prediction. Utilized elastic network models in the tertiary structure prediction to improve model resolution. Created Web servers for GOR V and CDM secondary structure prediction algorithms. Applied elastic network models to reveal the dynamics of protein structural mechanisms including ATP-binding proteins and compared the applicability of these models at multiple scales. Used elastic network models to predict bond breaking during titin stretching. Modeled bacterial ribosome to analyze the mRNA and nascent protein dynamics.

### *PEER REVIEWED PUBLICATIONS (underline indicates corresponding author)*

- Andorf CM, Lawrence CJ, Harper LC, Schaeffer ML, Campbell DA, Sen TZ (2010) The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. *Bioinformatics* 26(3):434-436
- Cannon EK, Birkett SM, Braun BL, Kodavali S, Jennewein DM, Yilmaz A, Antonescu V, Antonescu C, Harper LC, Gardiner JM, Schaeffer ML, Campbell DA, Andorf CM, Andorf D, Lisch D, Koch KE, McCarty DR, Quackenbush J, Grotewold E, Lushbough CM, Sen TZ, Lawrence CJ (2011) POPcorn: An Online Resource Providing Access to Distributed and Diverse Maize Project Data. *Int J Plant Genomics* 2011:923035
- Cheng H, Sen TZ, Jernigan RL, Kloczkowski A (2007) Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: combining GOR V and Fragment Database Mining (FDM). *Bioinformatics* 23(19):2628-2630

- Cheng H, **Sen TZ**, Kloczkowski A, Margaritis D, Jernigan RL (2005) Prediction of protein secondary structure by mining structural fragment database. *Polymer (Guildf)* 46(12):4314-4321
- Harper LC, Schaeffer ML, Thistle J, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Braun BL, Birkett SM, Lawrence CJ, **Sen TZ** (2011) The MaizeGDB Genome Browser tutorial: one example of database outreach to biologists via video. *Database (Oxford)* 2011:bar016
- Kurkuoglu O, Doruker P, **Sen TZ**, Kloczkowski A, Jernigan RL (2008) The ribosome structure controls and directs mRNA entry, translocation and exit dynamics. *Phys Biol* 5(4):046005
- 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* 2008:496957
- Rader AJ, Yennamalli RM, Harter AK, **Sen TZ** (2012) A rigid network of long-range contacts increases thermostability in a mutant endoglucanase. *J Biomol Struct Dyn*
- Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, **Sen TZ**, Lawrence CJ (2011) MaizeGDB: curation and outreach go hand-in-hand. *Database (Oxford)* 2011:bar022
- Sen TZ**, Andorf CM, Schaeffer ML, Harper LC, Sparks ME, Duvick J, Brendel VP, Cannon E, Campbell DA, Lawrence CJ (2009) MaizeGDB becomes 'sequence-centric'. *Database (Oxford)* 2009:bap020
- Sen TZ**, Cheng H, Kloczkowski A, Jernigan RL (2006) A Consensus Data Mining secondary structure prediction by combining GOR V and Fragment Database Mining. *Protein Sci* 15(11):2499-2506
- Sen TZ**, Feng Y, Garcia JV, Kloczkowski A, Jernigan RL (2006) The Extent of Cooperativity of Protein Motions Observed with Elastic Network Models Is Similar for Atomic and Coarser-Grained Models. *J Chem Theory Comput* 2(3):696-704
- Sen TZ**, Harper LC, Schaeffer ML, Andorf CM, Seigfried TE, Campbell DA, Lawrence CJ (2010) Choosing a genome browser for a Model Organism Database: surveying the maize community. *Database (Oxford)* 2010:baq007
- Sen TZ**, Jernigan RL, Garnier J, Kloczkowski A (2005) GOR V server for protein secondary structure prediction. *Bioinformatics* 21(11):2787-2788
- Sen TZ**, Kloczkowski A, Jernigan RL (2006) A DNA-centric look at protein-DNA complexes. *Structure* 14(9):1341-1342
- Sen TZ**, Kloczkowski A, Jernigan RL (2006) Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics* 7:355
- Sen TZ**, Kloczkowski A, Jernigan RL, Yan C, Honavar V, Ho KM, Wang CZ, Ihm Y, Cao H, Gu X, Dobbs D (2004) Predicting binding sites of hydrolase-inhibitor complexes by combining several methods. *BMC Bioinformatics* 5:205
- Sen TZ**, Kloster M, Jernigan RL, Kolinski A, Bujnicki JM, Kloczkowski A (2008) Predicting the complex structure and functional motions of the outer membrane transporter and signal transducer FecA. *Biophys J* 94(7):2482-2491
- Yennamalli RM, Rader AJ, Wolt JD, **Sen TZ** (2011) Thermostability in endoglucanases is fold-specific. *BMC Struct Biol* 11:10
- Yennamalli RM, Wolt JD, **Sen TZ** (2011) Dynamics of endoglucanase catalytic domains: implications towards thermostability. *J Biomol Struct Dyn* 29(3):509-526

**ISSUES OF CONCERN**

**Animal Care:** Not relevant

**Endangered Species:** Not relevant

**National Environmental Policy Act:** Not relevant

**Human Study Procedure:** Not relevant

**Laboratory Hazards:** Not relevant

**Occupational Safety and Health:** The work will be conducted in the office environment.

**Recombinant DNA Procedures:** Not relevant

**Homeland Security:** The data security against cyber attacks, accidents, and disasters is ensured through the implementation of a secure disaster recovery system. The disaster recovery system duplicates all the data in the working copy of the MaizeGDB database and the entire programming interface and transfers it in two different physical locations in Ames, IA (on separate servers in physically isolated buildings) and Columbia, MO.

**Intellectual Property Issues:** This research will be conducted to create resources maintained in the public domain. All informatics data will be made available to the scientific community via MaizeGDB. The MaizeGDB database is interoperable with Gramene, GrainGenes, MaizeSequence.org, PlantGDB, and various other resources, and thus will ensure that data is rapidly disseminated to the scientific community.

**EXISTING SPECIFIC COOPERATIVE AGREEMENTS:** None.

**APPENDICES**

**A. PAST ACCOMPLISHMENTS – CARSON M. ANDORF .....51**  
**B. LETTERS OF COLLABORATION.....53**  
**C. WORKING GROUP REPORT .....78**

## A. PAST ACCOMPLISHMENTS – CARSON M. ANDORF

### *EDUCATION*

**B.A.** Computer Science/Mathematics 1996-2000 Wartburg College, Waverly, IA  
**Ph.D.** Computer Science 2013 (anticipated) Iowa State University, Ames, IA

### *WORK EXPERIENCE*

2004-2008	Lead Bioinformatics Developer	NewLink Genetics, Ames, IA
2008- 2011	GS-9 IT Specialist	USDA-ARS CICGR, Ames, IA
2011-present	GS-11 IT Specialist	USDA-ARS CICGR, Ames, IA

### *ACCOMPLISHMENTS*

Developed a new machine learning method that uses generalized versions of Naive Bayes algorithm, k-order Markov Model, and a Support Vector Machine to use alternative representations of proteins for prediction of novel proteins. These prediction problems include function, structure, subcellular localization, kinetics, and protein-protein interactions. This method performs at high levels on a wide variety of different datasets (Andorf et al. 2007). This application was developed as both a standalone tool and as a web server. Head software developer on two commercial software projects VmatchNL (large scale genome-wide sequence matching) and GeneSeqerNL (spliced alignment gene structure prediction). Work at MaizeGDB through ARS funding has included (1) serving as the lead interface developer on the MaizeGDB website (Sen et al. 2009), (2) integrating the MaizeGDB database and interface with visualization tools (Cannon et al. 2011; Sen et al. 2010), (3) building tools and resources to integrate genomic and genetic data (Andorf et al. 2010), and (4) providing outreach to the maize community (Harper et al. 2011; Schaeffer et al. 2011) through technical support and as abstract coordinator for the Annual Maize Genetics Conference (2009 - 2013).

### *PEER REVIEWED PUBLICATIONS*

**Andorf C**, Dobbs D, Honavar V (2007) Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach. *BMC Bioinformatics* 8:284

**Andorf CM**, Lawrence CJ, Harper LC, Schaeffer ML, Campbell DA, Sen TZ (2010) The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. *Bioinformatics* 26(3):434-436

Cannon EK, Birkett SM, Braun BL, Kodavali S, Jennewein DM, Yilmaz A, Antonescu V, Antonescu C, Harper LC, Gardiner JM, Schaeffer ML, Campbell DA, **Andorf CM**, Andorf D, Lisch D, Koch KE, McCarty DR, Quackenbush J, Grotewold E, Lushbough CM, Sen TZ, Lawrence CJ (2011) POPcorn: An Online Resource Providing Access to Distributed and Diverse Maize Project Data. *Int J Plant Genomics* 2011:923035

Harper LC, Schaeffer ML, Thistle J, Gardiner JM, **Andorf CM**, Campbell DA, Cannon EK, Braun BL, Birkett SM, Lawrence CJ, Sen TZ (2011) The MaizeGDB Genome Browser tutorial: one example of database outreach to biologists via video. *Database (Oxford)* 2011:bar016

- Schaeffer ML, Harper LC, Gardiner JM, **Andorf CM**, Campbell DA, Cannon EK, Sen TZ, Lawrence CJ (2011) MaizeGDB: curation and outreach go hand-in-hand. Database (Oxford) 2011:bar022
- Sen T, **Andorf C**, Schaeffer M, Harper L, Sparks M, Duvick J, Brendel V, Cannon E, Campbell D, Lawrence C (2009) MaizeGDB becomes 'sequence-centric'. Database 2009:bap020
- Sen T, Harper L, Schaeffer M, **Andorf C**, Seigfried T, Campbell D, Lawrence C (2010) Choosing a genome browser for a Model Organism Database: surveying the maize community. Database 2010:baq007

**B. LETTERS OF COLLABORATION****Cold Spring Harbor Laboratory**

Doreen Ware, Ph.D.  
USDA ARS  
Adjunct Associate Professor

September 21<sup>st</sup> 2012

Dear Carolyn and Taner,

Thanks for sharing your Project Plan for MaizeGDB with me. Based on your draft and our existing collaborations, I see many areas where our work is complementary and look forward to continuing to work with your team. My specific areas of collaboration with your group are outlined below by Objective from your project plan. As you are aware from our existing collaboration, I have been a coPI on both the Maize Genome Sequencing Project and Diversity Project. My group has contributed to the development of the current maize B73 draft assembly and annotations, and as well as the recently released Maize HapMapII diversity panel. I am delighted to see that for Objective 1 you are looking to use community standards (from NCBI, the Genomic Standards Consortium, and Genome Reference Consortium) in forthcoming releases of the genome. In support of this objective, my group will continue to work with your group, Washington University, and NCBI to deliver the B73 RefGen\_v4 assembly and annotations to support adoption of maize by as part of the reference consortium. As you are aware updates to the assembly and annotations are a semi-automated process support by integration of data and we look forward to receiving the community data sets you are collecting. As in the past, my group will keep you informed on our timelines and anticipated deliverables. We appreciate MaizeGDB's help in keeping the community informed.

In Objective 2 you outline plans to curate biochemical pathway data and to expand the datasets that can be represented by implementing Cytoscape Web, both for genetic and interaction network analyses and to visualize relatedness among other entities including diverse inbred lines. My group looks forward to receiving the data that you and Mary Schaeffer will curate into MaizeGDB as part of the ongoing Gramene objectives to support the development of Plant Reactome in collaboration with Drs. Pankaj Jaiswal and Lincoln Stein.

Your group's Objective 3 is to implement a mechanism to query the MaizeGDB database directly. As you know, my group has deployed BioMart for Gramene and we would be happy to share with you technical expertise to create custom datasets for their genomic analyses. Please keep us posted on your software selection process.

I look forward to continuing to collaborate with you and the rest of the MaizeGDB Team.

Sincerely,

Doreen Ware

Cold Spring Harbor Laboratory, One Bungtown Road,  
Cold Spring Harbor, New York 11724  
Tel: 516/367-6979 Fax: 516/367-6851 Email:  
ware@cshl.edu



Department of Human Genetics

**Carolyn J. Lawrence**, PhD.,  
United States Department of Agriculture Research,  
Education and Economics Agricultural Research Service

Dear Carolyn,

Thanks for sharing the MaizeGDB Project Plan with me. I am pleased that your goals in Objective 1 involve our MAKER software to systemically improve maize structural annotations. MAKER is a portable, easy-to-use annotation engine designed for users with limited bioinformatics expertise. Its development has been supported by an NHGRI R01 entitled *Software for the creation and quality control of genome annotations*, (R01HG004694) and an NSF program grant, *Developing an Effective, Portable Annotation Engine for Plant Genomes* (IOS-1126998). MAKER identifies DNA repeats, performs EST, RNA-seq and protein alignments, predicts genes with both evidence-informed and *ab initio* methods, and automatically consolidates this information into gene annotations having evidence-based quality scores. Importantly, MAKER's annotation management and quality control functionalities allow us to gain an overview of the relationship of structural annotations to their evidence, and to begin to take steps to systematically improve them.

As you know, MAKER is currently being optimized for plant genomes. By collaborating with your group as well as Doreen Ware's, we have been able to coordinate well with ongoing endeavors as well as to make use of your group's knowledge of the unique biology of maize relative to other species. I am particularly interested to learn that your Project Plan involves working with the Genome Reference Consortium to document researchers' input to improve gene structures because these high-quality, hand-curated structures can be weighted heavily in determining quality scores to insure that well-characterized genes are well-annotated by MAKER.

To be specific, for maize we plan not only to improve the RefGen\_v3 annotations of the maize genome directly (currently underway in collaboration with your group as well as the Ware group), but to make available the pipelines we use so that they could be implemented directly for incremental updates to the maize reference genome in the future. This will definitely help to enable your group to support genome stewardship by creating well-documented methods and pipelines that would automate updates to genome annotations. The fact that MAKER outputs can be represented by Apollo visualization and curation software is also useful for your purposes given that WebApollo and its

**Eccles Institute of Human Genetics**  
15 North 2030 East, Room 2100  
Salt Lake City, Utah 84112-5330  
801-587-7707  
FAX: 801-585-3214  
myandell@genetics.utah.edu



future deployment via DNASubway looks like a good option for enabling community input and student participation in genome annotation improvement.

We look forward to continuing our collaboration.

Sincerely,



Mark Yandell  
Professor of Human Genetics  
H.A. & Edna Benning Presidential Endowed Chair  
Eccles Institute of Human Genetics  
University of Utah & School of Medicine

**Eccles Institute of Human Genetics**  
15 North 2030 East, Room 2100  
Salt Lake City, Utah 84112-5330  
801-587-7707  
FAX: 801-585-3214  
myandell@genetics.utah.edu

**DEPARTMENT OF BIOLOGY**

INDIANA UNIVERSITY  
College of Arts and Sciences  
Bloomington

To:

Dr. Carolyn J. Lawrence  
USDA-ARS  
1034 Crop Genome Informatics Laboratory  
Iowa State University  
Ames, IA 50011

October 9, 2012

RE: collaboration

Dear Carolyn,

I am glad to see that genome stewardship for B73 is listed as Objective 1 for your USDA-ARS 5-year project plan. Given that genome annotation has been an area of collaboration between us over the last 10 years, I look forward to continuing to work together on this aspect of your work.

As a part of the PlantGDB project, my group developed a tool called yrGATE (your Gene structure Annotation Tool for Ekaryotes) that enables researchers to contribute and visualize their own structural annotations. The yrGATE tool works as follows: if elements of a gene model are poorly supported or the gene is known to produce splice variants, researchers can use yrGATE to contribute alternate, user-contributed gene structure annotations that are made available to the public online. In collaboration with you at MaizeGDB, we released yrGATE for maize structural annotations nearly four years ago and have accepted more than 250 annotations to existing structures. These improvements are available via both PlantGDB's ZmGDB (the maize instance of xGDB, the eXtensible Genome Data Broker database that enables genome feature storage, display, and analysis from within their genomic context) and the MaizeGDB Genome Browser.

As described in your Project Plan, my group currently is funded by the NSF's Plant Genome Research Program "to implement and deploy pipelines that automate the xGDB pipeline for plant genome annotation and visualization such that any species could have an xGDB instance created fairly easily, (2) to enable gene structure

prediction using machine learning approaches, and (3) to enable researchers to structurally annotate an entire gene family across many plant species (termed "vertical" annotation)." We look forward to working with you at MaizeGDB to create a largely automated pipeline for genome annotation improvement and to potentially expand maize annotation through "vertical" annotation within the species as diverse inbred lines become sequenced.

Sincerely,

A handwritten signature in black ink, appearing to read "Volker Brendel", with a long horizontal flourish extending to the right.

Volker Brendel  
Professor of Biology and Computer Science  
Indiana University  
Department of Biology & School of Informatics and Computing  
Simon Hall 205C  
212 South Hawthorne Drive  
Bloomington, IN 47405-7003

Tel.: (812) 855-7074  
Email: [vbrendel@indiana.edu](mailto:vbrendel@indiana.edu)



Lawrence Berkeley National Laboratory  
Berkeley Bioinformatics Open Source Projects  
1 Cyclotron Road Mail Stop— 84R0171  
Berkeley, CA 94720

Carolyn J. Lawrence, Ph.D.  
*Research Scientist,  
USDA-ARS,  
Iowa State University,  
1034 Crop Genome Informatics Laboratory,  
Ames, Iowa 50011*

October 19, 2012

Dear Carolyn,

I thoroughly enjoyed talking with you at the HHMI meeting last June regarding the potential use of WebApollo as a tool to allow improvements to the maize genome assembly and annotation.

The desktop generation of Apollo was developed to enable biologists to inspect genome annotations closely and edit them. The inaccuracy of purely computational genome annotations necessitates an interactive tool to allow biological experts to evaluate and refine these approximations. Desktop Apollo was successfully adopted for use by a broad spectrum of biologists, but technology moves swiftly and applications must keep up with changing environments. Web-Apollo has all of the editing capabilities of the original, but now allows users direct access through any Web browser, and supports simultaneous edits from the entire community. Because your group plans to migrate from GBrowse to JBrowse (which Apollo shares a common codebase with), it seems that engaging with us on WebApollo would make it an easy transition, and will allow the maize community to contribute their improvements to the maize genome annotations.

Already Lisa Harper in your group has visited us to learn more from the perspective of a curator. Shortly thereafter, Gregg Helt in my group described WebApollo to the larger MaizeGDB Team and I understand that both curators and programmers in the group are very excited to work with WebApollo for maize genome assembly and annotation improvement. I look forward to continuing these exchanges of information and ideas and am excited that Dave Micklos at iPlant will work with us to support the project from the iPlant DNA Subway umbrella.

I would all like to reiterate my enthusiasm for a collaboration to ensure that Web-Apollo meets the needs of the maize community. I wish you the best of luck.

Sincerely,

Suzanna Lewis  
*Staff Scientist  
Lawrence Berkeley National Laboratory*



1 Bungtown Road  
Cold Spring Harbor, NY 11724  
Phone: (516) 367-5170  
Fax: (516) 367-5182  
Internet: www.dnalc.org  
Email: dnalc@csHL.edu

October 18, 2012

Carolyn J. Lawrence  
1034 Crop Genome Informatics Laboratory  
Iowa State University  
Ames, Iowa 50011

Dear Carolyn,

I enjoyed discussing the potential to adapt *DNA Subway* to make it more useful for maize genome annotation at the HHMI workshop in June. I am glad that you are including this collaboration as part of the MaizeGDB 5-year Project Plan for USDA-ARS. We are keen to help you deploy *DNA Subway* for improving the maize reference genome sequence assembly and annotations, as part of Objective 1 in your Project Plan.

*DNA Subway* bundles research-grade bioinformatics tools and databases into intuitive workflows, with the Red Line equipped for predicting and annotating genes in up to 150 kb of DNA sequence. *DNA Subway* currently has 3,537 registered and receives an average of 2,800 visits per month from a mixed group of researchers, faculty members, and students. The Red Line has been tested as a vehicle for community genome annotation with undergraduate students. Brent Buckner, at Truman State University, has used *DNA Subway* with nearly 100 undergraduate genetics students to annotate over 11 million bp of the maize genome. Advanced projects involved comparing syntenic regions in sorghum and examining presence-absence variation (PAV) in 30 inbred lines. Under the guidance of Jessica Garb, students at the University of Massachusetts, Lowell are using *DNA Subway* to annotate contigs from the black widow genome. This is part of the i5k project, which aims to sequence 5,000 insect and other arthropod genomes over five years.

*DNA Subway* uses the Apollo annotation editor, and we are anxious to continue our collaboration with Suzi Lewis at LBNL to deploy WebApollo as soon as practical. Your maize annotation project could provide the impetus to make this happen. DNALC computational biologist Sheldon McKay is a member of the GMOD Consortium, and is ready to work closely with Suzi's group on integration of WebApollo.

*DNA Subway* is part of the iPlant Collaborative, a major NSF program to develop cyberinfrastructure for biology research. As leads for Cold Spring Harbor Laboratory's involvement, Doreen Ware and I can provide your project easy access to broad expertise within iPlant that may be helpful to other aspects of your 5-year plan.

We look forward to working with you on the community annotation of the maize genome.

Best regards,

A handwritten signature in black ink that reads 'David Micklos'.

David Micklos  
Executive Director

iPlant Lead: Education, Outreach & Training

**Dolan DNA Learning Center**  
334 Main Street  
Cold Spring Harbor, NY 11724

**DNA Learning Center West**  
5 Delaware Drive - Suite 5  
Lake Success, NY 11042

**Harlem DNA Lab**  
2351 First Avenue at 120<sup>th</sup> Street  
East Harlem, NY 10035



DEPARTMENT OF HEALTH &amp; HUMAN SERVICES

Public Health Service

National Institutes of Health  
National Library of Medicine  
Bethesda, Maryland 20894

October 22<sup>nd</sup>, 2012

Carolyn J. Lawrence, Ph.D.  
USDA-ARS Research Geneticist  
1034 CGIL  
Iowa State University  
Ames, IA 50011

Dear Carolyn,

Thank you for contacting me to discuss the interest of your group to try out and evaluate the Genome Reference Consortium's data model and tools for maize genome assembly improvement.

The Genome Reference Consortium (GRC) was organized at the completion of the human genome sequencing project to oversee future updates to this reference assembly. The GRC has since become the group responsible for the maintenance and modernization of the human, mouse and zebrafish reference genome assemblies. We are comprised of The Genome Institute at Washington University, St. Louis (TGI), The Wellcome Trust Sanger Institute (WTSI), the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI). The addition of new organisms and member institutions to the GRC will require their funded commitment for reference genome improvement, community support for GRC membership and the review and approval of all current GRC members.

Within the consortium, NCBI develops and maintains the GRC database and software used for assembly improvement. Based on our discussions last week via conference calls involving Doreen Ware, Jack Okamura, and yourself, along with personnel in each group, we have decided to load our database with the necessary files for B73 RefGen\_v2 assembly once it is approved by GenBank and to allow the groups to work with the tools and determine whether and how our system might be applied to maize genome assembly stewardship. Our software was purposefully designed to permit groups such as yours to evaluate the extensibility of the GRC's tools and assembly model with respect to your own plans for assembly updates. Hopefully we can begin looking at this within the next month or two.

The assembly data model developed by the GRC (PLoS Biol. 2011 Jul;9(7):e1001091) has a specific mechanism for representing complex genomic variation and seems reasonable for storing the diversity represented by the maize genome, and we look forward to working with you.

Sincerely,

Valerie Schneider



United States Department of Agriculture

Research, Education, and Economics  
Agricultural Research Service

Carolyn J. Lawrence  
USDA-ARS  
1034 Crop Genome Informatics Laboratory  
Iowa State University  
Ames, IA 50011

Dear Carolyn and Taner,

A component of our Maize Diversity project is to find a long-term repository for our phenotypic, genotypic, and germplasm data. To this aim, we have formed and will continue a collaboration with MaizeGDB as the long-term repository for data coming out of the Maize Diversity project. This includes transitioning existing data from the Panzea database and directly depositing new data into MaizeGDB. The traditional relational databases utilized at MaizeGDB are not designed to handle the large-scale genotypic data being generated from the Maize Diversity project. Based on this need, we are already working together (In August 2012, Carson Andorf, a bioinformatics engineer at MaizeGDB, visited our lab at Cornell University to jump-start our collaboration efforts) to represent the data in a binary format and the Maize Diversity project's bioinformatics group, led by Dr. Qi Sun, will work with MaizeGDB to co-develop bioinformatics tools to store, retrieve, and visualize large-scale genotypic data. I would be happy to continue to provide exchanges in personnel between the groups both through in-person and teleconference meetings and am happy that you recognize the need, especially for maize, to consider how to represent diversity as a component of genome stewardship.

I also want to mention my group's interest in and expertise with breeding data and relatedness among lines across maize diversity. I see that for Objective 2, Goal 2b you will be working to deploy pedigree visualization tools as well as tools that would enable breeders to evaluate and visualize breeding networks based on sequence similarity. As the need arises, please feel free to contact me and individuals in my group to get our perspective on what tools are available that might be adopted (or adapted) to meet this need.

Sincerely,

Edward S. Buckler, Ph.D.  
USDA-ARS Research Geneticist  
Adjunct Professor of Plant Breeding and Genetics  
Institute for Genomic Diversity, Cornell University

ESB:sjm



North Atlantic Area § Robert W. Holley Center for Agriculture and Health  
538 Tower Road § Ithaca, NY 14853-2901  
Voice: 607 255 4520 § FAX: 607 254 6379 § E-mail: Edward.Buckler@ars.usda.gov  
An Equal Opportunity Employer

IOWA STATE UNIVERSITY  
Plant Sciences Institute

Center for Plant Genomics  
2035C Roy J. Carver Co-Laboratory  
Iowa State University  
Ames, Iowa 50011-3650  
515 294-7209  
FAX 515 294-5256

October 17, 2012

Carolyn J. Lawrence  
USDA-ARS  
1034 Crop Genome Informatics Laboratory  
Iowa State University  
Ames, IA 50011

Dear Carolyn,

I am glad to see that genome stewardship for B73 is listed as Objective 1 for your USDA-ARS 5-year project plan, and I am especially happy that you recognize the need for genome stewardship of diverse lines within the species *Zea mays*, not just the B73 reference genome.

As you know, maize exhibits levels of structural variation (SV) of non-repeat sequences that are unprecedented among higher eukaryotes. This SV includes hundreds of copy number variants (CNVs) and thousands of presence/absence variants (PAVs). Many of the PAVs contain intact, expressed, single-copy genes that are present in one haplotype but absent from another. The goal of our funded NSF Plant Genome Research Program project (P. Schnable, PI; B. Buckner, C. Lawrence and D. Nettleton, coPIs) is to test the hypothesis that differences in gene copy number (both gains and losses) contribute to the extraordinary phenotypic diversity and plasticity of maize. Already my group has worked with members of the MaizeGDB Team to transfer data generated to the community via MaizeGDB and to consider your group's plans on how to store and make available the large datasets of maize diversity currently being generated.

I endorse your perspective that any sequence-based information that can be represented at GenBank as the primary repository should be deposited there and that those data can be served alongside relevant maize data via MaizeGDB. This is both practical and a good use of resources. In addition, your plans to use the Genomic Standards Consortium model for genome representation (derived from their history working with human diversity) is important. Our research has shown that many chromosomal segments present in lines other than B73 genuinely do not exist in B73, meaning that mechanisms to represent the genome sequence of the species is not straight-forward and will be a complex undertaking, both for data storage and for deploying visualization tools that will enable researchers to make use of the diversity available for their work. Planning to adopt the same models developed for storing and visualizing human diversity makes sense and is a mechanism by which plants can springboard off lessons learned by the broad research community.

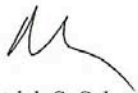


IOWA STATE UNIVERSITY  
Plant Sciences Institute

Center for Plant Genomics  
2035C Roy J. Carver Co-Laboratory  
Iowa State University  
Ames, Iowa 50011-3650  
515 294-7209  
FAX 515 294-5256

I look forward to continuing to work closely with your group, providing both data and expertise on how to work with and visualize diversity data.

Sincerely,



Patrick S. Schnable, Ph.D.  
Baker Professor of Agronomy  
Director, Center for Plant Genomics

UNIVERSITY OF CALIFORNIA, SAN DIEGO

UCSD

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO



SANTA BARBARA • SANTA CRUZ

Digitally enabled Genomic Medicine (DeGeM)  
 California Institute of Telecommunication and Information Technology (Calit2)  
 Associate Vice Chancellor, Research and Professor Office: 858-822-0885  
 9500 Gilman Drive, UCSD MC # 0466 Lab: 858-534-0443  
 La Jolla, CA 92093-5004 Fax: 858-822-1452

22 October 2012

Dr. Carolyn Lawrence  
 USDA-ARS  
 1034 CGIL, ISU  
 Ames, IA 50011

Dear Dr. Lawrence,

I write, as the PI of the NSF-funded collaborative effort, the Research Coordination Network for Genomic Standards Consortium (RCN4GSC), to confirm our interest in collaborating with your ARS MaizeGDB project during the next phase of your essential community resource. As someone building new bridges for genomics standards (and secondarily for other biological standards essential for effective communication and rapid scientific progress), I am particularly pleased to learn that the maize community has recognized the need to work with community standards, to ensure effective and fully accessible use of specifications and standards for genome assemblies and annotations. To date in the genomics standards effort, we have not had a deep engagement of plant biologists in these endeavors and it will be best for the entire biological sciences community to have experts like you and your colleagues so engaged. The only expectation to be part of any community standards efforts, and certainly all of our efforts, is to indicate and demonstrate a commitment to be involved, to have a voice, and to work to implement whatever aspects are relevant to your own work. The full participation of the leadership of MaizeGDB would be an important step forward. We too work closely with the INSDC (or NCBI, the USA partner of this world wide effort), and other major community database and knowledgebase resources, and recently have made tremendous progress in connecting traditional biodiversity standards to genomic standards, enhancing the emerging science of genomic biodiversity. I recognize you and your community have found interesting biodiversity for maize and I also hope to tap into your overall expertise in this case, as part of the knowledge and insight contained within the MaizeGDB. You have already become a member of the parent Genomic Standards Consortiums and the American chapter, the RCN4GSC, and we look forward to formalizing our interactions on a resource level as well, and to the continued community contributions of MaizeGDB.

Best regards,

A handwritten signature in blue ink that reads "John C. Wooley".

Professor John C. Wooley, Ph.D.  
[jwooley@ucsd.edu](mailto:jwooley@ucsd.edu)



October 11, 2012

Carolyn Lawrence and Taner Sen  
USDA-ARS  
Crop Genome Informatics Laboratory  
Iowa State University  
Ames, IA 50010

Re: Letter of Collaboration for MaizeGDB Project Plan

Dear Carolyn and Taner,

Thank you for sharing the MaizeGDB Project Plan with Nirav Merchant, Eric Lyons, and me to determine areas of mutual interest between MaizeGDB and the iPlant Collaborative.

Your plan's Objective 1 is to support genome stewardship for B73 and the diversity represented by maize as a species. Diversity measurements currently being generated include both genomic resequencing as well as information derived from expressed sequences (using, e.g., RNA-seq). As such, we are keen on sharing information between the MaizeGDB and iPlant teams to ensure that good solutions are arrived at as well as to minimize overlap where possible.

We at iPlant are also interested in supporting metabolite and gene network analysis and modeling (congruent with your Objective 2, goal 2a), as well as deploying data and tools that would support association studies and marker assisted/molecular breeding (in line with your Objective 2, goal 2b). Again, we are committed to not only keeping an open line of communication with your group on these topics, but on collaborating on data representation and tool deployment where synergy can be created by working together. Dr. Brett Tyler at Oregon State University has agreed to lead an iPlant working group on metabolite profiling and Basil Nikolau at Iowa State will be asked to participate in the working group. The efforts of this group will be coordinated with yours as described above.

Please keep in touch as projects develop. We look forward to working together to address these needs for the broad research community and for maize in particular.

Sincerely,

A handwritten signature in blue ink, appearing to read "Stephen A. Goff", written over a light blue circular stamp.

Stephen A. Goff  
Project Director, PI  
iPlant Collaborative

Thomas J. Keating Bioresearch Building, 1657 East Helen Street, Tucson, Arizona 85721



United States Department of Agriculture  
 Research, Education and Economics  
 Agricultural Research Service

October 10, 2012

Carolyn J. Lawrence  
 USDA-ARS  
 1034 CGIL  
 Iowa State University  
 Ames, IA 50011

Dear Carolyn and Taner,

As you know, I am a curator for MaizeGDB, and stationed at the Plant Genetics Research Unit in Columbia, MO. I regularly contribute to all aspects of MaizeGDB's endeavors, both by contributing to accomplishing deliverables and in guiding the work via weekly interactions on conference calls.

All aspects of the plan you outline will affect my work, described in the Columbia Project Plan entitled "Genetics and Genomics of Complex Traits in Grain Crops" and led by Mel Oliver. The objective I lead involves curating information on gene function into MaizeGDB, with the focus on relating experimentally confirmed relationships of important agronomic trait (e.g. drought response) and metabolic pathways (e.g. abscisic acid metabolism) to the maize reference genome sequence. This objective relies heavily on the infrastructure maintained by your group in Ames, both to facilitate data entry, and also to provide ready validation of the data, including checking for adherence to international data standards (e.g. the Plant Ontology and the Gene Ontology). Your Project Plan Objective 2, Goal 2a directly relates to this work.

I also can contribute guidance on developing the pedigree and breeding network tools (2b) and would be happy to offer guidance on developing tools that enable flexible and customized data query (Objective 3) by interacting with and asking for specific suggestions from the researchers in Columbia, both USDA and at the University of Missouri. I also do a fair amount of community support including Abstract Book preparation and serving on the Maize Genetics Conference Steering Committee, so those aspects of the Ames Project Plan are also items I plan to collaborate on.

Sincerely,

Mary Schaeffer (Polacco), PhD  
 Geneticist and Curator MaizeGDB



Midwest Area - Columbia Location - Plant Genetics Research Unit  
 203 Curtis Hall - University of Missouri - Columbia, MO 65211  
 Voice: 573-884-7873 - Fax: 573-884-7850 - E-mail: Mary.Schaeffer@ARS.USDA.GOV  
 An Equal Opportunity Employer



## SCIENTIFIC DEPARTMENTS

**Embryology**  
BALTIMORE, MARYLAND

**Geophysical Laboratory**  
WASHINGTON, DC

**Global Ecology**  
STANFORD, CALIFORNIA

**The Observatories**  
PASADENA, CALIFORNIA AND  
LAS CAMPANAS, CHILIF

**Plant Biology**  
STANFORD, CALIFORNIA

**Terrestrial Magnetism**  
WASHINGTON, DC

**Carnegie Academy for  
Science Education**  
WASHINGTON, DC

Carnegie Institution  
of Washington

260 Panama Street  
Stanford, CA 93305

650 325 1521 PHONE  
650 325 6857 FAX

## PLANT BIOLOGY

**Seung Yon Rhee, Ph.D.**

Staff Scientist

rhee@acoma.stanford.edu

(650) 325 1521 x251 FAX: (650) 325 6857

October 10, 2012

Carolyn J. Lawrence  
1034 Crop Genome Informatics Laboratory  
Iowa State University  
Ames, IA 50011

Dear Carolyn,

The collaboration between Plant Metabolic Network (PMN) and MaizeGDB has been very fruitful and valuable in developing a new computational pipeline for functional gene annotation and literature-based curation of CornCyc. After we released CornCyc 1.0, MaizeGDB curators contributed to the curation of proteins and pathways, including indole-3-acetic acid (IAA) biosynthesis I/VIII, DIMBOA, and GA12 biosynthesis. These meticulous curations enhanced the value of CornCyc, and consequently became part of the CornCyc 2.0 release. In addition, our collaboration with Dr. Taner Sen in upgrading our computational pipeline for more accurate enzyme function assignment is in the works and these assignments will form the backbone of next release of CornCyc.

We would be happy to continue our collaboration with MaizeGDB for future releases of CornCyc as well. Our project is funded until August 2014 and will be updating CornCyc on a regular basis during this period and any collaboration with MaizeGDB is absolutely welcome. When we create new versions of CornCyc, we will freely share these resources with you. If CornCyc is not funded after August 2014, PMN and MaizeGDB will decide about how to proceed to maintain CornCyc for the benefit of maize researchers.

As always, our group will be delighted to continue assisting you with the curational activities of maize metabolic networks.

Sincerely,

Sue Rhee

www.ciw.edu

# IOWA STATE UNIVERSITY

OF SCIENCE AND TECHNOLOGY

**Department of Genetics,  
Development and Cell Biology**  
1210 Molecular Biology Building  
Ames, IA 50011-3260  
TEL (515) 294-7322  
FAX (515) 294-7629

Date: 9 May 2013

From: Jack M. Gardiner, Ph.D.  
Curator-MaizeGDB  
Dept of Genetics, Development and Cell Biology  
Iowa State University, Ames, IA

*On location at:*  
School of Plant Sciences  
University of Arizona,  
Tucson, AZ 85721

To: Taner Z. Sen, Ph.D. and Carolyn J. Lawrence, Ph.D.  
USDA-ARS  
Crop Genome Informatics Laboratory  
Iowa State University  
Ames, IA 50011

Dear Taner and Carolyn,

Thank you for contacting me regarding the work I do as a curator for MaizeGDB. As you know, I work extensively with the maize research community to both recruit their data and assist them in to leverage the MaizeGDB database to further their research objectives. While I am involved in all curatorial activities at MaizeGDB, my particular area of expertise is microarray and RNA-Seq gene expression data. I developed this expertise both through working in a laboratory setting with various forms of expression data and via my previous experience as a project manager for Dr. Vicki Chandler's Maize Microarray Project.

In my current role as a curator for MaizeGDB, I am responsible for both recruiting new data sets and making them available to the maize community via a variety of software tools that allow a "visualize and analyze" approach to be taken towards hypothesis generation. As such, I have recruited a variety of atlas-level gene expression data sets (both microarray and RNA-Seq based). These resources are well used by the maize community and we continue to seek out new data sets that will expand our knowledge of the maize transcriptome. Currently, we are working on recruitment of an RNA-Seq data set that will allow a more detailed view of the maize root transcriptome.

I am enthusiastically and absolutely committed to working with the USDA/ARS-based MaizeGDB team to export all maize GO annotations into the R/Bioconductor format. I have prior experience with Bioconductor and found it to be a very powerful in analyzing gene expression data. I am also committed to making maize metabolic pathways available in the BioPAX pathway format, which will facilitate their utilization in Cytoscope and Bioconductor. These two popular visualization and analysis tools will allow researchers to effectively leverage genome-wide gene expression data sets. I look forward to the next five years as it is a great time to be a biologist and these tools and others like them, will change how biological questions are asked and addressed.

Best wishes,



Jack Gardiner



**Donnelly Centre**  
for Cellular + Biomolecular Research



**Gary Bader, Ph.D.**

Associate Professor, The Donnelly Centre for Cellular + Biomolecular Research  
Department of Molecular Genetics + Computer Science | University of Toronto

Sept 24, 2012

Taner Sen  
USDA-ARS  
Department of Genetics, Development and Cell Biology  
Iowa State University  
Ames, Iowa 50011

Dear Taner,

I am very excited that you are considering implementing Cytoscape Web at MaizeGDB. We designed Cytoscape Web for databases just like MaizeGDB to allow creating customizable views of interaction networks for their users.

Cytoscape Web is an open source software application, which is modeled on Cytoscape. It can be easily customized and incorporated into any website. Cytoscape Web is intended as a low overhead tool to add network visualization to a web application, and supports GraphML, XGMML, and SIF network formats. You will be happy to hear that we are currently developing a new version of Cytoscape Web using HTML5, which will make integrating Cytoscape Web into websites even easier.

By having access to Cytoscape Web, MaizeGDB will not only enable users to visualize entity relatedness and interaction data, but will give them the ability to manipulate visualizations interactively to better understand biological networks. Your users will be able to modify the node and edge positions and shapes, which will help display their network of interest and develop hypotheses. Cytoscape Web is being used by hundreds of biological websites, including some model organism databases. These provide great examples of how to implement Cytoscape Web usefully for researchers. Additionally, demos are available at <http://cytoscapeweb.cytoscape.org/demos> so that users will have a better understanding of the capabilities of Cytoscape Web.

We have extensive documentation to help Cytoscape Web programmers. In addition, Cytoscape has a very active developer community. If you chose to implement Cytoscape Web at MaizeGDB, we will be more than happy to assist you during its implementation and integration through e-mail and phone. If funds are available, we can host you in my group to guide you through the implementation process. I am enthusiastically looking forward to collaborating with you.

Sincerely,

Gary Bader  
Associate Professor  
University of Toronto

Donnelly Centre for Cellular + Biomolecular Research | 160 College Street Room 602 Toronto Ontario M5S 3E1 Canada  
T 416.978.3935 | F 416.978.8287 | [gary.bader@utoronto.ca](mailto:gary.bader@utoronto.ca) | [www.thedonnellycentre.utoronto.ca](http://www.thedonnellycentre.utoronto.ca) | [www.baderlab.org](http://www.baderlab.org)

**IOWA STATE UNIVERSITY**  
**Department of Agronomy**  
*Crop, Soil, and Environmental Sciences*

1204 Agronomy Hall  
Ames, Iowa 50011-1010  
515 294-5356  
FAX 515 294-3163  
[THOMASL@iastate.edu](mailto:THOMASL@iastate.edu)

Sept 27, 2012

Taner Sen  
USDA-ARS  
Department of Genetics, Development and Cell Biology  
Iowa State University  
Ames, Iowa 50011

Dear Taner,

I am very happy to hear that MaizeGDB is planning to implement pedigree and breeding similarity network visualization tools. Many plant breeders use MaizeGDB extensively research to understand complex traits and genetic diversity for breeding. With additional pedigree tools, MaizeGDB will become even more valuable to breeders. I would be more than happy to provide guidance on the needs of breeders, supply and direct you to pedigree information, and assist you in preparing surveys to gauge the needs of breeders that can be met through implementation of the types of tools you plan to deploy.

Sincerely



Thomas Lubberstedt  
Professor, K.J. Frey Chair in Agronomy  
Director, R.F. Baker Center for Plant Breeding



**IOWA STATE UNIVERSITY**  
OF SCIENCE AND TECHNOLOGY

Department of Agronomy  
Crop, Soil, and Environmental Sciences  
Ames, Iowa 50011-1010  
515 294-1360  
FAX 515 294-3163

October 14, 2012

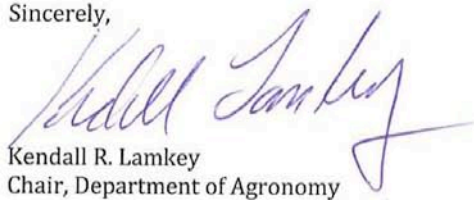
Taner Sen  
USDA-ARS  
Department of Genetics, Development and Cell Biology  
Iowa State University  
Ames, Iowa 50011

Dear Taner,

It was great talking to you and Carolyn about MaizeGDB's 5-year project plan. I fully support your plans to bring in breeding data and implement visualization tools to show pedigree information and haplotype relationships among maize lines. As I pointed out during our conversation, breeders are most interested in two broad questions: 1) what is the ancestry of lines going back to time zero, and 2) the relationships among lines. The more complete the answers to these questions, the more powerful our ability to build predictive models of hybrid performance. Visualizing line annotations showing their history and phenotypes will be highly useful for breeders.

It is very exciting that you will survey breeders to identify their needs for centralized data and its visualization, so that you can integrate such views at MaizeGDB. Our close proximity at Iowa State University is a major advantage. I would be more than happy to provide guidance about the most-sought breeder data, assist you in preparing the survey, and making connections in the maize breeding community.

Sincerely,



Kendall R. Lamkey  
Chair, Department of Agronomy



United States Department of Agriculture

Research, Education and Economics  
Agricultural Research Service

October 15, 2012

Taner Sen  
USDA-ARS  
1025 Crop Genome Informatics Lab  
Iowa State University  
Ames, Iowa 50011

Dear Taner,

Thank you for sharing MaizeGDB's 5-year Project Plan with me. I am very pleased to learn that you are planning to implement new breeding data and visualization tools at MaizeGDB. By adding breeding data and tools that are not otherwise readily accessible, it's more likely that breeders could use MaizeGDB in their research and that MaizeGDB could become genuinely useful to them.

The North Central Regional Plant Introduction (PI) Station serves as the germplasm repository for maize germplasm, including inbred lines, populations and landraces, and its wild relatives. In addition to the pedigree information in the GRIN database, several of our personnel have deep knowledge of the history of many breeding lines and populations. Many lines have complex histories and have been subjected to various breeding schemes, so working with you to be able to represent those complex relationships will be critical. We will be delighted to examine the tools you develop to assess whether known complex relationships are stored, and look forward to helping you formulate questions for a survey of breeders that will inform you of the types of data of interest to breeders and use cases.

I look forward to continuing a strong collaboration with MaizeGDB, and wish you much success in your efforts.

Sincerely,

A handwritten signature in cursive script that reads "Candice Gardner".

Candice Gardner  
Research Leader and Supervisory Plant Biologist

Cc: Carolyn Lawrence, PI, MaizeGDB  
Mark Millard, PIRU, and NCRPIS Maize Curator



Midwest Area § Plant Introduction Station  
G212 Agronomy Hall § Iowa State University § Ames, IA 50011  
Voice: 515-294-3255 § FAX: 515-294-4880 § E-mail: candice.gardner@ars.usda.gov  
An Equal Opportunity Employer



United States Department of Agriculture

Research, Education and Economics  
Agricultural Research Service

Dear Carolyn,

Your plan to develop and deploy tools to increase user-specified flexible queries in MaizeGDB (Objective 3) coincides closely with elements of our Objective 1 in SoyBase and the Legume Clade Database Project Plan. As the Lead Scientist of the SoyBase CRIS I write this to affirm our collaboration in these activities. The close proximity and collegiality of our groups will continue to ensure exchanges of ideas and mutual consultation.

I look forward to our continued interactions.

Sincerely,

A handwritten signature in black ink that reads "Randy C. Shoemaker". The signature is fluid and cursive, with a long horizontal flourish extending to the right.

Randy C. Shoemaker, Research Geneticist  
Senior Scientific Research Service



Midwest Area • Corn Insects and Crop Genetics Research Unit  
G401 Agronomy Hall • Iowa State University • Ames, IA 50011-1010  
Voice: (515) 294-6233 • FAX: (515) 294-2299 • E-mail: randy.shoemaker@ars.usda.gov



Cornell University

Institute for Biotechnology and Life Science Technologies  
**NYSTAR Designated Center for Advanced Technology**  
 130 Biotechnology Building  
 Ithaca, New York 14853-2703  
 t. 607.255.2300 f. 607.254.6379  
<http://www.biotech.cornell.edu>

October 12, 2012

Carolyn J. Lawrence  
 USDA-ARS  
 1034 Crop Genome Informatics Laboratory  
 Iowa State University  
 Ames, IA 50011

Dear Carolyn,

I am writing as Chair of the Maize Genetics Executive Committee (MGEC) to endorse your plans to support maize research community-building and community support activities (Objective 4). As a member of MGEC, you are aware that, "it is the mission of the Maize Genetics Executive Committee to identify both the needs and the opportunities for maize genetics, and to communicate this information to the broadest possible life science community. This community includes scientists, funding sources for scientists, and the end users for the accomplishments of maize genetics, from farmers to consumers" (Bennetzen MNL 75).

Some activities of the MGEC are managed by MaizeGDB. Personnel at MaizeGDB manage a listserv for members as well as the MGEC website, conduct elections of new members, and administer surveys of the community on behalf of MGEC. Other activities of MGEC are designed to support MaizeGDB and include approving community email requests, including MaizeGDB-specific sections in community survey content, and encouraging researchers to communicate with MaizeGDB personnel to help ensure the database remains relevant and useful.

We plan to continue this useful collaboration and appreciate that this aspect of the work conducted by MaizeGDB is specified in the Project Plan.

Sincerely,

Edward S. Buckler, Ph.D. (on behalf of the Maize Genetics Executive Committee)  
 USDA-ARS Research Geneticist  
 Adjunct Professor of Plant Breeding and Genetics  
 Institute for Genomic Diversity, Cornell University

*Current MGEC Membership: Tom Brutnell (Interim), William Tracy (2012), Sue Wessler (2012), Jeff Bennetzen (2013), Carolyn Lawrence (2014), Marja Timmermans (2015), Nathan Springer (2015), James Birchler (2016), Sarah Hake (2016).*

**IOWA STATE UNIVERSITY**  
OF SCIENCE AND TECHNOLOGY

Philip W. Bercraft  
Department of Genetics, Development &  
Cell Biology  
2116 Molecular Biology Building  
Ames, IA 50011-3260

Phone: 1-515-294-2903  
Fax: 1-515-294-6755  
Email: [bercraft@iastate.edu](mailto:bercraft@iastate.edu)

October 10, 2012

Carolyn J. Lawrence  
USDA-ARS  
1034 Crop Genome Informatics Laboratory  
Iowa State University  
Ames, IA 50011

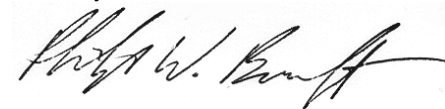
Dear Carolyn:

As current chair of the Maize Genetics Conference Steering Committee (MGCSC), I endorse your 5-year project plan as it pertains to the collaboration between the MGCSC and MaizeGDB. MaizeGDB and the Maize Genetics Conference (MGC) are both centrally important to the maize genetics community. Both serve to nucleate the community, providing valuable cohesion, and both act synergistically as critical avenues for the sharing of scientific information that facilitates maize improvement. The services provided by MaizeGDB are invaluable to the running of the conference. MaizeGDB hosts the conference website, and orchestrates online registration, abstract submission and printing of the program and abstract book. They also provide a searchable archive of abstracts from past conferences.

The MGC is one of the few major conferences to be organized on an ad hoc basis. The MGCSC is composed of volunteers from the maize community who each year work to decide on a venue, seek financial support, develop a program, and organize all the activities associated with the conference. These volunteer services help keep the costs of attending the MGC low, and graduate students can usually attend for free. This serves to greatly promote the education and training of the next generation of maize genetics and genomics researchers. Running the MGC on this ad hoc basis WOULD NOT BE POSSIBLE without the support of MaizeGDB.

It is impossible to express the gratitude of the MGCSC for the enabling services provided by MaizeGDB, and we look forward to continuing this fruitful collaboration long into the future.

Sincerely,



Philip W. Bercraft  
Chair, Maize Genetics Conference Steering Committee



UNIVERSITY OF  
MARYLAND

3125 Biomolecular Sciences Building #296  
College Park, Maryland 20742  
Voice: 301.405.5936 Fax: 301.314.1341  
[www.cbcb.umd.edu](http://www.cbcb.umd.edu)

CENTER FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY  
INSTITUTE FOR ADVANCED COMPUTER STUDIES

Carolyn J. Lawrence  
USDA-ARS  
1034 CGIL  
Iowa State University  
Ames, IA 50011

October 19, 2012

Dear Carolyn,

I am writing to you on behalf of the Maize Genetics and Genomics Database (MaizeGDB) Working Group to express our support for your group and to reconfirm our continued involvement in the activities of the MaizeGDB over the next 5 years. The MaizeGDB Working Group comprises scientists with broad expertise in the biology, genetics, genomics, informatics, and breeding of maize and other crops. Our group, thus, reflects the diversity of the community of stake-holders for MaizeGDB, both academic and industrial. The current membership includes Alice Barkan (U. of Oregon), David Jackson (CSHL), Anne-Francoise Lamblin (U. of Minnesota), Thomas Lubberstedt (Iowa State U.), Eric Lyons (UC Berkeley), Karen McGinnis (Florida State U.), Lukas Mueller (Cornell U.), Marty Sachs (USDA-ARS), Nathan Springer (U. of Minnesota), and myself.

Our role within the MaizeGDB community has been to serve as an advisory group to you and your colleagues in order to ensure that the continued maintenance and development of MaizeGDB serve the needs of the broad community it supports. Over the past six years, we have, and continue to serve our role through annual meetings where we evaluate the current state and development plans for MaizeGDB, and through direct informal interactions between Working Group members and the MaizeGDB staff. During the annual meetings we typically review a comprehensive status report from MaizeGDB as well as detailed plans for further development that you provide us. Our discussions start in an interactive session (either face-to-face or through phone conference) and often continue for several additional weeks through e-mail, allowing us to thoroughly consider the needs and priorities of the maize community. The conclusions of our meetings are provided to you in the form of a report, agreed to by the entire Working Group, which contains additional recommendations for your group. All the reports are publicly available from the MaizeGDB site ([http://www.maizegdb.org/working\\_group.php](http://www.maizegdb.org/working_group.php)), providing transparency, and allowing the rest of the maize community to provide feedback to us and to your group.



UNIVERSITY OF  
MARYLAND

3125 Biomolecular Sciences Building #296  
College Park, Maryland 20742  
Voice: 301.405.5936 Fax: 301.314.1341  
[www.cbcb.umd.edu](http://www.cbcb.umd.edu)

CENTER FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY  
INSTITUTE FOR ADVANCED COMPUTER STUDIES

Both myself and the rest of the working group are extremely pleased with the interaction between us and MaizeGDB. The last few years have been marked by the accelerated availability of maize data, made possible by new sequencing technologies and other technological advances. We feel our role has been particularly important during this period of rapid changes in helping prioritize the growth of MaizeGDB, and are delighted to see our recommendations implemented as MaizeGDB is adapting to the deluge of data that is becoming available.

Finally I would like to congratulate you and your colleagues for the tremendous resource you have put together and for the amazing progress the MaizeGDB has made over the past few years. The Working Group and myself are delighted to have been part of the growth of MaizeGDB, and are looking forward to continuing our fruitful collaboration over the next 5 years.

On behalf of the MaizeGDB Working Group

Sincerely,

A handwritten signature in black ink, appearing to read 'MP'.

Mihai Pop  
Associate Professor, Department of Computer Science, and  
Center for Bioinformatics and Computational Biology  
University of Maryland  
College Park, MD 20742

## C. WORKING GROUP REPORT

**Working Group Members:** Alice Barkan, Dave Jackson, Anne-Francoise Lamblin, Thomas Lübberstedt, Eric Lyons, Karen McGinnis, Lukas Müller, Mihai Pop (chair), Marty Sachs, Nathan Springer.

**Observers:** Jack Okamuro (USDA-ARS), Dave Marshall (USDA-ARS), Diane Okamuro (NSF), Craig Abel (USDA-ARS), Carl Simmons (Pioneer Hi-Bred Intl., Inc.), Bing-Bing Wang (Pioneer Hi-Bred Intl., Inc.)

This report summarizes the Working Group (WG) web/phone meeting held on March 30, 2012.

### CHARGE TO THE WORKING GROUP

During the March 30<sup>th</sup> meeting, the MaizeGDB team presented the current state of the database as well as their upcoming plans for development that will be summarized in a concept paper to be submitted to USDA-ARS. The concept paper outlines the framework in which MaizeGDB activities will be conducted over the next 5 years. The working group was charged with evaluating the scope of these activities and providing MaizeGDB with feedback.

### WORKING GROUP FINDINGS

The working group discussion started by evaluating the current status of MaizeGDB, then broadly followed the outline of the proposed project goals. Our findings are presented below according to this outline.

#### Current MaizeGDB Status and User Experience

The working group members have pointed out a number of limitations with the current MaizeGDB interface. A significant concern were performance issues that lead to slow response to both textual and BLAST queries. Also noted were difficulties in searching for specific identifiers (e.g., gene models do not appear to be searchable through the main search box), and the lack of sufficient cross-references between the data stored at MaizeGDB and the multiple existing databases (both other plant databases and general resources such as UniProt). The working group recognizes that many of these issues are easy to address and will likely be resolved once the redesigned MaizeGDB interface is deployed, however we want to stress the importance of ensuring an efficient and feature-rich interaction between end users and MaizeGDB.

Also noted in this discussion were several features that do not currently appear to be available, such as providing provenance information for the gene models and annotations displayed by MaizeGDB, and providing the ability to bulk download information through the interface.



**Concept paper objective 1: *Support stewardship of maize genome sequences.***

The working group recognizes this objective as key to the mission of MaizeGDB and acknowledges the importance of the proposed activities. Given the rapid growth in the availability of maize-related data, it is important for MaizeGDB to better define (e.g., through a public mission statement) what their role is within the maize community. More specifically, should the data stored and maintained at MaizeGDB be limited to gold standard datasets or include a broader collection of data available in the community? Also, what is entailed by ‘stewardship’ of the data stored at MaizeGDB, and how is this stewardship shared with other community members?

Achieving this objective will require effective partnerships with the other groups managing the underlying maize assembly, which is currently maintained by maizesequence.org and will be transitioned to Gramene later in the year. As the sequence itself, and the associated gene models and annotations are refined and updated through user input and MaizeGDB activities, it is important to record and present to the users data related to the accuracy and provenance of the entities stored in the database. Users should be able to know if changes to a sequence or gene model have been submitted even if the updates have not yet taken place, as well as to understand the information and computational analyses that underlie a specific gene call or functional annotation.

Working group members also outlined the difficulties that will be encountered by MaizeGDB as they start storing and displaying the rapidly accumulating information about the genomic diversity of maize. We recommend that MaizeGDB adopt or adapt approaches developed to cope with the storage and display of human diversity data (e.g. Hapmap). Also, pedigree information could be used to organize, and simplify access to the diversity data.

**Concept paper objective 2: *Deploy open, standardized access mechanisms to the MaizeGDB database that allow user-specified and flexible queries.***

The importance of improvements to the MaizeGDB interface was already stressed in our discussions and the working group strongly supports activities in this directions. Some of the feature requests outlined above (such as bulk download abilities) will likely be addressed by the proposed additions to MaizeGDB.

**Concept paper objective 3: *Create access to expanded datasets that describe gene function.***

The working group members agree with the importance of expanding the data related to gene function in MaizeGDB, especially as the types and volume of data are rapidly expanding. Network data (regulatory, metabolic, interaction, etc.), in particular, represent a valuable resource to scientists and MaizeGDB would benefit from storing, curating, and displaying such data.

At the same time, it is unrealistic to expect that MaizeGDB could serve as a repository or even a portal for all the data generated by the community. Instead we recommend that efforts be made to identify and prioritize specific datasets for inclusion into MaizeGDB. The prioritization should be based on both the breadth of the community that is served by the data, and also on whether these datasets could enable or enhance the analysis of the many resources being

generated within the community. MaizeGDB should play an active role in communicating with the groups generating and using the data in order to better understand how to best serve the needs of the community.

**Concept paper objective 4: *Provide community support services and increase documentation on responsiveness to community needs by coordinating annual meetings, conducting community elections and surveys, etc.***

The working group members unanimously and enthusiastically praised all the outreach and support activities undertaken by MaizeGDB staff and commented on the important role that MaizeGDB has played in bringing the maize community together. Scientific meetings such as the annual Maize Genetic Conference would not be possible without MaizeGDB. We feel that these outreach activities are more than simple educational/advertising campaigns and have had an important impact on maize research in general by fostering communication within the community, enabling collaborations, and helping educate the new generations of maize researchers.

These activities are an extremely important and valuable contribution of MaizeGDB and should be continued and supported in the future.

**EXECUTIVE SUMMARY AND RECOMMENDATIONS**

- The working group agrees with the general vision set out in the MaizeGDB Concept Paper, and finds the specific objectives adequately address the anticipated needs of the maize community over the coming years.
- At the same time, our group recognizes the fact that MaizeGDB's resources are limited and recommends that the specific developments be carefully prioritized according to community needs.
- The working group has noted the increased availability of datasets that could conceivably be incorporated within, or linked from, MaizeGDB. As these datasets are rapidly growing in number and size, it is imperative that MaizeGDB reevaluate the role they play within the community. The development of a clear mission statement, vetted by the community, would greatly help guide the prioritization of activities and ensure MaizeGDB continues to play a valuable role within the maize community.
- MaizeGDB's outreach efforts have been outstanding and have significantly benefited the maize research community. It is imperative that these efforts continue and are adequately supported.

## Mary Schaeffer Objectives and Milestones

2013-18 CRIS project: Genetics and Genomics of Complex Traits in Grain Crops

[with Mel Oliver (lead), Michael McMullen and Sherry Flint-Garcia]

[http://ars.usda.gov/research/projects/projects.htm?accn\\_no=424655](http://ars.usda.gov/research/projects/projects.htm?accn_no=424655)

**Objective 4:** Identify and curate key datasets that will serve to benchmark genomic discovery tools for key agronomic traits, especially response to biotic and abiotic environmental stressors.  
(non-hypothesis driven)

*Background:* The maize reference genome and the sequence diversity data represented at MaizeGDB become more useful if related to functional genomics data that range from levels of transcription in certain tissues and growth stages to associations with agronomically favorable traits. Cross-species data access is important because candidate gene functions in one species are very often inferred from other species, where functions have been experimentally determined (Nomura *et al.*, 2003; Chi *et al.*, 2012; Setter *et al.*, 2011). Cross-species genomics tools for gene function inference rely on machine-readable inputs that use structured and controlled vocabularies, otherwise known as ontologies. Both the Plant Ontology and the Gene Ontology are widely accepted standards for biological databases (Walls et al 2012; Ashburner et al 2000). MaizeGDB hosts the phenotypic data for the mutant collection of the Maize Genetics Cooperation Stock Center and for over 400 agronomic traits described in the literature (Schaeffer 2006, Sachs 2009). The challenges for this plan will be handling expanding data from maize genome sequencing projects and providing tools for functional analysis and genetic improvement. Thus this objective is an integral part of the project plan and one that ensures that our research delivers a product directly to the community and has an immediate impact on the field.

*Sub-objective 4.1:* Bring into The Maize Genome Database (MaizeGDB) the phenotypic data generated by critically important research endeavors including the Maize Diversity Project.

*Goal 4.1:* Provide facile access to phenotypic diversity associated to genotype.

*Action Plan:* The MaizeGDB team and the Maize Diversity Project members have initiated a pipeline directed to integrate the Maize Diversity Project data stored at Ithaca, NY into MaizeGDB, and to establish regular updates. Currently the bulk of phenotypic evaluations stored at the database served by [www.panzea.org](http://www.panzea.org) represent 154 traits for the 5,000 NAM lines. These data should become accessible from MaizeGDB in late 2012. Members of the MaizeGDB team will collaborate with Maize Diversity Project personnel to develop storage solutions and an interface to access these data. We will work with personnel at the North Central Regional Plant Introduction Station and personnel at GRIN to insure germplasm is adequately described.

To make the data accessible to other crop genome and germplasm/genebank resources, integration will involve current, emerging, and evolving ontologies to describe the data and collaboration with other crop databases. MaizeGDB will provide association files to external resources as requested. This builds on previous work at MaizeDB (the forerunner of MaizeGDB) that categorized over 400 agronomic traits with inheritance data. These associations were updated to harmonize with the Gramene Trait Ontology, and GRIN trait categories (Schaeffer *et al.*, 2006). This project also links with other members of MaizeGDB to insure that phenotype variation qualifiers are appropriate for both mutant and quantitative phenotype variations. Key datasets are mutants represented by the Maize Genetics Cooperation, Stock Center; additionally M.G. Neuffer (University of Missouri) is currently updating phenotypes for ~ 3,500 mutants. A phenotype-attribute ontology is being developed for all species (Eva Huala of The Arabidopsis Information Resource

## Mary Schaeffer Objectives and Milestones

2013-18 CRIS project: Genetics and Genomics of Complex Traits in Grain Crops

[with Mel Oliver (lead), Michael McMullen and Sherry Flint-Garcia]

[http://ars.usda.gov/research/projects/projects.htm?accn\\_no=424655](http://ars.usda.gov/research/projects/projects.htm?accn_no=424655)

(TAIR) is the plant coordinator).

*Sub-objective 4.2:* Curate maize metabolism and pathways data for release as a BioCyc database and as GO annotation files.

*Goal 4.2:* Annotate maize genome sequence with critical, experimentally confirmed, gene function.

*Action Plan:* Selected pathways with extensive maize data and important to this project will be curated using the BioCyc pathway tool suite, and GO annotation tools developed at MaizeGDB. The MaizeGDB team at Ames IA will provide access to curation tools, write scripts for updating the metabolism database, provide association files to the GO project, and coordinate with the MaizeGDB Working Group on priorities. We will engage community experts to evaluate and provide data. We have a substantial community contribution provided by Monika Frey, a summary of genes and enzymes in the DIMBOA pathway studied by Mike McMullen. Other relevant pathways to curate from the literature will include those implicated in responding to drought stress: ABA synthesis and regulation, carbon flux, and flowering (Setter *et al.*, 2011). We will use text-mining tools that are being developed for genome databases, in collaboration with the Protein Information Resource (PIR) project (<http://proteininformationresource.org/iPTMnet/iPTMnet.shtml>; Arighi *et al* 2011) to understand and predict post-translational modifications in signaling pathways.

*Contingencies:* (4.1) It is highly likely that there will be similar data generated by other groups that would expand the scale of information for curation. Descriptions of the process to import the diversity data of this project will be made available to educate other groups, thus enabling them to prepare their own data for upload to MaizeGDB. Additional types of functional data that relate to phenotypes will need to be accessed at or from the MaizeGDB website over the next 5 years. The MaizeGDB team, and its 10-member Working Group, will assist in setting priorities that may expand the scope of phenotype-focused curatorial activities. (4.2) Other genome-specific pathway tools may be developed that are superior to the BioCyc toolset. In that case, the GO annotations at MaizeGDB, with pathway representations stored in the BioCyc databases, would be transitioned for representation via the new tool.

### *Cooperation/Collaboration:*

(4.1) Within ARS: Carolyn Lawrence, Candy Gardner, and Randy Shoemaker, Ames, IA; Victoria C. Blake, Albany, CA.

External to ARS: Myron G. Neuffer, Columbia MO; Eva Huala, Stanford, CA; Pankaj Jaiswal, Corvallis, OR.

(4.2) Within ARS: Carolyn Lawrence, Ames, IA will provide interfaces that link phenotypic data for stakeholders and robust tools to support annotation.

External to ARS: C. Arighi, Newark, DE will coordinate a community-wide project to develop artificial intelligence tools to transfer gene functional data to genome databases.

Mary Schaeffer Objectives and Milestones  
 2013-18 CRIS project: Genetics and Genomics of Complex Traits in Grain Crops  
 [with Mel Oliver (lead), Michael McMullen and Sherry Flint-Garcia]  
[http://ars.usda.gov/research/projects/projects.htm?accn\\_no=424655](http://ars.usda.gov/research/projects/projects.htm?accn_no=424655)

Milestones copied from main documents

<b>National Program</b>		301, Plant Genetic Resources, Genomics and Genetic Improvement			
<b>Objective 4</b>		Identify and curate key datasets that will serve to benchmark genomic discovery tools for key agronomic traits, especially response to biotic and abiotic environmental stressors. (non-hypothesis driven)			
<b>Sub-objective 4.1</b>		Bring into MaizeGDB the phenotypic data generated by critically important research endeavors including the Maize Diversity Project. (non-hypothesis driven)			
<b>NP Action Plan Component</b>		Component 1: Crop Genetic Improvement.			
<b>NP Action Plan Problem Statement</b>		Problem Statement 1B: Innovative approaches to crop genetic improvement and trait analysis			
	<b>SY Team</b>	<b>Months</b>	<b>Milestones</b>	<b>Progress/ Changes</b>	<b>Products</b>
<i>Goal 4.1: Provide facile access to phenotypic diversity associated to genotype.</i>	<b>MLS</b>	<b>12</b>	2012 Maize Diversity Project Phenotypic data integrated at MaizeGDB.		Pipeline for integration of diversity phenotypic data.
	<b>MLS</b>	<b>24</b>	Update critical 2013 diversity trait data.		(1) Association files for Plant Ontology, Trait Ontology. (2) Data Files to support common statistical software.
	<b>MLS</b>	<b>36</b>	Update critical 2014 diversity trait data.		(1) Association files for Plant Ontology, Trait Ontology. (2) Data Files to support common statistical software.
	<b>MLS</b>	<b>48</b>	Update critical 2015 diversity trait data.		(1) Association files for Plant Ontology, Trait Ontology. (2) Data Files to support common statistical software.
	<b>MLS</b>	<b>60</b>	Update critical 2016 diversity trait data.		(1) Association files for Plant Ontology, Trait Ontology. (2) Data Files to support common statistical software.
<b>Sub-objective 4.2</b>		Curate maize metabolism and pathways data for release as a BioCyc database and as GO annotation files. (non-hypothesis driven)			
	<b>SY Team</b>	<b>Months</b>	<b>Milestones</b>	<b>Progress/ Changes</b>	<b>Products</b>
<i>Goal 4.2: Annotate maize genome sequence with critical, experimentally confirmed, gene function.</i>	<b>MLS</b>	<b>12</b>	ABA, auxin, select signaling pathways completed.		Updated BioCyc, GO files.
	<b>MLS</b>	<b>24</b>	Carbohydrate and photosynthesis pathways completed.		Updated BioCyc, GO files.
	<b>MLS</b>	<b>36</b>	Select Pest defense pathways completed.		Updated BioCyc, GO files.
	<b>MLS</b>	<b>48</b>	Select flowering related pathways completed.		Updated BioCyc, GO files.
	<b>MLS</b>	<b>60</b>	Select seed quality pathways completed.		Updated BioCyc, GO files.

## **6 – Presentations/Outreach**

### **National Corn Grower Association (NCGA) PodCasts (Gardiner)**

NCGA Communications Director Cathryn Wojcicki produced three Podcasts that have been disseminated through the NCGA outreach network. These describe the EFP browser, how SNPs are used, and what phenotypes are. These Podcasts are intended to inform farmers of how the science they support improves plants grown in their fields. To access these podcasts, look under the outreach tab in the upper right of the current front page at <http://www.maizegdb.org>.

### **Plant Genomes Resources Exhibit at PAG and ASPB (Schaeffer, Gardiner, Harper)**

MaizeGDB organizes an annual exhibit booth displaying plant genome database resources at the PAG and contributes to a booth at ASPB. This is done in collaboration with a number of other plant genome databases to both minimize costs and create a larger awareness of these resources.

### **MaizeGDB Redesign at Maize Meeting (Andorf)**

Andorf gave a talk on the MaizeGDB redesign at the Maize Genetics Conference (March 2013).

### **2012 International Sorghum Genomics Workshop (Lawrence)**

High-level description of MaizeGDB for consideration by Sorghum researchers as they determine how best to improve their resources with respect to informatics and community coordination.(November 14-16, 2012)

### **Corn Insects and Crop Genetics Research Unit Brown Bag Research Seminar Series (Lawrence and Andorf)**

Lawrence and Andorf described the MaizeGDB Project and the interface redesign to the research unit (December 2012).

### **MaizeGDB Tutorials at Maize Meeting**

Curator Lisa Harper and the MaizeGDB team conducted three MaizeGDB tutorials to a total of more than 60 attendees on March 15th at the annual Maize Genetics Conference in St. Charles, IL. These tutorials focused primarily on training researchers to use the newly redesigned website.

### **MaizeGDB Tutorials at Other Venues**

Curator Lisa Harper also conducted two similar MaizeGDB tutorials in February of 2013, one at the Genetics of Maize-Microbe Interactions Conference, Danforth Center MO, and a second at the University of Missouri at Columbia.

### **Corn Breeding Research Meeting (formerly NCCC167)**

Carolyn Lawrence described the MaizeGDB project to maize breeders who are hopeful that MaizeGDB can help them to coordinate their community events and enable access to more breeding-oriented data in the future.

### **PAG XXI PIR Workshop: Knowledge Mining Tools for Proteins, Complexes and PTMs**

Curator Mary Schaeffer presented the MaizeGDB plan for curation of metabolism, including the use GO annotations from data mining and community curation, and our collaborations with J Walsh and J Dickerson, the CornCyc and the MaizeCyc projects.

### **Crop Plant Trait Ontology Workshop**

Curator Mary Schaeffer presented an overview MaizeGDB use of the Plant Ontology to annotate phenotypes and tissues used for gene expression experiments. This was an international workshop convened by the Plant Ontology and Bill Gates funded CGIAR Crop Ontology projects towards developing a broader Trait Ontology. It included representatives from many plant genome databases and breeding projects.

## 7 – Peer-reviewed journal articles since March 2012

### **Maize Metabolic Network Construction and Transcriptome Analysis**

Monaco MK, Sen TZ, Dharmawardhana PD, Ren L, Schaeffer M, Naithani S, Amarasinghe V, Thomason J, Harper L, Gardiner J, Cannon EKS, Lawrence CJ, Ware D, Jaiswal P.  
The Plant Genome 2013 Mar 6(1):1-12. doi: 10.3835/plantgenome2012.09.0025. Epub 2013 Mar 6.

### **Maize chromosomal knobs are located in gene-dense areas and suppress local recombination.**

Ghaffari R, Cannon EK, Kanizay LB, Lawrence CJ, Dawe RK.  
Chromosoma. 2013 Mar;122(1-2):67-75. doi: 10.1007/s00412-012-0391-8. Epub 2012 Dec 9.

### **Predicting the binding patterns of hub proteins: a study using yeast protein interaction networks.**

Andorf CM, Honavar V, Sen TZ.  
PLoS One. 2013;8(2):e56833. doi: 10.1371/journal.pone.0056833. Epub 2013 Feb 19.

### **The plant ontology as a tool for comparative plant anatomy and genomic analyses.**

Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, Preece J, Athreya B, Mungall CJ, Rensing S, Hiss M, Lang D, Reski R, Berardini TZ, Li D, Huala E, Schaeffer M, Menda N, Arnaud E, Shrestha R, Yamazaki Y, Jaiswal P.  
Plant Cell Physiol. 2013 Feb;54(2):e1. doi: 10.1093/pcp/pcs163. Epub 2012 Dec 5.

### **An overview of the BioCreative 2012 Workshop Track III: interactive text mining task.**

Arighi CN, Carterette B, Cohen KB, Krallinger M, Wilbur WJ, Fey P, Dodson R, Cooper L, Van Slyke CE, Dahdul W, Mabee P, Li D, Harris B, Gillespie M, Jimenez S, Roberts P, Matthews L, Becker K, Drabkin H, Bello S, Licata L, Chatr-aryamontri A, Schaeffer ML, Park J, Haendel M, Van Auken K, Li Y, Chan J, Muller HM, Cui H, Balhoff JP, Chi-Yang Wu J, Lu Z, Wei CH, Tudor CO, Raja K, Subramani S, Natarajan J, Cejuela JM, Dubey P, Wu C.  
Database (Oxford). 2013 Jan 17;2013:bas056. doi: 10.1093/database/bas056. Print 2013.

### **Maize multiple archesporial cells 1 (mac1), an ortholog of rice TDL1A, modulates cell proliferation and identity in early anther development.**

Wang CJ, Nan GL, Kelliher T, Timofejeva L, Vernoud V, Golubovskaya IN, Harper L, Egger R, Walbot V, Cande WZ.  
Development. 2012 Jul;139(14):2594-603. doi: 10.1242/dev.077891. Epub 2012 Jun 13.

### **RCN4GSC Workshop Report: Managing Data at the Interface of Biodiversity and (Meta)Genomics, March 2011.**

Robbins RJ, Amaral-Zettler L, Bik H, Blum S, Edwards J, Field D, Garrity G, Gilbert JA, Kottmann R, Krishtalka L, Lapp H, Lawrence C, Morrison N, Tuama EÓ, Parr C, San Gil I, Schindel D, Schriml L, Vieglas D, Wooley J.  
Stand Genomic Sci. 2012 Oct 10;7(1):159-65. doi: 10.4056/sigs.3156511. Epub 2012 Jul 28.

### **Molecular simulation of fibronectin adsorption onto polyurethane surfaces.**

Panos M, Sen TZ, Ahunbay MG.  
Langmuir. 2012 Aug 28;28(34):12619-28. doi: 10.1021/la301546v. Epub 2012 Aug 16.

### **A rigid network of long-range contacts increases thermostability in a mutant endoglucanase.**

Rader AJ, Yennamalli RM, Harter AK, Sen TZ.  
J Biomol Struct Dyn. 2012;30(6):628-37. Epub 2012 Jun 26.

## APPENDIX: Known issues (examples; not an exhaustive list)

- Simple search box on top banner of new site must be made functional.
- Integration of links to appropriate tutorial video should be conducted across the entire MaizeGDB website.
- Updated tutorial videos (in preparation now) need to be available for the new site.
- Page load times and BLAST speed must be improved
- Some pages and features have yet to be moved over to new interface (e.g. Stock shopping cart)
- In addition to gene model record page links to metabolic pathway tools, list the pathway(s)
- Must enable researchers to download custom datasets – e.g., download the list of gene models for a specified region along with classical genes, predicted functions, motifs, metabolic pathways - with source (CornCyc, MaizeCyc), rice and Arabidopsis orthologs, insertional mutants, phenotypes, genetically mapped SNPs, Indels and SSRs; or do same from a list of gene models, with ability to pick columns of interest. Sources of data should be shown too.
- Must support diversity queries. E.g., “What SNPs are in a region are diverse for my inbreds of interest?” In output, include the B73 sequence flanking the SNP
- Update incongruency representation of genetic and B73 reference genome assembly to include denser genetic maps such as those from the Illumina MaizeSNP50 chip
- Must update some of the information about updated tracks in the Genome Browser (i.e., summaries under the ‘?’ are out of date, notably the summary of UniformMu data).
- Database is currently dependent on non-open source solutions (Oracle and VMware). The transition to OS solutions is underway and should enable eventual transition to iPlant infrastructure.
- Sorghum-B73 syntenic gene track -- display sorghum gene model names; currently display shows maize genes that have a syntelog in sorghum.
- Correction of gene assembly from the literature, e.g., nearly identical, multiple tandem repeats of *p1* in B73, see Goettel and Messing 2013. PMID 23266636
- BAC FPC contig links to Arizona are not working; likely also older links to databases that have changed.
- EC links from gene products to the enzyme nomenclature site BRENDA are not yet enabled.
- Pipeline of curated GO annotation from literature to the GO project not yet enabled.