



MaizeGDB STATUS REPORT

RECENT UPDATES, ACTIVITIES, AND NEW INITIATIVES

USDA-ARS
Project No. 3625-21000-045 (Ames, IA)
and
Project No. 3622-21000-027 (Columbia, MO)

Prepared by C. Lawrence

JANUARY 2008

Contact: C. Lawrence
USDA-ARS and Iowa State University
526 Science II
Ames, IA 50011
Email: Carolyn.Lawrence@ars.usda.gov
URL: <http://www.lawrencelab.org>
515-294-2265 (fax)
515-294-4294

TABLE OF CONTENTS

1 – Meeting Agenda and the Working Group’s Role	3
2 – Status	
Summary	4
Invited submission to the <i>International Journal of Plant Genomics</i>	5
MNL Volume 81, 2007	35
3 – Community Needs: the Allerton Report	38
4 – MaizeGDB Genome Browser	50
5 – Management Plan: The next five years <i>ARS Project Plan</i>	55
6 – Publications since last meeting	129
7 – Budget	130
8 – APPENDIX: MaizeGDB Genome Browser Survey	131

1 – Meeting Agenda and the Working Group's Role

08:30 a.m. Welcome	GO GET COFFEE IF YOU NEED IT!
09:00 a.m. Overview of past year and where we're headed	Carolyn
09:20 a.m. Data templates	Darwin
09:40 a.m. Map curation and PAG booth	Mary
10:00 a.m. Outreach and curation	Lisa
10:20 a.m. BREAK	
10:40 a.m. Development, community support, etc.	Trent
11:00 a.m. Genome Browser	Taner
11:30 a.m. 5-year plan, recap, and charge to Working Group	Carolyn
11:50 a.m. Working Group Executive Session	
12:30 p.m. Working Group summarizes for MaizeGDB Team	

The Working Group is tasked with evaluating MaizeGDB current status and recommending a course of action that should be taken to insure that the MaizeGDB project tracks the trajectory of maize research as closely as possible. The goal is to provide a 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.

2 – Summary

Since the last Working Group meeting late in 2006, the MaizeGDB team has sought to steadily improve the interfaces and data available at MaizeGDB and also to improve our own internal procedures so that we are better prepared for future growth. Outlined in the following invited submission to the *International Journal of Plant Genomics* are standard operating procedures in five broad areas:

- **Linking data information to genomic sequence information**
- **Methods of access**
- **Outreach**
- **Community support services**
- **MaizeGDB's utility for basic, translational, and applied research**

In addition, the team has (1) created a wiki to help us to communicate and document our procedures better internally, (2) begun to develop methods to take in large datasets efficiently, (3) created bulk data download mechanisms for all data types requested by cooperators, (4) submitted a grant proposal to the NSF's BD&I for centralizing maize information and (5) submitted MaizeGDB's ARS Project Plan (which includes a management plan; see p. 55) for review. The external review of the Project Plan is slated to occur any time, and we should get those reviews within the month.

Research Article

MaizeGDB: the Maize Model Organism Database for Basic, Translational, and Applied Research

Carolyn J. Lawrence,^{1,2,3,†} Lisa C. Harper,^{4,*} Mary L. Schaeffer,^{5,*} Taner Z. Sen,^{1,2,*}
Trent E. Seigfried,¹ and Darwin A. Campbell¹

¹ *USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, Iowa, USA*

² *Department of Genetics, Development and Cell Biology, Iowa State University, Ames,
Iowa, USA*

³ *Department of Agronomy, Iowa State University, Ames, Iowa, USA*

⁴ *USDA-ARS Plant Gene Expression Center, Albany, California, USA and Department of
Molecular and Biology University of California Berkeley, Berkeley, California, USA*

⁵ *USDA-ARS Plant Genetics Research Unit and Division of Plant Sciences, University of
Missouri Columbia, Columbia, Missouri, USA*

* *These authors contributed equally*

† *Communicating author*

Author email addresses: CJL – carolyn.Lawrence@ars.usda.gov, LCH –
ligule@nature.berkeley.edu, MLS – mary.schaeffer@ars.usda.gov, TZS –
taner.sen@ars.usda.gov, TES – trent.seigfried@ars.usda.gov, DAC –
darwin.campbell@ars.usda.gov

ABSTRACT

In 2001 maize became the number one production crop in the world with over 614 million tonnes produced (<http://faostat.fao.org>). Its success is due to the high productivity per acre in tandem with a wide variety of commercial uses: not only is maize an excellent source of food, feed, and fuel; its byproducts are used in the production of various commercial products. Maize's unparalleled success in agriculture stems from basic research, the outcomes of which drive breeding and product development. In order for basic, translational, and applied researchers to benefit from others' investigations, newly generated data must be made freely and easily accessible. MaizeGDB (<http://www.maizegdb.org>) is the maize research community's central repository for genetics and genomics information. The overall goals of MaizeGDB are to facilitate access to the outcomes of maize research by integrating new maize data into the database and to support the maize research community by coordinating group activities.

1. INTRODUCTION

Maize (*Zea mays* L.) is a species that encompasses the subspecies *mays* (commonly called ‘corn’ in the US) as well as the various teosintes that gave rise to modern maize. Maize is an important crop: not only is it one of the most abundant sources of food and feed for people and livestock the world over, it is also an important component of many industrial products. Maize byproducts are present in, e.g., glue, paint, insecticides, toothpaste, rubber tires, rayon, and molded plastics, among others. Maize is also currently the nation’s major source of ethanol, a major biofuel that is more environmentally friendly than gasoline and that may be a more economical fuel alternative in the long run. Although it is unlikely that ethanol production from maize directly will be sustainable long-term, maize’s suitability to serve as a model organism for developing fuelstock grasses is apparent [1]. Indeed, in addition to its value as a commodity, maize has been a premiere model organism for biological research for over 100 years. Many seminal scientific discoveries have first been shown in maize, such as the identification [2] and cloning [3] of transposable elements, the correlation between cytological and genetic crossing-over [4], and the discovery of epigenetic phenomena [5]. These exceptional characteristics of maize set this amazing plant apart: no other species serves as both a commodity and a leading model for basic research.

Today, with the accelerated generation of maize genetic and genomic information, the need for a centralized biological data repository is critical. MaizeGDB (the **Maize Genetics and genomics DataBase** [6]) is the Model Organism Database (MOD) for maize. Stored at MaizeGDB is comprehensive information on loci (genes and other genetically-defined genomic regions including QTL), variations (alleles and other sorts of polymorphisms), stocks, molecular markers and probes, sequences, gene product information, phenotypic images and descriptions,

metabolic pathway information, reference data, and contact information for maize researchers. Described in the Results and Discussion section are example workflows that could be followed by researchers to utilize the MaizeGDB resource for their research.

In addition to storing and making maize data available, the MaizeGDB team also provides services to the community of maize researchers and offers technical support for the Maize Genetics Executive Committee and the Annual Maize Genetics Conference. Also available at the MaizeGDB website as a service to the maize research community are bulletin boards for news items, information of interest to cooperators, lists of websites for projects that focus on the scientific study of maize, the Editorial Board's recommended reading list, and educational outreach items.

2. MATERIALS AND METHODS

2.1. Kinds of data in the database that link genetic and genome sequence information

MaizeGDB is the primary repository for the major genetic and cytogenetic maps and includes details about genes, mutants, QTL (quantitative trait loci), and molecular markers including 2,500 RFLPs (restriction fragment length polymorphisms), 4,625 SSRs (simple sequence repeats), 363 SNP (single nucleotide polymorphisms), 2,500 indels (insertion/deletion sites), and 10,644 overgos (**o**verlapping oligonucleotides). These data are described using 1.27 millions synonyms, 42,000 primer sequences, 16,394 raw scores from mapping based upon 16 panels of stocks, and 323,313 links to GenBank [7] accessions. GenBank accessions form the links between the genetic position on a chromosome, the sequence records at MaizeGDB, and the EST (expressed sequence tag) and GSS (genome survey sequence) contig assemblies at PlantGDB [8]

and Dana Farber (The Gene Indices at <http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=maize>, previously at TIGR [9]). All of the 3,520,247 sequences in MaizeGDB are accessible by BLAST [10] and can be filtered to report only mapped loci, including any SSRs and overgos that may not be mapped genetically, but via BACs (bacterial artificial chromosomes) in anchored contigs.

The inclusion of the public BAC FPC (Finger Print Contig) information [11] adds 439,449 BACs together with associated overgo, SSR, and RFLP markers, which are used to assemble the contigs and to link contigs onto genetic map coordinates. The order of loci on the BAC contigs is represented by over 27,000 sequenced-based loci on the IBM2 FPC057 maps (<http://www.maizegdb.org/cgi-bin/displaymapresults.cgi?term=ibm2+fpc0507>) in MaizeGDB, by links to contigs at both the Arizona FPC site (<http://www.genome.arizona.edu>) and the genome sequencing project (<http://www.maizesequence.org>). As the B73 genome sequence progresses, these BAC sequences are added to MaizeGDB along with links to the sequencing project, both from the BAC clones and from genetically mapped loci associated with a BAC.

The newest maps in MaizeGDB, IBM SNP 2007 (<http://www.maizegdb.org/cgi-bin/displaymapresults.cgi?term=ibm%20snp%202007>), are the first of a new generation of genetic maps from the Maize Diversity Project (<http://www.panzea.org>) kindly provided pre-publication by Dr. Mike McMullen. The SNP loci on these maps are associated with allelic sequences from a core set of maize and teosinte germplasm. Because the majority of the anticipated 1128 loci have been previously mapped onto BAC clones [11,12], these genetic maps tightly link sequence diversity to the B73 genome sequence.

2.2. Methods of access and database back end

The data stored at MaizeGDB are made available primarily through a series of interconnected Web pages, available at <http://www.maizegdb.org> (see Figure 1). These Web pages are dynamically generated and are written in PHP (the recursive abbreviation for PHP Hypertext Preprocessor [13]) and Perl [14]. Through this interface, each page shows detailed information on a specific biological entity (such as a gene) as well as basic information about data associated with it (genes are associated with maps, phenotypes, and citations, among others). These additional data types are linked to the gene page, enabling quick access to alternative data views. The site also includes links to related resources at other databases; genes, for example, are linked to Gramene [15].

One may access these individual data pages by using either (1) the search bar located at the top right of every page (Figure 1A), or (2) data type-specific advanced querying tools (accessible via the “Data Centers” links; Figure 1B) on the left side of the home page, or (3) the Bin Viewer tool (Figure 1C), which is located in the left margin of the home page or via a pull down labeled “Useful pages” (Figure 1D) accessible at the top of any MaizeGDB page. These tools allow researchers to easily find relevant data displays.

MaizeGDB's method of data delivery has three primary goals: placing information within the framework of its scientific meaning, making this information available to the researcher with minimal input (often only the relevant term), and requiring minimal effort from the researcher to comprehend the data displays. By focusing on biological context and ease of use as the primary focus of this interface (the “production” Web interface), the database is intended to be intuitive to the researcher as their click stream follows a logical path of biological associations.

The production Web interface, which most MaizeGDB users interact with, is only one component of the overall MaizeGDB infrastructure (Figure 2). Its data are typically updated on

the first Tuesday of each month and the Web interface only displays data that have been fully curated. Prior to being in that **Production Environment**, the raw data are prepared for public accessibility in a **Staging Environment**. In the Staging Environment, the most up-to-date information is available, new data are added to the database, and existing data are updated with new information. In addition to a Web interface that appears identical to the one in the Production Environment, the Staging Environment offers SQL (structured Query Language) read-only access to the community so that researchers interested in interacting with the data in a more direct and customized manner can have access to the most up-to-date information available.

Also available within the Staging Environment are Community Curation Tools to enable researchers to add small datasets to the database directly, as well as a set of Professional Curation Tools developed by Dr. Marty Sachs' group at the Maize Genetics Cooperation – Stock Center in Urbana-Champaign [16]. Whereas the Community Curation Tools have many safeguards to help researchers enter data step-wise and with enforced field requirements, the Professional Curation Tools allow MaizeGDB project members as well as Stock Center personnel to enter datasets in a more stream-lined and powerful fashion with fewer integrity enforcement rules (which slow down the data entry process considerably). It also should be noted that data added to the database via the Community Curation Tools are first marked as “Experimental” that must be “Activated” by professional curators at MaizeGDB. This ensures that only quality information is made publicly accessible. The availability of a Curation Web interface enables researchers to view the data as they will appear once they are uploaded to Production. If researchers wish to deposit complex or large datasets, it would not be reasonable to enter the data via the Community Curation Tools because those tools work via a “bottom-up”

approach whereby the records are (1) built based upon the most basic information included in the dataset and (2) entered one record at a time (i.e., not in bulk). For complex or large datasets, researchers are encouraged to submit data files to the curators at MaizeGDB. Those data are added to the database directly by curators and the database administrator.

To aid in the modeling of new types of data for inclusion in the MaizeGDB product and to enable programming to be tried out in a safe place, a **Test Environment** identical to the Staging Environment has been created. Note that three copies of the database exist, and that a **Disaster Recovery** system has been put in place whereby the Curation Database is backed up in a compressed format to a separate machine in Ames, Iowa daily. Once weekly, the Ames file is copied to Columbia, Missouri for off-site storage. While each environment and server has a specific purpose, all are configured such that they could serve a backup to each other. If any one server was to fail, either of the other two could provide full, unrestricted data access and site functionality. The curation database is backed up on a daily basis and is available for download (<http://goblin1.zool.iastate.edu/~oracle/>) for those who have Oracle RDBMS (Relational Database Management System) installed locally.

Each environment's server has a perpetual license and is supported by Oracle RDBMS powered by 2 x 2.0 GHz Xeon processors, 4 GB of RAM, 5 x 73 GB Ultra 320 10K RPM drives with Red Hat Advanced Server 2.1 operating system installed. The curation database, either partially or in its entirety, can be moved to MySQL, Microsoft Access, and nearly any other portable data format that a researcher would need. Requests to gain read-only SQL access to the curation database can be made via the feedback link that appears at the bottom of any MaizeGDB page. Data housed at MaizeGDB are in the public domain and are freely available for use without a license.

2.3. Outreach

One of the strengths of MaizeGDB is its responsiveness to community input, received either personally or by the feedback forms accessible at the bottom of each page (Figure 1E). To provide outreach and user support as well as to solicit input from researchers in a more active manner, several strategies are employed. The first is **tutorials and basic information** on MaizeGDB. The MaizeGDB Tutorial (Figure 1F) can be reached from the home page at the top of the left margin. A new user can go through this tutorial, and become familiar with how to use the site quickly. In addition, a "Site Tour" with an overview with examples can be found under the "Useful pages" pull down menu at the top of each page. More specific tutorial examples and other educational materials are available via the "Education" link, also within the "Useful pages" pull down menu. Also, on many of the "Data Center" pages (available from the left margin of the front page or via the "Useful pages" pull down) a discussion of the topic of the page that is suitable for the general public appears toward the bottom. Another form of outreach supported by MaizeGDB is **assistance at meetings and conferences**. Representatives from MaizeGDB attend and help researchers at the Annual Maize Genetics Conference (usually in March), the International Plant and Animal Genome Conference (January), and various other meetings through direct interaction in person. Finally, **researchers can request a MaizeGDB site visit**. About three times a year, an expert curator travels to various research locations and provides tutorials and support for maize researchers. For these visits, the local maize researchers are asked for a list of specific questions ahead of time. During the one to two day visits, researchers interact in groups and one-on-one with the traveling curator to learn how to utilize MaizeGDB for their research and to deposit data at MaizeGDB.

2.4. Community support services

MaizeGDB provides community support in several ways. Two members of the MaizeGDB team, MLS and TES, serve as *ex officio* members of the Maize Genetics Conference Steering Committee. They collect electronic abstracts for the Annual Maize Genetics Conference and handle the preparation and printing of the program for the conference. MaizeGDB personnel manage regular community surveys on behalf of the Maize Genetics Executive Committee. These surveys enable the Executive Committee to summarize the overall research interest of the maize community and to advise funding agencies on future research directions. MaizeGDB personnel also manage the Executive Committee's website (i.e., <http://www.maizegdb.org/mgec.php>) and conducts the Executive Committee's elections. MaizeGDB houses the mailing list for the annual Maize Newsletter and project personnel conduct semi-regular mailings to the maize community on behalf of interested researchers by maintaining an electronic list of researchers' contact information. Potential mailings to this list are vetted by the Executive Committee.

3. RESULTS AND DISCUSSION

To demonstrate how researchers utilize MaizeGDB, three example usage cases are presented here. Because researchers with very different goals can all utilize MaizeGDB to advance their work, the usage cases are classified by research type: basic, translational, and applied. See Figure 3 for examples of how these research types fit together. By enabling researchers to carry out workflows that support translational and applied research, MaizeGDB plays a part in influencing crop development directly. Although a single researcher might even include all of

these three aspects in his/her research simultaneously, here the researcher types are distinguished as follows: basic researchers investigate the fundamental biology of the organism, translational researchers work to determine the application of basic research outcomes for practical purposes [17], and applied researchers implement proven technologies to improve crops.

3.1. Basic

Many basic researchers work with mutants to understand the processes underlying biological phenomena. Once a new mutant is found, there are several standard methods used to elucidate normal gene functions. These efforts include determining whether the mutant represents an allele of a previously described gene, and if not, genetic mapping and cloning of the new gene. Information stored in MaizeGDB is useful in all of these steps.

In a large screen for mutations that change pericarp pigmentation from red to some other color, Researcher 1 has found a plant with a brownish-red pericarp coloration. She first wants search MaizeGDB to find all known mutants that have red pericarp phenotypes to determine whether this mutation represents a newly discovered gene. Because she does not know how others might have described the phenotype, she decides to browse existing phenotype terms and images. From the left margin of the MaizeGDB homepage, she selects "Mutant Phenotypes" under "Data Centers – Functional". On this page (<http://www.maizegdb.org/phenotype.php>), she selects "pericarp color" from the pull down menu labeled "Show only phenotypes relating to this trait" in the green search bar. A number of possible mutant phenotypes are returned, including "red pericarp". Clicking on the "red pericarp" phenotype link, she finds that the listed mutants are alleles of *p1* (*pericarp color1*; the vast majority) and *r1* (*colored1*). On this page (<http://www.maizegdb.org/cgi-bin/displayphenorecord.cgi?id=13818>), she scrolls to the bottom

and finds that there are many stocks that can be ordered from the Maize Genetics Cooperation – Stock Center that carry *Pl-rr* (an allele that causes red pericarp and red cob) or *Pl-rw* (red pericarp and white cob) as well as an *rl-cherry* stock that has red pigmentation in the pericarp. Having these stocks in hand will enable her to test whether the new mutant represents an allele of the *p1* or *rl* genes, so she decides to order a few for complementation analyses. Clicking on the stock links listed on the variation/allele page allows her access to a shopping cart utility (in the green right hand panel), and she orders seed from the Stock Center directly through the MaizeGDB interface. (Another way she could have found maize stocks that have red pericarp is the following: from the header of any page, select “Useful pages” and click “Stocks”. This pulls up the stock search page <http://www.maizegdb.org/stock.php>. In the green box, select stocks with the phenotype "red pericarp" from the pull down menu of all phenotype names and submit. A long list of stocks that contain alleles of *p1* with red pericarp and *rl-cherry* is returned. Alternatively, the Stock Center Catalog is also available from the Stocks Data Center page.)

Researcher 1 receives several appropriate stocks and performs allelism tests and determines that her mutant (which turns out to be recessive) is not allelic to *p1* or *rl*. She returns to MaizeGDB and again looks through “Mutant Phenotype” results using the "pericarp color" query. Listed there are brown pericarp, orange pericarp, white pericarp, and lacquer red pericarp phenotypes. She finds that there is no stock available for the brown pericarp phenotype (the *brown pericarp1* mutant has been lost), and all the others are alleles that confer colored pericarp in the dominant condition as a result of the presence of *Pl* alleles. To determine whether the new mutation could be an allele of *bp1*, she decides to map it genetically.

MaizeGDB houses the largest collection of publicly available genetic maps of maize (currently over 1,337 maps). These include maps of genes primarily defined by mutants with

morphological phenotypes ("Genetic 2005" is the most current), maps based on phenotypic molecular markers, and composite maps where various maps have been integrated. These maps can be easily accessed from the home page, via the left margin link to "Data Centers – Genetic – Maps" (<http://www.maizegdb.org/map.php>). This page not only allows various map search functions, but also provides information on the most popular maps and a handy reference to explain more about the various composite maps.

The maize genome is divided into genetic bins of approximately 20 centiMorgans each and that there are boundary markers with nearby SSRs that can be used for mapping (for further explanation see http://www.maizegdb.org/cgi-bin/bin_viewer.cgi). Researcher 1 decides to utilize SSRs to map her gene to bin resolution. To find the core markers from the home page, she clicks on "Tools – Bin Viewer" in the left margin of the home page. This provides a list of the core bin markers and a link to purchase relevant primers to screen her mapping population. She generates a mapping population, performs PCR experiments using the polymorphic markers, and maps her mutant to bin 9.02.

To see what genes are located in bin 9.02, she goes back to the Bin Viewer (from the homepage), and holds the cursor over the image of chromosome 9 until she sees "bin 9.02", then clicks. The result is a long list of genes, other loci, sequences, EST contigs, SSRs, BACs, and other data relating to bin 9.02. Searching through this data, she sees that *bp1* is listed under "other loci" in bin 9.02. This is a "lapsed locus" meaning that the stock has been lost, but perhaps she has found a new allele!

To see more specific genetic mapping data on *bp1*, she goes to the search bar along the top green bar of every page, selects "loci" from the pull down menu, types "bp1" into the field provided, and clicks the button marked "Go!" This brings her to the *bp1* locus page

(<http://www.maizegdb.org/cgi-bin/displaylocusrecord.cgi?id=61563>) where she can see that *bp1* is placed on three genetic maps. Clicking on each map, Researcher 1 learns that in 1935, *bp1* was mapped between *sh1* and *wx1* (*shrunk1* and *waxy1*), two well-studied genes. To search for molecular markers suitable for fine structure mapping, she visits “Data Centers – Genetic – Maps” from the link on the home page. In the green Advanced Search box, she enters *sh1* and *wx1* separately in the "Show only maps containing this locus" lines. This returns only genetic maps that contain both genes. She selects the map with the most markers – IBM2 2005 Neighbors 9 (with 2,488 markers). She finds *sh1* at position 80.30, and *wx1* at 185.00. To choose among several molecular markers, Researcher 1 follows the available links leading her to information about suitable primers, a number of variations (which can help to decide if there may be a polymorphism in her mapping populations), gel patterns, and any available GenBank accession numbers for sequences as well as sequenced BACs. She finally selects markers and performs fine structure mapping. As she finds markers closer and closer to the gene, she can proceed with positional cloning to determine whether the position is consistent with *bp1* (nice examples of how this is done can be found in [18-20]).

3.2. Translational

Research to understand the metabolic pathways that produce pigmentation (like those outlined in section 3.1) are well studied in maize [21]. One example of a well-characterized gene that confers pigmentation is *p1*, which encodes a transcription factor that regulates synthesis of flavones such as anthocyanins [22]. The *p1* gene along with its adjacent duplicate *pericarp color2* (*p2*) control pericarp and cob coloration, and cause silks to brown when cut. One flavone produced by the pathway is maysin, a compound which has been shown to be antinutritive to the

corn ear worm at concentrations above 0.2% fresh weight if husks limit access to the ear such that feeding on silks is required for the insect to enter [23]. Many QTL for resistance to corn earworm map near loci in the flavone synthesis pathway that are either regulatory genes (such as *p1* and *p2*), or at rate-limiting enzymatic steps, such as *c1* (*chalcone synthase1*) that contribute maysin accumulation in silks [24]. Understanding how maysin functions and how this information could be used for production agriculture is Researcher 2's area of expertise.

Researcher 2 has investigated maysin synthesis for some time, and has decided to clone an uncharacterized maysin QTL near *umc105a*, in the bin 9.02, which is bounded by *bz1* and *wx1* [24]. He believes that the QTL may be a previously described, but lost, *bp1* mutant thought to be involved in maysin synthesis. In the first step, he must first find molecular markers to more finely map the region (his preference would be to use SSRs, since members of the lab are already using them successfully). He plans to follow the strategy of chromosome walking to narrow down the region of interest [18-20] followed by association mapping to identify the actual QTL sequence [25,26]. Knowing this sequence would enable plant breeders to track the QTL for marker assisted selection.

To find SSR data for mapping to a bin region, Researcher 2 goes to the MaizeGDB home page and clicks on "Data Centers – Genomic – Molecular Markers/Probes" in the left margin, then clicks the "SSR" link at the top of the page (the link is located in "*Specific information is available on BACs, ESTs, overgos, and SSRs.*") Scrolling down to the green "Set Up Criteria" box, he then selects bin 9.02 and submits a search request. A report is returned that lists the available SSRs for bin 9.02, complete with primers, gel patterns for different germplasm, and related maps. By going back to the SSR page, he also downloads tabular reports of map locations of all SSRs on chromosome 9, including those that have been anchored to a BAC

contig. Using this information in the laboratory, members of his research group perform mapping experiments using several SSRs in bin 9.02 along with some others in the more distal part of bin 9.03. They discover that the mid-region peak for the QTL is very near an SSR for *bnlg1372*, which is anchored to a BAC contig.

To find sequenced BACs that may harbor the earworm resistance QTL, Researcher 2 uses the search bar at the top of each MaizeGDB page to find the locus *bnlg1372*. At the top of the *bnlg1372* page, he follows a link near the top of the page to the contig 373 display at the Maize Sequencing Project site (<http://www.maizesequence.org>). This is a rather large contig with many sequenced BACs and assigned markers. At the Maize Sequencing Project site, he uses the export function (a button at the left margin) to view a text list of all the markers and sequenced BAC clones that are available on the Finger Print Contig physical map. He finds that *bnlg1372* is assigned to the region “19742100,1974700”, encompassed by the sequenced BAC clone, c0324E10. This information provides coordinates for viewing the region on a large contig associated with *bnlg1372*, the sequence of BAC c0324E10, and any other BACs nearby. Researcher 2 sequences candidate regions in diverse germplasm and conducts association analysis using silk maysin levels as a trait. This may require other information about nearby markers, which also are accessible via MaizeGDB [27,28].

Although these investigations may require the development of further sequenced-based markers, Researcher 2 hopes that useful markers already exist and decides to explore MaizeGDB for any other sequences or primer-based markers already assigned to the region of interest including SNPs and indels. To do this from the locus page for *bnlg1372*, he clicks on the link to the most current IBM neighbors map listed, then explores the “sequence” and “primer” view versions of the map by clicking on the relevant links at the top of the page just under the map

name. The primer view shows primers associated with mapping probes along with the name of the probes – just what he needs to get going with the association mapping work.

3.3. Applied

Interested in breeding plants for organic sweet corn production, Researcher 3 has decided to use molecular markers to select for high maysin content, which would increase resistance to the corn earworm – a cause of significant damage to sweet corn [29]. Although plants could be genetically modified to carry the genes that confer high maysin levels in silks (e.g., see [30]), Researcher 3's farming clients require that their product be certified as both organic and "GMO-free". To meet the producers' needs, he has decided to pursue a marker-assisted selection program to create high maysin sweet inbred lines, which he will use to generate single-cross hybrids. To get started with the work, he searches MaizeGDB to find references, markers, and stocks for the project. Described here are the details on how he could use MaizeGDB to (1) access stocks known to have high maysin content directly and (2) locate relevant stocks based upon associated data with no prior knowledge of which stocks he wants to find. An outline of how he uses MaizeGDB to identify relevant selectable markers for tracking the various QTL associated with maysin accumulation also is described.

In the instance of looking for particular stocks, the researcher has identified GT114 as a high maysin line from [24]. Using the green search bar at the top of any MaizeGDB page, he searches "stocks" for "GT114". At that page, he sees a brief annotation stating that GT114 is a poor pollen producer and makes a note of that observation and plans to cross by IA453 and IA5125, sweet lines that produce pollen well, to ameliorate this potential difficulty. Clicking the link to GT114, he sees that it is an inbred line derived from GT-DDSA (DD Syn A) in Georgia,

and it is made available via the National Plant Germplasm Systems's Germplasm Resources Information Network (GRIN; <http://www.ars-grin.gov/>). Selecting the link for GRIN, a page opens at that site (<http://www.ars-grin.gov/cgi-bin/npgs/html/search.pl?PI+511314>). Listed there are the *Crop Science* Registration data, availability (noted as currently unavailable, but a call to Mark Millard, maize curator at the maintenance site indicates that he could access that stock in limited quantities if current resources allow), and an image of bulk kernels among other information. The image of bulked kernels is especially revealing: the kernels are yellow and the cob fragments appear red. Aware that a red cob would be unacceptable for breeding sweet corn (the red pigment could cause quite a mess for those cooking and eating corn on the cob), he decides to search MaizeGDB for other available high maysin stocks.

After a literature search of breeding stocks with a white cob that might still produce maysin in the silks, Researcher 3 starts searching stocks for those known to carry the *PI-wwb* allele, a dominant allele of the *p1* locus that confers white pericarp, white cob, and browning silks. By clicking the "Data Centers – Genetic – Stocks" link from the MaizeGDB homepage, he arrives at the Stocks Data Center page (which is also accessible via the "Useful pages" pull down at the top of every MaizeGDB page). He uses the Advanced Search box to limit the query by variation to those stocks associated with the allele *PI-wwb*. A number of the stocks returned on the results page have been evaluated for silk maysin accumulation (per associated publications) and could be further investigated as potential breeding stocks.

Although the *p1* gene accounts for much of the variability in maysin accumulation [31], association and QTL analyses for candidate genes for maysin accumulation also have identified *anthocyaninless1* (*a1*), *colorless2* (*c2*), and *white pollen1* (*whp1*) as contributing significantly [31,32]. Researcher 3 can track the dominant *PI-wwb* allele visually by selecting for browning

silks given that the sweet lines he will be using in the breeding program have silks that do not brown, but tracking the other factors will require the use of molecular markers. To find molecular markers to select for desirable alleles of, e.g., *a1*, Researcher 3 uses the search menu at the top of any page at MaizeGDB to find “loci” using the query “a1”. The results page (<http://www.maizegdb.org/cgi-bin/displaylocusresults.cgi?term=a1>) lists many loci with a1 as a substring, but shows the exact match (the *a1* locus) at the top of the list. Clicking on that link shows the *a1* locus page (<http://www.maizegdb.org/cgi-bin/displaylocusrecord.cgi?id=12000>), which lists useful information including six probes/molecular markers that could be used for tracking useful *a1* alleles. Using the same process, he also finds markers for the *c2* and *whp1* loci and sets to work determining which markers to use for his selections.

4. CONCLUSIONS

Because MaizeGDB stores and makes accessible data of use for a variety of applications, it is a resource of interest to maize researchers spanning many disciplines. The fact that basic research outcomes are tied to translational and applied data enables all researcher types to utilize the MaizeGDB resource to further their research goals, and connections to external resources like Gramene, NCBI, and GRIN make it possible for researchers to find relevant resources quickly, irrespective of storage location.

At present, maize geneticists are at the cusp of a milestone: the genome of the maize inbred B73 is being sequenced in the U.S., with anticipated completion in 2008. In addition, scientists working in Mexico at Langebio (the National Genomics for Biodiversity Laboratory) and Cinvestav (Centro de Investigacion y Estudios Avanzados) have announced through a press release (July 12, 2007) that they completely sequenced 95% of the genes with 4X coverage in a

native Mexican popcorn called palomero, though the data have not yet been released and the quality of the data is unknown (see

http://www.bloomberg.com/apps/news?pid=20601086&sid=aO.Xj8ybAExI&refer=latin_america

a). At present and as more maize sequence becomes available, relating sequences to the *existing* compendium of maize data is the primary need that must be met for maize researchers in the immediate future. Creating and conserving relationships among the data will enable researchers to ask and answer questions about the structure and function of the maize genome that previously could not be addressed. To address this need, MaizeGDB personnel will create a “genome view” by adopting and customizing a Genome Browser that could be used to integrate the outcomes of the Maize Genome Sequencing Project. For genome browser functionality, basic researchers have an interest in visualizing genome structure, gene models, functional data, and genetic variability. Translational researchers would like to be able to assign values to genomic and genetic variants (e.g., the value of a particular allele in a given population) and to view those values within a genomic context. Applied researchers are interested in tagging variants for use as selectable markers and retrieving tags for particular regions of the genome. To best meet these researchers’ needs, the “genome view” will allow researchers to visualize a gene within its genomic context and a soon to be created “pathway view” will enable the visualization of a gene product within the context of relevant metabolic pathways annotated with Plant Ontology

(<http://www.plantontology.org/>) [33] and Gene Ontology

(<http://www.geneontology.org/index.shtml>) [34] terms. By making sequence information more easily accessible and fully integrated with other data stored at MaizeGDB, it will become possible for researchers to begin to investigate how sequence relates to the architecture of the maize chromosome complement. How are the chromosomes arranged? Is it possible to relate

the genetic and cytological maps to the assembled genome sequence? Are there sequences present at centromeres that signal the cell to construct kinetochores, the machines that ensure proper chromosome segregation to occur, at the correct site? MaizeGDB aims to enable researchers to discover answers to such queries that will enhance the quality of basic maize research and ultimately the value of maize as a crop. It will become possible to interrogate the database to find answers to these and other complex questions, and the content of the genome can better be related to its function, both within the cell and to the plant as a whole.

Convergence of traditional biological investigation with the knowledge of genome content and organization is currently lacking, and is a new area of research that will open up once a complete genome sequence and a method for searching through the whole of the data are both in place. It is the ability to investigate and answer such basic research questions that will serve as the basis for devising sound methods to breed better plants. Once the relationships among sequence data and more traditional maize data like genotypes, phenotypes, stocks, etc. have been captured, it is important that those data be presented to researchers in a way that can be easily understood without requiring that they have any awareness of how the data are actually stored within a database. It is these needs – creating connections between sequence and traditional genetic data, improving the interface to those data, and determining how sequence data relate to the overall architecture of the maize chromosome complement – that the MaizeGDB team seeks to fulfill in the very near future.

ACKNOWLEDGMENTS

We are indebted to the community of maize researchers and the MaizeGDB Working Group (Drs. Volker Brendel, Ed Buckler, Karen Cone, Mike Freeling, Owen Hoekenga, Lukas Mueller, Marty Sachs, Pat Schnable, Tom Slezak, Anne Sylvester, and Doreen Ware) for their continued enthusiasm, help, and guidance. We are grateful to Dr. Bill Beavis for giving us the idea to highlight MaizeGDB's utility for the three user types. We thank Drs. Mike McMullen, Jenelle Meyer, Bill Tracy, and Tom Peterson for helpful discussions concerning *p1* and maysin research as well as Dr. Damon Lisch for suggestions on seminal discoveries in maize and Mark Millard at the USDA-ARS North Central Regional Plant Introduction Station for samples of corn with red cobs.

REFERENCES

- [1] C. J. Lawrence, V. Walbot. "Translational Genomics for Bioenergy Production from Fuelstock Grasses: Maize as the Model Species," *Plant Cell*. 2007.
- [2] B. McClintock. "The origin and behavior of mutable loci in maize," *Proc Natl Acad Sci U S A*, 36 (6):344-355. 1950.
- [3] N. Fedoroff, S. Wessler, M. Shure. "Isolation of the transposable maize controlling elements Ac and Ds," *Cell*, 35 (1):235-242. 1983.
- [4] H. B. Creighton, B. McClintock. "A Correlation of Cytological and Genetical Crossing-Over in Zea Mays," *Proc Natl Acad Sci U S A*, 17 (8):492-497. 1931.
- [5] E. H. Coe. "The Properties, Origin, and Mechanism of Conversion-Type Inheritance at the B Locus in Maize," *Genetics*, 53 (6):1035-1063. 1966.
- [6] C. J. Lawrence, M. L. Schaeffer, T. E. Seigfried, D. A. Campbell, L. C. Harper. "MaizeGDB's new data types, resources and activities," *Nucleic Acids Res*, 35 (Database issue):D895-900. 2007.
- [7] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler. "GenBank," *Nucleic Acids Res*, 35 (Database issue):D21-25. 2007.
- [8] Q. Dong, C. J. Lawrence, S. D. Schlueter, M. D. Wilkerson, S. Kurtz, C. Lushbough, V. Brendel. "Comparative plant genomics resources at PlantGDB," *Plant Physiol*, 139 (2):610-618. 2005.
- [9] J. Quackenbush, F. Liang, I. Holt, G. Pertea, J. Upton. "The TIGR gene indices: reconstruction and representation of expressed gene sequences," *Nucleic Acids Res*, 28 (1):141-145. 2000.

- [10] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res*, 25 (17):3389-3402. 1997.
- [11] F. Wei, E. Coe, W. Nelson, A. K. Bharti, F. Engler, E. Butler, H. Kim, J. L. Goicoechea, M. Chen, S. Lee, G. Fuks, H. Sanchez-Villeda, S. Schroeder, Z. Fang, M. McMullen, G. Davis, J. E. Bowers, A. H. Paterson, M. Schaeffer, J. Gardiner, K. Cone, J. Messing, C. Soderlund, R. A. Wing. "Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History," *PLoS Genet*, 3 (7):e123. 2007.
- [12] J. Gardiner, S. Schroeder, M. L. Polacco, H. Sanchez-Villeda, Z. Fang, M. Morgante, T. Landewe, K. Fengler, F. Useche, M. Hanafey, S. Tingey, H. Chou, R. Wing, C. Soderlund, E. H. Coe, Jr. "Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization," *Plant Physiol*, 134 (4):1317-1326. 2004.
- [13] R. Lerdorf, P. MacIntyre, K. Tatroe, Safari Tech Books Online. Programming PHP. Sebastopol, Calif.: O'Reilly, 2006.
- [14] L. Wall, T. Christiansen, J. Orwant. Programming Perl. Beijing ; Cambridge [Mass.]: O'Reilly, 2000.
- [15] D. H. Ware, P. Jaiswal, J. Ni, I. V. Yap, X. Pan, K. Y. Clark, L. Teytelman, S. C. Schmidt, W. Zhao, K. Chang, S. Cartinhour, L. D. Stein, S. R. McCouch. "Gramene, a tool for grass genomics," *Plant Physiol*, 130 (4):1606-1613. 2002.
- [16] R. Scholl, M. M. Sachs, D. Ware. "Maintaining collections of mutants for plant functional genomics," *Methods Mol Biol*, 236:311-326. 2003.

- [17] S. Carpenter. "Science careers. Carving a career in translational research," *Science*, 317 (5840):966-967. 2007.
- [18] E. Bortiri, G. Chuck, E. Vollbrecht, T. Rocheford, R. Martienssen, S. Hake. "ramosa2 encodes a LATERAL ORGAN BOUNDARY domain protein that determines the fate of stem cells in branch meristems of maize," *Plant Cell*, 18 (3):574-585. 2006.
- [19] E. Bortiri, D. Jackson, S. Hake. "Advances in maize genomics: the emergence of positional cloning," *Curr Opin Plant Biol*, 9 (2):164-171. 2006.
- [20] H. Wang, T. Nussbaum-Wagler, B. Li, Q. Zhao, Y. Vigouroux, M. Faller, K. Bomblies, L. Lukens, J. F. Doebley. "The origin of the naked grains of maize," *Nature*, 436 (7051):714-719. 2005.
- [21] E. Coe, M. G. Neuffer, D. Hosington. "The genetics of corn", In: Sprague G. F., Dudley J. W., editors. *Corn and Corn Improvement*. Madison, WI, pp. 81-258.
- [22] E. Grotewold, B. J. Drummond, B. Bowen, T. Peterson. "The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset," *Cell*, 76 (3):543-553. 1994.
- [23] B. R. Wiseman, M. E. Snook, D. J. Isenhour. "Maysin content and growth of corn earworm larvae on silks from first and second ears of corn," *J. Econ. Entomol.*, 86:939-944. 1993.
- [24] P. F. Byrne, M. D. McMullen, M. E. Snook, T. A. Musket, J. M. Theuri, N. W. Widstrom, B. R. Wiseman, E. H. Coe. "Quantitative trait loci and metabolic pathways: genetic control of the concentration of maysin, a corn earworm resistance factor, in maize silks," *Proc Natl Acad Sci U S A*, 93 (17):8820-8825. 1996.

- [25] J. M. Thornsberry, M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, E. S. t. Buckler. "Dwarf8 polymorphisms associate with variation in flowering time," *Nat Genet*, 28 (3):286-289. 2001.
- [26] M. Yano, T. Sasaki. "Genetic and molecular dissection of quantitative traits in rice," *Plant Mol Biol*, 35 (1-2):145-153. 1997.
- [27] S. A. Flint-Garcia, A. C. Thuillet, J. Yu, G. Pressoir, S. M. Romero, S. E. Mitchell, J. Doebley, S. Kresovich, M. M. Goodman, E. S. Buckler. "Maize association population: a high-resolution platform for quantitative trait locus dissection," *Plant J*, 44 (6):1054-1064. 2005.
- [28] S. Salvi, G. Sponza, M. Morgante, D. Tomes, X. Niu, K. A. Fengler, R. Meeley, E. V. Ananiev, S. Svtashev, E. Bruggemann, B. Li, C. F. Hainey, S. Radovic, G. Zaina, J. A. Rafalski, S. V. Tingey, G. H. Miao, R. L. Phillips, R. Tuberosa. "Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize," *Proc Natl Acad Sci U S A*, 104 (27):11376-11381. 2007.
- [29] W. F. Tracy. "Sweet Corn", In: Hallauer A. R., editor. *Specialty Corns, 2nd Edition*: CRC Press, pp. 155-198, 2001.
- [30] E. T. Johnson, M. A. Berhow, P. F. Dowd. "Expression of a maize Myb transcription factor driven by a putative silk-specific promoter significantly enhances resistance to *Helicoverpa zea* in transgenic maize," *J Agric Food Chem*, 55 (8):2998-3003. 2007.
- [31] J. D. Meyer, M. E. Snook, K. E. Houchins, B. G. Rector, N. W. Widstrom, M. D. McMullen. "Quantitative trait loci for maysin synthesis in maize (*Zea mays* L.) lines selected for high silk maysin content," *Theor Appl Genet*, 115 (1):119-128. 2007.

- [32] S. J. Szalma, E. S. t. Buckler, M. E. Snook, M. D. McMullen. "Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks," *Theor Appl Genet*, 110 (7):1324-1333. 2005.
- [33] K. Ilic, E. A. Kellogg, P. Jaiswal, F. Zapata, P. F. Stevens, L. P. Vincent, S. Avraham, L. Reiser, A. Pujar, M. M. Sachs, N. T. Whitman, S. R. McCouch, M. L. Schaeffer, D. H. Ware, L. D. Stein, S. Y. Rhee. "The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant," *Plant Physiol*, 143 (2):587-599. 2007.
- [34] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, 25 (1):25-29. 2000.

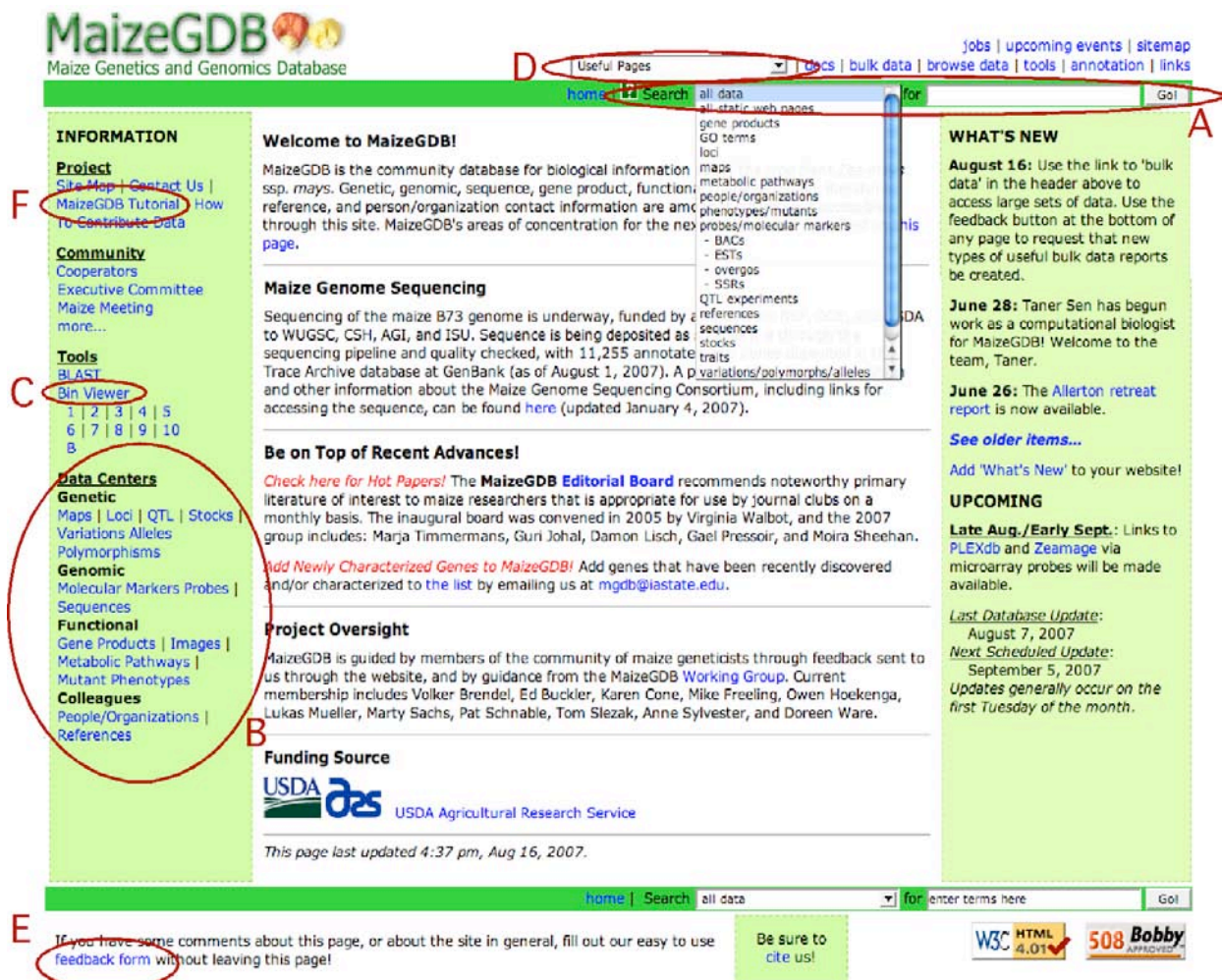


FIGURE 3: The MaizeGDB home page. The most commonly utilized search functionality for MaizeGDB is the search bar (A), which is available within the header of any MaizeGDB page. To browse data and to search specific data types using specific limiters, the Data Centers (B) are also quite useful. Also available is a Bin Viewer (C), which allows for a view of lots of data types within the context of their chromosomal location. To enable access to the Data Centers and other displays of interest from any MaizeGDB page, a pull-down menu for “Useful pages” (D) is accessible on the header of any MaizeGDB page. The footer of all MaizeGDB pages contains a context-sensitive “feedback form” link (E). Researchers use the feedback form to report errors, ask questions, and to contact the MaizeGDB team directly. For newcomers to the site, the MaizeGDB Tutorial (F) can help them to get a jump start on how to use the site.

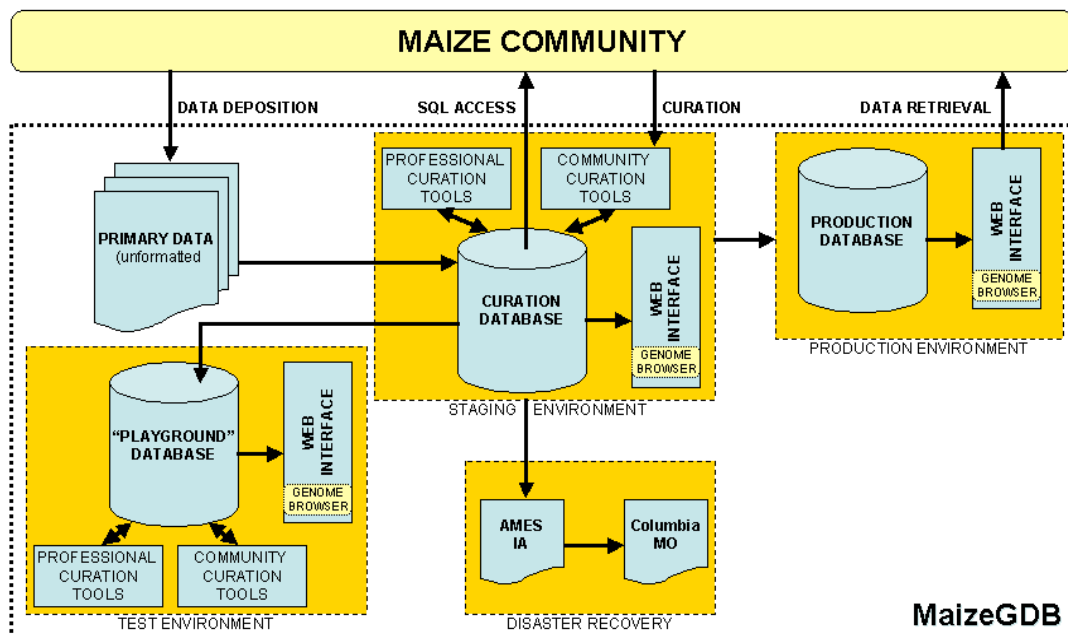


FIGURE 2: Simplified infrastructure of MaizeGDB. The community of maize researchers can add data to the database (downward-facing arrows from the uppermost yellow box) via direct data deposition (upper left) and via a set of Community Curation Tools that interacts with the Curation Database (upper center). Researchers are also allowed access to maize data (upward-facing arrows from the lower dashed box) via a Web interface that can be accessed at <http://www.maizegdb.org> (upper right) and by way of SQL access to the Curation Database, which houses the most up-to-date data available (upper center). These functionalities are supported by two of the three environments: Production and Staging, respectively (upper dashed orange boxes). Available for use by MaizeGDB personnel to facilitate data modeling and trial programming manipulations is a third environment called Test (lower left dashed orange box), which is identical to the Staging Environment. To ensure that the most up-to-date copy of the database is backed up, a Disaster Recovery process has been instituted (lower center dashed orange box) whereby a compressed copy of the database is backed up to a separate machine in Ames, Iowa daily, and to a server in Columbia, Missouri weekly.

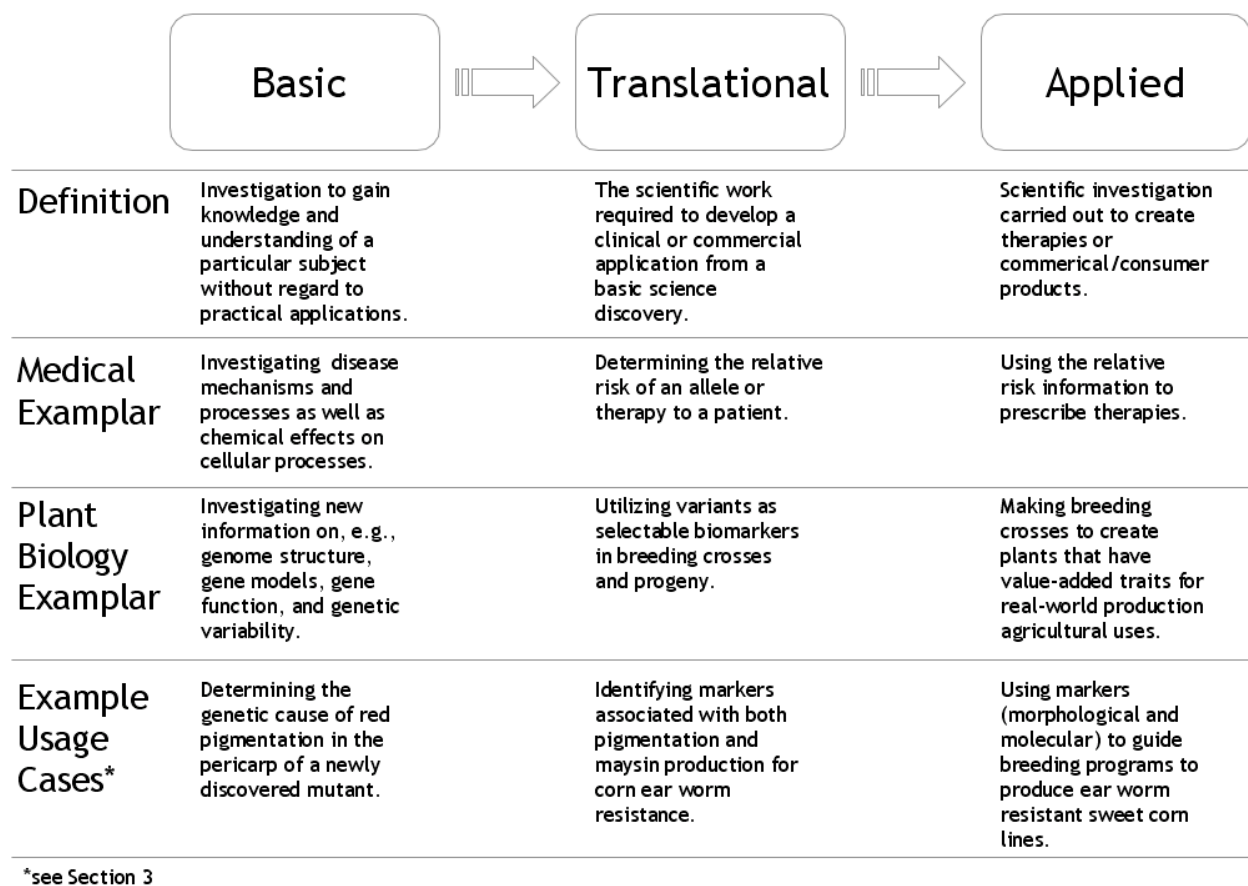


FIGURE 3: Three types of biological research. Research can be divided into three categories: Basic, Translational, and Applied. Outcomes from basic research feed into translational predictions, and developed uses for these findings constitute the basis for developing real-world applications that benefit humanity and the world. Listed below the flow of research are definitions for each research type as well as a medical and plant biological models for how the different divisions are interrelated. Also shown are overviews of the example usage cases presented in Section 3.

V. MAIZE GENETICS AND GENOMICS DATABASE (www.maizegdb.org)

New Personnel in 2007

Lisa Harper USDA-ARS Plant Gene Expression Center, Albany, CA

Feb 2007, Part-time Curator and Outreach Coordinator

In her first year on staff, Lisa plans to visit 3 cooperator sites: University of Florida, University of Georgia, and University of Arizona. One of her first curation tasks is to better integrate data from the RescueMu and Maize Inflorescence Architecture Projects with the rest of MaizeGDB so that these datasets can be searched via the site's integrated mechanisms.

Taner Sen USDA-ARS at Iowa State University, Ames, IA

To begin June 2007, Computational Biologist

Early on, Taner will be working to incorporate a genome browser into MaizeGDB to display the B73 sequence and to serve as a basis for representing gene models. Be on the lookout for inquiries from Taner on your preferences for genome browsing capabilities!

Data Improvements

MaizeGDB has added and facilitated the addition of a wide variety of new data, along with incrementally improving the existing data through regular manual and automated updating. Some of our most noteworthy newer initiatives in this area are described below.

Sequence Pipeline

Public sequence data for all of the *Zea* species are updated from Volker Brendel's PlantGDB on a monthly basis and linked with relevant manually-curated data within MaizeGDB. Individual sequences are also linked to contigs generated by external projects that include PlantGDB and the Dana Farber Cancer Institute. The Maize Genome Sequencing Consortium's B73 sequences are associated to BACs on a monthly basis from the data releases posted at maizesequence.org.

Editorial Board

We have initiated and currently maintain an Editorial Board whose members contribute a paper each month to be highlighted at MaizeGDB. Perhaps most exciting are reports that the Editorial Board has directly led to the founding of journal clubs on various campuses! Students and faculty alike download the recommended papers and meet to discuss them. The 2006 Editorial Board was made up of: Tom Brutnell (chair), Surinder Chopra, Karen McGinnis, Wojtek Pawlowski, and Jianming Yu. The 2007 Board consists of: Marja Timmermans (chair), Guri Johal, Damon Lisch, Gael Pressoir, and Moira Sheehan.

Data Additions – Larger Sets

TILLING: We have worked extensively with Cliff Weil's team at Purdue to include the output of the Maize TILLING project in MaizeGDB. This includes integrated primer, probe, locus, variation, and gene product data, along with an integrated interface for ordering stocks from the TILLING project. The current schedule (see http://www.maizegdb.org/data_schedule.php) is to update TILLING data twice yearly.

New maps: The Maize Mapping Project and a number of community members have volunteered a number of new maps for inclusion in MaizeGDB. These include new QTL maps, continued refinements of the IBM and IBM Neighbors maps, and maps that describe the structure of the AGI physical maps. The current schedule is to update maps once each spring.

Contributing your data to MaizeGDB

You may contribute data in a number of ways to MaizeGDB. The easiest is very like a 'wiki', where you simply add a comment using the annotation tool. You will first need to register, using the menu item 'annotation' on the top menu bar of the homepage. Once registered, every time you access MaizeGDB, you will be able to annotate any page. Annotations will appear in the monthly updates of the database. A second way is to use the community curation tools. Inquire at mgdb@iastate.edu for access.


If you are developing a project that will generate large datasets and that you would like to submit to MaizeGDB, you need to contact Carolyn Lawrence before you submit the proposal.

New Tools

We have continued our commitment to providing a consistent and clean interface, continued maintenance and improvement of that interface, and integration of new interface options where appropriate. Some noteworthy changes include new map displays and a stand-alone tool to compare cytological and genetic maps.

Map Display Update: One major interface addition is the inclusion of new map displays designed with the aid of commentary from a number of maize community members. We have added three new options that enable interesting new ways of viewing maps without

Figure 1. This is a map view of UMC 98, arrived at by clicking on UMC 98 on the *tub1* locus record. There are two things to note here. First, right below the name of the map there is a line with "summary view" in bold and links to "sequence view" "primer view" and "score view."


[summary view](#) | [sequence view](#) | [primer view](#) | [score view](#)

[Jobs](#) | [upcoming events](#) | [sitemap](#)

[Maize Genetics and Genomics Database](#)

 | [reports](#) | [browse data](#) | [tools](#) | [annotation](#) | [links](#)

[Home](#) | [Search](#)

UMC 98 1 w/ Primers

[summary view](#) | [sequence view](#) | [primer view](#) | [score view](#)

Download this data in a tab-delimited text file!

Primer Type	Primer	Probe	Locus	Map Coordinate
N/A	N/A	p-csu804	csu804b(dnp)	0
N/A	N/A	p-rpp c654	rgpc654	6.4
N/A	N/A	p-csu738	csu738	11
N/A	N/A	p-tub1	tub1	11
N/A	ACTTGCCTTGGCTGCCCTTAC	p-phi056	tub1	11
N/A	CGGACACGACTTCCGAGAA	p-phi056	tub1	11
N/A	TGCTTCACATTCAGTCCACGTCAG	p-phi097	tub1	11
N/A	CCACGACGATGATTAACGACG	p-phi097	tub1	11
Left End	ACGTGGATCAGATGGAGTTCACTG	CL2242_8_ov	tub1	11
Right End	ATATTGCTTCACGCTCAAGTGAAC	CL2242_8_ov	tub1	11
N/A	N/A	CL2242_8	tub1	11
Left End	CGCGAGGGTTTTCCGACATCAAGAC	p-umc94	umc94a	11.9
Right End	AGCGGATACCAATTTCCACAGGA	p-umc94	umc94a	11.9
N/A	N/A	p-bn18.05	bn18.05a	12
N/A	N/A	p-csu589	csu589	12
Left End	GTCCTGGGTTTTTGGCTCCCTATTT	csu589_PCR	csu589	12
Right End	CTCGATAAAGACGCAATTTGTGTC	csu589_PCR	csu589	12
N/A	N/A	p-csu896	fus6	12
Left End	AGGCGGCGAATCTTACACATCTCC	CL32446_1_ov	fus6	12
Right End	AAGCAAGCGGACCAAGGAAGATG	CL32446_1_ov	fus6	12

Figure 3. Clicking on the “primer view” link takes you to a map view that has four columns: primer, probe, locus, and coordinate. This table identifies probes that detect each locus on the map and also notes those that have primers available.

Morgan2McClintock: The Morgan2McClintock Translator was developed through our continued collaboration with Hank Bass and a new collaboration with Lorrie Anderson. The tool utilizes the maize Recombination Nodule map (Anderson et al., 2003 and 2004) to calculate approximate cytological positions for loci given a genetic map, and to calculate approximate genetic positions for loci given a cytological map (Lawrence et al., 2006). Morgan2McClintock is a stand-alone tool and can be run on any machine enabled to serve PHP. You can use it online at MaizeGDB: from the home page, choose "maps", then choose Recombination Nodule Map to arrive at <http://www.maizegdb.org/RNmaps.php>). Alternatively, go to: <http://www.lawrencelab.org/Morgan2McClintock>.

Maize Community Support

The MaizeGDB team offers support to the maize community in a variety of fashions. This support aids the annual Maize Genetics Conference, provides community addresses for mailings, an abstract submission interface, assembly and printing of the program, and integrates the abstracts into MaizeGDB. It supplies address lists for this Newsletter, and hosts the Newsletter, with links to the database. We facilitate community interaction with the Maize Genetics Executive Committee, including community surveys, elections and community-wide messaging on important issues. We also maintain a community job board (which has had dozens of job postings and has significantly aided at least ten job placements since its initiation), as well as a community calendar of upcoming events that may be of interest to the larger community.

Copies and Schema of MaizeGDB

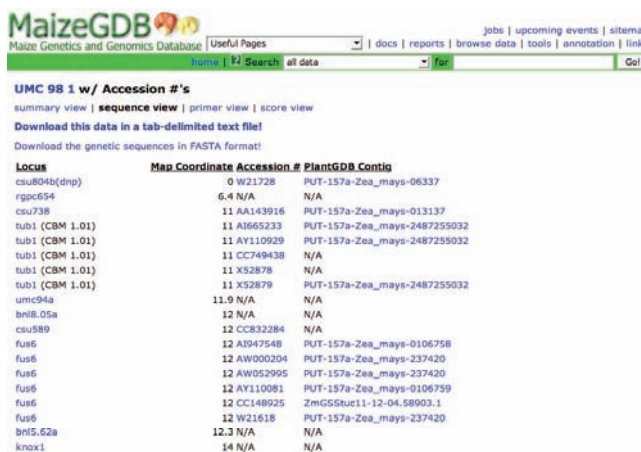


Figure 2. If you click on the “sequence view,” you are shown columns for: the locus name, the map coordinate, an accession number, and a PlantGDB contig.

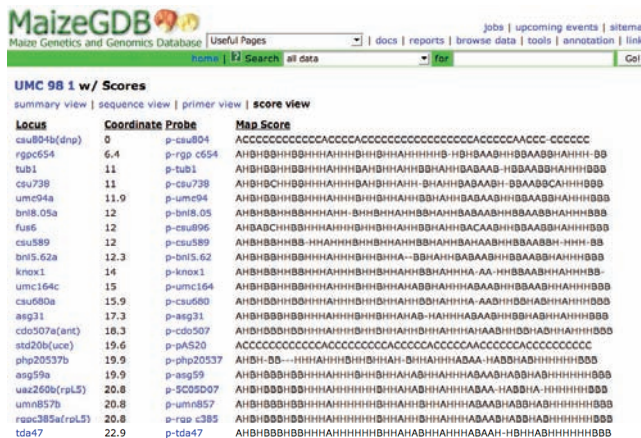


Figure 4. Clicking on “score view” allows you to see the markers used to generate a particular map along with associated map scores, enabling you to review the raw mapping data for the experiment. Note that not all maps in MaizeGDB have associated scores; if you see the “score view” option, you’re in luck!

Full copies of the database as well as individual tables and custom-formatted dumps are provided to individuals who make requests to the MaizeGDB team at mgdb@iastate.edu. Copies support Oracle, MySQL, and Microsoft Access. The current MaizeGDB schema can be accessed at <http://www.maizegdb.org/MaizeGDBSchema.pdf>.

Five-Year Plan

We are in the process of drafting our five-year plan for the USDA-ARS. Objectives were developed with input from the MaizeGDB Working Group and are available online at <http://www.maizegdb.org/objectives.php>.

Acknowledgements

MaizeGDB is guided by members of the community of maize geneticists through feedback sent to us through the website, and by guidance from the MaizeGDB Working Group. Current membership includes Volker Brendel, Ed Buckler, Karen Cone, Mike Freeling, Owen Hoekenga, Lukas Mueller, Marty Sachs, Pat Schnable, Tom Slezak (chair), Anne Sylvester, and Doreen Ware.

Citing MaizeGDB

MaizeGDB may be cited using any or all of these references:

Lawrence CJ, Schaeffer ML, Seigfried TE, Campbell DA, Harper LC, 2007. MaizeGDB's new data types, resources and activities. *Nucleic Acids Res.* 35:D895-900.

Lawrence CJ, Seigfried TE, Brendel V, 2005. The maize genetics and genomics database. The community resource for access to diverse maize data. *Plant Physiol.* 138:55-58.

Lawrence CJ, Dong Q, Polacco ML, Seigfried TE, Brendel V, 2004. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.* 32: D393-397.

Submitted by the MaizeGDB team May 8, 2007
Mary Schaeffer
Lisa Harper
Trent Seigfried
Darwin Campbell
Carolyn Lawrence

***The Future of Maize Genetics
Planning for the Sequenced Genome Era***

**A Maize Genetics Community Retreat
Allerton Park and Conference Center
March 20 - 22, 2007**

***Support provided by the USDA-CSREES-NRI-Plant Genome and the NSF Plant Genome
Research Programs and the University of Illinois.***

Executive Summary	page 2
Detailed Report	3
Introduction	3
Challenging Unanswered Questions	3
Current Resources and Future Needs	4
<i>Sequencing and annotation</i>	5
<i>Centralized databases</i>	5
<i>Transformation improvements</i>	6
<i>Reverse genetics resources</i>	7
<i>Expression profiling</i>	8
<i>Phenomics</i>	9
<i>Community membership</i>	10
Summary and Next Steps	12

Executive Summary: *The Future of Maize Genetics - Planning for the Sequenced Genome Era* A Maize Genetics Community Retreat at Allerton, March 20 - 22, 2007

Leaders in the maize community met for a two-day retreat to discuss the strengths, challenges, and initiatives that define the future of maize research. To guide strategic planning, the community first identified key questions in biology that can be best answered using maize as a model system. These biological questions were considered relative to the overarching goal of understanding the genetic basis of traits in maize. The research directions defined below and the plans to achieve the goals serve as the foundation for basic research and provide the tools for improving food, fuel, and crop yields in a changing environment.

Key biological issues define our research goals and directions:

- Maize is the pre-eminent model for studying genome evolution and trait variation due to its unsurpassed natural diversity, genome duplication history and range of adaptations.
- Because adaptation is critical to agriculture, maize research will continue to be a model for understanding the basis of genetic interactions with the environment.
- Study of maize heterosis will provide key information about how genes and alleles interact.
- Maize will continue to lead in the area of epigenetics. Imprinting, paramutation and transposons were discovered in maize and are readily studied with color markers.
- Maize is positioned as a leading model for developing cellulosic biofuels for the future.
- Maize is a model for the study of development and physiology of unique traits such as C4 photosynthesis, a persistent endosperm, inflorescence structure, etc.
- Maize cytogenetics is highly advanced and continues to provide tools for understanding mechanisms of meiosis and for developing the potential of chromosome manipulation.

Community resources will help achieve our research goals:

To advance these research areas, community resources must be created or strengthened. Short and long term planning will help leverage the sequenced maize genome and will position maize as a model for tool development and for hypothesis-driven and translational research.

Short Term Planning (Expect substantial progress in the next one to three years):

1. Current sequencing/annotation will be completed and additional map-based sequencing efforts initiated.
2. Centralized databases with increased funding are needed now.
3. Indexed reverse genetic resources need to be finalized and will accelerate many areas of research. Current mutagenesis libraries should be indexed with new technologies.
4. Expression platforms and informatic tools should be selected and developed.
5. Transformation capacity should be enhanced in the public sector. Continuous support mechanisms for public transformation need to be resolved.

Long Term Planning (Start now, with sustained efforts over the next decade)

1. Databases and stock center capacity will be enhanced, coordinated and supported.
2. Map-based sequences of other inbreds, races, and teosintes will be available.
3. A phenomics initiative will be underway, with large scale and multi-dimensional phenotyping capabilities for multiple environments available for the entire community.
4. The maize scientific community will be large, diverse, well-trained, and interactive.

Detailed Report: *The Future of Maize Genetics - Planning for the Sequenced Genome Era*, A Maize Genetics Community Retreat at Allerton, March 20 - 22, 2007

Introduction

The maize genetics community held a two-day retreat to discuss the future of maize research in the post-genomics era. The meeting included principal investigators representing approximately 60 labs from universities, colleges, industry, and USDA-ARS. Participants were primarily from the U.S., but representatives from the United Kingdom, France, and Germany were also present. Staff from ARS, NRI, NSF, DOE, and NCGA participated in discussion on the second day of the retreat.

The goal of the meeting was to develop a strategic plan for the future of maize genetics research. To guide strategic planning, the community first identified important questions in biology that can be best answered using maize as a model system. These biological questions were considered within the context of an overarching goal to understand the genetic basis of traits in maize -- traits that are the foundation for improving food, fuel, and fiber crop yields in a changing environment. Second, the community considered the current research capacity for answering these biological questions and also explored how to translate the answers to practical outcomes. It was noted that recent research accomplishments have opened many new avenues of investigation. For example, advances in genetic mapping technology have enhanced functional genomics so that gene functions can be discovered using a single population of plants and inexpensive sequencing technology. This groundwork will allow rapid establishment of productive genomics programs in related agronomically important species and will advance research in biofuels. Following the plan presented here, maize researchers will be able to accelerate the rate of discovery considerably, delivering an expanding knowledge base for the needs of breeders, for biotechnology industry and for continued basic research.

Challenging unanswered questions in biology best addressed by research on maize

Discoveries in the current genomic era of biology have generated new questions and enabled new approaches to long-standing questions. The maize community discussed these issues in broad terms and then focused on the subset of questions that could clearly be addressed best in maize due to its unique development and biology, its genetic and evolutionary history and its genome architecture.

- How is genomic diversity maintained, and how does it change during evolution?
- What is the underlying molecular genetic basis for specific traits in a species?
- Can we use maize to predict what genes will regulate plant growth in related species?
- Can natural variation provide information to develop novel breeding traits?
- What drives genome evolution, and how are these processes impacted by interaction with the environment?
- What is the genetic, molecular and physiological basis of hybrid vigor (heterosis)?
- What are the impacts of sequence-independent inheritance (epigenetics) on the growth development and evolution of maize?
- How does cytogenetic variation impact genome architecture, agronomic traits, and plant breeding efforts?

Maize is particularly useful to answer questions about genome evolution, genetic diversity and selection because allelic variation in maize is greater than in any other eukaryote. Also, due to its duplication history, maize is a model monocot for understanding evolutionary mechanisms that surround genome duplication events. Gene content and order varies considerably between maize lines, reflecting local transposon-mediated rearrangements and tandem duplications. This great genetic diversity translates into differences in phenotype and variation in how maize interacts with the environment. The genetic diversity also provides a rich toolset for the study of heterosis. With the ability to compare maize alleles with those of wild *Zea* accessions (teosintes), maize also provides an excellent species to study selection. An understanding of selection will allow researchers to harness existing diversity for advancing biological understanding and crop improvement.

Maize is well suited to study epigenetics because of the accessible phenomena associated with transposition, imprinting, and paramutation, three processes that were first identified in maize. The study of epigenetics is facilitated by the separate male and female flowers, which simplifies the process of conducting controlled pollinations. Maize has a rich collection of active transposons in the genome and color markers that are simple to score. The large size of the plant allows researchers to sample specific tissues at distinct time points from a single individual. Furthermore, epigenetic events confer heritable phenotypes, and can therefore provide direct information for crop improvement, placing maize at the forefront of translating basic research for the agronomic community.

Maize is a genetic model for other grasses with its rich collection of mutants, genetic diversity and ease of moving between phenotype and genotype. Information from maize can be easily translated to other important, less tractable members of the grass family. For example, maize is a member of the Andropogoneae, and thus is closely related to other energy crops such as *Miscanthus*, switchgrass, sorghum, and sugarcane. Knowledge of cell wall synthesis and degradation can be obtained in maize and then transferred to these potential crops for which few genetic resources are available.

Arabidopsis thaliana has been a model for understanding principles behind growth and development, but some key developmental, cellular, and physiological processes do not occur in *Arabidopsis*. Biologically and economically important features such as C4 photosynthesis, a persistent endosperm, phase dependent epidermal differentiation, complex inflorescence structure, and sex determination are best studied in maize.

Finally, maize has been central to research in cytogenetics and continues to provide cutting edge information about genome architecture. Tools developed cytogenetically will be useful for future chromosome manipulations, which can benefit both basic and applied research. The combination of easily analyzed chromosomes, meiotic mutants, well-studied segregation phenomena, and increasingly sophisticated cytogenetic tools continues to position maize as a model system for this area of research.

Current Tools, Resources and Future Needs

A major goal of the Allerton Retreat was to assess the current state of research capabilities as the maize B73 genome becomes available. It is clear that fully sequenced genomes have revolutionized the corresponding research communities. With long term planning, we can learn from these past experiences and develop the tools and capacity to optimize and fully leverage the value of a sequenced maize genome. Topics that were

considered most imperative to achieve this goal are summarized here and represent the starting point for further discussion.

Annotation of a fully sequenced B73 genome and additional genome sequencing is the foundation for future research

The maize B73 genome is currently being sequenced using a minimal tiling BAC approach with full display at <http://www.maizesequence.org> (a project site that will exist during the sequencing project's funding period). The sequenced genome promises to revolutionize maize research. With annotation, it will be the foundation upon which complementary resources and activities such as reverse genetics and phenomics will be built. To achieve this promise, a fully annotated, accessible, and centralized sequence database will be essential because all additional resources depend upon robust integration of sequence information. The sequence project site must be transitioned into a community-based permanent platform that will have robust long-term support. The community expressed the desire that **MaizeGDB should become the centralized sequence resource soon after the genome is complete (2009-2010).**

The maize community recognizes that nearly all relevant questions posed here will need sequence information beyond the annotated B73 genome. **Genomic sequence from additional maize lines is essential to advance crop improvements and to exploit maize for its unparalleled strength as a model system and as a fuel, food and fiber resource worldwide.** Developing map-based sequence information of additional genomes was identified as a priority due to the unique duplication history of the maize genome and due to the exceptional haplotype variability among inbred lines. A physical map from a second inbred line, and ultimately multiple lines, was considered important. Post-Allerton follow-up discussion will allow for continued assessment about the most efficient and effective way to accomplish synthesis of sequence information from multiple genomes.

Databases need to be centralized

Continued assessment and coordination of data deposition is essential to all advances in maize research. Currently, various types of plant database resources exist and are utilized by maize researchers, including Model Organism Databases (MODs), Clade Oriented Databases (CODs), Automatic Annotation Shops (AA), Static Repositories, and Laboratory Information Management Systems (LIMS; a category that includes coordinated project databases). A (non-exhaustive) list of databases used by maize researchers includes:

Resource Type	Funding Agency	Resource	Website
MOD	USDA-ARS	MaizeGDB*	http://www.maizegdb.org
MOD	NSF	TAIR	http://www.arabidopsis.org/
COD	NSF, USDA-ARS	Gramene	http://www.gramene.org
COD	USDA-ARS	GrainGenes	http://wheat.pw.usda.gov
COD/AA/LIMS	NSF	PlantGDB	http://www.plantgdb.org
COD/AA/LIMS	NSF, USDA-ARS	PLEXdb	http://plexdb.org
AA	NSF	TIGR	http://www.tigr.org/
AA/LIMS	NSF/USDA/DOE	MGSC's Maize Genome Browser*	http://www.maizesequence.org
AA/LIMS	NSF	MAGI*	http://www.plantgenomics.iastate.edu/maize/
AA/LIMS	NSF	FPC-maize*	http://www.genome.arizona.edu/fpc/maize/

Static	NIH	NCBI	http://www.ncbi.nlm.nih.gov/
Static	NIH	UniProt	http://www.pir.uniprot.org/
Static/LIMS	USDA-ARS	GRIN	http://www.ars-grin.gov/
LIMS	NSF	Panzea*	http://www.panzea.org/
LIMS	NSF	ChromDB	http://www.chromdb.org/

* indicates a maize-specific resource

MaizeGDB is of particular interest because it is the MOD for maize. The MaizeGDB website serves biological information about the crop plant *Zea mays* ssp. *mays*. Genetic, genomic, sequence, gene product, functional characterization, literature reference, and person/organization contact information are among the datatypes accessible through MaizeGDB. Based upon community evaluation and input, MaizeGDB will continue to focus on the following areas of concentration over the next five years: 1) integration of new maize genetic and genomic data into the database, including expansion of phenotype data and tools, 2) expansion of structural and genetic map sets, 3) access to gene models calculated by leading gene structure prediction groups through the MaizeGDB interface, and 4) support of community services such as coordinating the Maize Meeting, MGEC Elections, Polls, etc.

Long-term support for MaizeGDB from USDA-ARS was recognized; however, **new and creative funding mechanisms are required now to provide sufficient resources to exploit a fully sequenced genome**. The fact that maize will become a model genome for other complex grass genomes necessitates even more careful planning. Coordination with Gramene is critical to success. With additional resources, MaizeGDB should be able to integrate project data from diverse studies, keep gene function data current, oversee community curation, and carry out gene and plant ontology as well as metabolic pathway curation. To improve access to maize sequence data, resources that integrate various gene models and annotation sets must be made available to MaizeGDB. Complex datasets from federally funded projects should be deposited into MaizeGDB. However, collaborations should be established between MaizeGDB and the researchers who develop these complex datasets to insure efficient and cost-effective data flow directly into MaizeGDB. For the other database resources listed in the above table, recommendations from this group are in agreement with those cited by the Plant Database Working Group (see <http://www.maizegdb.org/PDBNeeds.pdf>).

Transformation technology needs to be advanced and costs reduced

Improved maize transformation resources remain one of the highest priorities for the community. A sequenced maize genome will continue to drive research hypotheses that require direct testing in transgenic plants. Furthermore, transformation capabilities will bridge the gap between basic and applied research. To achieve these goals, **several critical needs were identified: an increased capacity for public sector maize transformation, improved transformation of diverse lines and reduced transformation costs**. Improved regulatory transfer would also facilitate progress and communication among researchers. The community recommends cohesive action to evaluate ways to improve regulatory compliance within current and changing Federal guidelines.

Public researchers currently produce transgenic maize primarily by outsourcing to the Plant Transformation Facility (PTF) at Iowa State University or through facilities at their own institutions. These centers constitute a critical and reliable resource. The success and demand on the PTF clearly validates the ever-increasing need for maize transformation in

the public sector. Costs remain higher than industry, however, reflecting both industry technologies that are unavailable to the public and differences in production scale. Thus, increased transformation capacity, properly implemented, will correlate with reduced costs per transgenic event.

There are several major limitations to capacity building in the public sector. First, more trained transformation experts are essential to insure quality outcomes. Second, more facilities, particularly greenhouses, are also necessary to grow transgenics to seed. Increased funding would be required to devote more resources to existing facilities, either concentrated in one main location or in multiple, collaborating centers that would allow for standardization of genotypes and transformation vectors. Third, reliable transformation of multiple genotypes is needed to reduce the time frame for post-transformation analysis by one or two years for every project.

In addition to improving the production pipeline, continued research is essential to advance methods of transformation. Ideally, this should be facilitated through dialogue with industry to address any bridgeable gaps that might exist between public and private sector methodologies. **Major investments should be made in training and facility improvement. A transformation task force that includes academic and industry representatives should be formed to facilitate this goal.**

Efforts to streamline the APHIS notification process would be beneficial. Such efforts could be accomplished within Federal guidelines. For example, the community could develop a standardized operating procedure (SOP) for transgenic lines so that users can quickly and consistently provide the required information. In particular, common notification requests for frequently used transgenic resources could be standardized by the community, then communicated to APHIS. This cooperation between the maize community and APHIS regulators would help researchers for whom regulatory compliance can be prohibitive, such as researchers at smaller institutions or researchers who experiment infrequently with transgenic maize.

Similar to the challenge of regulatory compliance, many researchers do not have the infrastructure to grow transgenic events to seed. Furthermore, in the future integrated genomics world, it is essential that all researchers should be able to navigate between *Arabidopsis* and maize to conduct transgenic experiments. Most *Arabidopsis* researchers lack both experience and facilities to carry out the intensive aspects of maize transformation. Multiple centralized field sites across the US dedicated to growing transgenic plants would both facilitate compliance and enable more researchers from diverse institutions to use this critical technology.

The community noted the importance of communication with the public sector to publicize the value of transgenic maize research. One mechanism to do this would be for qualified representatives to communicate directly with reporters or media outlets that are in place at most institutions, to widely publicize our message.

Reverse genetics resources need to be expanded

A sequence-indexed collection of mutations is essential for researchers to exploit the genome sequence fully. It was noted that multiple mutagens are necessary to insure broad coverage of the genome and generate a range of allelic lesions. These would include transposon insertions, small deletions, and point mutations. **It is imperative that these lines be accessible through a community web browser to facilitate dissemination of the**

resource. Training in the use of the resource should be an essential and embedded component of dissemination. This collection should be searchable by BLAST, browsable, and linked to readily available seed stocks. **The sequence-indexing of transposon collections needs to be on validated germinal alleles so that seed are available for the community to advance the study of identified mutations.**

To date, several populations have been developed for forward and reverse genetics in maize inbred lines. The use of inbred materials greatly facilitates phenotypic analysis in near-isogenic lines and should be given strong consideration in population development. This is particularly relevant for maize, where a long generation time limits most researchers to propagating at most two generations/year. Large Uniform*Mu*, *Ac/Ds* and TILLING populations have been developed in the W22 inbred. The existence of these W22 populations provides a case for sequencing the W22 genome, which potentially could be one of the choices for a second physical map. The maize community considers that further discussion is needed to come to consensus about sequencing plans after the first-stage completion of B73.

TILLING populations have also been developed in B73. *Mutator* and *Ac/Ds* populations are nearly completed in B73 as an effort to exploit the genome sequence and provide greater accessibility for researchers across a broader geographic distribution. To achieve the goal of near-saturation mutagenesis (95% chance of a disruption in any given gene) additional line development is essential.

A number of approaches were discussed for chemical, radiation and insertional mutagenesis. There was much excitement over the potential for 454 and Solexa sequencing technologies to deliver quickly a near-saturation collection of *Mutator* insertions. Several *Mutator* populations exist with high copy number germinal *Mu* insertions. It was estimated that over 300,000 *Mu* insertions could be rapidly sequenced from Uniform*Mu* from the McCarty lab, and additional populations from the Schnable and Martienssen labs could provide similar levels of coverage. It was noted that a minimal input of resources could help accomplish the task of generating these essential resources.

The possibility of increasing *Ac/Ds* populations and developing fast neutron populations was also discussed to complement the non-transgenic *Mutator* populations. For instance, over 30% of maize genes are represented in tandem duplications, suggesting that a large number of potentially redundant paralogs are present in the maize genome; because many single gene mutations cause a phenotype, detailed analysis of locally duplicated genes in maize will address a key general question in biology, namely the mechanisms that permit subfunctionalization of duplicated genes. The task of recombining single gene insertions in tightly linked paralogs to create double mutant stocks is daunting and unlikely to succeed without a strong genetic selection. *Ac/Ds* can be used to sequentially mutagenize tandem gene clusters providing a resource to define functions for a sizeable fraction of the maize genes. Similarly, fast neutron mutagenesis programs will result in a range of deletion sizes, some of which will encompass multiple adjacent genes. Detailed genetic analysis of a locus is greatly facilitated by using an allelic series of mutants wherever possible. **Increasing effort in generating, expanding and integrating the data from these populations now will pay huge dividends in the coming years.**

Better access to gene expression profiling tools and datasets is needed

A number of platforms presently exist in the maize community for expression analysis. The 44K Agilent and 46K NSF-Arizona long oligo arrays, shoot apical meristem (SAM) cDNA arrays, and a first-generation 18K Affymetrix GeneChip are publicly available. However, the fully sequenced maize genome offers the opportunity to begin conducting gene expression profiling experiments using “all genes” platforms. Moreover, recent advances in cost effective deep sequencing (e.g. Solexa, NextGen) might yet provide another alternative for expression profiling in different tissues and variants including their microRNAs and alternative splice products. The maize community will continue discussion about which platforms will be best and what toolsets need to be developed to insure long-term utility of datasets generated by expression profiling. Tools must be developed that allow datasets to be browsed, queried, visualized, meta-analyzed and linked to the physical and genetic maps of maize. Development of cost-effective platforms is also paramount. **These ambitious goals will require substantial database efforts and funding, but are absolutely critical to optimize use of these cost-intensive datasets.**

Consistent with the emergence of the genome sequence, two “all genes” platforms are currently being designed by collaborative efforts of industry, maize biologists and informaticians, including an Affymetrix 100K GeneChip and an Agilent 105K *in situ* synthesized glass slide array. **Community input into the design of expression profiling platforms continues to be a high priority to the maize community.** The Agilent arrays will allow community input via customization, which is facilitated by its flexible format. A database of sequences, customizable formats and designs will be maintained by Agilent to allow results to be compared across experiments conducted on the various versions of Agilent arrays. The Affymetrix GeneChip will be developed in consultation with the community. It is predicted to include ~70K B73 gene models, allowing the remaining ~30K sequences to be used for evaluating allele-specific expression and the abundance of sRNA, transposon and retrotransposon transcripts according to community input. Each platform provides complementary approaches that together reduce bias associated with sequence variability among alleles.

These “all genes” arrays will be valuable for annotation of the genome, particularly for those sequences that were not represented as ESTs. It will be essential that all platforms adopt a common nomenclature and should be able to retrofit with updated annotations. These features will be a key to the longevity and widespread utility of the “all genes” arrays. The group considered that both of these platforms are cost-effective choices in the current climate and that competition will accelerate the improvement of database support and will lower costs. **Efforts should focus on incorporating community input, developing data integration tools and maintaining accessibility so diverse groups of researchers can benefit from public investment in profiling experiments and database tools.** The maize community also looks toward developing profiling platforms in the future that will accommodate advances in systems biology.

A major phenomics effort will contribute to basic and translational research

Understanding the function of genes and networks is a central research goal both currently and also in the post-genomics era. Phenotyping is one of the major strengths of the maize community. The maize community envisions carrying this capacity to the next level by **developing large scale and multi-dimensional phenotyping capabilities.** The group

recognized that understanding adaptation and applying the information to improve agriculture can be best achieved through in-depth phenotyping in diverse environments for numerous traits.

The major goals of this recommended effort are to: 1) harness genetic diversity to assign biological functions to sequences, i.e., associating traits with genes, and 2) enable predictive biology via an iterative process of discovery and validation. This effort will involve broad community involvement to collect and analyze phenotypes in great depth and breadth on a common set of diverse genotypes. It likely will also require common center(s) for production and quality control of seed, shared planting locations and protocols for collecting phenotypes, and centralized quality control for experimental design and data analysis. A series of phenotypes, including agronomic, morphological, cellular and molecular traits will be measured. Genotypes will include combinations of the natural variants, transformants and mutagenized populations developed by the research community. This effort will require development of new high throughput analytical tools, *e.g.*, remote sensing and image analysis. There will be very wide dissemination of collected data and efforts to coordinate sharing of results.

A unified phenotyping effort proposed here will require new scales of coordination within the community, will require continued advances in cyberinfrastructure, and further development of centralized databases for analysis, synthesis and dissemination of phenomics data. At the outset, researchers will guide the effort by establishing consistency of phenotyping language and by maintaining quality of experimental design and well-designed reporting mechanisms. First and foremost, database resources must be planned to provide for the integration and dissemination of data. Planning for large-scale phenotyping efforts should include industry, if mechanisms for shared and public access of data generated can be resolved unambiguously.

The maize community needs to broaden and strengthen its membership

The maize genome can be best leveraged by increasing the diversity of participants and strengthening the depth of training. The maize research community is committed to training creative, independent and collaborative scientists who conduct hypothesis-driven research, tool development and research translatable to agriculture. This can be achieved by **strengthening and diversifying graduate education and post-doctoral training**. It is clear from the experience of other research communities that establishing effective partnerships, nationally and internationally, is also essential to exploit a genome fully. The maize community considered types of partnerships that need to be developed and strengthened in the coming years including industry, international contacts, a wider spectrum of US-based researchers involved in maize research as well as other plant and non-plant biologists in general.

Enhance public - private interactions

Academic-industry partnerships have historically been strong for the maize community. A number of ideas were considered to strengthen our relationships further with industry. One idea was to provide mechanisms for industry to donate resources to the community, such as the EST database that was made accessible through an MTA, and funding of undergraduate summer internship programs. Another idea was to help young scientists make a transition to industry from academia by setting up a partnership with the

private sector to facilitate tours and visits and by providing talks for web viewing on ‘how to get trained for an industry position’. Issues related to public accessibility of data from new industry-academic partnerships need to be resolved. Adding a private sector member to the MGEC could be considered, as a mechanism to help facilitate interactions.

Enhance international dimension

International efforts should be coordinated to avoid duplication of effort and to foster dialogue. Ideas to enhance international collaboration included inviting one international speaker who has not been to the Annual Maize Genetics Conference to present each year. The community can also encourage and recognize greater efforts by PIs to become involved in Developing Country Collaboration supplements to NSF grants. Fellowship opportunities for graduate students abroad are currently lacking. Furthermore, to encourage the growth of maize research internationally, more meetings could be conducted outside of the US, such as in Mexico, South America, Europe, Asia, and Africa. International researchers might become more engaged in maize research if meetings are more accessible periodically. Independent or satellite workshop/short courses, such as the 2004 CIMMYT workshop conducted before the Annual Maize Genetics Conference, would further attract international participants.

Enhance local and national participation: Forming partnerships as outreach

The maize community is committed to improving science training in the US by reaching out to: small colleges and universities, traditionally under-funded research institutions, undergraduate institutions, minority-serving institutions, community colleges, tribal colleges, the K-12 system, and the general public. It was noted that maize genetics is an excellent “hook” for attracting new participants to science, because corn is such a familiar food item in the US and so many genetics tools are available. The goals of outreach activities should be 1) to integrate research and education and 2) to provide for a mutually beneficial partnership between members of the maize research community and new participants. These goals can be achieved by emphasizing relevance for all involved. Educational partnerships should be logical to the researcher’s expertise/interest as well as to the recipients’ needs and environment. **Communication among researchers with active and successful outreach programs should be strengthened to avoid duplication of effort.** To avoid continuous reinvention of methods, it might be useful to organize outreach advisory boards that can help guide new programs. Such advisory boards could be established and centralized through the current plant genome research outreach portal (plantgdb.pgrop). Best practices could also be highlighted at the Annual Maize Genetics Conference at the new designated poster session on Outreach and Training. The annual meeting should also be a central venue to bring new students and researchers from outside the maize field together with the current maize research community, thus enticing them to maize research. The Maize Genetics Meeting Steering Committee should continue to develop innovative ways to fund fellowships to new participants.

Summary and Next Steps

The sequenced maize B73 genome holds great promise for contributing to basic and translational research. To take full advantage of that promise we need to 1) make MaizeGDB the centralized sequence resource, 2) make plans and implement next level map-based sequencing efforts, 3) provide increased capacity and lower costs for maize transformation technology, 4) increase sequence-based reverse genetics, 5) coordinate expression platforms so all data are easily shared, 6) conduct a major phenomics effort that is effectively integrated, and 7) increase participation in maize research. A timeline for completing these goals was discussed as follows:

Activity	Timeline for completion
Convert MaizeGDB to a sequence-oriented database	2011 (three years)
Implement additional sequencing efforts. Generate physical maps from other genomes and anchor their sequences to their maps	2013 (3-5 years)
Establish high capacity transformation facilities	2013 (5 years)
Establish a near-saturation reverse genetics resource	2013 (3-5 years)
Standardize expression platforms	2013 (5 years)
Phenomics project underway	2016 (8-10 years)
Increase and diversify the maize research community	2018 (10 years)

Allerton Retreat participants agreed on several first steps to begin to implement the long-term plan described here. **First**, this planning document will be disseminated to the broader maize research community by the MGEC via MaizeGDB. **Second**, this document, combined with an executive summary, will be presented to guests at the Allerton Retreat including representatives of the funding agencies (NSF, DOE, USDA) and the NCGA. **Third**, an article will be submitted to *The Plant Cell* to inform the broader plant biology community of the future directions of maize research in the genomic era. **Fourth**, taskforces will be formed to focus on solutions to particular research bottlenecks, including transformation, and to help shape future new research efforts, such as phenomics. Through MGEC guidance, we anticipate that taskforce action plans will help maintain dynamic assessment of progress and will guide maize research into the future.

4 – MaizeGDB Genome Browser

INTRODUCTION

Based upon the 2006 Working Group Report (available at http://www.maizegdb.org/working_group.php) and the Allerton Report, it has become evident that the focus of MaizeGDB must be shifted to better accommodate a sequence-centric paradigm. In order to (1) show how the data at MaizeGDB relate to the maize genome, (2) relate MaizeGDB's current sequence data with other sequence information as it becomes available, (3) become the keeper of maize's 'official' set of gene models (which will enforce proper nomenclature), and (4) create a way to compare the various assemblies and annotations simultaneously, the feasibility of implementing of a genome browser at MaizeGDB was investigated and various available software were evaluated.

Choosing a genome browser to address the maize community needs has its challenges as there are several browsers available. The most widely used browsers are Ensembl, GBrowse, NCBI Mapviewer, and UCSC (in alphabetical order). They all have intrinsic strengths and weaknesses. For example, among the browsers listed above, Ensembl provides the best tools for comparative genomics. However, because the Ensembl attempts to provide every iota of available data on a given species, it is slower than the other browsers in generating webpages and important pieces of data can become buried among other less useful information. Each genome browser software has such strengths and weaknesses, so determining which software best suits the needs of maize geneticists is a task that requires careful analysis.

When it comes to choosing a genome browser for MaizeGDB, a very important challenge for the team was whether to go with Ensembl as our genome browser because the Maize Genome Sequencing Consortium has been publishing genome data using Ensembl (see <http://www.maizesequence.org>) and the Ware group offered to collaborate with us by allowing the MaizeGDB team access to the browser for data update and display purposes. This would likely be a very good use of funds in the short term. However, although choosing a browser other than Ensembl would require extra effort at the beginning, providing extensive links between MaizeGDB's browser and MaizeSequence.org/Gramene might enhance the maize research by offering an alternative view of the maize genome to cooperators. It became our goal to decide whether to work with the Ware group on Ensembl or to implement another browser.

Another issue for MaizeGDB was how much input we wanted from the maize community. Model Organism Databases usually make an executive decision when it comes to software selection. Because the maize community communicates well, has a clear vision of their research problems, and has good ideas on how best visualize a sequenced maize genome (and because we believe that the next generation genome browsers should not be specified by a secluded group of programmers), we prepared a survey (see APPENDIX) to gauge cooperators' impressions of existing software and to find out what sorts of functionalities they would like to have in a maize genome browser.

PREPARATION OF THE SURVEY

The MaizeGDB team worked with the Working Group and Maize Genetics Executive Committee to prepare and disseminate a survey.

Eliminating bias

After we prepared the initial draft of the survey, it was sent to the Working Group and Maize Genetics Executive Committee for suggestions. After we incorporated their suggestions, we sent it to Assistant Professor Patrick Armstrong in the Department of Psychology at Iowa State University who made suggestions on how to eliminate bias in the survey. His recommendations were:

- 1) Instead of asking people to circle a specific website that uses a specific browser, put the websites in alphabetical order so that people will be able to find the websites they are using very easily without thinking which genome browser the website is using.
- 2) When people take surveys, they sometimes tend to be very careful (or complete) in the first questions, and then less enthusiastic near the end. So put the most important questions in the beginning.
- 3) Combine ranking of features with rating (basically assigning weights to specify how important a specific feature is to the user)

We implemented the first two of his recommendations. For the third recommendation, we decided to rephrase the rating by including the question “which features are most indispensable to you?”

Choosing the survey takers

We sent the survey to all the “maize cooperators”. For MaizeGDB, maize cooperators are attendees of maize meetings, researchers publishing frequently on maize, or people who specifically request to be considered a maize cooperator. We realize that this definition does not cover every maize researcher, but it provides a large pool of people with an interest in maize. Overall, 1241 surveys were sent out.

We ensured the privacy of cooperators by sending to each cooperator a link to the survey along with a randomly generated unique key that cannot be tracked by individual and enforces the rule that each cooperator can take the survey only once.

SURVEY STATISTICS

The raw survey results can be found at <http://shrimp1.gdcb.iastate.edu/browser-survey/analyze.php>. If you prefer tabulated results, you can direct your Internet browser to <http://shrimp1.gdcb.iastate.edu/browser-survey/analyze-tab-delimited.php>.

The Number of Respondents

Among the 1241 cooperators surveyed, 99 responded. Although the number is small, it is useful to compare it with the response to the last MGEC election where 234 of the 1190

surveyed cast a ballot. Because the Genome Browser Survey requested detailed answers to the researchers' needs and because not every maize researcher needs a genome browser (or he/she does not feel knowledgeable enough to answer a detailed survey on genome browser preferences), this level of response did not seem unreasonable.

Time Spent Accessing Maize Online

37% percent of the survey takers spend an hour or two each week online to access maize data online. 39% percent spend between two and five hours. 15% spend more than 5 hours online to access maize data. Only 8% of the survey takers did not use online maize data resources.

Genome Browsers Used

66% of the respondents use Maizesequence.org and the number is the same for Gramene users. A total of 75% use either MaizeSequence or Gramene. Although both sites use Ensembl as a genome browser, only 26% of the respondents acknowledged they are using Ensembl. This result shows that the users may not be aware of the underlying software for browsers that the various websites are using.

TAIR, MAGI, and PlantGDB are being used by 54% of the respondents (but not always by the same people). 42% use NCBI's Map Viewer. As above, although 45% use TAIR, only 22% are aware that it is using GBrowse.

Note that instead of choosing genome browser names using drop-down menus, a few respondents chose to write the names of the sites under, "Other Genome Browsers". We have not included those statistics into our analysis.

Feature Rankings

The features are sorted as follows (rankings are shown in parentheses where a lower number indicates more support): Ease of use (1.9), visuals (2.6), speed (3.2), cross-species comparison (3.7), multiple gene selection (4.1), differentiation between computational and experimental data (4.1), and ontologies (5.1). Clearly, the respondents want an intuitive genome browser that allows researchers locate the needed data in the most accessible and fastest fashion.

Desired Features

The most desired feature section of the survey is very helpful to guide genome browser developers in the creation of new features. The users want to reach specific data in the most intuitive tools. They also want downloadable data sets in various formats. The respondents are in need of enhanced cross-referencing between different websites. They desire the most current data and the tools that are easy to learn and to apply for their specific research needs. In short, the users want the minimized hassle and effort in reaching the needed maize data.

Bad Genome Browser Examples

Among 29 comments left in "Bad genome browser examples", 19 of them cite either Maizesequence.org or Gramene (66%), which use Ensembl as a genome browser. The

reason might be that Maizesequence.org or Gramene is the most used browser for the maize cooperators (75% of the respondents uses either site), but the high percentage of those discontent hints that real issues with Ensembl may need to be addressed. The respondents usually cite the slowness of the website as the major (and sometimes the only) problem. Another cited problem is, as one respondent says, “many, many non-intuitive steps to get information”. It seems that Gramene will benefit if it provides training to the community on how to use their Genome Browser most efficiently and enhances links between various types of maize data (e.g., linking expression data with “EST/genes”).

CONCLUSION

Based upon results of the Genome Browser Survey, we support the use of GBrowse rather than Ensembl or other browser for the following reasons:

- 1) In the “Feature ranking”, the three most desired features are listed as: ease of use, visuals, and speed. Indications from the data are that cooperators do not consider Ensembl to be easy to use, and it is definitely not fast when compared to the other software available. Cross-species comparison capabilities (where Ensembl shines) is only ranked 4th, and Gbrowse now has such capabilities available (Synbrowse, CMap, etc.).
- 2) As indicated in the “Indispensable features”, cooperators would like to see specific tool development in the genome browser to enhance their research (e.g., finding genes between two markers). Therefore, a genome browser chosen by MaizeGDB should allow high flexibility in terms of code, tools, and community involvement. The flexibility of tool development is intrinsic feature of GBrowse that allows customizable plug-in architecture because it defines itself as a community-based open source project. In the case of Ensembl, the code development is primarily done by a group in UK and *ad hoc* tool development is carried out by research groups for their specific needs. This tool development is specific to a particular Ensembl version, and for each new version of Ensembl, the tool must be modified or re-written.
- 3) MaizeSequence.org/Gramene is already providing maize sequence information using Ensembl. Providing this information using GBrowse would allow researchers to use different genome browsers for different applications. For example, when a cross-species comparison across many clades is necessary, Ensembl is very powerful; however, when it comes to developing customizable visualization and analysis solutions for maize-specific research problems, GBrowse would stand out. Offering these two browsers to use by maize researchers will facilitate answering different research problems and will enhance agricultural research overall.
- 4) Interlinks can be provided between different genome browsers, so choosing another browser than Ensembl that maizesequence.org uses will enrich the research tools with minimal cost. In the case of MaizeGDB, we will provide links to data sets in Gramene.

ROADMAP FOR IMPLEMENTING THE BROWSER

We plan to start implementing the Genome Browser right away. We will obtain a copy of the database from Maizesequence.org as well as copies of the various maize genome assemblies currently available so that we can start working on how the sequence data will fit within the database schema of GBrowse.

We plan to choose 5 people to provide guidance (we will make sure that they are from academia, industry, and outside of the US), and 10 people for beta testing among the cooperators who agreed at the end of the survey to be a part of the Genome Browser implementation.

We plan to write proposals to include new analysis of the maize genome in structural genomics and systems biology that includes protein structure models and pathway representations. Note that for pathways, a collaboration with Gramene would be reasonable given their current focus on pathway data and tools.

Project Plan

NP 301 – Plant Genetic Resources, Genomics and Genetic Improvement

Panel Review: September - December 2007

Old ARS Research Project Number

3625-21000-045-00D

Research Management Unit

Corn Insects and Crop Genetics Research Unit

Location

Ames, Iowa

Project Title

The Maize Genetics and Genomics Database

Investigator(s)

C.J. Lawrence (Lead Scientist)...	1.00
T.Z. Sen.....	1.00
L.C. Lewis.....	0.03

Scientific Staff Years

2.03

Planned Duration

60 months

Pre-Peer Review Signature Page

**(SIGNATURE AND DATES MUST BE COMPLETE PRIOR TO DISTRIBUTING THIS PROJECT PLAN TO
PEER REVIEWERS)**

[Lawrence, 3625-21000-045-00D and The Maize Genetics and Genomics Database]

This project plan was found to meet the peer review criteria, to be in compliance with the Project Plan Instructions and Format, and demonstrate how the research team will conduct research in a manner appropriate for this area of research. The funds committed toward this project are sufficient to support the planned research.

Leslie C. Lewis /s/
Research Leader

08/06/07
Date

This project plan was prepared by a qualified research team and demonstrates the research team's best effort towards achieving the assigned research objectives. All internal review and approval requirements have been met. This project plan is relevant to the Agricultural Research Service's National Program [enter NP # and title] Action Plan and was prepared in accordance with the outlined objectives, experimental approach, and project duration previously agreed to by the National Program Team and Research Team. To validate the plan's readiness for implementation and gain recommendations for improvement, the project plan is now available for peer review.

Area Director

Date

These officials have not performed a scientific merit peer review. Their statements do not necessarily require expertise in the scientific subjects associated with this research. The approval to implement this project plan cannot be made without scientific peer review coordinated by the Office of Scientific Quality Review, ARS, USDA.

TABLE OF CONTENTS

PROJECT SUMMARY	4
I. OBJECTIVES	5
II. NEED FOR RESEARCH.....	5
III. SCIENTIFIC BACKGROUND	9
IV. APPROACH AND RESEARCH PROCEDURES	14
OBJECTIVE 1: INTEGRATE NEW MAIZE GENETIC AND GENOMIC DATA INTO THE DATABASE.....	14
• <i>Sub-objective 1.A. Expand mutant and phenotype data and tools.....</i>	14
• <i>Sub-objective 1.B. Expand structural and genetic map sets.....</i>	18
• <i>Sub-objective 1.C. Provide access to gene models calculated by leading gene structure prediction groups through the MaizeGDB interface.....</i>	21
• <i>Sub-objective 1.D. Compile and make accessible at MaizeGDB the annual Maize Newsletter.</i>	22
OBJECTIVE 2: PROVIDE COMMUNITY SUPPORT SERVICES, SUCH AS LENDING HELP TO THE COMMUNITY OF MAIZE RESEARCHERS WITH RESPECT TO DEVELOPING AND PUBLICIZING A SET OF GUIDELINES FOR RESEARCHERS TO FOLLOW TO ENSURE THAT THEIR DATA CAN BE MADE AVAILABLE THROUGH MAIZEGDB; COORDINATING ANNUAL MEETINGS; AND CONDUCTING ELECTIONS AND SURVEYS.	23
V. PHYSICAL AND HUMAN RESOURCES.....	25
VI. PROJECT MANAGEMENT AND EVALUATION	27
VII. MILESTONES AND EXPECTED OUTCOMES	29
VIII. ACCOMPLISHMENTS FROM PRIOR PROJECT PERIOD	34
IX. LITERATURE CITED	36
X. PAST ACCOMPLISHMENTS OF INVESTIGATORS – CAROLYN J. LAWRENCE.....	38
PAST ACCOMPLISHMENTS OF INVESTIGATORS – TANER Z. SEN.....	40
PAST ACCOMPLISHMENTS OF INVESTIGATORS – LESLIE C. LEWIS	42
XI. HEALTH, SAFETY, AND OTHER ISSUES OF CONCERN	45
APPENDIX - LETTERS AND WORKING GROUP REPORT	45

LIST OF FIGURES AND TABLES

FIGURE 1: THE BIOINFORMATICS FOOD CHAIN	10
TABLE 1. DATABASES UTILIZED BY MAIZE RESEARCHERS	11
FIGURE 2: SIMPLIFIED INFRASTRUCTURE OF MAIZEGDB	12
FIGURE 3: METHOD FOR COMMUNITY-DRIVEN DATA INPUT.....	15
FIGURE 4: THEMES SHARED AMONG OBJECTIVES.....	25

PROJECT SUMMARY

In 2001 maize became the number one production crop in the world. Its success is largely due to the plant's high productivity per acre in tandem with its wide variety of commercial uses: not only is maize an excellent source of food, feed, and fuel, but also its byproducts are used in the production of paint, soap, rubber, plastics, and various other commodities. Maize's unparalleled success in agriculture stems from basic research, the outcomes of which drive breeding and product development.

In order for applied researchers to benefit from basic biological investigation, generated data must be made freely and easily accessible. MaizeGDB (<http://www.maizegdb.org>) is the maize research community's central repository for genetics and genomics information. The overall goal of our work is to create and maintain unified public resources that facilitate access to the outcomes of maize research. We will attain this goal by addressing two objectives. The first is to integrate new maize genetic and genomic data into MaizeGDB and the second is to support the maize research community by coordinating group activities.

To accomplish the first objective, we will expand mutant and phenotype datasets as well as structural and genetic maps. In particular, we will integrate genetic maps with the B73 genome sequence and create methods of access and presentation that convey the substantial variation in maize genome using a Genome Browser as the basis for display. To achieve the second objective, we will conduct community support services such as coordinating annual meetings and conducting elections and surveys.

I. OBJECTIVES

The long-term objectives of this project are to synthesize, display, and provide access to maize genomics and genetics data for the research community and applied users. The challenges to be met over the next five years will be handling expanding data from maize genome sequencing and providing tools for functional analysis and genetic improvement. To address these challenges, we will focus on the following objectives:

Objective 1: Integrate new maize genetic and genomic data into the database.

- **Sub-objective 1.A.** *Expand mutant and phenotype data and tools.*
- **Sub-objective 1.B.** *Expand structural and genetic map sets emphasizing the:*
 - *Integration of the IBM genetic maps with the B73 genome sequence (This will be a shared objective with Mary Schaeffer, Columbia, Missouri, on ARS Project no. 3622-21000-027-00D, entitled “Genetic Mechanisms and Molecular Genetic Resources for Maize.”);*
 - *Creation of views that convey the substantial variation in maize genome structure; and*
 - *Integration of the next-generation genetic map being generated by the Maize Diversity Project into a genomic view to enable its effective use by plant breeders (This will be a shared objective with Mary Schaeffer, Columbia, Missouri, on ARS Project no. 3622-21000-027-00D, entitled “Genetic Mechanisms and Molecular Genetic Resources for Maize.”).*
- **Sub-objective 1.C.** *Provide access to gene models calculated by leading gene structure prediction groups through the MaizeGDB interface.*
- **Sub-objective 1.D.** *Compile and make accessible at MaizeGDB the annual Maize Newsletter.*

Objective 2: Provide community support services, such as lending help to the community of maize researchers with respect to developing and publicizing a set of guidelines for researchers to follow to ensure that their data can be made available through MaizeGDB; coordinating annual meetings; and conducting elections and surveys.

II. NEED FOR RESEARCH

II.A Description of the Problem to be Solved:

Maize (referred to commonly as corn or by its botanical name *Zea mays* L. spp. *mays*) is an important crop. Not only is it one of the most abundant sources of food and feed for people and livestock the world over, but also it is an important component of many items where its content is less apparent. Maize is used in the manufacture of diverse commodities including 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 economical fuel alternative in the long run. 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. In the 1920's Emerson recognized that the rapidly expanding compendium of maize data needed organization,

publication, and curation. To meet that need, he began circulating a Maize Newsletter, which is still published to this day. **Emerson realized that unless scientific findings were stored and organized in an accessible manner, the products of costly research would be effectively lost.** Today, with the accelerated generation of maize genetic and genomic information, the need for a centralized repository is critical. Because the depth and breadth of maize data increased steadily over the years, it became desirable to create the ability to query across data and to find new associations among divergent sets of data. This functionality could only be realized by storing the data in a relational database to ensure dynamic associations between various types of data generated by different researchers. It was the aim of MaizeDB, the first generation maize genetics database, to meet that need. Subsequently, maize researchers developed the need to pose complex queries on maize data, similar to those that could be accommodated at other high-profile biological databases like The Arabidopsis Information Resource (TAIR; Rhee et al., 2003) and GenBank (Benson et al., 2007). MaizeGDB (the **Maize Genetics and Genomics Database**; Lawrence et al., 2007) superseded MaizeDB in 2003 to overcome the challenges presented by new types of data and the need to respond to complex queries. MaizeGDB is the Model Organism Database (MOD) for maize and is strongly supported by the maize community, as demonstrated in the 2007 Allerton Report (see <http://www.maizegdb.org/AllertonReport.doc>), the MaizeGDB Working Group's 2006 Report (see Appendix and also at <http://www.maizegdb.org/2006WorkingGroup.pdf>), and the Maize Genetics Executive Committee (see letter of support and collaboration).

At present, maize geneticists are at the cusp of yet another milestone: the genome of the maize inbred B73 is being sequenced in the U.S., with anticipated completion in 2008. In addition, scientists working in Mexico at Langebio (the National Genomics for Biodiversity Laboratory) and Cinvestav (Centro de Investigacion y Estudios Avanzados) have just announced through a press release (July 12, 2007) that they completely sequenced 95% of the genes with 4 X coverage in a native Mexican popcorn called palomero, though the data have not yet been released and the quality of the data is unknown. In any case, relating sequences to the *existing* compendium of maize data is the primary need that must be met for maize researchers in the immediate future. **Creating and conserving relationships among the data will enable researchers to ask and answer questions about the structure and function of the maize genome that previously could not be addressed.**

By making sequence information more easily accessible and fully integrated with other data stored at MaizeGDB, it will become possible for researchers to begin to investigate how sequence relates to the architecture of the maize chromosome complement. How are the chromosomes arranged? Is it possible to relate the genetic and cytological maps to the assembled genome sequence? Are there sequences present at centromeres that signal the cell to construct kinetochores, the machines that ensure proper chromosome segregation to occur, at the correct site? MaizeGDB aims to enable researchers to discover answers to such queries that will enhance the quality of basic maize research and ultimately the value of maize as a crop. It will be possible to interrogate the database to find answers to complex questions, and the content of the genome can be related to its function, both within the cell and to the plant as a whole. **Convergence of traditional biological investigation with the knowledge of genome content and organization is currently lacking, and is a new area of research that will open up once a complete genome sequence and a method for searching through the whole of the data are**

both in place. It is the ability to investigate and answer such basic research questions that will serve as the basis for devising sound methods to breed better corn plants. Once the relationships among sequence data and more traditional maize data like genotypes, phenotypes, stocks, etc. have been captured, it is important that those data be presented to researchers in a way that can be easily understood without requiring that they have any awareness of how the data are actually stored within a database.

It is these needs – creating connections between sequence and traditional genetic data, improving the interface to those data, and determining how sequence data relate to the overall architecture of the maize chromosome complement – that the MaizeGDB team seeks to fulfill.

II.B.Relevance to ARS NP 301 Action Plan:

National Program (NP) 301, Plant Genetic Resources, Genomics, and Genetic Improvement, supports research that expands, maintains, and protects our genetic resource base, increases our knowledge of genes and genomes, and, through novel tools and approaches, manages and delivers vast amounts of genetic and phenotypic information. The ultimate goals for the preceding efforts are to improve the production efficiency and the health and value of our nation's crops.

The research described in this project plan will contribute to NP 301 Component 2: Crop Informatics, Genomics, and Genetic Analyses as follows:

- *Problem Statement 2A: Genome Database Stewardship and Informatics Tool Development.* MaizeGDB team members will work to keep the resource's data and methods for data access up-to-date to maintain its utility and relevance and will support stakeholders by providing informatic support services. Data curation focus areas are described in Sub-objectives 1.A, 1.B, 1.C, and 1.D. Incorporation of existing software tools into the resource as well as the creation of unique tools to meet researchers individual needs are described in Sub-objectives 1.A, 1.B, and 1.C. Genome Database Stewardship also requires ensuring the transfer of generated data to MaizeGDB, which necessitates a concerted effort and facilitated communication with the maize community as described in Objective 2.
- *Problem Statement 2B: Structural Comparison and Analysis of Crop Genomes.* Because the maize genome shows within-species sequence and structural variations commensurate with the within-family variations known for other plants (Fu and Dooner, 2002), the Genome Browser to be made available at MaizeGDB (described in Sub-objectives 1.A, 1.B, and 1.C) will be enabled to support within-species comparative genomics activities.
- *Problem Statement 2C: Genetic Analyses and Mapping of Important Traits.* As described in Sub-objective 1.A, expansion of mutant and phenotype data and tools at MaizeGDB will enable researchers to identify the genes that contribute to traits of value. By expanding structural and genetic map sets as described in Sub-objective 1.B and linking those data to the sequence annotations described in Sub-objective 1.C, the

locations of identified genes can be determined thus allowing researchers to devise breeding schemes that take advantage of nearby markers for stock selections.

II.C. Potential Benefits Expected by Attaining Objectives:

Our work will facilitate the utilization (analysis and application) of maize genetic and genomic data for the research community and applied users. The generated centralized resource for maize data will enable researchers to leverage comprehensive knowledge on the nature of genetic factors that affect yield and their molecular mechanisms. This will lead to increased yield and enhancement for direct improvement in the conventional sense via targeted use of genetic engineering and will redirect research in such a way that novel uses of maize for food, feed, fuel, and other industrial applications can be devised.

II.D. Anticipated Products of the Research:

The MaizeGDB Team will work together to generate actively curated and reliable genetic, genomic, and phenotypic description data sets: (1) annotated gene sequences will be assigned to candidate gene locations, (2) genetic, physical, and cytogenetic maps will be integrated, (3) single points of access (portals) to multiple databases will be developed and made available, (4) software and data analysis tools that enable the analysis of genetic and genomic data sets will be created, and (5) databases will become better interconnected and interoperable through the use of controlled vocabularies, ontologies, and context-sensitive links. Because many of these anticipated products involve or are connected to sequence data, sequences will become more easily accessible and fully integrated with other data stored at MaizeGDB, which will add tremendous functionality. Downstream, this will allow more efficient tools and more rapid research progress for all maize researchers. The integration of maize sequence to MaizeGDB will allow: (1) the investigation of how sequence relates to maize chromosome architecture and cellular function; (2) improved throughput for gene cloning; (3) the development of streamlined workflows that enable researcher to traverse from phenotype to sequence and sequence to seed (and visa-versa); (4) discovery of additional upstream and downstream regulatory sequences including microRNAs; and the list goes on.

II.E. Customers of the Research and Their Involvement:

Customers include geneticists, plant breeders, the Midwest associations of corn producers, the National Corn Growers Association, organic food producers, consumers, and students in the agricultural sciences. Input from these customers/stakeholders will be taken through the MaizeGDB Working Group (see Appendix) and via feedback mechanisms embedded in the MaizeGDB resource itself (<http://www.maizegdb.org>). The Maize Genetics Executive Committee (MGEC) and the Maize Genetics Conference Steering Committee (MGCSC) are two groups that 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, which 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. Interactions with the MGEC are made possible by including members of the MGEC in the MaizeGDB Working Group. At present, 5 of the 13 MGEC members also serve on the MaizeGDB Working Group, and one MGEC member (M. Schaeffer) is also a member of the MaizeGDB team directly. Two of the 14 MGCSC members serve on the MaizeGDB Working Group, and two ex officio members (T. Seigfried and M. Schaeffer) are also members of the

MaizeGDB team directly. Including the activities of the MGEC and MGCSC as MaizeGDB services allows the MGEC and MGCSC to rely upon the MaizeGDB staff to handle their elections, polls, abstract collection, and other operational and information gathering activities. This service enables the MaizeGDB staff to bring useful statistics and data into MaizeGDB in addition to emphasizing the fact that MaizeGDB is the central resource for maize data as well as maize community information. Hence, these joint activities are mutually beneficial.

In past years, the Annual Maize Genetics Conference and International Plant and Animal Genome Conference have been highly productive forums where MaizeGDB staff members have received broad input from maize researchers and workers from the U.S. and abroad. Input from attendees of those conferences is expected to continue to be useful in guiding database and tool development in the future. New customers may emerge as the maize genome becomes more fully sequenced and as comparative genomics resources increase linkages to MaizeGDB.

III. SCIENTIFIC BACKGROUND

III.A. Types of Data Repositories:

Databases storing genomic information fall into various categories based upon their role within a larger context (see Figure 1, next page). A **Laboratory Information Management System (LIMS)** is the most basic sort of database and interface solution, and can be as simple as a spreadsheet stored on a computer in a particular laboratory. Complex systems where the LIMS is made up of various data pipelines and/or laboratories are generally highly customized and are created to support an individual research group's shared data management needs. Data stored within a LIMS environment represent the group's working information and generally are not made available for use outside of the group that generated the data. **Static Repositories (SRs)** are those resources where data are deposited for long-term storage. The data generally are not changed over time, hence the moniker 'static.' The most well known SR for biological data is GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>; Benson et al., 2007), the federally funded resource that stores sequence data for all species. An **Automatic Annotation Shop (AAS)** harvests data from SRs and runs those data through analysis pipelines to create products that have added value for use by researchers. TIGR (The Institute for Genomic Research; <http://www.tigr.org/>) is an AAS that provides value-added sequence-based products including genome assemblies and repeat databases based upon the sequence set stored at GenBank (Chan et al., 2006). **Model Organism Databases (MODs)** are generally species specific, and have been created for maize (MaizeGDB; <http://www.maizegdb.org>), soybean (SoyBase; <http://www.soybase.agron.iastate.edu>), *Arabidopsis* (TAIR; <http://www.arabidopsis.org>), and various other species including *Drosophila*, *C. elegans*, mouse, zebrafish, *E. coli*, etc. These databases are built and maintained by teams of information technology specialists and curators, and represent highly curated products that recapitulate the biology of a particular species by storing species-specific data types and making available specialized tools for analyzing those data within their specialized biological context. **Clade-Oriented Databases (CODs)** store and make accessible those data that can be leveraged by researchers to enable comparative biological analyses including sequence similarity and genomic synteny information. CODs are especially important for communities working on groups of species simultaneously like potato, tomato, and pepper (SGN; <http://www.sgn.cornell.edu>; Mueller et al., 2005). The various database types function in a coordinated fashion, and data flow from the LIMS to the MODs and CODs as

shown in Figure 1 (below). A (non-exhaustive) list of the various databases used by maize researchers is shown in Table 1 (next page).

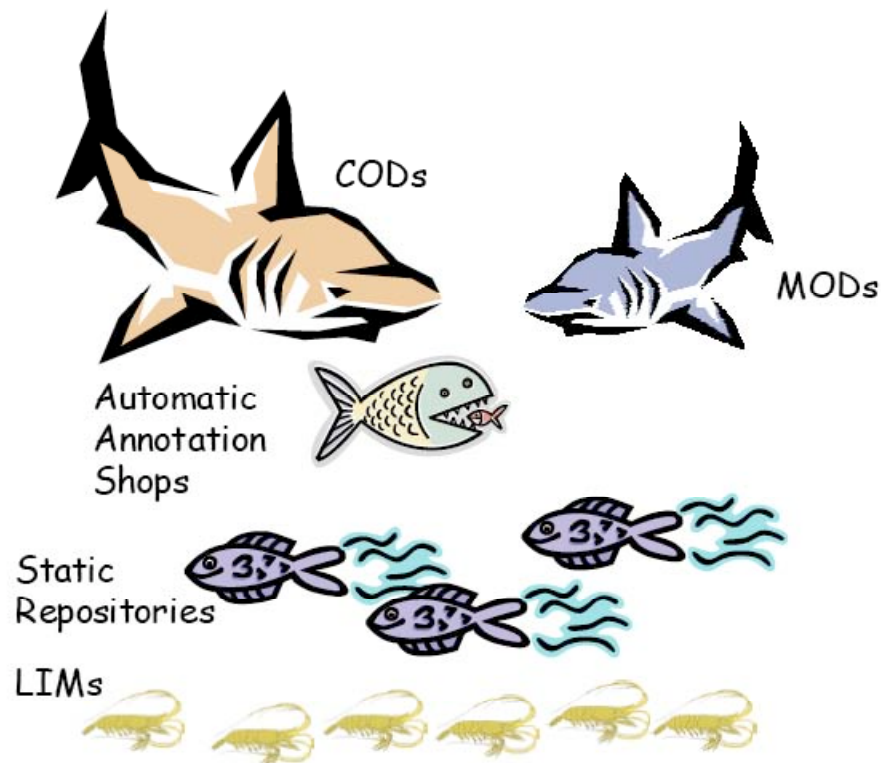


FIGURE 1: THE BIOINFORMATICS FOOD CHAIN.

At the bottom of the food chain are Laboratory Information Management Systems (**LIMS**). Next in the food chain are the static data repositories that are responsible for providing long-term storage for the information generated by the data providers. The information produced by automatic annotation shops is in turn taken up by Model Organism Databases (**MODs**). These are community databases focused on a single species. MODs take the information provided by automatic annotation shops, enhance it with hand curation, integrate it with information from the literature, and relate it to other data sets and resources. The Clade-Oriented Databases (**CODs**) are multi-species databases that usually have a clade-specific emphasis. They integrate information from the static data repositories, annotation shops, and MODs into a single integrated database designed expressly for making comparisons among species. [Image reproduced here with permissions granted by Dr. L. Stein on behalf of the Plant Genome Databases Working Group (Beavis et al., 2005).]

TABLE 1. DATABASES UTILIZED BY MAIZE RESEARCHERS. Please see text for abbreviations.

RESOURCE TYPE	RESOURCE	WEBSITE
MOD	MaizeGDB*	http://www.maizegdb.org
MOD	TAIR	http://www.arabidopsis.org
COD	Gramene	http://www.gramene.org
COD	GrainGenes	http://wheat.pw.usda.gov
COD/AA/LIMS	PlantGDB	http://www.plantgdb.org
COD/AA/LIMS	PLEXdb	http://plexdb.org
AA	TIGR	http://www.tigr.org
AA/LIMS	MGSC's Maize Genome Browser*	http://www.maizesequence.org
AA/LIMS	MAGI*	http://www.plantgenomics.iastate.edu/maize
AA/LIMS	FPC-maize*	http://www.genome.arizona.edu/fpc/maize
SR	NCBI/GenBank	http://www.ncbi.nlm.nih.gov
SR	UniProt	http://www.pir.uniprot.org
SR/LIMS	GRIN	http://www.ars-grin.gov
LIMS	Panzea*	http://www.panzea.org
LIMS	ChromDB	http://www.chromdb.org

* indicates a maize-specific resource

III.B. MaizeGDB's Role:

MaizeGDB is the MOD for maize. Stored at MaizeGDB are loci (genes and other genetically-defined genomic regions including QTL), variations (alleles and other sorts of polymorphisms), stocks, molecular markers and probes, sequences, gene product information, phenotypic images and descriptions, metabolic pathway information, reference data, and contact information for maize researchers. In addition to storing and making available maize data, the MaizeGDB team also provides services to the community of maize geneticists and technical support for the MGEC and the Annual Maize Genetics Conference. Bulletin boards for news items, information of interest to cooperators, lists of websites for projects that focus on the scientific study of maize, an editorial board's recommended reading list, and educational outreach items are among the webpages made available through the MaizeGDB site.

The Web interface at <http://www.maizegdb.org>, which most researchers utilize to retrieve data from MaizeGDB, is only one component of the overall MaizeGDB infrastructure (Figure 2; next page). Prior to being in that **Production Environment**, the raw data are prepared for public accessibility in a **Staging Environment**. In the Staging Environment, the most up-to-date information is available, new data are added to the database, and existing data are updated with new information. In addition to a Web Interface that appears identical to the one in the Production Environment, the Staging Environment offers SQL read-only access to the community so that researchers interested in interacting with the data in a more direct and customized manner can have access to the most up-to-date information available. Also available within the Staging Environment are Community Curation Tools to enable researchers to add small datasets to the database directly, as well as a set of Professional Curation Tools developed by Dr. M. Sachs' group at the Maize Genetics Cooperation – Stock Center in Urbana-Champaign (Scholl et al., 2003). Whereas the Community Curation Tools have many safeguards to help researchers enter the data step-wise and with enforced field requirements, the Professional Curation Tools allow MaizeGDB project members as well as Stock Center personnel to enter

datasets in a more stream-lined and powerful fashion with fewer integrity enforcement rules (which slow down the data entry process considerably). It also should be noted that data added to the database via the Community Curation Tools are first marked as “Experimental” data that must be “activated” by professional curators at MaizeGDB. This ensures that only quality information is made publicly accessible. The availability of Curation Web Interface enables researchers to view the data as they will appear once they are uploaded to Production. If researchers wish to deposit complex or large datasets, it would not be reasonable to enter the data via the Community Curation Tools because those tools work via a “bottom-up” approach whereby the records are (1) built based upon the most basic information included in the dataset and (2) entered one record at a time (i.e., not in bulk). For complex or large datasets, researchers are encouraged to submit data files to the curators at MaizeGDB. Those data are added to the database directly by curators and the database administrator. To aid in the modeling of new types of data for inclusion in the MaizeGDB product and to enable programming to be tried out in a safe place, a **Test Environment** identical to the Staging Environment has been created. Note that three copies of the database exist, and that a **Disaster Recovery** system has been put in place whereby the Curation Database is backed up in a compressed format to a separate machine in Ames, IA daily. Once weekly, the Ames file is copied to Columbia, MO for off-site storage.

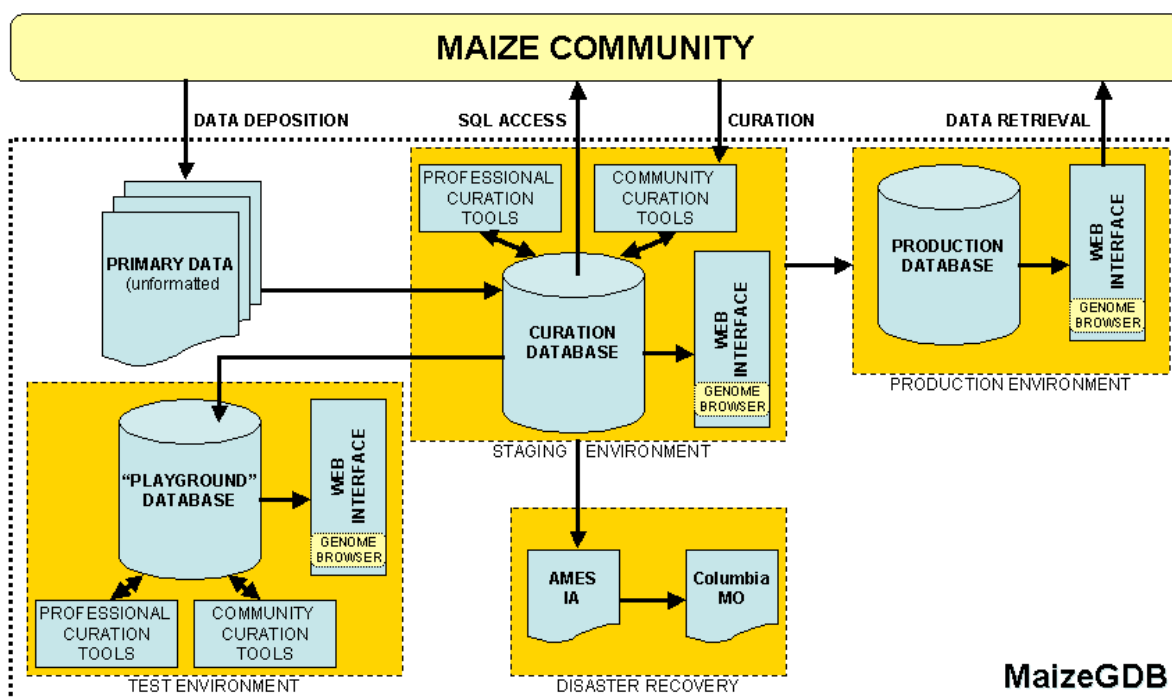


FIGURE 2: SIMPLIFIED INFRASTRUCTURE OF MAIZEGDB.

The community of maize researchers can add data to the database (downward-facing arrows from the uppermost yellow box) via direct data deposition (upper left) and via a set of Community Curation Tools that interacts with the Curation Database (upper center). Researchers are also allowed access to maize data via a Web interface that can be accessed at <http://www.maizegdb.org> (upper right) and by way of SQL access to the Curation Database, which houses the most up-to-date data available (upper center). These functionalities are supported by two of the three environments: Production and Staging, respectively (upper dashed

orange boxes). Available for use by MaizeGDB personnel to facilitate data modeling and trial programming manipulations is a third environment called Test (lower left dashed orange box), which is identical to the Staging Environment. To ensure that the most up-to-date copy of the database is backed up, a Disaster Recovery process has been instituted (lower center dashed orange box) whereby a compressed copy of the database is backed up to a separate machine in Ames, Iowa daily, and to a server in Columbia, Missouri weekly.

III.C. Other Related CRIS Projects

There are currently 12 active CRIS projects on maize genetics and genomics that relate to data storage and accessibility as well as to long-term plant databases, **three of which currently support MaizeGDB directly**: (1) 3625-21000-045-00D, this project, formerly called “Database of Maize Genome Information,” Lead Scientist Dr. C. Lawrence; (2) 3622-21000-026-00D, “Maize Genome Database,” Lead Scientist Dr. M. Schaeffer; and (3) 3622-21000-027-00D, “Genetic Mechanisms and Molecular Genetic Resources for Maize,” Lead Scientist Dr. M. McMullen (note that 36222-21000-026-00D expires soon and is to become a part of 3622-21000-027-00D).

CRIS projects of relevance to this research include: (1) “Conservation and Utilization of Maize Genetics Stocks” (3611021000-019-00D), Lead Scientist Dr. M. Sachs and (2) “Dissecting Complex Traits in Maize by Applying Genomics, Bioinformatics, and Genetic Resources” (1907-2100-021-00D), Lead Scientist Dr. E. Buckler. Also of interest are the GrainGenes and SoyBase projects – (3) “Database and Bioinformatic Resources for Small Grains Research and Crop Improvement” (5325-21000-010-00D), Lead Scientist Dr. O. Anderson and (4) “Curation and Development of SoyBase and Its Integration with Other Plant Genome Databases” (3625-21000-038-00D), Lead Scientist Dr. R. Shoemaker, respectively, as well as related work carried out under the project (5) “Comparative Genomic Analyses, Bioinformatics and Resource Development for Cereal Genomes” (1907-21000-023-00D), Lead Scientist Dr. D. Ware. As noted in the collaborators portions of section IV. APPROACH AND RESEARCH PROCEDURES, we collaborate with these groups currently and plan to continue these very useful interactions.

Other relevant database projects include: (1) the Maize Genome Sequencing Consortium’s project to sequence maize B73’s data release site <http://www.maizesequence.org> with informatics project led by Dr. D. Ware), (2) Gramene (<http://www.gramene.org>; Dr. D. Ware), PlantGDB (<http://www.plantgdb.org>; Dr. V. Brendel), (3) PLEXdb (<http://www.plexdb.org>; Dr. R. Wise), (4) MAGI (<http://www.plantgenomics.iastate.edu/maize>; Dr. P. Schnable), (5) TIGR (<http://www.maize.tigr.org>; Dr. R. Buell), and (6) the Dana-Farber Cancer Institute’s Gene Indices project (<http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=maize>; Dr. J. Quackenbush). All of these research projects complement our proposed work in such a way that these projects are focused on subsets of maize data, whereas MaizeGDB is a general repository for all types of maize genetics and genomics information. As noted in the collaborators portions of section IV. APPROACH AND RESEARCH PROCEDURES, Lead Scientist Dr. C. Lawrence is aware of these projects and is often an active participant and/or collaborator.

One additional project that will become important for creating a collaboration is the Plant Science Cyberinfrastructure Collaborative (PSCIC), which is soon to be funded by the National

Science Foundation. As stated in the request for proposals in 2006, PSCIC aims to create a “new type of organization – a cyberinfrastructure collaborative for plant science” and presents an exciting opportunity to strengthen the collaboration and data integration between plant researchers. Such collaborative will produce novel data, and, once the PSCIC grant has been awarded, we will actively pursue collaborations with that project team to integrate to-be generated maize data into MaizeGDB and to better integrate MaizeGDB with other online resources.

IV. APPROACH AND RESEARCH PROCEDURES

One strength of the MaizeGDB project is steady stakeholder interest and input. The following objectives were developed in concert with the input from the MaizeGDB Working Group, a group of eleven maize geneticists and bioinformatics researchers who have interest in and expertise that is of use to the MaizeGDB project. Guidance from that group (the MaizeGDB Working Group Report from late 2006) is included in the attached Appendix. Note also that this past March (2007), a retreat was held for maize research investigators at the Allerton Conference Center in Illinois near the University of Illinois (Urbana-Champaign). Objectives listed here will begin to address a number of the concerns voiced by researchers in that meeting’s final report (available online at <http://www.maizegdb.org/AllertonReport.doc>).

Objective 1: Integrate new maize genetic and genomic data into the database.

- **Sub-objective 1.A.** *Expand mutant and phenotype data and tools.*

Hypothesis 1.A – This is not hypothesis-driven research.

Experimental Design: Gene function prediction and inference drive genetics research outcomes to promote applied research. The insights that can be gained about a gene’s function given a phenotype are unmatched by all other forms of functional characterization. Value-enhancing phenotypes are generated by natural, targeted, or random mutations to a gene or its regulatory elements directly, and by methods of modulating gene expression (e.g., RNA interference). In traditional “forward” genetics approaches, the usual route is to identify an interesting phenotype and pursue methods to determine the gene that causes the phenotype, while “reverse” genetics approaches identify a gene of interest then mutate that gene and observe the induced phenotypes. Although a wealth of information exists to utilize existing phenotype data to advance agricultural research, such information must be centralized or at least centrally available to facilitate effortless access. Access to and analysis of phenotype data requires that (1) storage of data be handled consistently with integration to related data types and (2) intuitive and consistent methods of access be implemented.

New phenotypic data, in addition to those available in peer-reviewed literature, are being generated from high-throughput projects, both in the past and currently underway. Curation of such maize data is essential to create a consistent, reliable, and accurate data resource. Thus far, MaizeGDB has been highly successful in the timely handling of various types of maize data (for example, MaizeGDB is the leading resource for published mutants of maize). Because it will be difficult for MaizeGDB’s small staff to

maintain this status without help from the community of maize geneticists, alternative approaches of data curation will be promoted via outreach: (1) the use of a set of **Community Curation Tools** will be encouraged to enable researchers to deposit their own small datasets into the database directly, and (2) **Automated Methods to Mine Literature** will be developed. Two **Large-scale Phenotypic Datasets** also will be prepared for inclusion in MaizeGDB by personnel working at MaizeGDB in collaboration with the groups that generated the data (see below).

Community Curation

The infrastructure of MaizeGDB is shown in Figure 2. The two methods that researchers can use to deposit data include direct data deposition (i.e., submission of datasets to personnel at MaizeGDB) and direct curation by use of the Community Curation Toolset. Shown in Figure 3 is a flow chart that demonstrates the process by which researchers decide which of the available methods of data deposition best suit their needs.

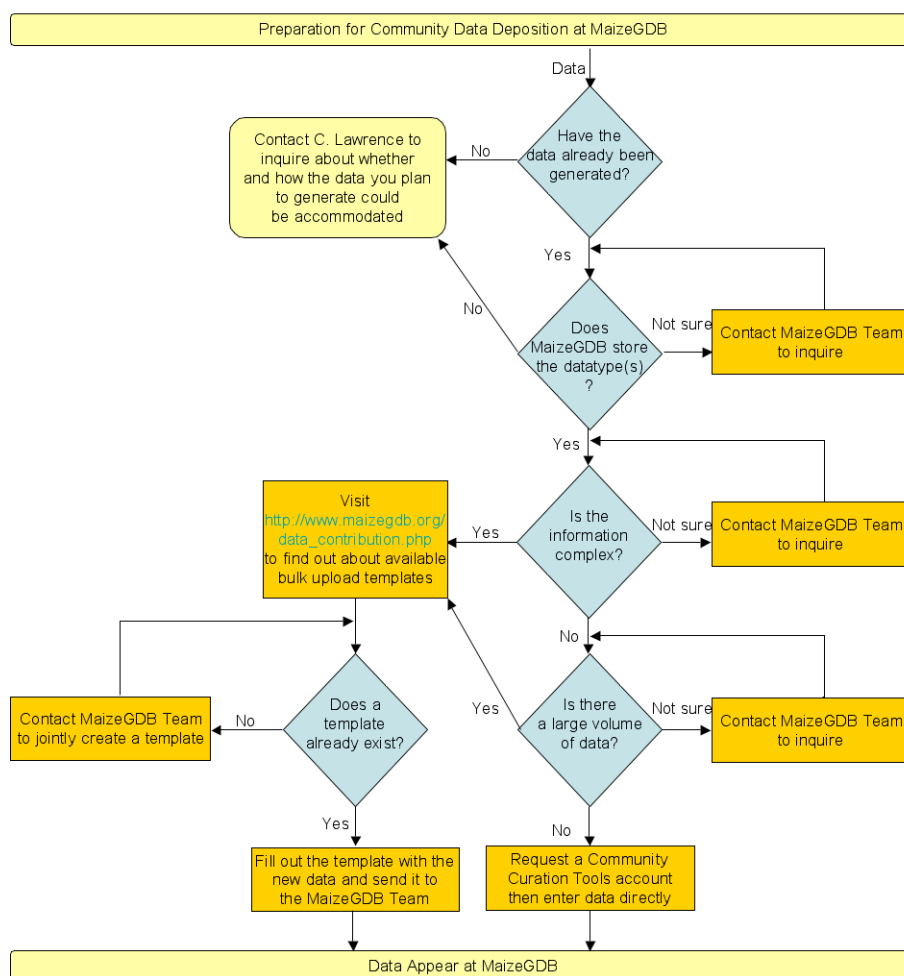


FIGURE 3: METHOD FOR COMMUNITY-DRIVEN DATA INPUT.

When researchers prepare to input data into MaizeGDB, a logical process can be followed to determine which methods of data deposition best suit the requirements for a particular dataset. Gold rounded rectangles represent the start and end points, blue diamonds represent decision points, and orange boxes represent processes. By following the decision tree shown, researchers can decide how to proceed at any given point in the data generation and deposition cycle.

Maize research community members will be trained in how to decide by which venue to deposit their data and how to use the Community Curation Tools via outreach training sessions. Each year, MaizeGDB personnel will visit no fewer than two sites where data

deposition training sessions will be conducted. Not only will this enable the tools to become more widely publicized, it will allow MaizeGDB personnel to work with researchers directly to ensure that the tools are appropriate for their needs.

To further encourage researchers to deposit their data into MaizeGDB, we will seek guidance from the MaizeGDB Working Group about inquiring with the editors of relevant journals (e.g., *Genetics*, *The Plant Cell*, *Crop Science*, etc.) regarding the possibility of working together to request that all maize data must be deposited into MaizeGDB prior to an article's publication. This is already enforced by nearly all journals to ensure that sequences are deposited into GenBank – a model that the editors might be willing to extend. Currently, NSF encourages grant-seekers who work on maize to transfer their generated data into MaizeGDB to ensure their long-term data preservation. A similar requirement from the journals will facilitate the dissemination of maize data through a centralized model organism database. Other methods for encouraging researchers to deposit their data into MaizeGDB via Community Curation also will be considered and pursued as resources allow.

Automated Methods to Mine Literature

With current personnel, we will explore the feasibility of adapting tools used by other groups to automatically mine the literature for data that should be curated. One example of software application for consideration is 'Textpresso' (Müller et al., 2004), which was developed for WormBase (<http://www.wormbase.org>; Bieri et al., 2007) but is adaptable to other MODs. Other approaches would not involve direct mining of literature, but would leverage the annotation in PubMed records where sequences are listed to references, but in this case, a curator likely would be required to make additional links to genes and to update the evidence codes for electronic Gene Ontology (GO) annotations associated to the gene models.

Large-Scale Phenotypic Datasets

In addition, current phenotypic datasets including (but not limited to) the Maize Inflorescence Architecture Project's EMS (Ethyl MethaneSulfonate) mutagenesis program (<http://gremlin1.gdcb.iastate.edu/MIP/EMSphenotypeDB/>) and the Maize Gene Discovery Project's RescueMu resources (<http://gremlin1.gdcb.iastate.edu/Mu/project/RescueMu/zmdb/RescueMuPhenotypeDB/>; Lunde et al., 2003), which are not present within MaizeGDB will be integrated into the database. To enable this integration, the datasets must be annotated with appropriate descriptors (i.e., the various Plant Ontologies' terms) and remodeled to fit the MaizeGDB schema. To enhance curational efforts in the future, we will work in concert with maize researchers as they devise methods for initial data storage to reduce the need for secondary curation once the data are migrated to the MaizeGDB resource for long-term storage.

Several high throughput "forward genetics" mutagenesis programs (utilizing both transposon and chemical lesion induction) are underway. As stated in the MaizeGDB Working Group's 2006 Guidance document (see Appendix), it is in the community interest that these studies be well-curated at MaizeGDB. Because MaizeGDB's staff is

not sufficient to curate all these projects' data directly, we will generate standards for data deposition and define file formats for automated inputs of these large datasets. This will be done in collaboration with the Stock Center to ensure that data comply with any distribution requirements, such as for those required for transgenics. This will enable researchers to be responsible for submitting their own data and will assist us to leverage their expertise for data curation. To inform researchers about when certain datasets will become available, we will develop and provide timelines for integration of key datasets.

Viewing Phenotypic Data

To enable the creation of biological connections by researchers via integrated views, mutant analyses and phenotype-genotype relationships will be made accessible via two types of interfaces: a "genome view" (described more fully under Sub-objectives 1.B and 1.C) and a "pathway view." While the "genome view" will allow researchers to visualize a gene within its genomic context, the "pathway view" will enable the visualization of a gene product within the context of relevant metabolic pathways annotated with Plant Ontology and Gene Ontology terms. Relevant microarray data also will be made accessible via linkages to PLEXdb, the Plant Expression Database (<http://plexdb.org/>; formerly BarleyBase; Shen et al., 2005), GEO, the Gene Expression Omnibus database (<http://www.ncbi.nlm.nih.gov/geo/>; Barrett et al., 2007), and the Maize Oligonucleotide Array Project's Zeamage database (<http://www.maizearray.org>; Gardiner et al., 2005) to enable researchers to probe the relationship between the expression levels and metabolic pathways. Because Zeamage's funding ends within the year, those data also are being migrated to MaizeGDB directly for long-term storage and display.

Stocks

The availability of phenotype data is especially useful in a research program if the seed stocks that demonstrate the phenotype are also easy to obtain. Maize has a tremendous number and quality of genetic stocks, and an excellent stock center (directed by Dr. Marty Sachs, ARS Urbana, Illinois; Scholl et al., 2003). In collaboration with Dr. Sachs (an ex officio member of the MaizeGDB Working Group), continued curation of stock data will be a top priority for the MaizeGDB staff as well as for the stock center personnel who will use the existing set of Professional Curation tools for data entry.

Contingencies: If the community of maize geneticists fails to adopt and utilize the Community Curation Tools for entering their data directly into the database, recently published phenotypic data will not be added to the database at a rate that would enable MaizeGDB to maintain its status as the leading resource for published mutants of maize. Because the funding agencies are now requiring that their awardees place their data into MaizeGDB, this contingency is not considered to be likely. However, if researchers fail to utilize the Community Curation Tools, new venues for showcasing the tools to researchers for their use will be sought out, and/or tools that researchers believe would be more useful for their purposes will be adopted. In instances where new types of phenotype data that do not fit into the current MaizeGDB schema are contributed, those data will only be incorporated into the database as dictated by the availability of resources.

Collaborations: Dr. M. Schaeffer, ARS, Columbia, Missouri, on integration of the various Plant Ontologies' terms with current and future mutant/phenotype records at MaizeGDB and preparation of association files for the Plant Ontology site, <http://www.plantontology.org>; Dr. C.R. Shyu, University of Missouri, Columbia, Missouri, on matching current descriptions with Plant Ontology terms and with potentially implementing the use of his image search tool that is currently under development; Drs. S. Hake and L. Harper, ARS, Albany, California, on integrating the EMS and RescueMu mutant phenotype descriptions with those that currently are present in the database; Dr. M. Sachs, ARS, Urbana, Illinois, on connecting phenotypes to stock data; Dr. R. Wise, ARS, Ames, Iowa, on connections to PLEXdb; Dr. R. Buell, The Institute for Genome Research on Zeamag; and Dr. O. Anderson, ARS, Albany, California, and Dr. R. Shoemaker, ARS, Ames, Iowa, on managing crop genome databases for small grains and legumes.

- **Sub-objective 1.B.** *Expand structural and genetic map sets emphasizing the:*
 - *Integration of the IBM genetic maps with the B73 genome sequence (This will be a shared objective with Mary Schaeffer, Columbia, Missouri, on ARS Project no. 3622-21000-027-00D, entitled “Genetic Mechanisms and Molecular Genetic Resources for Maize”);*
 - *Creation of views that convey the substantial variation in maize genome structure; and*
 - *Integration of the next-generation genetic map being generated by the Maize Diversity Project into a genomic view to enable its effective use by plant breeders (This will be a shared objective with Mary Schaeffer, Columbia, Missouri, on ARS Project no. 3622-21000-027-00D, entitled “Genetic Mechanisms and Molecular Genetic Resources for Maize”).*

Hypothesis 1.B – This is not hypothesis-driven research.

Experimental Design: MaizeGDB plays a key role in documenting, hosting, and curating many of the widely utilized maize genetic and cytogenetic maps that have been created over the last decade. While considerable genetic mapping will continue both for enhancing the representations of agronomically important traits and refining the assembled B73 genomic sequence, gene location information in MaizeGDB needs to evolve from a lower resolution genetic map-centric description to a higher resolution sequence-centric description of the gene's precise position on sequenced pseudomolecules. The main focus of MaizeGDB during this time of transition will be on linking relevant datasets, especially the centrally important maps such as (1) IBM2, (2) its neighbors that include approximate map locations for most mapped loci, and (3) the new maize diversity map to the genome sequence. This will enable the transition from a map-centric to a sequence-centric paradigm to occur seamlessly. To address this need, MaizeGDB personnel will create a “genome view” (mentioned also in Sub-objective 1.A) by adopting and customizing a Genome Browser (likely Gbrowse, Ensembl, or the UCSC browser) that could be used to integrate the outcomes of the Maize Genome Sequencing Project with existing and forthcoming map data.

The maize genome has an especially high level of DNA sequence polymorphism and extended regions of non-homology between inbred lines. Hence the diversity represented by the maize gene pool is unparalleled in both a phenotypic and molecular sense. This provides a unique vehicle to explore questions in evolution, domestication, development, trait expression, functional allelic diversity and the interrelated processes that shape such events and their outcomes. The potential for significant discovery via translational genomics exists through the application of new technologies and bioinformatic tools coupled with thorough phenotypic evaluation for useful traits and molecular characterization of diverse maize germplasm. The identification and evaluation of functional and evolutionarily important allelic variation needs to be turned into a comprehensive genomics activity. But such a goal is dependent on being able to associate diverse information in a uniform manner. To enable this activity, the MaizeGDB Genome Browser will display diversity data (like single nucleotide polymorphisms), knob locations, and other data that could be mapped onto the B73 sequence and map backbones. To aid integration of phenotypic and diversity data, and to make this information useful to applied researchers, the next-generation genetic map being generated by the Maize Diversity Project also will be incorporated into the Genome Browser to enable its effective use by plant breeders. This map will require representing some 15 million SNPs for several thousand genes, and ~1000 QTL candidate genes across 26 different inbreds, teosintes and landraces. In addition QTL for many agronomic traits are being defined for these lines, using B73 as the reference.

The choice of any particular genome browser software depends upon the needs of the maize community, which is made up of at least three distinct groups: basic, translational, and applied researchers. Meeting only the needs of basic researchers would be a necessary and useful first step. Planning to include views that will support translational and applied researchers will ensure that this tool will play a part in influencing crop development more directly. Although a single researcher might even include all of these three aspects in his/her research simultaneously, here we intentionally distinguish translational researchers (those working to determine the application of basic research outcomes for practical purposes) from applied researchers (who implement proven technologies to improve crops).

For genome browser functionality, basic researchers have an interest in visualizing genome structure, gene models, functional data, and genetic variability. Translational researchers would like to be able to assign values to genomic and genetic variants (e.g., the value of a particular allele in a given population) and to view those values within a genomic context. Applied researchers are interested in tagging variants for use as selectable markers and retrieving tags for particular regions of the genome. To further define these researchers' genome browser needs and preferences, we will seek community input through personal contacts, electronic mailing, and polls. Because each browser offers different strengths and weaknesses, we will seek to determine the community's consensus and work towards meeting their needs.

One major issue for the choice of the Genome Browser is the syntenic differences between maize inbred lines. Fu and Dooner (2002) recently reported gene colinearity

violations between the maize lines McC and B73. An interesting study in itself, Fu and Dooner's work inspires us to compare and analyze maize inbred lines not only individually, but also collectively using comparative genomics tools. The MaizeGDB Genome Browser should therefore contain a rich toolbox to allow sequence-based and gene-based comparisons between different inbred lines.

The three most popular and powerful genome browsers, Gbrowse, Ensembl, and the UCSC browser, will be evaluated for use by the MaizeGDB project. All three have genomic analysis and comparison tools for individual species as well as for comparative genomics views. These browsers offer slightly different capabilities. (1) *Gbrowse* (Stein et al., 2002) is an open source software suit written with Bioperl and BioSQL. Although it has an attractive visual display, its comparative genomics tools are limited to CMap (for comparison of physical and genetic maps), Sybil (multi-organism synteny viewer), and SynBrowse (Pan et al., 2005; synteny, homologous genes, conserved elements). (2) *The Ensembl Genome Browser* (Stalker et al., 2004) is equipped to contain and display assembled sequence, cross-species synteny, genes, transcripts, restriction enzymes, proteins, dot-plots, protein domains, and orthologs. It specifically gives links to gene/protein families in SCOP and PFAM, a functionality missing in other browsers. Overall, the Ensembl browser is highly integratable and portable as its open source code is written in Perl and MySQL, but it is not as fast as the UCSC browser. (3) In an attractive and compact display, *The UCSC browser* (Kuhn et al., 2007) provides annotation tracks, assembly, genomic sequences, predicted genes, cDNA, repetitive sequences, variation, expression, SNP, HapMap, as well as cross-species comparisons for alignment, conservation, and homologies. It also allows custom tracks. Most importantly, because the code is efficiently written in C++, it enables the fastest data visualization among all three browsers.

Contingencies: The choice of software to be adopted and customized for MaizeGDB's Genome Browser will be determined based upon results of community polls to determine desired functionalities, what data are made available by the Maize Genome Sequencing Consortium, and, more specifically, what data and software Dr. Ware's group makes publicly available as an outcome of their maize genome sequence analyses (see <http://www.maizesequence.org>). If Dr. Ware's group's implementation of Ensembl as a Genome Browser were found to be both sustainable and extensible for MaizeGDB's purposes, we would work to adopt their Genome Browser and adapt it to our purposes or to work with Dr. Ware's group to make linkages to <http://www.maizesequence.org> or <http://www.gramene.org> (depending upon that group's long-term funding situation). This would ensure that direct outcomes from the Maize Genome Sequencing Project would be utilized for MaizeGDB's purposes as much as possible. Whether and how quickly the Maize Diversity Project's next-generation genetic map could be incorporated into the Genome Browser will depend in part on the format of their data release and the extent that SNP data are submitted to GenBank. By working with Drs. Buckler and McMullen and others to ensure that the data can be readily integrated with other maps and with the maize genome sequence, we will minimize the lag time for the incorporation of these data into the database for viewing them through the Genome Browser.

Collaborations: Dr. D. Ware, ARS, Ithaca, New York (Cold Spring Harbor Laboratory) on maize sequence data integration and visualization; Dr. S. Clifton (Maize Genome Sequencing Project Coordinator), Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri on the progress of the maize sequencing project; Drs. M. Schaeffer, M. McMullen, ARS, Columbia, Missouri, and E. Buckler, ARS, Ithaca, New York on integrating map and diversity data into the Genome Browser; and Drs. O. Anderson, ARS, Albany, California, and R. Shoemaker, ARS, Ames, Iowa, on managing crop genome databases for small grains and legumes.

- **Sub-objective 1.C.** *Provide access to gene models calculated by leading gene structure prediction groups through the MaizeGDB interface.*

Hypothesis 1.C – This is not hypothesis-driven research.

Experimental Design: Raw genomic sequence data become more useful to biologists after careful annotation of information that gives the data functional biological meaning. A key first step toward annotating genomes is the *de novo* prediction of gene locations, and subsequently the prediction of individual gene structures (also referred to as ‘gene models’). Gene structure prediction can be accomplished by a variety of methods. Most often, gene models are predicted for newly sequenced genomes using only computational pattern-finding algorithms trained to recognize regions that ‘look like genes’ based upon their sequence features and/or genomic context. Next, transcript evidence (cDNA or EST information) is used to decisively discriminate gene structures (determination of precise transcription and exon and intron boundaries). In general, for those sequenced organisms that have a MOD, that database becomes the keeper of the ‘official’ set of gene models to which researchers can most easily refer.

At present, there are several groups with independent gene model annotations for maize. (1) Emerging BACs are annotated by both Dr. D. Ware’s group at the Cold Spring Harbor Laboratory (as a part of the Maize Genome Sequencing Project; <http://www.maizesequence.org/index.html>) and Dr. V. Brendel’s PlantGDB group at Iowa State University (<http://www.plantgdb.org/ZmGDB/>). (2) Genome Survey Sequence assemblies are generated and made available by Dr. Brendel’s group, Dr. P. Schnable’s group at Iowa State University (<http://magi.plantgenomics.iastate.edu/>), and Dr. R. Buell’s group at the Venter Institute/TIGR (<http://maize.tigr.org/release4.0/assembly.shtml>). These groups’ assemblies are referred to as GSS contigs (Dong et al., 2005), MAGIs (Fu et al., 2005), and AZMs, respectively. (3) PlantGDB, TIGR, and Dr. J. Quackenbush’s group at the Dana Farber Cancer Institute (<http://biocomp.dfci.harvard.edu/tgi/>) also provide EST assemblies (referred to as PUTs, TIGR Gene Indices, and The Gene Indices, respectively). Each of these groups has computed a slightly different method of sequence assembly and annotation. Currently, no group presents all assemblies and annotations within the context of the same Genome Browser for comparative purposes.

To address the need for a single “genome view” that includes all major genome assemblies and predicted gene structures, MaizeGDB will store the leading gene structure

prediction groups' maize genome annotations, and the MaizeGDB Genome Browser will be equipped to enable the display of those groups' data simultaneously. This will make it possible for the assemblies and annotations to be compared directly and will set the stage for MaizeGDB, the MOD for maize, to become the resource where the 'official' set of gene models for maize could reside. Based on the existing trend for the development of gene models of Arabidopsis, we expect that the gene models for maize will gradually improve over time, as more data and analyses become available. MaizeGDB is committed to include those additions/deletions in the official set of gene models and to continually enhance genomic representations of maize. These changes will be incorporated into yearly releases of major versions of Gene Models after year three (listed as Gene Models versions 2.0, 3.0, and 4.0 in the milestones table). If necessary, several minor updates will be released throughout the year (e.g., Gene Models version 3.2, 3.3, etc.).

Contingencies: Given that no funding agency has yet named a group to be the official maize genome annotators, it is possible that the list of collaborators would have to be expanded to include others. If that were to happen, new tactics to ensure collaboration with the official group would need to be devised. Alternatively, if many groups were to decide to work on this problem and various different annotations were generated and made available, it would not be possible for the MaizeGDB team to store and display representations from all groups given MaizeGDB's current resources and staff. In this case, the MaizeGDB Working Group would be asked to decide which groups' annotations should be represented at MaizeGDB. The release of versions of the gene models is also subject to the funding of genome annotators; however, as in the case of Arabidopsis, longer funding for improved annotations is highly likely when the genome of a model organism is made available.

Collaborations: Drs. D. Ware, ARS, Ithaca, New York (Cold Spring Harbor Laboratory) on maize sequence data integration and visualization; S. Clifton (Maize Genome Sequencing Project Coordinator), Genome Sequencing Center, Washington University School of Medicine, St. Louis, Missouri on the progress of the maize sequencing project; Drs. V. Brendel, Iowa State University, Ames, Iowa, P. Schnable, Iowa State University, Ames, Iowa, R. Buell, Venter Institute/The Institute for Genomic Research, Rockville, Maryland, and J. Quackenbush, Dana-Farber Cancer Institute/Harvard School of Public Health, Boston, Massachusetts on maize sequence annotation and gene model prediction; and Drs. O. Anderson, ARS, Albany, California, and R. Shoemaker, ARS, Ames, Iowa on managing crop genome databases for small grains and legumes.

- **Sub-objective 1.D.** *Compile and make accessible at MaizeGDB the annual Maize Newsletter.*

Hypothesis 1.D – This is not hypothesis-driven research.

Experimental Design: In the late 1920s it was recognized by the community of maize geneticists that the data they were recording needed organization, publication, and curation. To this end, R. A. Emerson and others began compiling the Maize Genetics

Cooperation-Newsletter (MNL) in 1929. The MNL is a compendium of notes and information on working research intended to be shared throughout the maize research community, and it is published to this day. Bits of information communicated through the MNL range from updates to a researcher's ongoing work which are not yet of the breadth or quality for peer-reviewed publication to notes on which field microscopes work best for gauging pollen viability. Articles submitted to the MNL may not be used as references without the consent of the authors, thus enabling the maize community to share their cutting-edge research with each other without fear of being quoted (if, for instance, it were to turn out that communicated ideas and findings were not fully developed). By its very nature, the MNL encourages cooperation – a hallmark of the maize research community.

At present, an endowment at the University of Missouri at Columbia supports the MNL's compilation and distribution to the more than six hundred researchers and libraries in hard copy each year. Its current co-editors are Dr. Schaeffer (ARS, Columbia) and Dr. Birchler (University of Missouri, Columbia). In addition to distribution in hard copy, volumes from 1977 onward are currently available online at MaizeGDB and linked to the Reference citations within MaizeGDB. In the coming years, not only will new volumes of the MNL be added to MaizeGDB, copies back to 1929 will be scanned and made available at MaizeGDB as well. In addition to making the pdf images of the past MNL available, character recognition software will be used to create a product that can be indexed for text-searching. Most of the past MNL notes are already incorporated as references in MaizeGDB, with links to MaizeGDB records, but not with links to the notes themselves.

Contingencies: If the community of maize researchers decides that newer technologies for communicating maize research accomplishments and issues should be adopted, MaizeGDB personnel will work with the community to develop and support newly defined communication needs.

Collaborations: Drs. M. Schaeffer, ARS, Columbia, Missouri, and J. Birchler, University of Missouri, Columbia, Missouri are co-editors of the Maize Newsletter. They will compile the print copy of the newsletter and send to Ames an electronic version to be formatted and made available electronically through MaizeGDB.

Objective 2: Provide community support services, such as lending help to the community of maize researchers with respect to developing and publicizing a set of guidelines for researchers to follow to ensure that their data can be made available through MaizeGDB; coordinating annual meetings; and conducting elections and surveys.

Hypothesis 2 – This is not hypothesis-driven research.

Experimental Design: The success of the maize research community in their endeavors to act 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. In the 2006

MaizeGDB Working Group Report (see Appendix), MaizeGDB is cited as playing “a central role in conducting central maize genetics community functions (i.e., with annual meetings, votes, surveys, and the Maize Newsletter).” The Working Group further states that “this role should continue as it is critical to the success and cohesion of the research direction for the community.”

In the coming years, personnel at MaizeGDB will provide the infrastructure to enable researchers to elect new members to the MGEC. This entails inviting researchers to nominate others for the ballot, confirming nominees’ willingness to run, 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, managing reports that enable the MGCSC to choose speakers from submissions, and creating the Conference Program.

In addition, we will work with researchers directly to jointly develop and subsequently publicize a set of guidelines for them to follow to ensure that their data can be made available through MaizeGDB. These guidelines will include descriptions on what types of data can be taken in directly, links to standards for data deposition and defined file formats for automated inputs (as mentioned in Sub-objective 1A), and a set of timelines for how long it will take to incorporate data into the database that will include a list of focus data types for curation activities by month.

Finally, the MaizeGDB Working Group will meet once yearly, and the MaizeGDB Team will create a document for the Working Group that outlines the previous period’s progress. Any reports or guidance from the Working Group will be considered based on the availability of resources, and the MaizeGDB responses to their guidance will be drafted for dissemination.

Contingencies: If the MGCSC or the MGEC were to 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 develop solutions for the MaizeGDB Working Group’s “Medium Priority” tasks (see the Working Group’s November 2006 Guidance document at http://www.maizegdb.org/working_group.php), which include improving the representation of diversity data and the development of a feedback mechanism for researchers to improve upon gene model predictions. Another possible avenue to explore in such a case would be to look into including proteomic data in MaizeGDB and finding ways to connect proteomic data with genetic data with data representations made available via the Genome Browser.

Collaborations: The MGEC (see <http://www.maizegdb.org/mgec.php> for current membership and affiliations) on keeping abreast of current needs communicated by the community, The MGCSC (see the most recent conference site listed at http://www.maizegdb.org/maize_meeting/ for current membership and affiliations) on facilitating Maize Meeting planning and operations, and Drs. O. Anderson, ARS, Albany,

California, and R. Shoemaker, ARS, Ames, Iowa on managing crop genome databases for small grains and legumes.

Shared Features among Objectives

Throughout this project plan, shared features emerge among the various tasks described. The interrelationships of our sub-objectives are shown in Figure 4. A primary objective is to link experimentally observed functional genome annotation such as phenotypes (Sub-Objective 1A) with electronic expert-defined gene models and electronic annotations (Sub-Objective 1C); and with tools for breeders, such as genetically mapped markers and traits (Sub-Objective 1B). Much of these data are available in an early form in the notes for the MNL (Sub-Objective 1D), or the abstracts provided for the Maize Meeting (Objective 2).

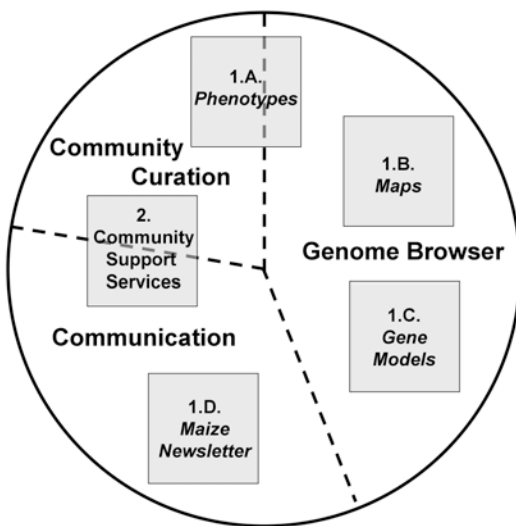


FIGURE 4: THEMES SHARED AMONG OBJECTIVES.

Three major features of the Project Plan include (clockwise from top right): the need to implement a multifunctional Genome Browser, the need to facilitate Communication among maize researchers, and the need to encourage and facilitate Community Curation. Overlaid onto these features are gray boxes that represent the Objectives and Sub-objectives described above. By identifying this contextual framework that underlies the tasks to be accomplished, the importance of these major areas of concentration becomes evident. Awareness of these three themes will drive the development of resources to meet researchers' needs and helps to unite these initiatives by creating common overarching goals.

V. PHYSICAL AND HUMAN RESOURCES

Cat 1 personnel:

Ames, IA

Dr. Carolyn Lawrence, Lead Scientist (Research Geneticist; under the supervision of **Dr. L. Lewis**)

Non-Cat 1 personnel funded by ARS:

Ames, IA

Dr. Taner Sen, SY (Cat 4 Computational Biologist; under the supervision of Dr. L. Lewis)

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

Trent Seigfried, Bioinformatics Engineer (Information Technology Specialist; under the supervision of Dr. C. Lawrence)

Columbia, MO

Dr. Mary Schaeffer, SY (Cat 4 Curator/Geneticist; under the supervision of Dr. M. Oliver) is 1.0 FTE on ARS Project no. 3622-21000-027-00D, entitled "Genetic Mechanisms and

Molecular Genetic Resources for Maize.” Dr. M. Schaeffer’s contributions to Objective 1 are jointly listed on that Project Plan as well as in a letter of collaboration from her and Lead Scientist Dr. M. McMullen along with her Milestones and previous accomplishments.

Albany, CA

Dr. Lisa Harper (Cat 3 Curator/Geneticist; located at the Plant Gene Expression Center, Albany, CA; under the supervision of Dr. C. Lawrence in collaboration with Dr. S. Hake) is 0.5 FTE on this CRIS. Her contributions to Objective 1 are shown in this Project Plan and are confirmed in an attached joint letter of collaboration with Dr. S. Hake.

Offices:

Ames, IA

Ca. 800 ft² located in room 526 on 5th floor, Science II at Iowa State University, a floor shared with members of the Genetics, Developmental, and Cell Biology department at ISU. Room contains workstation equipment for five workers, plus additional space for one of the current MaizeGDB servers.

Ca. 300 ft² located in room 1565 on 1st floor, Agronomy Hall at Iowa State University, a floor shared with members of the Agronomy Department’s Plant Breeding and Genetics Panel.

Within the next 6 months, the MaizeGDB group in Ames (i.e., Lawrence, Sen, Campbell, and Seigfried) as well as the SoyBase (Dr. R. Shoemaker; ARS), PLEXdb (Dr. R. Wise; ARS), and PlantGDB (Dr. V. Brendel; Iowa State University) groups will be moving to the Crop Genome Informatics Laboratory, a building on the Iowa State University campus that is currently under renovation. These groups will share ca. 6,000 ft² of space including offices and a state-of-the-art server room.

Columbia, MO

Primary office is about 240 ft² in 203 Curtis Hall. Additional Office space in 210 Curtis is used for part-time staffing, file cabinets, scanning, and printing. It contains three workstations used for this project and several scanners.

Albany, CA

Shared office space (cubicles) that house the administrative staff for the Plant Gene Expression Center includes one cubicle (ca. 70 ft²) for MaizeGDB personnel.

Server Rooms:

Ames, IA

Ca. 150 ft² climate controlled room in Molecular Biology Building, shared with Dr. V. Brendel (collaborator). Contains two production MaizeGDB servers. Upon relocation to the Crop Genome Informatics Laboratory, approximately 180 ft² in additional server space will become available for use.

Columbia, MO

A 135 ft² climate controlled room in Curtis Hall that is shared with other ARS investigators in the Plant Genetics Research Unit. It contains servers equipped with mapping softwares and

databases used in mapping data intake, and is equipped with a UPS. A second climate controlled room in the TelCom building, equipped with UPS and a generator, houses a server for the staging copy of the Maize Newsletter and the community mapping service and also a server for Ames Disaster Recovery (see below). The server for backups of Curtis Hall data is maintained at the TelCom building. The server for backups of data at the TelCom building is maintained in Curtis Hall.

Machines:**Ames, IA**

Three Dell rack mount servers. PowerEdge 2850, Processor: Dual 2.8GHz/2x2MB Cache Xeon 800MHz FSB Memory: 4GB DDR2 400MHz 4-146Gb 10K rpm hard drive. Oracle Database Oracle license is perpetual for the life of the project, and annual support contract fees are budgeted yearly.

The Apple Workgroup Cluster is a 16 node Dual Processor G5 Xserve Apple Work Group Cluster for Bioinformatics. Each cluster node consists of two 64-bit G5 2.0 GHz processors, 2 GB of RAM and an 80 GB hard drive. The controlling Xserve has 4 GB RAM, two 250 GB internal drives and 2TB of attached storage in an Xserve Raid. The installed iNquiry software package has over 170 bundled bioinformatics applications and a convenient web interface.

A Network Attached Storage device has roughly 800GB of usable space and is the Ames Disaster Recovery storage location for a compressed copy of the MaizeGDB database.

Columbia, MO

A Dell Power Vault 705N server, with 200 GB drive at Columbia, MO is available for use by the Ames IT team for the purpose of Disaster Recovery. A compressed copy of the Curation Database is copied to the Columbia off-site server on a weekly basis. Backups of Columbia, MO data, including Desktops, are on a Dell Power Edge 750 with Fedora Core 4 operating system.

VI. 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 Sen, Schaeffer, Harper, Campbell, and Seigfried to accomplish Objectives 1 and 2. For Objective 1A, 1C, and 1D, Campbell will create efficient database representations and Seigfried will develop the Web interface code to ensure proper outcomes to complex queries. Seigfried will also be responsible for generating Web-based polls, gathering data, and presenting them in a user-friendly format. Campbell and Seigfried will work under the guidance of Lawrence to accomplish these tasks. For Objective 1B, Schaeffer and Harper will continue the ongoing curation and integrate IBM genetic maps with the B73 genome sequence in coordination with Lawrence. For Objective 1B, Sen will communicate and receive input from the maize community and implement a Genome Browser that will communicate with the MaizeGDB databases and Web interface in coordination with the MaizeGDB team. Lawrence will be responsible for coordinating all 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 Working Group, and the Executive Committee. 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 between Ames, IA, Columbia, MO, and Albany, CA to assess progress. 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 Working Group once a year formally, and communicates with other maize researchers informally numerous other times at conferences and through ad hoc phone calls and e-mail.

VII. MILESTONES AND EXPECTED OUTCOMES

Project Title ^a		The Maize Genetics and Genomics Database		Project No. ^b	3625-21000-045-00D <i>See also letter of collaboration from McMullen and Schaeffer for 3622-21000-027-00D</i>
National Program ^c		301, Plant Genetic Resources, Genomics, and Genetic Improvement			
Objective ^d		1. Integrate new maize genetic and genomic data into the database			
Subobjective ^e		1.A. Expand mutant and phenotype data and tools			
NP Action Plan Component ^f		2: Crop Informatics, Genomics, and Genetic Analyses			
NP Action Plan Problem Statement ^g		2A: Genome Database Stewardship and Informatics Tool Development 2B: Structural Comparison and Analysis of Crop Genomes 2C: Genetic Analyses and Mapping of Important Traits			
Hypothesis ^h	SY Team ⁱ	Months	Milestones ^j	Progress/Changes ^k	Products ^l
This is not hypothesis-driven research	CL	12	Visit no fewer than 2 sites where researchers will be trained in the use of Community Curation Tools		
	TZS	12	Evaluate Textpresso and other softwares for mining literature		
	CL	24	Visit no fewer than 2 sites where researchers will be trained in the use of Community Curation Tools		
	CL	24			RescueMu and EMS data fully incorporated into and available via MaizeGDB
	CL	36	Visit no fewer than 2 sites where researchers will be trained in the use of Community Curation Tools		
	CL	36	Implement automated literature mining software		
	CL, TZS	36	Evaluate existing software for creating Pathway View		
	CL, TZS	48	Pathway View software selected		
	CL	48	Visit no fewer than 2 sites where researchers will be trained in the use of Community Curation Tools		
	CL	60	Visit no fewer than 2 sites where researchers will be trained in the use of Community Curation Tools		
	CL, TZS	60	Pathway View released at MaizeGDB		Publicly accessible MaizeGDB Pathway Viewer
	CL	60			Report functionality of MaizeGDB Pathway Viewer in a peer-reviewed journal

Project Title ^a		The Maize Genetics and Genomics Database		Project No. ^b		3625-21000-045-00D <i>See also letter of collaboration from McMullen and Schaeffer for 3622-21000-027-00D</i>	
National Program ^c		301, Plant Genetic Resources, Genomics, and Genetic Improvement					
Objective ^d		1. Integrate new maize genetic and genomic data into the database					
Subobjective ^e		1.B. Expand structural and genetic map sets					
NP Action Plan Component ^f		2: Crop Informatics, Genomics, and Genetic Analyses					
NP Action Plan Problem Statement ^g		2A: Genome Database Stewardship and Informatics Tool Development 2B: Structural Comparison and Analysis of Crop Genomes 2C: Genetic Analyses and Mapping of Important Traits					
Hypothesis ^h	SY Team ⁱ	Months	Milestones ^j	Progress/ Changes ^k		Products ^l	
This is not hypothesis-driven research	TZS, CL	12	Select software for MaizeGDB Genome Browser				
	TZS, CL	24	Release first version of Genome Browser to display the Maize Sequencing Project's B73 sequence annotations			Publicly accessible MaizeGDB Genome Browser	
	TZS, CL	36	Add functionality to show diversity data, microarray probes, and ontology terms				
	CL, TZS	36	Utilize maize genome sequence for creating hypotheses about the maize genome architecture and test hypotheses using bioinformatic techniques				
	CL	48				Report findings on genome architecture findings in a peer-reviewed journal	
	CL	48				Report functionality of MaizeGDB Genome Browser in a peer-reviewed journal	

Project Title ^a		The Maize Genetics and Genomics Database		Project No. ^b		3625-21000-045-00D	
National Program ^c		301, Plant Genetic Resources, Genomics, and Genetic Improvement					
Objective ^d		1. Integrate new maize genetic and genomic data into the database					
Subobjective ^e		1.C. Provide access to gene models calculated by leading gene structure prediction groups through the MaizeGDB interface					
NP Action Plan Component ^f		2: Crop Informatics, Genomics, and Genetic Analyses					
NP Action Plan Problem Statement ^g		2A: Genome Database Stewardship and Informatics Tool Development 2B: Structural Comparison and Analysis of Crop Genomes 2C: Genetic Analyses and Mapping of Important Traits					
Hypothesis ^h	SY Team ⁱ	Months	Milestones ^j	Progress/ Changes ^k		Products ^l	
This is not hypothesis-driven research	TZS, CL	24	Provide access to Maize Genome Sequencing Project's predicted gene structures via the MaizeGDB Genome Browser			Predicted gene structures represent maize's 'Official Gene Models, version 1.0'	

	TZS, CL	24	Set up collaboration with the group(s) who are officially annotating the maize genome		
	TZS, CL	36	Equip Genome Browser to show multiple groups' genome assemblies		Release Gene Models version 2.0
	TZS, CL	48	Equip Genome Browser to show multiple groups' improvements in genome assemblies		Release Gene Models version 3.0
	TZS, CL	60	Equip Genome Browser to show improvements in multiple groups' genome assemblies		Release Gene Models version 4.0

Project Title ^a		The Maize Genetics and Genomics Database		Project No. ^b		3625-21000-045-00D See also Letter of collaboration from McMullen and Schaeffer for 3622-21000-027-00D	
National Program ^c		301, Plant Genetic Resources, Genomics, and Genetic Improvement					
Objective ^d		1. Integrate new maize genetic and genomic data into the database					
Subobjective ^e		1.D. Compile and make accessible at MaizeGDB the annual Maize Newsletter					
NP Action Plan Component ^f		2: Crop Informatics, Genomics, and Genetic Analyses					
NP Action Plan Problem Statement ^g		2A: Genome Database Stewardship and Informatics Tool Development					
Hypothesis ^h		SY Team ⁱ	Months	Milestones ^j	Progress/ Changes ^k		Products ^l
This is not hypothesis-driven research	CL	12	Add the electronic version of the MNL vol. 83 to MaizeGDB			MNL vol. 83 available online	
	CL	24	Add the electronic version of the MNL vol. 84 to MaizeGDB			MNL vol. 84 available online	
	CL	36	Add the electronic version of the MNL vol. 85 to MaizeGDB			MNL vol. 85 available online	
	CL	48	Add the electronic version of the MNL vol. 86 to MaizeGDB			MNL vol. 86 available online	
	CL	60	Add the electronic version of the MNL vol. 87 to MaizeGDB			MNL vol. 87 available online	

Project Title ^a	The Maize Genetics and Genomics Database			Project No. ^b	3625-21000-045-00D
National Program ^c	301, Plant Genetic Resources, Genomics, and Genetic Improvement				
Objective ^d	2. Provide community support services				
Subobjective ^e					
NP Action Plan Component ^f	2: Crop Informatics, Genomics, and Genetic Analyses				
NP Action Plan Problem Statement ^g	2A: Genome Database Stewardship and Informatics Tool Development				
Hypothesis ^h	SY Team ⁱ	Months	Milestones ^j	Progress/ Changes ^k	Products ^l
This is not hypothesis-driven research	CL	12	Conduct annual MGEC elections		

08/03/07

301 Lawrence 3625-21000-045-00D PrePlan

	CL	12			Submit database update manuscript to Nucleic Acids Research, a peer-reviewed journal
	CL	12	Conduct annual Working Group meeting		Report and response documents compiled for the Working Group and available at MaizeGDB
	CL	12	Manage Maize Meeting Abstract submission and compilation		Book of abstracts for Maize Meeting
	CL	12	Guidelines for depositing data at MaizeGDB completed		Guidelines will be updated at http://www.maizegdb.org/data_contribution.php
	CL	24	Conduct annual MGEC elections		
	CL	24	Conduct annual Working Group meeting		Report and response documents compiled for the Working Group and available at MaizeGDB
	CL	24	Manage Maize Meeting Abstract submission and compilation		Book of abstracts for Maize Meeting
	CL	36	Conduct annual MGEC elections		
	CL	36			Submit database update manuscript to Nucleic Acids Research, a peer-reviewed journal
	CL	36	Conduct annual Working Group meeting		Report and response documents compiled for the Working Group and available at MaizeGDB
	CL	36	Manage Maize Meeting Abstract submission and compilation		Book of abstracts for Maize Meeting
	CL	48	Conduct annual MGEC elections		
	CL	48	Conduct annual Working Group meeting		Report and response documents compiled for the Working Group and available at MaizeGDB
	CL	48	Manage Maize Meeting Abstract submission and compilation		Book of abstracts for Maize Meeting
	CL	60	Conduct annual MGEC elections		
	CL	60			Submit database update manuscript to Nucleic Acids Research, a peer-reviewed journal
	CL	60	Conduct annual Working Group meeting		Report and response documents compiled for the Working Group and available at MaizeGDB
	CL	60	Manage Maize Meeting Abstract submission and compilation		Book of abstracts for Maize Meeting

The goal of the table is to present a summary of the project in a form that is easily used to link to Annual Report of Progress (421's) and Performance Plans for each scientist. The intent of the table is to be a dynamic representation of the project that captures over the project life cycle the important progress and products derived from the project.

Note: Table can be expanded by copying any section below the project title line.

Explanation of Footnotes

^a **Project Title** from the project plan

^b **Project Number** from the ARS-416

^c Number and name of the primary **National Program**

^d **Objective** from the project plan

^e **Subobjective** from project plan (*if used, if not this line can be deleted*)

^f **Component(s)** from the National Program Action Plan that can be used to identify the component being addressed for each objective or subobjective

^g **Problem Statement(s)** from the National Program Action Plan that can be used to identify the problem being addressed for each objective or subobjective

^h A statement of the **hypothesis** for the objective, if appropriate. Otherwise the non-hypothesis statement

ⁱ Initials of the **project team members** contributing expertise to the specific hypothesis and significant collaborators (if a vacancy exists on the project, identify this position within the table)

^j **Milestones** for the specific months of the project, be as specific as possible as to the measurable milestones

^k The **Progress/Changes** section is completed at the end of each year by the project team as part of the Project Management and Evaluation process and a summary of these are entered into the table. When there is a revised milestone or hypothesis this is entered for the next period of the project plan.

^l Specific **products** of the project for each hypothesis line.

VIII. ACCOMPLISHMENTS FROM PRIOR PROJECT PERIOD

1. **Terminating ARS Research Project Number:** 3625-21000-045-00D

2. **Title:** Database of Maize Genome Information

3. **Project Period:** April 3, 2004 to October 31, 2008

4. **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 (new hire June 2007)
 Elisabeth C. Harper – 0.50 FTE Cat 3 Geneticist (new hire February 2007)
 Darwin A. Campbell – 1.00 FTE Cat 6 Information Technology Specialist
 Trent E. Seigfried – 1.00 FTE Cat 6 Information Technology Specialist
 Leslie C. Lewis – 0.03 FTE Cat 1 Supervisory Research Entomologist (Research Leader)

Collaborating in Columbia, Missouri:

Mary L. Schaeffer – 1.00 FTE Cat 4 Geneticist

5. **Project Accomplishments and Impact:**

Successfully combining the data previously stored at MaizeDB and ZmDB into a single resource has been the single most important accomplishment made by the MaizeGDB team over the course of the project.

Other important accomplishments include: (1) creation of the Web interface to the database (accessible at <http://www.maizegdb.org>), (2) creation of multiple pipelines to facilitate interaction with the community of maize geneticists, enabling the development of data manipulation tools designed by community members for all to use, (3) development of multiple sets of tools to enable curators and the general public to modify and update data, (4) development and implementation of the sequence pipeline that populates the MaizeGDB sequence tables with sequences derived from PlantGDB on a monthly basis, and (5) management of the various training and informative opportunities that have been provided to the wider community of plant biologists including students, through meetings including the International Plant and Animal Genome Conference, the Annual Maize Genetics Conference, and one-time workshops such as the Maize Genetics, Genomics, and Bioinformatics Workshop held at CIMMYT, Mexico in 2004.

Relative to the Maize Genome Sequencing Project that is now underway, MaizeGDB personnel have: (1) created a Maize Genome Sequencing Information Center (<http://www.maizegdb.org/genome/>), which served as a clearinghouse for data pertinent to the maize genome sequencing request for proposals, (2) met with sequencing project personnel just prior to the funding announcement to plan methods of collaboration and integrated accesses to data, (3) traveled to work with collaborator D. Ware's group to coordinate data accesses to <http://www.maizesequence.org> via MaizeGDB, (4) created a sequencing project information page (accessible at

http://www.maizegdb.org/sequencing_project.php), and (5) provided datasets to be overlaid onto the assemblies at <http://www.maizesequence.org>.

Other services we have provided that have been much appreciated by stakeholders include: (1) the creation and management of a MaizeGDB Editorial Board webpage (http://www.maizegdb.org/editorial_board.php), which hosts a journal club-like list of stellar publications for researchers to access, (2) creation of election forms and management of Maize Genetics Executive Committee elections, (3) conducting the Maize Genetics Executive Committee's survey of maize geneticists' top priorities using forms created by MaizeGDB personnel, and (4) coordination of informatics tools and services to support the Annual Maize Genetics Conference (including an abstract submission form as well as facilitating interactions among the MGCSC members).

In addition to the MaizeGDB resource itself, MaizeGDB personnel created and released the Morgan2McClintock Translator, which converts genetic map locations to physical (cytological) map positions and vice versa. Use of this tool enables researchers to merge various maps and map types, helping to generate an integrated overview of the maize genome. The tool is available online at <http://www.lawrencelab.org/Morgan2McClintock/>. Its availability allows researchers to move between map types. For example, many of the chromosomal breakpoints for deficiency and translocation stocks are only mapped cytologically and are not associated with gene locations. The Morgan2McClintock Translator permits researchers to find the location of the breakpoints relative to genes and to use those stocks to test their functional genomics hypotheses. This has a direct impact on research reducing the number of lines to be tested (thus enabling them to spend time on other work) by minimizing the size of genomic regions that must be evaluated experimentally. The tool also may be useful in determining the size of gaps in maize genome assemblies.

6. How Past Objectives and Accomplishments Relate to the Proposed Objectives:

The proposed work to be pursued in the next five years will substantially build upon the tools that were developed over the course of the past five years, such as the existing MaizeGDB infrastructure, public Web interface, and curation tools. Past work to integrate maps and other data with sequence information will serve as a basis for current efforts, and we will continuously track maize community needs by communicating with maize researchers, and specifically by cooperating closely with the Maize Genome Sequencing Consortium, MaizeGDB Working Group, MGCSC, and MGEC. By responding to their inputs and feedback, personnel working at MaizeGDB will be able to maintain the project's agility and ability to meet stakeholders' needs as they evolve.

IX. LITERATURE CITED

- Barrett, T, Troup, DB, Wilhite, SE, Ledoux, P, Rudnev, D, Evangelista, C, Kim, IF, Soboleva, A, Tomashevsky, M, Edgar, R (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucl. Acids Res.*, 35, D760-D765.
- Beavis, WD, Gessler, DD, Rhee, SY, Rokhsar, DS, Main, D, Mueller, LA, Stein, LD, Huala, E, Lawrence, CJ (2005) Plant Biology Databases: A Needs Assessment (Advisory Whitepaper for the NSF, DOE, and USDA). <http://www.maizegdb.org/PDBNeeds.pdf>.
- Bennetzen, J (2001) The Maize Genetics Executive Committee (MGEC). *Maize Genetics Cooperation Newsletter*, 75:v-vi.
- Benson, DA, Karsch-Mizrachi, I, Lipman, DJ, Ostell, J, Wheeler, DL (2007) GenBank. *Nucl. Acids Res.*, 35, D21-D25.
- Bieri, T, Blasiar, D, Ozersky, P, Antoshechkin, I, Bastiani, C, Canaran, P, Chan, J, Chen, N, Chen, WJ, Davis, P, Fiedler, TJ, Girard, L, Han, M, Harris, TW, Kishore, R, Lee, R, McKay, S, Muller, HM, Nakamura, C, Petcherski, A, Rangarajan, A, Rogers, A, Schindelman, G, Schwarz, EM, Spooner, W, Tuli, MA, Van Auken, K, Wang, D, Wang, X, Williams, G, Durbin, R, Stein, LD, Sternberg, PW, Spieth, J (2007) WormBase: new content and better access. *Nucl. Acids Res.*, 35, D506-D510.
- Chan, A, Perte, G, Cheung, F, Lee, D, Zheng, L, Whitelaw, C, Pontaroli, A, SanMiguel, P, Yuan, Y, Bennetzen, J, Barbazuk, WB, Quackenbush, J, Rabinowicz, PD (2006) The TIGR Maize Database. *Nucl. Acids Res.*, 34, D771-D776.
- Dong, Q, Lawrence, CJ, Schlueter, SD, Wilkerson, MD, Kurtz, S, Lushbough, C, and Brendel, V (2005) Comparative Plant Genomics Resources at PlantGDB. *Plant Physiol.*, 139, 610-618.
- Fu, H. and Dooner, H (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA*, 99, 9573-9578.
- Fu Y, Emrich SJ, Guo L, Wen TJ, Ashlock DA, Aluru S, Schnable PS (2005) Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. *Proc. Natl. Acad. Sci. U S A*. 102(34):12282-12287.
- Gardiner, JM, Buell, CR, Elunwh, R, Galbraith, DW, Henderson, DA, Iniguez, AL, Kaeppler, SM, Kjni, JJ, Liu, J, Sndth, A, Zheng, L, Chandler, VL (2005) Design, production, and utilization of long oligonucleotide microarrays for expression analysis in maize. *Maydica* 50(3-4):425-435.
- Kuhn, RM, Karolchik, D, Zweig, AS, Trumbower, H, Thomas, DJ, Thakkapallayil, A, Sugnet, CW, Stanke, M, Smith, KE, Siepel, A, Rosenbloom, KR, Rhead, B, Raney, BJ, Pohl, A, Pedersen, JS, Hsu, F, Hinrichs, AS, Harte, RA, Diekhans, M, Clawson, H, Bejerano, G, Barber, GP, Baertsch, R, Haussler, D, Kent, WJ (2007) The UCSC genome browser database: update 2007. *Nucl. Acids Res.*, 35, D668-D673.
- Lawrence, CJ, Schaeffer, ML, Seigfried, TE, Campbell, DA, Harper, LC (2007) MaizeGDB's new data types, resources, and activities. *Nucl. Acids Res.*, 35, D895-D900.
- Lunde, CF, Morrow, DF, Roy, LM, Walbot, V (2003) Progress in maize gene discovery: a project update. *Funct. Integr. Genomics*, 31(1-2):25-32.
- Mueller, LA, Solow, TH, Taylor, N, Skwarecki, B, Buels, R, Binns, J, Lin, C, Wright, MH, Ahrens, R, Wang, Y, Herbst, EV, Keyder, ER, Menda, N, Zamir, D, Tanksley, SD (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol.*, 138(3), 1310-1317.

- Müller, HM, Kenny, EE, Sternberg, PW (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* 2(11):e309.
- Pan, X, Stein, L, Brendel, V (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, 21(17), 3461-3468.
- Rhee, SY, Beavis, W, Berardini, TZ, Chen, G, Dixon, D, Doyle, A, Garcia-Hernandez, M, Huala, E, Lander, G, Montoya, M, Miller, N, Mueller, LA, Mundodi, S, Reiser, L, Tacklind, J, Weems, DC, Wu, Y, Xu, I, Yoo, D, Yoon, J, Zhang, P (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucl. Acids Res.*, 31(1), 224-228.
- Scholl, R, Sachs, M, Ware, D (2003) Maintaining collections of mutants for plant functional genomics. *Methods Mol. Biol.*, 236, 311-326.
- Shen, L, Gong, J, Caldo, RA, Nettleton, D, Cook, D, Wise, RP, Dickerson, JA (2005) BarleyBase--an expression profiling database for plant genomics. *Nucl. Acids Res.*, 33, D614-618.
- Stalker, J, Gibbins, B, Meidl, P, Smith, J, Spooner, W, Hotz, HR, Cox, AV (2004) The Ensembl Web site: mechanics of a genome browser. *Genome Res.*, 14, 951-955.
- Stein, LD, Mungall, C, Shu, S, Caudy, M, Mangone, M, Day, A, Nickerson, E, Stajich, JE, Harris, TW, Arva, A, Lewis, S (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, 12, 1599-1610.

X. PAST ACCOMPLISHMENTS OF INVESTIGATORS – CAROLYN J. LAWRENCE

EDUCATION

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

WORK EXPERIENCE

2003-2005	Postdoctoral Research (Bioinformatics)	Iowa State University, Ames, IA
2005-present	GS-12 Cat 1 Research Geneticist	USDA-ARS CICGR, Ames, IA
	Collaborator/Assistant Professor	Department of Genetics, Development and Cell Biology, Iowa State Univ.

ACCOMPLISHMENTS

Previous ARS and competitive funding has been used to improve data accessibility and analysis tools for flowering plants. We have: (1) created novel methods for maize data access and analysis via MaizeGDB (Lawrence et al., 2007; Lawrence et al., 2005; and Lawrence et al., 2004), (2) generated the Morgan2McClintock Translator to enable integration of genetic and cytological maps for both maize and tomato (Lawrence et al., 2006), and (3) enabled comparative genetics via the creation and continued development of the PlantGDB resource (Dong et al., 2005). In addition, maize's appropriateness for use as a model for biofuel grass development also was investigated, confirmed, and reported (Lawrence and Walbot, in press). Recently C. Lawrence has consulted on the planned updates to the ARS Germplasm Resource Information Network (GRIN) project's eminent migration to a new platform to support the storage of molecular data and leads a joint ARS/NSF initiative to provide plant genomics outreach to Native Americans (see <http://www.lawrencelab.org/Outreach/2006/home.html>).

PEER REVIEWED PUBLICATIONS (underline indicates corresponding author)

Lawrence, C.J. and Walbot, V. Maize as a model for bioenergy production from fuelstock grasses. *The Plant Cell*, in press.

Lawrence, C.J., Schaeffer, M.L., Seigfried, T.E., Campbell, D.A., and Harper, L.C. MaizeGDB's new data types, resources, and activities (2007). *Nucleic Acids Research* 35(Database issue):D895-900.

Stein, L.D., Beavis, W.D., Gessler, D.D., Huala, E., **Lawrence, C.J.**, Main, D., Mueller, L.A., Rhee, S.Y., and Rokhsar, D.S. Save our data! (2006). *The Scientist* 20(4):24-25.

Lawrence, C.J., Seigfried, T.E., Bass, H.W., and Anderson, L.K. Predicting chromosomal locations of genetically mapped loci in maize using the Morgan2McClintock Translator (2006). *Genetics* 172(3):2007-2009.

Dong, Q., **Lawrence, C.J.**, Schlueter, S.D., Wilkerson, M.D., Kurtz, S., Lushbough, C., and **Brendel, V.** Comparative plant genomics resources at PlantGDB (2005). *Plant Physiology* 139:610-618.

- Lawrence, C.J.**, Seigfried, T.E., and Brendel, V. The Maize Genetics and Genomics Database. The community resource for access to diverse maize data (2005). *Plant Physiology* 138:55-58.
- Lawrence, C.J.**, Dawe, R.K., Christie, K.R., Cleveland, D.W., Dawson, S.C., Endow, S.A., Goldstein, L.S.B., Goodson, H.V., Hirokawa, N., Howard, J. Malmberg, R.L., McIntosh, J.R., Miki, H., Mitchison, T.J., Okada, Y., Reddy, A.S.N., Saxton, W.M., Schliwa, M., Scholey, J.M., Vale, R.D., Walczak, C.E., and Wordeman, L. A standardized kinesin nomenclature (2004). *The Journal of Cell Biology* 167(1):19-22.
- Lawrence, C.J.**, Zmasek, C.M., Dawe, R.K., and Malmberg, R.L. LumberJack: a heuristic tool for sequence alignment exploration and phylogenetic inference (2004). *Bioinformatics* 20(12):1977-1979.
- Baran, S.B., **Lawrence, C.J.**, and Brendel, V. PGROP – a gateway to plant genome research outreach programs (2004). *Plant Physiology* 134(3):889.
- Lawrence, C.J.**, Dong, Q., Polacco, M.L., Seigfried, T.E., and Brendel, V. MaizeGDB: the community database for maize genetics and genomics (2004). *Nucleic Acids Research* 32(Database issue):D393-397.
- Lawrence C.J.**, Malmberg, R.L., Muszynski, M.G., and Dawe, R.K. Maximum likelihood methods reveal conservation of function among closely related kinesin families (2002). *Journal of Molecular Evolution* 54(1):42-53.
- Lawrence C.J.**, Morris, N.R., Meagher, R.B., and Dawe, R.K. Dyneins have run their course in plant lineage (2001). *Traffic* 2(5):362-363.
- Lawrence C.** and Holaday, A.S. Effects of mild night chilling on respiration of expanding cotton leaves (2000). *Plant Science* 157(2):233-244.

PAST ACCOMPLISHMENTS OF INVESTIGATORS – TANER Z. SEN

EDUCATION

B.S.	Chemical Engineering	1992-1996	Bogazici University, Istanbul, Turkey
M.S.	Chemical Engineering	1996-1998	Bogazici University, Istanbul, Turkey
Ph.D.	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	USDA-ARS CICGR, Ames, IA
	Collaborator/Assistant Professor	Department of Genetics, Development and Cell Biology, Iowa State University

ACCOMPLISHMENTS

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. This finding shows that the protein interaction network not only contains a vertical fractal pattern at different hierarchical levels of network architecture, but also a horizontal pattern among the proteins at the same level. Harnessed the presence of this horizontal interaction pattern 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. Predicted the tertiary structure of FecA ferric citrate signaling and transport protein in yeast and determined its dynamics using elastic network models. Improved secondary structure prediction accuracy by data mining protein structure fragments in the PDB. 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. Used the transfer matrix method to create protein lattice models of various rectangular shapes and probed how the strength of hydrophobic interactions influences thermodynamically favorable structures. Extended the transfer matrix method to the hexagonal lattice systems to provide a more realistic picture of proteins. 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)

Cheng, H., Sen, T.Z., Jernigan, R.L., Kloczkowski, A. Consensus Data Mining (CDM) Protein Secondary Structure Prediction Server: Combining GOR V and Fragment Database Mining (FDM), *Bioinformatics*, in press.

- Myron, P., **Sen, T.Z.**, Jernigan, R.L., Kloczkowski, A. Generation and enumeration of compact conformations on the 2D triangular and 3D fcc lattices, *Journal of Chemical Physics*, in press.
- Sulkowska, J.I., Kloczkowski, A., **Sen, T.Z.**, Cieplak, M., Jernigan, R.L. Predicting the order in which contacts are broken during single molecule protein stretching experiments, *Proteins*, in press.
- Sen, T.Z.**, Kloczkowski, A., Jernigan, R.L. A DNA-centric look at protein-DNA complexes (2006). *Structure* 14(9):1341-1342.
- Sen, T.Z.**, Cheng, H., Kloczkowski, A., Jernigan, R.L. The Consensus Data Mining (CMD) secondary structure prediction method by combining GOR V and Fragments Database Mining (2006). *Protein Science* 15:2499-2506.
- Sen, T.Z.**, Kloczkowski, A., Jernigan, R.L. Functional clustering of yeast proteins from the protein-protein interaction network (2006). *BMC Bioinformatics* 7:355.
- Sen, T.Z.**, Feng, Y., Garcia, J.V., Kloczkowski, A., Jernigan, R.L. The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models (2006). *Journal of Chemical Theory and Computation* 2:696-704.
- Fernandez, A., Tawfik, D.S., Berkhout, B., Sanders, R., Kloczkowski, A., **Sen, T.Z.**, Jernigan, R.L. Protein Promiscuity: Drug Resistance and Native Functions -- HIV-1 Case (2005). *J. Biomol. Struct. Dyn.* 22(6):615-624.
- Sen, T.Z.**, Jernigan, R.L., Garnier, J., Kloczkowski, A. The GOR V server for protein secondary structure assignment (2005). *Bioinformatics*, 21(11):2787-2788.
- Cheng, H., **Sen, T.Z.**, Kloczkowski, A., Margaritis, D., Jernigan, R.L. Prediction of protein secondary structure by mining fragments database (2005). *Polymer* 46:4314-4321.
- Sen, T.Z.**, Kloczkowski, A., Jernigan, R.L., Yan, C., Honavar, V., Ho, K.-M., Wang, C.-Z., Ihm, Y., Cao, H., Gu, X, Dobbs, D. Predicting binding sites of hydrolase-inhibitor complexes by combining several methods (2004). *BMC Bioinformatics* 5:205.
- Varshney, V., Dirama, T.E., **Sen, T.Z.**, Carri, G.A. A Minimal Model for the Helix-Coil Transition of Worm-like Polymers. Insights from Monte Carlo Simulations and Theoretical Implications (2004). *Macromolecules* 37:8794-8804.
- Kloczkowski, A., **Sen, T.Z.**, Jernigan, R.L. The transfer matrix method for lattice proteins-an application with cooperative interactions (2004). *Polymer* 45:707-716.
- Konuklar, F.A.S., Aviyente, V., **Sen, T.Z.**, Bahar, I. Modeling the deamidation of asparagine residues via succinimide intermediates (2001). *Journal of Molecular Modeling* 7(5):147-160.

PAST ACCOMPLISHMENTS OF INVESTIGATORS – LESLIE C. LEWIS

EDUCATION

B.S.	Agriculture	1961	University of Vermont, Burlington, VT
M.S.	Dairy Nutrition	1963	University of Vermont, Burlington, VT
Ph.D.	Entomology	1970	Iowa State University, Ames, IA

WORK EXPERIENCE

1967-1990: Research Entomologist, USDA-ARS, Ankeny, IA

1990-Present: Research Leader and Supervisory Research Entomologist, USDA-ARS, Ankeny/Ames, IA

ACCOMPLISHMENTS

Dr. Lewis' contributions to agricultural science are many, beginning with a program to mass rear the European corn borer. This program included development of a diet and techniques to suppress *Nosema pyrausta* in laboratory colonies. Prior to development of a laboratory diet, research on the European corn borer could be conducted only when the insect was active in nature.

In cooperation with private industry, Dr. Lewis led a team of scientists who developed a granular formulation of *Bacillus thuringiensis* for European corn borer control. This research was the earliest to define efficacy of *B. thuringiensis* against the European corn borer and was sought by industry as they developed *B. thuringiensis* transgenic maize.

He developed the concept that beneficial insects, insect pathogens, and chemical insecticides can function in synchrony to control multiple corn insect pests. Previously, most scientists believed that either an insect pathogen, a beneficial insect, or a chemical could be an effective control agent, but never simultaneously. Dr. Lewis showed that the egg parasite of corn borer, *Trichogramma brassicae*, was compatible with SLAM, a feeding stimulant/adulticide for corn rootworm, and that all combinations could be used as successful components of a management program.

Dr. Lewis defined a unique tri-trophic relationship among an entomopathogenic fungus, *Beauveria bassiana*, the corn plant, and the European corn borer. *Beauveria bassiana* forms an endophytic relationship within the plant and suppresses larval populations of the European corn borer throughout the growing season. This was the first report of an insect-killing fungus or any insect pathogen establishing endophytism within a green plant. This pioneering research was the impetus for other scientists to search for endophytic relationships that have been documented in coffee, tomato, and cocoa.

Dr. Lewis determined that non-target organisms are not significantly affected by *Bt*-transgenic corn by assembling and co-directing a team that investigated the impact of *Bt*-transgenic corn on several insect families common to a corn field, and larvae of the monarch butterfly. This was a concerted effort between ARS, Land Grant Institutions, private industry, and citizen groups to study the impact of transgenic corn on non-target organisms. This provided the EPA with the data to make science-based decisions regarding the registration of Bt corn to manage the European corn borer.

PEER REVIEWED PUBLICATIONS

- Lewis, L.C.** 2007. Ecological considerations for the use of entomopathogens in IPM, in Ecologically-based integrated pest management. Koul, O., and Cuperis, G.W. (eds.) Taylor & Francis, UK. p. 249-268.
- Prasifka, P.L., Hellmich, R.L., Prasifka, J.R., and **Lewis, L.C.** 2007. Effects of Cry1Ab-expressing corn anthers on the movement of Monarch butterfly larvae. *Environ. Entomol.* 36:228-233.
- Lewis, L.C.**, Sumerford, D.V., Bing, L.A., and Gunnarson, R.D. 2006. Dynamics of *Nosema pyrausta* in natural populations of the European corn borer, *Ostrinia nubilalis*: a six-year study. *BioControl* 51:627-642.
- Coates, B.S., Sumerford, D.V., Hellmich, R.L., and **Lewis, L.C.** 2005. Sequence variation in the cadherin gene of *Ostrinia nubilalis*: A tool for field monitoring. *Insect Biochem. Mol. Biol.* 35(2):129-139.
- Coates, B.S., Hellmich II, R.L., and **Lewis, L.C.** 2005. Polymorphic CA/GT and GA/CT microsatellite loci for *Ostrinia nubilalis* (Lepidoptera: Crambidae). *Mol. Ecol. Notes* 5(1):10-12.
- Lewis, L.C.**, Gunnarson, R.D., and Robbins, J.C. 2005. *Trichogramma brassicae* (Hymenoptera: Trichogrammatidae) and SLAM®, an integrated approach to managing European corn borer and corn rootworms. *BioControl* 50:729-737.
- Coates, B.S., Hellmich, R.L., and **Lewis, L.C.** 2005. Two differentially expressed *Ostrinia nubilalis* ommochrome-binding protein-like genes (obp1 and obp2) in larval fat body. *J. Insect Sci.* 5:19.
- Lopez, M.D., Prasifka, J.R., Bruck, D.J., and **Lewis, L.C.** 2005. Utility of ground beetle species as indicators of potential non-target effects of Bt crops. *Environ. Entomol.* 34:1317-1324.
- Prasifka, J.R., Hellmich, R.L., Dively, G.P., and **Lewis, L.C.** 2005. Assessing the effects of pest management on non-target arthropods: the influence of plot size and isolation. *Environ. Entomol.* 34:1181-1192.
- Arnold, A.E., and **Lewis, L.C.** 2005. Ecology and evolution of fungal endophytes and their roles against insects, in Insect-Fungal Associations, Ecology and Evolution. Vega, F.E., and Blackwell, M. (eds.) Oxford University Press, New York, New York. p. 74-96.
- Anderson, P.L., Hellmich, R.L., Sears, M.K., Sumerford, D.V., and **Lewis, L.C.** 2004. Effects of Cry1Ab-expressing corn anthers on monarch butterfly larvae. *Environ. Entomol.* 33:1109-1115.
- Reardon, B.J., Hellmich II, R.L., Sumerford, D.V., and **Lewis, L.C.** 2004. Growth, development, and survival of *Nosema pyrausta*-infected European corn borers (Lepidoptera: Crambidae) reared on meridic diet and Cry1Ab. *J. Econ. Entomol.* 97:1198-1201.
- Coates, B.S., Sumerford, D.V., Hellmich, R.L., and **Lewis, L.C.** 2004. Partial mitochondrial genome sequences of *Ostrinia nubilalis* and *Ostrinia furnicalis*. *Intl. J. Biol. Sci.* 1:13-18.
- Lewis, L.C.**, Bruck, D.J., and Gunnarson, R.D. 2002. On-farm evaluation of *Beauveria bassiana* for control of *Ostrinia nubilalis* in Iowa, USA. *BioControl* 47:167-176.
- Coates, B.S., Hellmich, R.L., and **Lewis, L.C.** 2002. Nuclear small subunit rRNA group I intron variation among *Beauveria* spp. provide tools for strain identification and evidence of horizontal transfer. *Curr. Genet.* 41:414-424.

- Bruck, D.J., and **Lewis, L.C.** 2002. Rainfall and crop residue effects on soil dispersion and *Beauveria bassiana* spread to corn. *Appl. Soil Ecol.* 20:183-190.
- Lewis, L.C.**, Bruck, D.J., and Gunnarson, R.D. 2002. Measures of *Bacillus thuringiensis* persistence in the corn whorl. *J. Invertebr. Pathol.* 80:69-71.
- Bruck, D.J., and **Lewis, L.C.** 2002. Whorl and pollen-shed stage application of *Beauveria bassiana* for suppression of adult western corn rootworm. *Entomol. Exp. Appl.* 103:161-169.
- Bruck, D.J. and **Lewis, L.C.** 2002. *Carpophilus freemani* (Coleoptera: Nitidulidae) as a vector of *Beauveria bassiana*. *J. Invertebr. Pathol.* 80:188-190.
- Morjan, W.E., Pedigo, L.P., and **Lewis, L.C.** 2002. Fungicidal effects of glyphosate and glyphosate formulations on four species of entomopathogenic fungi. *Environ. Entomol.* 31(6):1206-1212.

XI. HEALTH, SAFETY, AND OTHER 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.

Homeland Security: The data security against cyberattacks, accidents, and disasters is ensured through the implementation of a secure DISASTER RECOVERY system (shown in Figure 2). The DISASTER RECOVERY duplicates all the data distributed in the Curation Database and the entire programming and transfers it in two different physical locations in Ames, IA (on separate servers in other buildings from where the MaizeGDB Curation Database is located) 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 SCAs: None.

APPENDIX – LETTERS OF COLLABORATION



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

July 19, 2007

Carolyn Lawrence
Corn Insects and Crop Genetics Research Unit
USDA Agricultural Research Service
526 Science II
Iowa State University
Ames, IA 50011

Dear Carolyn:

I am very happy to continue my collaboration with MaizeGDB in the areas of data curation with priority to the roles we have identified where I would be most useful. I have asked Mike McMullen to co-sign this letter, because he will not only collaborate on attaining Sub Objective 1b but he is also the lead scientist for *Project no. 3622-21000-027-00D, entitled "Genetic Mechanisms and Molecular Genetic Resources for Maize"* to which I am assigned as an SY.

In more detail, and listed by objectives on your pre-plan draft July 12, 2007

Objective 1: *Integrate new maize genetic and genomic data into the database.*

Sub Objective 1a. *Expand mutant and phenotype data and tools.*

I will contribute to this objective by integrating the terms used in the various Plant Ontologies with the current and future phenotype records at MaizeGDB. As needed, I will request new terms from the Plant Ontology Consortium.

On an annual basis:

- I will review the current use in the database and provide any needed refinements
- I will update the terms and relationships in MaizeGDB to accommodate any refinements in the central ontologies relating to phenotypes.
- I will provide association files to the Plant Ontology database, currently maintained at Cold Spring Harbor Laboratories. Associations are currently made to genes, alleles and Stocks in MaizeGDB, based on phenotype evidence and form links from the Plant Ontology repository to MaizeGDB.

Sub-objective 1.B. *Expand structural and genetic map sets emphasizing the:*

- *Integration of the IBM genetic maps with the B73 genome sequence*



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

On an annual basis: I will provide updated consensus maps based on the current IBM. At this time these maps include the IBM2 neighbors, and cIBM maps, with refinements based on the physical map. Refinements of the bins maps will be based on the IBM Neighbors. IBM Neighbors currently relies on incorporation of newly public genetic maps with documentation that will include map scores and probe details. One such map will be the new INDEL (insertion-deletion) maps of Pat Schnable based on the IBM panel of stocks and the IBM SNP (single nucleotide polymorphism) map from the maize diversity project (see also below).

- o *Integration of the next-generation genetic map being generated by the Maize Diversity Project into a genomic view to enable its effective use by plant breeders*

This objective will be met in collaboration with Mike McMullen. Deliverables to MaizeGDB will include NAM maps, genotype and phenotype information for 26 mapping panels of stocks, encompassing 26 US, CIMMYT tropical lines and wild relatives of maize and some 1200 new single nucleotide polymorphism loci (SNP).

Sub-objective 1.D. Compile and make accessible at MaizeGDB the annual Maize Newsletter

On an annual basis, I will supply an electronic copy to Ames in a format suitable for integration into MaizeGDB. As you know, Jim Birchler is co-editor for the newsletter, and there is a separate fund to cover costs of layout, redaction, and mailing. I am preparing electronic images (pdf) of early MNLs, from hard copies and in a resolution suitable for electronic conversion to text, that will be useful for electronic indexing. These will also be delivered to Ames, and in the format for integration into MaizeGDB.

Proposed Milestones are attached along with a CV formatted to the OSQR standard.

Most sincerely yours,



Mary (Polacco) Schaeffer, PhD
Research Geneticist/Curator MaizeGDB



Michael McMullen
Research Geneticist USDA-ARS

Attachments:

Schaeffer Milestones
Schaeffer Brief CV and accomplishments

Milestones – Mary Schaeffer (MS)

Project Title^a	Genetic Mechanisms and Molecular Genetic Resources for Maize			Project No.^b	3622-21000-027-00D in collaboration with 3625-21000-045-00D
National Program	NPI 301 Plant Genetic Resources, Genomics and Genetic Improvement				
Objective 1 3625-21000-045-00D	Integrate new maize genetic and genomic data into MaizeGDB				
Subobjective 1a 3625-21000-045-00D	Expand mutant and phenotype data and tools				
Approach	SY Team	Months	Milestones	Progress/ Changes	Products
Integrate Plant Ontologies' terms with the current and future phenotype records at MaizeGDB. As required, add new terms via the Plant Ontologies Consortium and MaizeGDB. Rationale: to improve phenotype searching across all angiosperms, especially where there are full genome sequences to aid plant breeders discover candidate genes for desired traits.	MS	preproject	All Plant Ontology terms in MaizeGDB. All phenotypes associated with the Stock Center annotated.		FY08 Association files released to Plant Ontology Consortium
	MS	FY09	All phenotypes, including those associated with germplasm in maize diversity project annotated.		FY09 Association files released to Plant Ontology Consortium
	MS	FY10	New phenotypes annotated.		FY10 Association files released to Plant Ontology Consortium
	MS	FY11	New phenotypes annotated.		FY11 Association files released to Plant Ontology Consortium
	MS	FY12	New phenotypes annotated.		FY12 Association files released to Plant Ontology Consortium
	MS	FY13	New phenotypes annotated.		FY13 Association files released to Plant Ontology Consortium
Subobjective 1b CL project number	Expand structural and genetic map set emphasizing the integration of the IBM genetic maps with the B73 genome sequence; integration of the next generation genetic maps being generated by the Maize Diversity Project into a genomic view				
Approach	SY Team	Months	Milestones	Progress/ Changes	Products
Incorporate genetic maps each year with full documentation. Integrate onto IBM Neighbors all key genetic maps and refine order based on the B73 genome sequence. Rationale is two fold: (1) MaizeGDB is the primary repository for genetic map documentation; (2) using a single genetic coordinate system facilitates candidate gene sequence discovery by plant	MS	preproject	Neighbors includes all major maps released in 2007.		IBM 2007 Neighbors maps released.
	MS/MDM	FY09	Neighbors includes SNP maps and refined by 2008 released B73 genome sequence.		IBM 2008 Neighbors maps released.
	MS/MDM	FY10	Neighbors includes NAM maps from maize diversity project		IBM 2009 Neighbors maps released.
	MS/MDM	FY11	Neighbors includes all major maps released in 2010 and is refined by any updates to the B73 sequence		IBM 2010 Neighbors maps released.
	MS	FY12	Neighbors includes all major maps released in 2011 and is refined by any updates to the B73 sequence.		IBM 2011 Neighbors maps released.

breeders	MS	FY13	Neighbors includes all major maps released in 2012 and is refined by any updates to the B73 sequence		IBM 2012 Neighbors maps released.
Subobjective 1d^{CL} project number	Compile and make accessible at MaizeGDB the annual Maize Newsletter				
Approach	SY Team	Months	Milestones	Progress/Changes	Products
Edit and produce MNL with co-editor Jim Birchler; Scan images of past MNL, as separated notes, and convert to text for indexing. Integrate all with MaizeGDB as individual notes linked to the citations. Rationale. Much genetic information is summarized or documented as notes in the MNL and is more accessible to all if electronic.	MS	preproject	MNL 82 at MaizeGDB All past issues have optical images, as		Hard copy MNL v. 82
	MS	FY09	MNL 83 at MaizeGDB _ of past issues converted to text		Hard copy MNL v. 83
	MS	FY10	MNL 84 at MaizeGDB _ past issues converted to text		Hard copy MNL v. 84
	MS	FY11	MNL 85 at MaizeGDB _ past issues converted to text		Hard copy MNL v. 85
	MS	FY12	MNL 86 at MaizeGDB All past issues converted to text		Hard copy MNL v. 86
	MS	FY13	MNL 87 at MaizeGDB		Hard copy MNL v. 87

^{a,b} Milestones will be associated with *Project no. 3622-21000-027-00D, entitled "Genetic Mechanisms and Molecular Genetic Resources for Maize"* where Mary Schaeffer is assigned as an SY.

Objectives and sub-objectives are listed here according to the 2007 Pre-Plan for the Project entitled *"The Maize Genetics and Genomics Database"* with Carolyn Lawrence as Lead Scientist.

PAST ACCOMPLISHMENTS OF MARY L. SCHAEFFER (POLACCO)

Contact

203 Curtis Hall
University of Missouri, Columbia, MO 65211

Phone: (573) 884-7873 Fax: (573) 884-7850
E-mail Mary.Schaeffer@ARS.USDA.GOV

Education

AB

Chemistry

Swarthmore College, Swarthmore, PA

PhD

Biochemistry

Duke University, Durham, NC

Work Experience

1972-1974	Assistant Professor	Department of Biochemistry, University del Valle, Cali, Colombia
1976-1979	Postdoctoral; NIH Postdoctoral Fellow	Departments of Molecular Biophysics and Biochemistry; Biology Department Yale University, New Haven, CT
1979-1995	Research Assistant Professor	Biochemistry Department, Agronomy Department, University of Missouri, Columbia, MO
1995 - current	GS-13 Cat 4 Geneticist Adjunct Associate Professor	USDA-ARS PGRU, Columbia, MO Division of Plant Sciences, University of Missouri, Columbia, MO

Accomplishments

In the early development of the USDA-ARS MaizeDB (1990-93): I contributed to design the robust relational schema (Sybase) that is now used with minor modifications at MaizeGDB (Oracle, Lawrence *et al.* 2007). From 1990-1998, I curated data for new genes and gene products; mutants from the literature and from collections of Neuffer and Martienssen; new genetic maps, probes and other documentation; and external database identifiers required to establish reciprocal links with those databases. In 1995, MaizeDB had operating WWW reciprocal links with SwissProt; at this time, most databases, including GenBank, were just beginning to develop a WWW interface. Since 1999, and with competitive funding, I integrated genetic project data towards creation of a genetic and physical map resource (Coe *et al.* 2002; Cone *et al.* 2002) that has been used for candidate gene discovery, and that under girds the B73 genome sequencing project (Gardiner *et al.* 2004). I also integrated these data into Maize[G]DB. To facilitate anchoring of BAC contigs to maize chromosomes (Wei *et al.* 2007), I created an algorithm to compute a single map order, Neighbors maps, for loci on any of the various genetic maize maps. It is based on the high resolution IBM2 maps; and it represents statistical uncertainties on contributing maps. After refinement with cDNA probes ordered on BAC contigs (IBM2 FPC0507 maps), the Neighbors maps include over 30,000 empirically defined loci, documented with primers, map scores, GenBank sequence accessions and BACs. In 2003, working with Stanford Baran, Volker Brendel, and members of the current MaizeGDB team, I designed and tested the community curation interface for MaizeGDB, adding a QTL module in 2006. At Missouri, I provide a community mapping service, (<http://www.maizemap.org/cimde.html>), with software, developed with competitive funding, and which supports computation of map locations from raw data supplied by users. This service is in active use. As an early member of the Plant Ontology Consortium, and working with a team funded with competitive funding, I developed consensus anatomy, growth and developmental terms for angiosperms and that would be useful for phenotype and gene expression annotation in genome databases (Vincent *et al.* 2002; Jaiswal *et al.* 2005; Pujar *et al.* 2006; Ilic *et al.* 2007). I incorporated these into MaizeGDB along with the Plant Ontology accession identifiers. I have been a member of the Maize Nomenclature Committee since its founding; I am also a member of the Maize Executive Committee and, *ex officio*, of the Maize Meeting Steering Committee. I am co-editor, with Jim Birchler, of the Maize Genetics Cooperation Newsletter.

PEER REVIEWED PUBLICATIONS (underline indicates corresponding author; Mary Schaeffer was formerly known as Mary Polacco)

1. Wei, F., Coe, E., Nelson, W., Engler, F., Bharti, A., Butler, E., Kim, H.R., Goicoechea, J.L., Lee S., Fuks, G., Sanchez-Villeda, H., Schroeder, S., Fang, Z., McMullen, M., Davis, G., Bowers, J.E., Paterson, A.H., **Schaeffer, M.**, Gardiner, J., Cone, K., Messing, J., Soderlund, C. and Wing, R.A. The Physical and Genetic Structure of the Maize (*Zea mays* cv. B73) (2007) *PloS Genetics* (*in press*)
2. Ilic, K., Kellogg, E.A., Jaiswal, P., Zapata, F., Stevens, P.F., Vincent, L.P., Avraham, S., Reiser, L., Pujar, A., Sachs, M.M., Whitman, N.T., McCouch, S.R., **Schaeffer, M.L.**, Ware, D.H., Stein, L.D. and Rhee, S.Y. The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant (2007) *Plant Physiol* 143:587-599.
3. Lawrence, C.J., **Schaeffer, M.L.**, Seigfried, T.E., Campbell, D.A., and Harper, L.C. MaizeGDB's new data types, resources, and activities (2007) *Nucleic Acids Research* 35(Database issue):D895-900.
4. Shyu, C.R., Green, J., Lun, D.P.K., Kazic, T., **Schaeffer, M.L.** and Coe, E.H. Image Analysis for Mapping Immeasurable Phenotypes in Maize (2007) *IEEE Signal Processing Magazine* 24: 116-119.
5. Yim, Y.-S., Moak, P., Sanchez-Villeda, H., Musket, T., Close, P., Klein, P.E., Mullet, J.E., McMullen, M.D., Fang, Z., **Schaeffer, M.L.**, Gardiner, J.M., Coe, E.H. and Davis, G. L. A BAC pooling strategy combined with PCR-based screenings in a large, highly repetitive genome enables integration of the maize genetic and physical maps (2007) *BMC Genomics* 8: 47.
6. Coe, E.H. and **Schaeffer, M.L.** Uncaging mutants: moving from mutants to menageries to ménages (2006) *Maydica* 51:263-267.
7. Pujar, A., Jaiswal, P., Kellogg, E.A., Ilic, K., Vincent, L., Avraham, S., Stevens, P., Zapata, F., Reiser, L., Rhee, S.Y., Sachs, M.M., **Schaeffer, M.**, Stein, L., Ware, D. and McCouch, S. Whole-plant growth stage ontology for angiosperms and its application in plant biology (2006) *Plant Physiol* 142:414-428.
8. **Schaeffer M.**, Byrne, P. and Coe, E.H. Consensus quantitative trait maps in maize: a database strategy (2006) *Maydica* 51: 357-367.
9. Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., Rhee, S.Y., Sachs, M.M., **Schaeffer, M.**, Stein, L., Stevens, P., Vincent, L., Ware, D. and Zapata, F. Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages (2005) *Comp Funct Genomics* 6:388-397.
10. Gardiner, J., Schroeder, S., **Polacco, M.L.**, Sanchez-Villeda, H., Fang, Z., Morgante, M., Landewe, T., Fengler, K., Useche, F., Hanafey, M., Tingey, S., Chou, H., Wing, R., Soderlund, C. and Coe, E.H. Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization (2004) *Plant Physiol* 134:1317-1326
11. Lawrence, C.J., Dong, Q., **Polacco, M.L.**, Seigfried, T.E., and Brendel, V. MaizeGDB: the community database for maize genetics and genomics (2004) *Nucleic Acids Research* 32(Database issue):D393-397.

12. Fang, Z., Cone, K., Sanchez-Villeda, H., **Polacco, M.**, McMullen, M., Schroeder, S., Gardiner, J., Davis G., Haverman, S., Yim Y., Vroh Bi, I. and Coe, E.H iMap: a database-driven utility to integrate and access the genetic and physical maps of maize (2003) *Bioinformatics* 19:2105-2111.
13. Fang, Z., **Polacco, M.**, Chen, S., Schroeder, S., Hancock, D., Sanchez-Villeda, H. and Coe, E. cMap: the Comparative Genetic Map Viewer (2003) *Bioinformatics*, 19:416 – 417.
14. Kazic, T., Coe, E.H., **Polacco, M.L.**, and Shyu, C.R. Whither Biological Database Research (2003) *Omics* vol 7:61-66
15. Sanchez-Villeda, H., Schroede, S., **Polacco, M.**, McMullen, M., Havermann, S., Davis, G., Vroh-Bi, I., Cone, K., Sharopova, N., Yim, Y., Schultz, L., Duru, N., Musket, T., Houchins K., Fang, Z., Gardiner, J. and Coe, E. Development of an integrated laboratory information management system for the maize mapping project (2003) *Bioinformatics*. 19:2022-2030.
16. Coe, E., Cone, K., McMullen, M., Chen, S.S., Davis, G., Gardiner, J., Liscum, E., **Polacco, M.**, Paterson, A., Sanchez-Villeda, H., Soderlund, C., and Wing, R. Access to the Maize Genome: An Integrated Physical and Genetic Map (2002) *Plant Physiol*. 128:9-12.
17. Cone, K.C., McMullen, M.D., Vroh Bi, I., Davis, G.L., Yim, Y.S., Gardiner, J.M., **Polacco, M.L.**, Sanchez-Villeda, H., Fang, Z., Schroeder, S.G., Havermann, S.A., Bowers, J.E., Paterson, A.H., Soderlund, C.A., Engler, F.W., Wing, R.A. and Coe, E.H. Genetic, physical and informatics resources for maize. On the road to an integrated map (2002) *Plant Physiol* 130: 1686-1696.
18. **Polacco, M.**, Coe, E., Fang, Z., Hancock, D., Sanchez-Villeda, H. and Schroeder, S. MaizeDB – a functional genomics perspective (2002) *Comp Funct Genom* 3:128-131.
19. Sharopova, N., McMullen, M.D., Schultz, L., Schroeder, S., Sanchez-Villeda, H., Gardiner, J., Bergstrom, D., Houchins, K., Melia-Hancock, S., Musket, T., Duru, N., **Polacco, M.**, Edwards, K., Ruff, T., Register, J.C., Brouwer, C., Thompson, R., Velasco, R., Chin, E., Lee, M., Woodman-Clíkeman, W., Long, M.J., Liscum, E., Cone, K., Davis, G. and Coe, E.H. Development and mapping of SSR markers for maize (2002) *Plant Mol Biol* 48:463-481.
20. Vincent, L., Bruskiewich, R., Coe, E., Jaiswal, P., McCouch, S., **Polacco, M.**, Stein, L. and Ware, D. The Plant Ontology™ Consortium and plant ontologies (2002) *Comp Funct Genomics* 3:137-142.



Department of Computer Science
College of Engineering
University of Missouri-Columbia

Medical and Biological
Digital Library Lab
238 Engineering Building West
Columbia, MO 65211-2060 Phone (573) 884-3534
<http://medbio.cecs.missouri.edu> Fax (573) 882-8318



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

July 10, 2007

Dear Carolyn,

I am writing to express my interest in continuing to work with your group to offer access to our web-based phenotypic information management system, VPhenoDBS (<http://www.PhenomicsWorld.org>), to maize researchers via MaizeGDB and to make maize mutant images stored at MaizeGDB searchable via query images and ontological terms.

Experimentation with mutant maize plants is an effective method for understanding the roles of specific genes as well as for visualizing the phenotypic effects of these mutations. For visually observed phenotypic effects, annotations are made by scientists to document the physical state of the mutated plant; however, the language used to describe the mutations can be vague, especially in terms of color, texture, and size (e.g. the leaf is pale green, the kernel is variegated, the plant is short). Color descriptions are further complicated by the fact that 'light green' to one person may be described as 'yellow green' by another. To combat this vagueness or uncertainty in mutant descriptions, image processing and computer vision algorithms can be developed to quantify these types of visual features, eliminating the subjective component of human perception in these kinds of descriptions.

We are continuing to develop VPhenoDBS, which will use these features to allow biologists to perform complex queries (query by image example, query by text annotation/ontology, and query by physical and genetic map information) on maize images. The web-based system will be publicly accessible to the plant community, particularly for the maize community for the initial stage. We propose simple standards to capture phenotypic images for various body parts and development stages using commercially available digital cameras, color palettes, rulers, and homogeneous background settings under a consistent lighting condition. All images deposited to the VPhenoDBS using the simple standards, along with their corresponding text annotations, will be searchable and cross referenced to various maps with a unique

Innovative ideas to improve human health and life sciences computationally

an equal opportunity and ADA institution

visualization tool. The ultimate goal of this research is to allow a geneticist to submit phenotypic and genomic information on a mutant to a knowledge base and ask, "What genes or environmental factors cause this visually observed phenotype?"

Because our work together already has been so fruitful, I look forward to working with Mary Schaeffer and others in your group to continue this very interesting collaboration!

Sincerely,



Chi-Ren Shyu
Shumaker Associate Professor
126 Engineering Building West
University of Missouri
Columbia, Missouri 65211-2060
TEL: 573-882-3884
FAX: 573-882-8318



25 July 2007

Carolyn Lawrence, Ph. D.
USDA-ARS Research Geneticist
Corn Insects and Crop Genetics Research Unit
USDA-ARS
526 Science II
Iowa State University
Ames IA USA 50011

Dear Carolyn,

We are writing to express our enthusiasm to collaborate with you as outlined in MaizeGDB's Project Plan. We look forward to continued collaboration on the integration of mutant phenotype description data and tools generated by large projects into MaizeGDB. These projects include the Maize Inflorescence (EMS) Project and the Maize Gene Discovery Project (RescueMu), and we anticipate continued successful collaborations with you on additional projects of this type in the future.

Sincerely,

A handwritten signature in cursive script, reading "Sarah Hake".

Sarah Hake, Ph.D. Director, Plant Gene Expression Center,

A handwritten signature in cursive script, reading "Lisa Harper".

Lisa Harper, Ph.D. Geneticist (Plants)



Pacific West Area - Plant Gene Expression Center
800 Buchanan Street Albany, CA 94710-1105
Sarah Hake, Ph.D., Voice: 510.559.5907 Fax: 510.559.5878 E-mail: maizesh@nature.berkeley.edu
Agricultural Research - Investing in Your Future



United States
Department of
Agriculture

Agricultural
Research
Service

Midwest Area
Soybean/Maize
Germplasm, Pathology &
Genetics Research Unit

Martin M. Sachs

S108 Turner Hall
1102 S. Goodwin Ave.
Urbana, IL 61801-4730

(217) 244-0864 [phone]

(217) 333-8064 [fax]

msachs@uiuc.edu [e-mail]

June 25, 2007

Carolyn J. Lawrence, Ph.D.
USDA-ARS Corn Insects and Crop Genetics Research Unit
1565 Agronomy Hall
Iowa State University
Ames, IA 50011

Dear Carolyn,

I am writing to convey my group's enthusiasm to continue collaborating with you and others working at MaizeGDB. We will continue enter data about our collection at the Maize Genetics Cooperation • Stock Center into MaizeGDB. This information describes the maize genetics stocks and related data.

As you know, I direct Maize Genetics Cooperation • Stock Center. My group maintains and distributes maize genetic stocks and provides information about this collection to the community of maize researchers. Instead of our maintaining a separate database for this information, we've been entering it directly into MaizeGDB.

As in the past, Maize COOP personnel will work closely with your group to connect phenotypes to stock data and to enter our group's genetic stock (and associated) data into MaizeGDB. Indeed, because most of these data are being entered into the database directly by us at the MGCSC via our customized tools, committing to this collaboration is very easy for us!

In addition, our collaboration has been useful to our shared stakeholders in that genetic data stored at MaizeGDB often are useful only if the stocks bearing the genetic variations are available for further research. Conversely, the stocks are only useful if the data that describes their genetic constituency is made accessible. This collaboration serves both of our groups very well.

Sincerely,

Marty Sachs
Research Geneticist (Plants)
Director, Maize Genetics Cooperation • Stock Center



United States Department of Agriculture

Research, Education and Economics
Agricultural Research Service

June 26, 2007

SUBJECT: Letter of Collaboration for MaizeGDB 5-year plan

TO: Carolyn J. Lawrence, Ph.D.
1565 Agronomy Hall
Iowa State University
Ames, IA 50011FROM: Roger Wise, Research Geneticist
Corn Insects and Crop Genetics Research Unit
Department of Plant Pathology, Iowa State University
Ames, IA 50011

Dear Carolyn,

I am writing this letter to confirm our intent to collaborate with you and colleagues at MaizeGDB in linking with expression data at PLEXdb. The most relevant linkages would come under **Objective 1: Integrate new maize genetic and genomic data into the database and Sub-objective 1.A. Expand mutant and phenotype data and tools.**

Julie Dickerson, Volker Brendel, and I have recently been funded by NSF ([NSF-BD&I #0543441](#)) to continue our work on PLEXdb (<http://plexdb.org>), a unifying resource for the analysis of plant and plant pathogen expression data. Within the context of PLEXdb, MaizePLEX currently supports the two Affymetrix maize GeneChips as well as the NSF-funded oligonucleotide array.

Consistent with the emergence of the maize genome sequence, I will be spearheading an effort to develop a new Affymetrix 100K "all genes" GeneChip. As highlighted in the Expression Profiling section of the **2007 Maize Genetics Community Retreat (Allerton)** report, "tools must be developed that allow datasets to be browsed, queried, visualized, meta-analyzed and linked to the physical and genetic maps of maize". Our collaborative efforts should be focused on development of these tools, which are critical to optimize the use of these cost-intensive datasets.

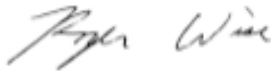
Midwest Area - Corn Insects and Crop Genetics Research Unit
Department of Plant Pathology 411 Bessey Hall Iowa State University Ames IA 50011-1020
Voice: 515-294-9756 FAX: 515-294-9420 E-mail: rwise@iastate.edu Web: <http://wiselab.org/>

An Equal Opportunity Employer

Maize sequences can be aligned with microarray probe sets, and thus, linked to all tools and expression data in PLEXdb, including the linkage of maize probe sets to the Gramene protein pages and rice synteny browser in PlantGDB. BLASTX results using UniProt provide users better annotation including GO terms, IUBMB enzyme nomenclature, domains and other key information, to understand a protein and its function. Cross-platform analysis using TBLASTX results will be used to retrieve "orthologous" probe sets from different GeneChips. This will greatly facilitate user searches to classify predicted functions in association with specific metabolic or biochemical pathways. These enhanced linkages will enable meta-analyses involving multiple data sets to facilitate future comparative and functional analyses of cereal genes.

Our linkage with MaizeGDB will provide these integrated expression resources to the maize community. I look forward to working with you on this project.

Best regards,



Roger Wise, Ph.D.



Midwest Area - Corn Insects and Crop Genetics Research Unit
Department of Plant Pathology 411 Bessey Hall Iowa State University Ames IA 50011-1020
Voice: 515-294-9756 FAX: 515-294-9420 E-mail: rwise@iastate.edu Web: <http://wiselab.org/>

An Equal Opportunity Employer



United States Department of Agriculture

Research, Education and Economics
Agricultural Research Service

July 13, 2007

Dr. Carolyn J. Lawrence
USDA-ARS Research Geneticist
Genetics Bldg
Iowa State University
Ames, IA, 50011

Dear Carolyn,

This is confirm the plan for interactions between MaizeGDB and GrainGenes, the ARS databases for maize and the small grains, respectively. Although our respective projects focus on different crops, most of the data types, problems, and resource development issues will have common foundations. The interaction of our two projects will be an asset for both, and we will look for avenues of cooperation in addressing data issues, developing related database and analysis tools, and serve as interactive resources for considering how to handle delivery of the best bioinformatics products for our respective user communities. As long-term coordinator of the GrainGenes database, I and rest of the GrainGenes staff have considerable experience managing crop databases and interacting with user communities. We will provide whatever advice and insights we have whenever called upon by yourself or other MaizeGDB staff. In turn, GrainGenes will interact with the MaizeGDB staff whenever developments in MaizeGDB might be adapted for the GrainGenes project and we will call on MaizeGDB for advice where MaizeGDB experience will be an asset to GrainGenes development. Naturally the same will hold true in reverse; i.e., an developments by GrainGenes will always be available to MaizeGDB. Let me know whatever assistance GrainGenes can be to MaizeGDB.

Sincerely,

A handwritten signature in cursive script, reading "Olin D. Anderson".

Dr. Olin D. Anderson
Supervisory Research Geneticist
Genomics and Gene Discovery Research Unit

Pacific West Area - Western Regional Research Center
800 Buchanan Street, Albany, CA 94710-1105
Voice: 510.559.5773; Fax: 510.559.5818; E-mail: oandersn@pw.usda.gov

Agricultural Research - Investing in Your Future



United States Department of Agriculture

Research, Education and Economics
Agricultural Research Service

22 June 2007

Dr. Carolyn J. Lawrence
USDA-ARS Research Geneticist
Department of Agronomy
Ames, Iowa 50011

Dear Dr. Lawrence:

This letter is to confirm my willingness and desire to collaborate with you and your staff during the operation of your project "The Maize Genetics and Genomics Database". I and scientists on my projects will be pleased to offer discussion and advice on the management of crop genome databases. In return, we will receive good ideas and management plans from you and your staff.

Sincerely,

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

Randy C. Shoemaker
USDA-ARS Research Geneticist
Ames, Iowa 50011



July 10, 2007

Dr. Carolyn Lawrence
USDA-ARS Research Geneticist
Corn Insects and Crop Genetics Research Unit
Iowa State University
Ames, IA 50011

Dear Carolyn,

We are writing to confirm our agreement to collaborate on your project entitled "The Maize Genetics and Genomics Database." This partnership is a natural extension of our existing collaboration to exchange data as part of the "Gramene: A Resource for Comparative Grass Genomics" (Ware) and "Sequencing the Maize Genome" (Ware and Clifton), and is a natural for your Objective 1: Integrate new genetic maps and genomic data into the database.

A major objective of our research (from both the Gramene and Maize Genome Sequencing Consortium's projects) is to develop and use computational approaches to integrate agronomic and genotypic data, allowing rapid analysis of plant genomes. To this aim, the Ware group, as resources allow, will: provide annotation of genes from maize BAC sequences, display these sequences with in a genome browser to the maize community, integrate diversity analyses and results with Gramene website, map the next generation sequencing maize SNPs on the maize genome, and collaborate to integrate phenotypic data with the whole genome sequencing. Clifton, outreach manager for the Maize Genome Sequencing Consortium's B73 genome sequencing project, also will continue to deliver to MaizeGDB updates on that project's progress over time.

As with our existing collaboration, we would be very happy to provide exchanges between personnel in the groups through phone call or in-person meetings. We look forward to our future discussions and exchanges on this topic. Please contact us if we can provide anything additional that may help you meet the goals of the MaizeGDB project.

Sincerely,

Doreen Ware, USDA-ARS
Adjunct Assistant Professor
Cold Spring Harbor Laboratory

Sandra Clifton, Research Assistant Professor
Assistant Director
Washington University Genome Sequencing Center



North Atlantic Area * Plant, Soil & Nutrition Laboratory
Tower Road * Ithaca, NY 14853-2901
Cold Spring Harbor Laboratory * 1 Bungtown Road * Cold Spring Harbor, NY 11724
Voice: 516 367 6979 * FAX: 516 367 8389 * E-mail: ware@cshl.edu
An Equal Opportunity Employer



United States Department of Agriculture

Research, Education, and Economics
Agricultural Research Service

Carolyn J. Lawrence
USDA-ARS
Corn Insects and Crop Genetics Research Unit
Ames, IA 50011

Dear Carolyn,

I look forward to working with you and other members of the MaizeGDB group to design ways to integrate QTL and diversity data into the Genome Browser you plan to make available via the MaizeGDB website. As you know, this fits well with my own plans to develop bioinformatic tools to mine and present functional allelic variation in maize.

Making connections between genomics and germplasm diversity is the most straightforward way to apply the discoveries of genomics to plant breeding. Over the last decade, QTL linkage and association mapping studies have helped to initiate these connections, but most raw data from these studies never reached public databases. This is due to the fact that major public plant databases are not well equipped to handle diverse marker segregation data and the associated quantitative and qualitative phenotypic data. Additionally, there are few database (DB) aware analysis tools for diversity data. We plan to (1) work with public databases to develop flexible data models for storing and accessing genotypic and phenotypic data, and (2) develop database-aware analysis tools for germplasm diversity. Together, these should stimulate public sharing, storage and web-based analysis and access to diversity data.

For our collaboration, I am most interested in working to develop ways to display all gene QTL effects on the gene models which will be made available at MaizeGDB, and to devise useful ways to generally display diversity data via the MaizeGDB Genome Browser.

Sincerely,

A handwritten signature in black ink, appearing to read "Ed Buckler".

Edward S. Buckler
Research Geneticist, USDA-ARS
Email: esb33@cornell.edu
Voice: (607)255-4520
Fax: (607) 255-6249



North Atlantic Area • Plant, Soil & Nutrition Laboratory Research Unit
US Plant, Soil & Nutrition Laboratory
Tower Road • Ithaca, NY 14853-2901
Voice: 607-255-4520 • Fax: 607-255-6249 • E-mail: esb33@cornell.edu
An Equal Opportunity Employer

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Department of Genetics,
Development and Cell B
1210 Molecular Biology
Ames, Iowa 50011-3260
515 294-7322
FAX 515 294-6755

Dr. Volker Brendel
Bergdahl Professor of Bioinformatics

To:

Carolyn J. Lawrence, Ph.D.
MaizeGDB
1565 Agronomy Hall
Iowa State University
Ames, IA 50011

June 28, 2007

RE: MaizeGDB Project Plan

Dear Carolyn:

Thank you for sharing your MaizeGDB 5-year project plan with me. I am impressed with how well MaizeGDB has developed under your leadership, and I strongly support the vision for more sequence-centric views in the coming years as genome sequences will become available.

My group is committed to making all the maize data stored and generated at PlantGDB/ZmGDB (www.plantgdb.org and www.plantgdb.org/ZmGDB/) completely available, and we'll be happy to provide maize sequence annotation and gene model predictions to MaizeGDB for inclusion in the database and within the genome browser you are planning to make available within the next year. I think integration of the various plant databases and resources will remain a prominent topic in the next few years. It will be good to continue our successful collaborations on MaizeGDB and PlantGDB and set examples of what good integration should look like, both from a user perspective and from the technical side of database management.

With best wishes,



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

21 June 2007

Carolyn J. Lawrence, Ph.D.
USDA-ARS Research Geneticist

Dear Carolyn,

I was pleased to learn that your new USDA Project Plan includes efforts to incorporate data from the MAGI website into MaizeGDB. Over the last several years the MAGI site has become an important community resource. We have also provided GFF files from this site to a number of groups that wished to include our annotation on their local sites. We would be pleased to do so for MaizeGDB.

We would ask that you include a counter (that we have developed) within MaizeGDB to track the numbers of users who interact with our data. The resulting data provide information that is useful to our funding agencies.

Sincerely,



Patrick S. Schnable
Professor, Maize Genetics
Director, Center for Plant Genomics

J. Craig Venter

I N S T I T U T E

June 22, 2007

Carolyn J. Lawrence, Ph.D.
USDA-ARS Research Geneticist

Dear Carolyn,

This letter is to state my intent to collaborate with you and the MaizeGDB staff on integrating our maize array data into the overall efforts of MaizeGDB. We are happy to provide for you our maize data including AZM, repeats, gene annotation, etc. for inclusion in MaizeGDB. Our data can be viewed at www.maizearray.org. We are also excited that within the next year, this data will be linked to the genome browser you are planning to develop.

As the current funding for our project ends September 30, 2007, it will be essential that we provide you with this data this summer so that all needed quality control and other assessments can be completed before the project ends. I am pleased that MaizeGDB can provide such a service as it is important that all genomic data be archived and accessible regardless of the life span of any specific project.

Please do not hesitate to contact me or someone on the maize team here via maizearray_only@tigr.org.

Sincerely,



C. Robin Buell
Associate Investigator



Department of Biostatistics
and Computational Biology and
Department of Cancer Biology
Dana-Farber Cancer Institute

Professor of Biostatistics
Harvard School of Public Health

44 Binney Street
Boston, Massachusetts 02115-6084
617.632.3012 tel 617.632.2444 fax

July 12, 2007

Carolyn J. Lawrence, Ph.D.
Depts. of Agronomy and Genetics, Development, and Cell Biology
1565 Agronomy Hall
Iowa State University
Ames, IA 50011

Dear Carolyn,

Thank you for taking the time the other day to share your plans for continuing to expand MaizeGDB. I was very impressed with your vision to provide access to gene models calculated by the various gene structure prediction groups side-by-side through the MaizeGDB interface. I was honored to hear that you are using our Maize Gene Index (ZmGI) as part of MaizeGDB and that you are interested in displaying the ZmGI alongside the MAGI and PlantGDB predicted gene models.

As we discussed, I am willing to help in any way possible to help you achieve your goal. We recently implemented a web services portal to all of our Gene Index databases but would be more than willing to set up other data transfer and update protocols, including specialized flat files or database table dumps, necessary for you to use our resources to advantage. We would also be happy to work with you to test and debug the genome browser you are developing and to host you or someone from your group if you are interested in coming to work directly with us for a few days to get the data transfer process working.

I believe MaizeGDB's genome browser is going to be a wonderful resource for the Maize Research Community and I want to thank you for the opportunity to participate. My group and I are looking forward to working with you.

Best,

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

John Quackenbush
Professor of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
Professor of Cancer Biology, Dana-Farber Cancer Institute
Professor of Computational Biology and Bioinformatics, Harvard School of Public Health

IOWA STATE UNIVERSITY

Plant Sciences Institute

Center for Plant Genomics
2035C Roy J. Carver Co-Laboratory
Iowa State University
Ames, Iowa 50011-3050
515 294-7209
FAX 515 294-5256

2 July 2007

Carolyn J. Lawrence, Ph.D.
USDA-ARS Research Geneticist

Dear Carolyn,

On behalf of the Maize Genetics Executive Committee, I am writing to convey our continued enthusiasm to work with MaizeGDB. We appreciate the functionalities provided by your group in support our committee's activities, including administering community polls, organizing and managing the MGEC's elections, and sending out notifications to the community of maize researchers. We look forward to continuing to collaborate with your group in the coming years.

As you know, this past March immediately prior to the maize genetics conference leaders of the maize genetics community met for a retreat at Allerton to discuss the strengths, challenges, and initiatives that will define the future of maize research. The Allerton Report (<http://www.maizegdb.org/AllertonReport.doc>) is the community's assessment of the key biological issues that will define our community's research goals and the community resources needed to help us achieve these goals. Key to many listed needs is a requirement for data to be centralized and for MaizeGDB to become a more sequence-oriented resource.

We are happy that your Project Plan for the next five years addresses some of these needs. It is our hope that in the near future, resources will be allocated such that MaizeGDB could grow in both scope and depth to evolve into a more all-encompassing resource.

Sincerely,



Patrick S. Schnable
For the Maize Genetics Executive Committee:

Mary Alleman, Tom Brutnell, Ed Buckler, Karen Cone, Sarah Hake, Jo Messing, Mary Schaeffer, Jean-Philippe Vielle-Calzada, Marty Sachs, Pat Schnable (chair), Anne Sylvester, and Virginia Walbot



Lawrence Livermore National Laboratory

Tom Slezak
Associate Program Leader, Informatics & Assays
Chem/Bio National Security Program
Lawrence Livermore National Laboratory
7000 East Ave. Mail Stop L-174
Livermore, CA 94550
slezak@llnl.gov, 925-422-5746

June 24, 2007

Carolyn Lawrence
MaizeGDB

Dear Carolyn:

I am writing this letter in my capacity as chair of the MaizeGDB Working Group to affirm that this group would like to continue to serve as an external review panel for the efforts that you lead under USDA-ARS funding.

The Working Group has evolved with MaizeGDB over the past several years and we have seen that the Working Group has helped keep MaizeGDB abreast of current needs communicated by the community of maize researchers. Although the constituency of the Working Group has changed over time, and will in the future, the members are dedicated to assisting you to provide the most relevant resources for the maize research community.

We wish you the best of success in your project planning efforts.

Sincerely,

A handwritten signature in blue ink that reads "Thomas R. Slezak".

Thomas R. Slezak
Chair, MaizeGDB Working Group

Maize Genetics Steering Committee



Thomas Brutnell (*Chair*), Steve Moose (*Co-Chair*), Jorge Nieto-Sotelo, Erin Irish, Mei Guo, Peter Rogowsky, Mike Muszynski, Elizabeth Kellogg, Giuseppe Gavazzi, Pablo Rabinowicz (*ex officio*), Karen Cone (*ex officio*), Marty Sachs (*ex officio*), Mary Schaeffer (*ex officio*), Trent Seigfried (*ex officio*)

25 June 2007

Dr. Carolyn Lawrence
MaizeGDB Director
USDA/ARS, Iowa State University
Ames, IA

Dear Carolyn,

On behalf of the Maize Genetics Conference Steering Committee, it is my pleasure to provide you with a letter of support for your MaizeGDB's Project Plan. As you state: "In the 2006 Working Group Report (available online at http://www.maizegdb.org/working_group.php), MaizeGDB is cited as playing, "a central role in conducting central maize genetics community functions (i.e., with annual meetings, votes, surveys, and the Maize Newsletter)." The Working Group further states that, "this role should continue as it is critical to the success and cohesion of the research direction for the community."

We couldn't agree more. We rely on MaizeGDB to post announcements, organize the printed program for the meeting, host electronic abstract submission pages, post meeting schedules, itineraries and travel information, and display .pdf copies of presentations given at the maize genetics conference (http://www.maizegdb.org/maize_meeting/). At our last Committee meeting in March 2007, we discussed the importance of transparency for committee discussions and decisions and elected to have our minutes posted at MaizeGDB. This year, we are moving our registration deadline forward and are depending on MaizeGDB to provide us with email lists that have broad reach across our community to promote our meeting in a timely fashion.

The Steering Committee recognizes the importance of a centralized resource that will not only allow us to better communicate within the maize research community, but to reach out to the broader scientific community.

We wish you the best of luck with your proposal.

Sincerely,

A handwritten signature in black ink, appearing to read "T. P. Brutnell", with a stylized flourish at the end.

Thomas P. Brutnell, Chair Maize Genetics Steering Committee

APPENDIX – MAIZEGDB WORKING GROUP REPORT

CONTEXT

MaizeGDB personnel receive feedback from project stakeholders (the community of maize researchers) through messages sent through the Web site, via emails sent directly to the group, by way of personal contact, and by guidance from the MaizeGDB Working Group. The MaizeGDB Working Group is a panel composed of eight regular members, two *ex officio* members (our close collaborators Marty Sachs and Volker Brendel), and one chair (Tom Slezak). The panel gives MaizeGDB personnel individual suggestions throughout the year and meets once yearly as a group to evaluate the progress the MaizeGDB team has made as well as to suggest changes that would enable the team to best track the needs of the community of maize researchers. The Working Group replaces the Steering Committee, which guided the MaizeDB to MaizeGDB transition, and decides membership and membership tenure rather than leaving this task up to members of the MaizeGDB team. This appendix is a verbatim copy of the MaizeGDB Working Group's recommendations from the September 2006 meeting, which served in large part to guide the development of the Project Plan.

WORKING GROUP REPORT 2006

Volker Brendel, Ed Buckler, Karen Cone, Mike Freeling, Owen Hoekenga, Lukas Mueller, Marty Sachs, Pat Schnable, Tom Slezak, Anne Sylvester, and Doreen Ware

MaizeGDB is one of the most important resources for maize researchers, and it provides much of the glue that holds this international collection of scientists together. Excellent data has been well curated into MaizeGDB, and it is much more accessible than it was several years ago at MaizeDB. Overall, the MaizeGDB group has done a great job on limited resources. However, the next few years will provide major challenges to MaizeGDB, as the maize genome is sequenced and maize functional genomics research generates orders of magnitude larger datasets. MaizeGDB must play a central role in this revolution, to continue to serve the needs of the growing maize community. The working group identifies three central needs that will insure future MaizeGDB success:

(1) MaizeGDB needs greater resources. This year 10.9 billion bushels of corn will be produced with a farm gate value of \$37.5 billion. Additionally, the value for ethanol production could double this value over the next decade. Currently, there are 3.5 employees devoted to the efforts of MaizeGDB, which is 0.0015% of the value of the corn crop (0.03% of the R&D effort). This is meager support for a database that is the central clearing house for genomic and genetic data for such an important crop. These resources will not allow the maize community to synthesize the results of genomics adequately and will impede the ultimate transfer of knowledge from the basic plant biologists in the lab out to the applied community and into the marketplace.

The working group noted that TAIR has 6 times as many employees as MaizeGDB, while the Drosophila database (flybase.org) has 9 times as many. Database improvement and maintenance are labor intensive and ongoing tasks requiring continued infusions of resources. Given the importance of corn production to the US economy as a whole, MaizeGDB needs a level of investment at least commensurate with Drosophila and Arabidopsis given the crucial mission to providing linkage of data for agronomic improvement.

Additional resources can come from at least three sources. First, federal interagency cooperation could provide increased resources. Second, the MaizeGDB group needs to take an active role in funding their research through competitive grants, as they are now beginning to do. Third, MaizeGDB needs to insure that any future federally-funded projects that propose to make use of MaizeGDB for lasting curation of valuable data must include budgetary support for any data submissions, curation activities that MaizeGDB is asked to provide.

At least doubling or tripling of funding through various sources would greatly enhance the effort by MaizeGDB at Ames, Iowa. An additional two programmers and three curators at Ames would greatly accelerate the research. There are other bioinformatics groups around the country that can and are contributing to maize bioinformatics, but it is critical that MaizeGDB remain the vigorous leading group that synthesizes bioinformatics and makes it accessible to the entire research and agronomic community.

(2) MaizeGDB needs to prioritize goals and future plans. The MaizeGDB project has given careful thought to its future plans and asked for working group input on prioritization. The working group presents high and medium priority areas for consideration by MaizeGDB:

High Priority Areas

- Community service should continue.
 - o Currently, MaizeGDB plays a central role in conducting central maize genetics community functions (i.e., with annual meetings, votes, surveys, and the maize newsletter). This role should continue as it is critical to the success and cohesion of the research direction for the community. However, the working group recommends a careful self-evaluation of whether all actions are most efficient. For example, to optimize Trent Seigfried's time and make use of his skills, the group suggests that the Maize Genetics Meeting functions could be automated or otherwise not wholly dependent upon one of the two developers. It is important that MaizeGDB team continue to play this role, however, whenever possible they should get the curation resources from the community rather than providing them.
- Gene function prediction is important. Over the next few years, we will know the sequence of over 50,000 maize genes, but connecting genes to phenotypes is key to making genomics useful. MaizeGDB needs to play a leading role in curating, displaying, and analyzing the mutagenesis efforts in maize that will provide tools for functional analysis.
 - o Curation of published analyses – Currently, MaizeGDB is the leading resource for published mutants of maize, which has nearly a 100 year history of research. These efforts take considerable curation effort and the working group encourage the MaizeGDB team to promote alternative approaches (eg., via outreach) and to develop automated approaches to mine literature (examples coming from the honey bee community).
 - o Curation of stocks – Maize has a tremendous number and quality of genetic stocks. Continued curation of data regarding these stocks should be a top priority.
 - o Several high throughput mutagenesis programs (both transposon and chemically-induced) are underway. It is definitely in the community interest that these studies be well curated at MaizeGDB. Many of these projects were funded through competitive funding, but funds were not requested for MaizeGDB curation. It may be necessary for these projects, MaizeGDB, and funding agencies to add resources for appropriate curation of these projects. Additionally, MaizeGDB needs to define the file formats for inputs, essentially reusable standards for data deposition. These standards should be readily accessible from the website.
 - o Integrated views and analysis of mutagenesis and phenotypic effects to gene are needed. These views need to be both from a genome view and from a pathway view. Competitive grants could provide resources to support these views, since the research questions are fundamental to basic functional genomics.
 - o MaizeGDB needs to develop timelines for integration of the key datasets, and then publicize the timeline so that the research community knows when outcomes will be accessible.

- Maize Maps (Structural/Genetic/Cytogenetic) should continue to expand– MaizeGDB plays a key role in hosting and curating many of the maize genetic and cytogenetic maps that have been created over the last decade. However, as the sequenced genome for maize becomes available, the central maps will be changing to a B73 sequence. The main focus of MaizeGDB should be on linking relevant datasets to this sequence.
 - o MaizeGDB should take a leading role in integrating the IBM genetic maps with the B73 genome.
 - o When map data from the maize diversity study becomes available MaizeGDB should work with these researchers to develop a next generation genetic map.
 - o Since there is substantial variation in maize genome structure, additional genomes will be sequenced and where possible coherent views need to be built in MaizeGDB.
 - o MaizeGDB should focus on 2 major views - genetic and physical for now. Cytogenetic view although interesting and important has less general interest. Currently, there are often too many options, and users do not know what are the preferred maps.

Medium Priority Areas

- Gene structure prediction. As the genome is sequenced, many research groups around the nation are trying to predict gene structure and apply automated annotation of the genome. We do not believe that MaizeGDB should focus on doing this initial annotation but rather focus on:
 - o Integrating, recording, and presenting the leading gene models (currently three). They should ensure their software can relate these gene models to the sequence centric genome views.
 - o Organize experts in the communities to comment on the models (if funded as part of grant on this topic).
- Natural Diversity. Maize is the most diverse crop in the world, and USDA-ARS is responsible for the maintenance of this natural diversity. MaizeGDB needs to play a role presenting maize diversity and eventually helping plant breeders make use of this diversity. There are three other existing efforts ongoing in this area (Panzea, GRIN, Gramene), and MaizeGDB needs to work with these groups to develop an efficient display of information. The working group suggests the following approaches to deal with the intense curation effort required.
 - o MaizeGDB should only curate QTL studies for germplasm that is held by the Maize Stock Center. Currently, the only mapping population with repeatable data held by the stock center is the IBM population.
 - o The IBM QTL studies should be curated with their phenotypic score data into the GDPDM schema that is currently used by the maize diversity group, wheat, rice, and sorghum groups. With these phenotypic scores, it will be possible to reanalyze the data as computational and genetic methodologies improve in the future.
 - o Several groups will begin mapping QTL down to the gene level over the next two years. These gene level QTL –trait associations should be integrated with the MaizeGDB gene function prediction work.

(3) A Management Plan is needed. MaizeGDB is a young ARS group responsible for synthesizing, displaying and coordinating the outputs of the maize genetics/genomics

community. Management of competing needs and setting clear objectives is the key to developing an ever-improving MaizeGDB. It is essential that the MaizeGDB team set priorities to deal with the plethora of extant and emergent data. In particular, there needs to be coordination between the Iowa and Missouri teams in order to distribute the time efforts more efficiently to meet common goals. The working group suggests that the unified MaizeGDB team articulate together their goals in the short, mid and long term. Most importantly, the team needs to articulate how and when these priority goals will be met. The working group recognizes that the priority goals suggested in this report require far more resources than are currently available. We suggest that two management plans could be devised: one that can be carried out within the budget limits of the current USDA-ARS commitment and a second that would be possible if appropriate resources were available. In this way, it will be transparent what is and is not possible within the resource constraints.

Short-term actions need to serve the long-term goals of MaizeGDB. The working group suggests that activities within the team need to be driven by a broader vision of the purpose and goals of MaizeGDB. For example, it seems that current action decisions are made based on urgency and the most persistent requests from the community. The team needs to take a leadership role in defining how, when and which datasets should be curated. If there are commonly agreed short and long-term actions, then all action decisions can be made based on helping to move MaizeGDB towards the long-term goals.

In summary, the working group makes the following suggestions for developing a management plan:

- Datasets to be curated and displayed need to be prioritized. Participant actions and timelines must be established in a systematic fashion.
- Numerous groups would like to have their data displayed at MaizeGDB, but MaizeGDB should feel free to say no if these data do not match the long-term goals.
- As in any scientific endeavor, it is important to devote substantial time to development of the newest views or approaches rather than just putting out fires.
- The PoPcorn portal was very interesting, and we encourage MaizeGDB to develop an independent and diverse portfolio of competitive grant ideas and funded projects.
- The management plan should be in written form and consistent with realistic expectations.

Publications (†Invited *Peer Reviewed)

- *† **Lawrence, C.J., Harper, L.C., Schaeffer, M.L., Sen, T.Z., Seigfried, T.E., and Campbell, D.A.** MaizeGDB: The Maize Model Organism Database for Basic, Translational, and Applied Research *International Journal of Plant Genomics*. (under review).
- * Lushbough, C., Bergman, M.K., **Lawrence, C.J.**, Jennewein, D., and Brendel, V. BioExtract Server – An Integrated Workflow-enabling System to Access and Analyze Heterogenous, Distributed Biomolecular Data *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (under review).
- * Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M.D., **Lawrence, C.J.**, Lushbough, C., and Brendel, V. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Research* (in press).
- * The Gene Ontology Consortium (**Seigfried, T.E.** is a member of the Consortium) The Gene Ontology project in 2008. *Nucleic Acids Research* (in press).
- † **Harper, L.C., Sen, T.Z., and Lawrence, C.J.** Plant Cytogenetics in Genome Databases in *Plant Cytogenetics Volume 1: Genome Structure and Chromosome Function* H. Bass and J. Birchler (Editors) Springer (in press).
- † **Lawrence, C.J.** and Ware, D. Databases and Data Mining in *The Maize Handbook*, S. Hake and J. Bennetzen (Editors) Springer (in press).
- † **Lawrence, C.J.** and Walbot, V. Reply: specific reasons to favor maize in the U.S. (2007) *Plant Cell* 19(10):2973.
- *† **Lawrence, C.J.** and Walbot, V. Translational Genomics for Bioenergy Production from fuelstock Grasses: Maize as the Model Species (2007) *The Plant Cell* 19(7):2091-2094.
- * **Lawrence, C.J., Schaeffer, M.L., Seigfried, T.E., Campbell, D.A., and Harper, L.C.** MaizeGDB's new data types, resources, and activities (2007) *Nucleic Acids Research* 35(Database issue):D895-900.
- † **Lawrence, C.J.** MaizeGDB, the Maize Genetics and Genomics Database in *Plant Bioinformatics*, D. Edwards (Editor) for the series *Methods in Molecular Biology* (2007) Humana Press, Totowa, New Jersey, USA.
- * Yim, Y.S., Moak, P., Sanchez-Villeda, H., Musket, T.A., Close, P., Klein, P.E., Mullet, J.E., McMullen, M.D., Fang, Z., **Schaeffer, M.L.**, Gardiner, J.M., Coe, E.H. Jr, and Davis, G.L. A BAC pooling strategy combined with PCR-based screenings in a large, highly repetitive genome enables integration of the maize genetic and physical maps (2007) *BMC Genomics* 8:47.
- * Ilic, K., Kellogg, E.A., Jaiswal, P., Zapata, F., Stevens, P.F., Vincent, L.P., Avraham, S., Reiser, L., Pujar, A., Sachs, M.M., Whitman, N.T., McCouch, S.R., **Schaeffer, M.L.**, Ware, D.H., Stein, L.D., and Rhee, S.Y. The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant (2007) *Plant Physiology* 143(2):587-99.
- * Shyu, C.R., Harnsomburana, J., Green, J., Barb, A.S., Kazic, T., **Schaeffer, M.**, and Coe, E. Searching and mining visually observed phenotypes of maize mutants (2007) *Journal of Bioinformatics and Computational Biology* 5(6):1193-213.
- * Wei, F., Coe, E., Nelson, W., Bharti, A.K., Engler, F., Butler, E., Kim, H., Goicoechea, J.L., Chen, M., Lee, S., Fuks, G., Sanchez-Villeda, H., Schroeder, S., Fang, Z., McMullen, M., Davis, G., Bowers, J.E., Paterson, A.H., **Schaeffer, M.**, Gardiner, J., Cone, K., Messing, J., Soderlund, C., and Wing, R.A. Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History (2007) *PLoS Genetics* 3(7):e123

MaizeGDB Project in Ames, IA (3625-21000-045-00D)

Total available: **\$567,364**
 \$467,364 permanent
 \$100,000 temporary transfer on yearly basis from Columbia, MO.

Item	Sub-items	\$	%
Salaries and Benefits	Total Salaries (Lawrence, Sen, Seigfried, Campbell, Harper)	363443	64
†PGECH (Harper)	Supplies	3000	0.5
	Infrastructure	7000	1
Travel	PAG (x5), Maize Meeting (x5), ISMB (x3), Workshops (x3), Working Group Meeting (variable), and incidental trips (variable)	54500	10
Contracts	Oracle support	8100	1
RSA	ISU on-campus supplies spending	3222	0.5
All Other (Includes Supplies/Materials)	Computer hardware, software, furniture, office supplies, publication costs, move to CGIL, etc.	128099	23
	<i>Total</i>	<i>567364</i>	<i>100</i>

New NSF Subcontracts

1. Construction Of Comprehensive Sequence Indexed Transposon Resources For Maize – Don McCarty, PI

July 1, 2009 thru June 30, 2010

\$73,885 to ISU to curate stock links for ordering, the annotated location of each uniquely identified mapped insertion in the maize genome that the line contains, and any phenotype information available for that line as well as to create methods to store and a pedigree index number and will work toward the creation of methods to display pedigree relationships to other UniformMu lines. Methods for accessing *Mu* data will be made possible by new bulk download mechanisms.

2. The grass regulome initiative: Integrating control of gene expression and agronomic traits across the grasses – Erich Grotewold, PI

September 1, 2009 thru August 31, 2010

\$26,206 to ISU to store the data generated by the project personnel's experimental characterization of transcription factors (TFs) and their direct targets, as well as the *cis*-regulatory elements (CRE) that a select group of TFs recognize in the corresponding target genes.

Genome Browser Survey

Dear Cooperator,

The primary mission of MaizeGDB is to serve the maize community, and the MaizeGDB team is very fortunate that maize researchers are dedicated to guiding that mission

A main goal for MaizeGDB is to facilitate access to the outcomes of maize research in an intuitive format. In that spirit, MaizeGDB is investigating the possibility of adopting and customizing a genome browser as the basis for displaying the annotated maize sequence as it becomes available. While it should be noted that MaizeGDB currently does not have the resources or expertise on staff to annotate the maize genome directly, displaying all available structural (gene model) and functional annotations simultaneously within the context of a genome browser is possible, as is making all relevant data available for download and off-site manipulation. If an outside group is funded to officially annotate the maize genome, MaizeGDB will pursue collaboration to house the official annotations long-term and will seek to create mechanisms for continued community annotation once the primary annotators' project has ended.

As you are probably aware, there are many different genome browsers, each having strengths and weaknesses. We are looking for maize researchers (students, postdocs, faculty, and others) who are willing to help choose a genome browser software for MaizeGDB by taking the following survey. Even if you don't have any experience with genome browsers, please give us a few minutes of your time to contribute your thoughts. You will be the ones using the genome browser regularly, so your input is of the utmost importance for planning purposes.

Feedback to the MaizeGDB team is very important for getting the genome browser software selection and implementation right. At the end of the survey you have the option to reveal your identity if you so desire. This would enable you to work with the team to make sure that the tools to be adopted and developed meet your needs.

Thanks very much for your time and interest,
The Maize Genetics Executive Committee

Please fill out as many of the questions below as you can. If you don't have an appropriate answer for a question, don't worry -- just continue to the next question.

How many hours per week do you personally access maize data from any web source?

0 hours

Please list the websites you use most for your maize research. Please rank them in order of the most time spent, with #1 being the site you spend the most time on (e.g., ZmGDB, MaizeSequence.org, and MAGI, etc.).

1	
2	
3	
4	
5	
6	
7	

What types of **genome browsers** have you used? Please check the browsers below that you have experience with. Please write in any that are not listed (browsers for any species are welcome). Links open in a new window, so you won't lose the survey by clicking.

- ☐ Ensembl (<http://www.ensembl.org>)
- ☐ FlyBase (<http://flybase.org/>)
- ☐ GBrowse (<http://www.gmod.org/wiki/index.php/GBrowse>)
- ☐ Gramene (<http://www.gramene.org/>)
- ☐ JGI (<http://www.jgi.doe.gov/>)
- ☐ MAGI (<http://magi.plantgenomics.iastate.edu/>)
- ☐ MaizeSequence.org (<http://www.maizesequence.org/>)
- ☐ NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4577)
- ☐ PlantGDB (<http://www.plantgdb.org/>)
- ☐ Pombe Browser (<http://pombe.ncbi.nih.gov/genome/>)
- ☐ Sigenae Contig Browser (<http://public-contigbrowser.sigenae.org/>)
- ☐ SynBrowse (<http://www.synbrowse.org/>)
- ☐ TAIR (<http://www.arabidopsis.org/>)
- ☐ T-DNA Express (<http://signal.salk.edu/cgi-bin/tdnaexpress>)
- ☐ TIGR (Craig Venter Institute) (<http://www.tigr.org/>)
- ☐ UCSC Browser (<http://genome.ucsc.edu/>)
- ☐ VEGA (<http://vega.sanger.ac.uk/>)
- ☐ WormBase (<http://www.wormbase.org/>)
- ☐ Others (please list: _____)

What **genome browsers** are or would be the most useful for your research? (You can give examples of web sites of other species you enjoy using.)

Which aspects of your favorite **genome browser(s)** do you like? Please rank them according to your preference (1 is the most important) and provide genome browser names as examples. If you don't see an aspect mentioned but you think it is important, please include them as the last rows of your response.

--

Speed (give examples, if any:

)

--

Ease of use (give examples, if any:

)

--

Visually informative (give examples, if any:

)

--

Allows cross-species (or multiple inbred line) comparisons (give examples, if any:

)

--

Ability to highlight or visualize multiple genes (give examples, if any:

)

--

Distinction between experimental and computationally-derived data (give examples, if any:

)

--

Links to Gene Ontology/Plant Ontology/etc terms (give examples, if any:

)

--

Other (give examples and/or comments, if any:

)

Among these features, which ones are the most indispensable for your research?

What specific features would you like to see in a MaizeGDB **Genome Browser**? (even if the feature does not exist in currently available genome browsers)

Can you provide examples of features of **genome browsers** that are frustrating or not useful (please list the specific web sites)?

Name and/or Web Address	

List here any other comments about **genome browsers** and your preferences. Please be as specific as you can

Would you be willing to provide more feedback?

Yes

Would you be willing to serve on a guidance panel?

Yes

Would you be willing to be a beta tester?

Yes

If you answered **yes** to any of the last three questions, please provide:

2 of 3

2008 MaizeGDB Working Group Report

1/8/08 11:17 AM
Page 132 of 133

your name:

your email:

This information will *not* be associated with your survey responses!

Submit

MaizeGDB



Maize Genetics and Genomics Database

This image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and extend across the width of the page. There is no text or other markings on the paper.

MaizeGDB



Maize Genetics and Genomics Database

This image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and extend across the width of the page. There is no text or other markings on the paper.

MaizeGDB



Maize Genetics and Genomics Database

This image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and extend across the width of the page. There is no text or other markings on the paper.