MaizeGDB
Maize Genetics and Genomics Database

MaizeGDB STATUS REPORT

RECENT UPDATES, ACTIVITIES, AND NEW INITIATIVES

USDA-ARS
Project No. 3625-21000-051 (Ames, IA)
and
Project No. 3622-21000-027 (Columbia, MO)

Prepared by: The MaizeGDB Team

Carson M. Andorf, Lou Butler, Darwin A. Campbell, Ethalinda (Ethy) Cannon, Lisa C. Harper, Carolyn J. Lawrence, Mary L. Schaeffer, and Taner Z. Sen

JANUARY 2010

Contact: C. Lawrence
USDA-ARS
1034 Crop Genome Informatics Laboratory
Iowa State University
Ames, IA 50011
Email: Carolyn.Lawrence@ars.usda.gov
URL: http://www.maizegdb.org
515-294-8280 (fax)
515-294-4294

TABLE OF CONTENTS

1 – Meeting Agenda **(Shown as Pacific Time)**

3:45 p.m. Dial in to the conference call
1-888-282-8355 PASSCODE 13781

4:00 p.m.	Welcome and overview	Carolyn
	Genome Browser	Taner
	Interface development	Carson
	Curation and outreach	Mary & Lisa
	POPcorn	Ethy
	Database	Darwin
	Recap and charge to the Working Group	Carolyn
5:00 p.m.	Working Group Executive Session	
5:50 p.m.	Working Group Summarizes for MaizeGDB Team	
7:00 p.m.	Optional dinner in Old Town: meet at the train station	

Working Group's Role

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.

Working Group Membership

Volker Brendel*, Mike Freeling, Owen Hoekenga, Anne-Francoise Lamblin, Thomas Lübberstedt, Karen McGinnis, Lukas Mueller, Mihai Pop (chair), Marty Sachs*, Pat Schnable, Anne Sylvester, and Doreen Ware.

Additions: Thomas Lübberstedt (unable to attend)

- These close collaborators are considered *ex officio* members of the Working Group.

2 – Status

Executive Summary

Since the Working Group Meeting in May of 2009, our focus on data curation has been reinvigorated with: further development of the *Maize Gene Review*, literature curation based upon Editorial Board input, the addition of Lou Butler to incorporate and improve upon M.G. Neuffer's lesion mutant data descriptions, and the addition of a 2/3 curator (currently vacant) to integrate CIMMYT data as well as GRASSIUS and Uniform*Mu* resources. The MaizeGDB Genome Browser now is available in two versions: a BAC-based browser and a newly released B73 RefGen_v1 browser. Links out to other views are available, data including centromeres are mapped, and additional data tracks are planned. With the Locus Lookup Tool in place and plans to show inferred centiMorgan distances within the browser, MaizeGDB is on the cutting edge of integrating genetic information with sequence data. The POPcorn resource has continued to develop despite our problems with recruiting a programmer. The goal of sequenced-indexed searching is being addressed with two utilities: *POPcorn BLAST*, which permits researchers to BLAST against maize sequence at multiple sites, and *Sequence-Indexed Data Search*, which carries out multi-step searches of sequence-indexed data. Efforts to create a virtual server environment are underway with anticipated full virtualization planned for Spring of 2010. The next major endeavor will be to create a Pathway View, as outlined in the MaizeGDB Project Plan. Current plans are to work with Gramene to further develop their implementation of MaizeCyc as a joint venture.

Personnel:

Federal

Carolyn Lawrence, research geneticist and lead scientist, USDA-ARS in Ames, IA
Taner Sen, computational biologist, USDA-ARS in Ames, IA
Carson Andorf, IT specialist (bioinformatics engineer), USDA-ARS in Ames, IA
Darwin Campbell, IT specialist (database administrator), USDA-ARS in Ames, IA
Mary Schaeffer, geneticist (curator), USDA-ARS in Columbia, MO
Lisa Harper, geneticist (curator), USDA-ARS in Albany, CA

State

Ethalinda (Ethy) Cannon, POPcorn solution/application architect, ISU in Ames, IA
Lou Butler, Program coordinator (term curator), ISU in Tucson, AZ
VACANT, POPcorn programmer, ISU in Ames, IA
VACANT, 2/3 curator, ISU – location to be determined, funds from USAID via CIMMYT and NSF via GRASSIUS and UniformMu

Noteworthy:

Lawrence, **Sen**, **Andorf**, and **Campbell** were members of the group awarded the USDA-ARS Midwest Area's Outreach, Diversity, and Equal Opportunity Award in 2009 for Outreach to American Indians (led by C. Lawrence and C. Gardner; funded by NSF)

Campbell was awarded the USDA-ARS Midwest Area's Research Support Award in 2009.

Response to May 2009 Working Group guidance
Report by Carolyn Lawrence in response to "May 2009 Guidance" (accessible online at http://www.maizegdb.org/working_group.php)

- BLAST queries to proteins are now supported.
- Linkage to remote sites from the MaizeGDB Genome Browser are now in place.
- Microarray probes mapped to the BAC-based Genome Browsers are in place. As other datasets become available and are mapped to the B73 RefGen_v1 sequence, they will be incorporated.
- Comparative tools are planned, but thus far are not yet implemented. Linkages into CoGe are being incorporated currently.
- Release of RefGen_v1 has allowed for immediate data consistency. Continued use of a particular genome assembly will be a requirement from now on. Sequence version information soon will be more transparent and the GenBank versions of sequence data will be noted.
- Curation is taking center stage and we are evolving curation partnerships with Gramene (MaizeCyc), CIMMYT (moving their molecular characterizations into MaizeGDB), UniformMu, GRASSIUS and others to leverage existing resources and identify new resources to support curational efforts.

The MaizeGDB Genome Browser
Report by Taner Z. Sen

In continuing our efforts to develop the MaizeGDB Genome Browser, our main goal is to create a data-consistent view by displaying genomic data mapped to the same release of the pseudomolecule sequence (B73 RefGen_v1). To fulfill this aim, we obtained pseudomolecule-mapped data from various research groups, integrated the new data with our database, and created new Web views.

Our biggest challenge was to encourage maize researchers to align their data to the pseudomolecule in time for our release in November. Although the pseudomolecule sequence was available since Spring 2009, most groups were not able to map their data to the pseudomolecule before MaizeSequence.org's release in November. Even now, some maize data mapped previously to BACs are still not mapped to the pseudomolecule (for example JGI's Mo17 SNP data).

Our accomplishments can be summarized as follows:

1) We kept the pre-pseudomolecule version of the MaizeGDB Genome Browser intact. We renamed it as **the "BAC-based" browser**, and created interlinks within the new pseudomolecule browser.
2) Our aim when serving the pseudomolecule-mapped data in the MaizeGDB Genome Browser was to ensure **data consistency**. We therefore created tracks only for data mapped to the B73 RefGen_v1. We currently have the following data in the pseudomolecule-based genome browser: the B73 RefGen_v1 sequence; centromere regions and anti-CENH3 ChiP data from the Jiang/Presting groups; MAGI data from the Schnable lab; ESTs, cDNAs, and *Ac/Ds* from PlantGDB; and MIPS repeats and filtered gene set from MaizeSequence.org.
3) We included **outgoing links** ("buttons") at the top of the pseudomolecule-based MaizeGDB Genome Browser to allow users to view the same genomic regions from the BAC-based context and via the Maizesequence.org and PlantGDB resources.
4) Quality of the Filtered Gene Set from MaizeSequence.org are now shown as a color-coded track served via DAS from PlantGDB, as is a track showing community annotations of gene structure.

In the future,

1) To assist maize cooperators to align their data to the current and future versions of a B73 genome assembly, we are in the process of creating a **BLAST service** that will enable users to align large datasets to the pseudomolecules. This BLAST service will allow nucleotide and protein queries.
2) We will integrate **more data** into the MaizeGDB Genome Browser in the near future: PlantGDB's assembly of GSSs and PUTs, UniformMu, and IBM 2008 markers aligned by the Maize Genome Sequencing Consortium.
3) When available, new tracks will be created for the following data: Microarray probes from PLEXdb, Mo17 indels from JGI, SNPs from Ed Buckler from the Diversity Project, SNPs from Pat Schnable's group, and Dana-Farber TCs from JGI.
4) Currently the browser notes how many base pairs are represented in a given screen. Soon, the approximate number of centiMorgans also will be noted based upon stored genetic maps.

Interface and tool development
Report by Carson Andorf

Since May of 2009, I have made changes to 193 web pages on their production site. This includes HTML pages, record pages (pages that displays data from the MaizeGDB database), and tool pages (e.g., Genome Browser, Locus Lookup). These changes do not include static files such as images and documents. They changes include new pages, functionality upgrades, bug fixes, content changes, and integration with new data.

One of the major goals of the past six months is the integration of MaizeGDB interface with the recently released Maize B73 RefGen_v1 pseudomolecule sequence. Currently we have created or integrated the following pages and tools with the new data:

1. Genome Browser
2. Chromosome image viewer
3. Locus Lookup
4. Locus Pair Lookup
5. BLAST tool
6. EST record pages
7. cDNA record pages
8. BAC record pages

In addition to the integration with the pseudomolecule sequence, the Locus Lookup and the BLAST tools' functionality have been upgraded. The Locus Lookup allows for more flexible searching. The BLAST tool offers faster searching and additional visualization displays.

Highlights of other examples of new or modified MaizeGDB pages can be found below:
Newly characterized maize gene page: This page now displays all newly updated maize genes sorted by the most recent updated gene. An "updated" gene is defined as a gene that has had a recent reference, a new gene product, a new variation, or has been recently recommended by a maize scientist.

Mo17 SNP page: This page has been upgraded to show a detailed view of the B73 and Mo17 sequence centered on a specified Mo17 SNP. This region will show any nearby insertions, deletions, or substitutions. The user can change the region size to 81, 241, 401, or 561 base pairs. An integrated tool will also "predict" the Mo17 sequence for the specified region.

Maize Meeting 2010 website: The Maize Meeting website has been redesigned by MaizeGDB this year. The Missouri site and the MaizeGDB site have been consolidated and a new interface has been added.

MaizeGDB curation and outreach

Report by Mary Schaeffer

Community curation - via the ***maize gene review***, sponsored by the Maize Genetics Cooperation Newsletter, J Birchler, M Schaeffer eds. This online journal has peer review, and will be assigning DOI accessions. See www.maizegenereview.org.

The goal is to engage community experts in providing succinct descriptions of empirically confirmed locus functions to MaizeGDB. Authors provide one or two paragraph summaries of a gene with reference citations and an unpublished image with a caption; information about key alleles; regulation of/by other genes or genetic elements; gene product, function, pathways (GO terms welcome); and any other information, e.g. paralogous loci. Each contribution counts as a reference citation in the 'maize gene review', and 'gold stars' at MaizeGDB. Some 24 reviews have been published online, and in the 2009 print copy of the MNL Some 50 others have been promised and many of these are in process. 'In process' includes entering genes and sequence accessions, with the reference citation. Volunteered submissions, genes published in papers recommended by the editorial board and any newly sequenced genes take priority in this process.

As a MaizeGDB curator, I transfer this information to MaizeGDB, along with GenBank links for each allele, and, where possible, link each review to the genome browser using the cDNA accession. Formatting of the journal website and links to MaizeGDB is aided by part-time graduate students in the School of Journalism.

MNL back issues. Almost all of the past MNL, which go back to early 1930's, are being double checked for linking to the MaizeGDB citations.

Working with MG Neuffer and Lou Butler towards updating Neuffer's text descriptions of mutant images at MaizeGDB and summary articles for the *maize gene review*.

Genetic Maps Curation

In progress.

A Consensus IBM genetic map (Schnable et al.) with some new loci and aligned to the B73 RefGen_v1 pseudomolecules.

In the future.

The only large maps on the horizon may come from the extensive SNP genotyping of the 5000 RILs from the NAM project.

Community Outreach

Organized a new workshop, Sunday morning, Jan 10, at PAG, on Plant Phenotypes. This year speakers represent Oryzabase, TAIR, Soybase, CGIAR, SGN, LemnaTec. The introductory remarks by Chi-Ren Shyu will show his image lookup tool, which uses images and text from MaizeGDB, along with the Plant Ontologies. A goal of the workshop is to discuss strategies for representing phenotype and related 'omics data; a panel discussion is planned for last 30 minutes of the workshop.

Hosted the Plant Genome Databases Outreach Exhibit Booth at PAG 2010 meetings. This year there are 13 databases.

Curation and outreach (continued)
Report by Lisa Harper (half time)

Literature curation: The current policy is for our five-member Editorial Board to nominate about 5 papers per month, which are curated (that is, the data is entered into MaizeGDB). The 2009 Board was: Mike Scanlon, Jane Dorweiler/Lyudmila Sidorenko, Peter Balint-Kurti, Randy Wisser and Cliff Weil. The 44 papers entered June-Dec 2009 can be viewed at:
http://www.maizegdb.org/cgi-bin/editorial_board.cgi#

Papers are also added to MaizeGDB by request from individual researchers, and where it appears to curators that the data will be high-impact (about 20 such papers in 2009).

Editorial Board members serve one year, and in Decemeber 2009, each nominated several replacements. The new Board for 2010 is: Erik Vollbrecht, Pat Brown, Mario Arteaga-Vazquez, Nick Lauter and Paula McSteen.

Phenotype curation: As of December 2009, I have received and reviewed expert help for the revision of approximately 200 phenotypes – about 17% of all phenotypes we currently list. These revisions are written, but not yet uploaded. To collect information from experts in the phenotype area I contact researchers by email, phone, and in person. Each time I contact someone, I gently explain the great benefits of having a controlled vocabulary, and I have gotten full and enthusiastic cooperation from most everyone. As a long time researcher in the maize community, I understand that asking maize researchers to use controlled vocabulary is "difficult".

Tutorials and Outreach:
The tutorial movie showing caveats of the BAC-based Genome Browser was reviewed by many people, including the Maize Sequencing Consortium, and released in June 2009. It was generally well received and is thought to be informative based on email, verbal feedback, and online ratings (4.5/5 stars).

A new B73 RefGen_v1 pseudomolecule-based genome browser tutorial movie is in progress and will be released at the end of January 2010.

Curation and outreach (continued)
Report based on the POPcorn supplement proposal authored by Lou Butler and Carolyn Lawrence.

A one-year supplement from NSF to the POPcorn grant was requested to enhance and expand the data in MaizeGDB for M.G. Neuffer's (MGN's) extensive maize mutant collection. It is well known that maize is an excellent system for basic research based on determining gene function through mutation, and MGN's collections of mutants and his careful analysis of them is unparalleled in the maize community. Documenting these treasures in MaizeGDB for others to leverage these mutants for their own research purposes is the focus of the work described here.

The purpose of this collaboration is to 1) improve existing mutant phenotypic data for maize mutants already posted within MaizeGDB and 2) further characterize and standardize new data that have been recently generated so that the new mutations are consistent with the format of the original dataset. The new mutants are a collection of approximately 200 EMS-induced dominant mutants and a smaller collection of disease lesion mutants generated by MGN.

Much of the data to be documented already are available and simply must be added to the MaizeGDB resource. The data are being entered by Lou Butler (LB). She prepared and entered all the original datasets and prepared all the images for MGN's mutants into MaizeDB and assisted MGN in the preparation of *Mutants of Maize*. Because LB was already familiar with the terms and general layout of MaizeGDB, little to no training time or direction was required. LB has been given full curation privileges to MaizeGDB and collects data directly from MGN via Skype in regular sessions. This allows MGN to gather manageable groups of data from original sources in his office in Missouri, and give the information to LB (who lives in AZ) over a series of phone sessions; she enters this information directly into MaizeGDB. After each session, she reviews the work completed and determines whether crosslinking is necessary and maintains communication with MaizeGDB personnel regarding progress and to make requests for technical assistance. This arrangement allows MGN to focus on gathering, organizing, and confirming data and images while LB focuses on entering the data appropriately into MaizeGDB.

Quality of existing mutant images is being assessed by MGN and LB as they are processed for upload into MaizeGDB. Where image quality is determined to be inferior, new ones are being generated. In addition, for newly discovered lesion mutants digital images are being captured and optimized for the first time.

POPcorn

PI: Carolyn Lawrence, Co-PI: Taner Z Sen, Solution/Application Architect: Ethy Cannon
Report by Ethy Cannon

As of December 2009 there are 71 projects and 112 resources in the searchable project database. This database contains maize projects (funded or non-funded research efforts) and resources (online data, tools, and/or information, typically but not always, associated with a research project). For each project and resource we collect a short description, lists of institutions and investigators, and project funding information. Projects and resources are tagged by one or more categories (e.g. "sequencing", "mutation", "breeding"). Where applicable, resources are associated with related projects and projects with each other. The site is available here: http://www.maizegdb.org/POPcorn/.

The goal of sequenced-indexed searching is being addressed with two utilities: *POPcorn BLAST*, which permits researchers to BLAST against maize sequence at multiple sites, and *Sequence-Indexed Data Search*, which carries out multi-step searches of sequence-indexed data.

As of December, 2009, *POPcorn BLAST* can BLAST against target sequence at MaizeGDB, PlantGDB, GRASSIUS, and NCBI.

As of December, 2009, *Sequenced-Indexed Data Search* permits searching for seed stock containing a specific mutation (*Ac*/*Ds*, Uniform*Mu*, TILLING), matching loci, phenotypes, and PlantGDB PUT identifiers. Results contain summary information about the match and links to the original source for more information.

An example of a sequenced-indexed data search is locating Uniform*Mu* seed stock for a given sequence:

1. BLAST query sequence against NCBI GSS with entrez search term "*UniformMu*"
2. Parse definition line to get locus name
3. Get LOCUS record at MaizeGDB
4. Get VARIATION records for that LOCUS at MaizeGDB
5. Get STOCK_GENOTYPIC_VAR records for each VARIATION at MaizeGDB
6. Get STOCK records for each STOCK_GENOTYPIC_VAR at MaizeGDB
7. Construct links to MGC Stock Center for each stock found.

POPcorn carries out all of these steps on behalf of the researcher.

Two risks of the sequenced-based searching are obscuring the provenance of the data and a poor interpretation of the results. To this address both concerns, POPcorn is collecting attribution and citation information, links to collaborator sites, and short descriptions are to be displayed prominently in the POPcorn pages. Curation pages will allow project PIs to maintain this information themselves.

Prototype versions of *POPcorn BLAST* and the *Sequence-Indexed Data Search* will be released for community input in February 2010.

Collaborators during this period include MaizeGDB (http://maizegdb.org), PlantGDB (http://plantgdb.org), GRASSIUS (http://GRASSIUS.org), and BioExtract (http://bioextract.org).

Database
Reported by Darwin Campbell

Data additions in the past year have resulted in 1.7M new records with the addition of 14 new tables supporting the needs of MaizeGDB and POPcorn development. The implementation of the MaizeGDB Genome Browser added 20 new tables and nearly 13M records. We maintain our weekly database backup process both locally and remotely (University of Missouri in Columbia, MO.)

Recent focus is on the implementation of a virtual server environment. A virtual server environment aligns with Executive Order *13423: Strengthening Federal Environmental, Energy, and Transportation Management, sec. 2(h); sec. 3(a, and (f))* by improving energy efficiencies and extending computer replacement lifecycles by reducing the number of physical computers while maximizing existing hardware capacity. Virtual servers interact with each other and users just like physical servers, without dedicating new or specific hardware to a specific purpose. New virtual servers allow the testing of new operating systems, software patches and application interactions with the ability to roll changes out of the testing procedure with out starting over.

After careful consideration and investigation of virtualization platforms and consultations with IT specialists who work with an installed virtual environment, we purchased 6 core licenses of VMware ESX 3.5 (enough for 3 servers with 2 cores per server). ESX is the physical server operating system that supports the creation/operation of virtual servers and one instance of vCenter Server; the VMware application that manages the physical host servers; virtual server load balancing and high availability.

The virtual environment is made up of three identical servers; each acts as a host. Two will be in the CGIL building and the third will be placed in another building to act as the "high availability" server. In the event of a failure of one of the physical hosts in CGIL, vCenter Server on the SAN device will sense the failure and activate a copy of the virtual server(s) on the spare physical server, which will pick up the services provided by the failing host with priority given to production virtual servers.

We plan on full implementation by the end of Spring 2010.

5 – Topics of requested input

1. What are some reasonable ways to encourage researchers to document their use of MaizeGDB in publications? Acknowledgements? Citations? Sending out a notice with each large dataset shared that its use should be documented in resulting publications? Any ideas here would be most welcome.

2. Please comment on the NSF supplement supporting Gerry Neuffer and Lou Butler. Is this the sort of thing that should be pursued again? Is it a good mechanism for collecting data from senior or retired members of our research community? Perhaps from active persons who didn't or don't have a component in their budget to transfer data to MaizeGDB as well?

3. MaizeGDB has evolved a paradigm of centralizing IT work (Ames) and locating curators where a large group of maize researchers is located (Berkeley/Albany, CA; Columbia, MO; and Tucson, AZ). Is this reasonable? Should this sort of thing be continued? Expanded?

4. As more lines are sequenced, what sorts of data analysis will be needed? More specifically and more urgently, what is the best way to represent the HAPMAP of maize and what sorts of analysis tools are going to be needed? Because the human data rely on populations, not inbreds, it is likely that their tools will not leverage this aspect of plant breeding.

5. Please comment on our additions and improvements in curation. Have we improved the areas of curation that were of concern to you? Are there specific areas of data curation we should be focues on?

3 – Appendix: peer-reviewed publications since last meeting

Published and available online at **http://database.oxfordjournals.org/cgi/content/full/2009/0/bap020**

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2009, Article ID bap020, doi:10.1093/database/bap020

Original article

MaizeGDB becomes 'sequence-centric'

Taner Z. Sen[1,2], Carson M. Andorf[1], Mary L. Schaeffer[3], Lisa C. Harper[4,5], Michael E. Sparks[2], Jon Duvick[2], Volker P. Brendel[2,6], Ethalinda Cannon[2], Darwin A. Campbell[1] and Carolyn J. Lawrence[1,2,*]

[1]USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, [2]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, [3]USDA-ARS Plant Genetics Research Unit and Division of Plant Sciences, University of Missouri, Columbia, MO 65211, [4]USDA-ARS Plant Gene Expression Center, Albany, CA 94710, [5]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720 and [6]Department of Statistics, Iowa State University, Ames, IA 50011, USA

*Corresponding author: Tel: +1 515 294 4294; Fax: +1 515 294 8280; Email: carolyn.lawrence@ars.usda.gov

Present address: Michael E. Sparks, USDA-ARS Bovine Functional Genomics Laboratory, Beltsville, MD, USA

Submitted 15 May 2009; Revised 24 June 2009; Accepted 11 July 2009

MaizeGDB is the maize research community's central repository for genetic and genomic information about the crop plant and research model *Zea mays* ssp. *mays*. The MaizeGDB team endeavors to meet research needs as they evolve based on researcher feedback and guidance. Recent work has focused on better integrating existing data with sequence information as it becomes available for the B73, Mo17 and Palomero Toluqueño genomes. Major endeavors along these lines include the implementation of a genome browser to graphically represent genome sequences; implementation of POPcorn, a portal ancillary to MaizeGDB that offers access to independent maize projects and will allow BLAST similarity searches of participating projects' data sets from a single point; and a joint MaizeGDB/PlantGDB project to involve the maize community in genome annotation. In addition to summarizing recent achievements and future plans, this article also discusses specific examples of community involvement in setting priorities and design aspects of MaizeGDB, which should be of interest to other database and resource providers seeking to better engage their users. MaizeGDB is accessible online at http://www.maizegdb.org.

Database URL: http://www.maizegdb.org

Introduction

Maize is one of very few species that serve both as an important research model and as a crop from which diverse products and resources are generated [reviewed in (1, 2)]. This breadth of scope is recapitulated by the wide variety of informatics needs expressed by the community of maize biologists—not only are tools for handling genetic and genomic information needed, support for translational and applied research is also of great interest [reviewed in (3)].

To better understand the broad needs of the research community and prioritize development goals, a Working Group (http://www.maizegdb.org/working_group.php) made up of maize geneticists and computational biologists meets annually to discuss the MaizeGDB project's status and to suggest how to further develop the MaizeGDB resource. In addition, the maize community periodically organizes meetings to gather information on key needs to move maize research forward. In March 2007, lab heads met at the Allerton Park and Conference Center in Monticello, IL, to discuss 'The Future of Maize Genetics' [meeting report available at http://www.maizegdb.org/AllertonReport.doc and (4)]. Guidance from the MaizeGDB Working Group and Allerton reports agree that two needs are of the utmost priority: improving access to the genome sequence of inbred line B73 (as well as other maize genome sequences as they become available) and creating tools to improve phenotype data collection, storage and analysis. With this in mind, sequence

data and phenotypes constitute much of the current MaizeGDB Project Plan, a document that outlines work to be accomplished by MaizeGDB over a 5-year period (2009–14). In brief, the goals are as follows:

(1) to integrate new maize genetic and genomic data into the database by

- expanding mutant and phenotype data and tools as well as structural and genetic map sets emphasizing the integration of the IBM genetic maps with the B73 genome sequence;
- creating views that convey the substantial variation in maize genome structure;
- integrating the next generation genetic map being generated by the Maize Diversity Project (5) into a genomic view to enable its effective use by plant breeders;
- providing access to gene models calculated by leading gene structure prediction groups through the MaizeGDB interface;
- compiling and making accessible the annual Maize Newsletter at MaizeGDB

and

(2) to 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.

MaizeGDB currently has a wide range of maize data including genetic maps, gene products, loci, alleles, phenotypes, stocks, sequences and markers. However, centralized access to currently ongoing maize projects that create sequence-indexed data (roughly 10–15 projects at any given time) is notably lacking. Reported here are some recent updates to MaizeGDB with emphasis on improving the handling and accessibility to sequence data, especially data generated by the Maize Genome Sequencing Project for B73 (6). Of particular note are (i) the new MaizeGDB Genome Browser (see 'Genomic sequence data display and integration with genetic maps' section), (ii) a new project ancillary to MaizeGDB called POPcorn, which currently serves as a portal to maize research projects with a centralized maize sequence similarity search resource coming soon and (iii) a recently launched project to involve the community of maize geneticists in genome annotation for B73 (outlined in 'Current endeavors' section).

MaizeGDB's standard operating procedures, machine architecture, accessibility and a description of how the databases are administered are described elsewhere (1, 2). Data made available via MaizeGDB are in the public domain.

Genomic sequence data display and integration with genetic maps

Genome browser

Based upon the 2006 MaizeGDB Working Group guidance (available at the bottom of http://www.maizegdb.org/working_group.php) and the Allerton meeting report (4), the MaizeGDB Team began development toward making MaizeGDB become more sequence-centric in early 2007. To this end, an initiative to implement a MaizeGDB Genome Browser was launched in early 2008 and completed in December 2008. The MaizeGDB Genome Browser enables MaizeGDB to become the long-term and centralized keeper of maize gene models (which ensures proper nomenclature) and serves as a way to compare various groups' assemblies and annotations simultaneously.

A variety of genome browser applications were evaluated via a survey prepared on behalf of the Maize Genetics Executive Committee (accessible online at http://www.maizegdb.org/blanksurvey.html) to gauge cooperators' impressions of existing software and to find out what functionalities they would like to have in a maize genome browser. A summary of the survey results is available online at http://www.maizegdb.org/genome_browser_survey.php.

Based upon results of the survey, GBrowse (7) was selected for the following reasons:

(i) The three most desired features reported were ease of use, visuals and speed. Cross-species comparison capabilities [where Ensembl (8) shines] was ranked fourth, and GBrowse has such capabilities available (SynBrowse (9), GBrowse_syn [http://gmod.org/wiki/GBrowse_syn], CMap [reviewed in (10)], etc.).

(ii) Cooperators would like to see specific tool development to enhance their research (e.g. finding all the genes between two given markers). Flexibility of code and tool development is an intrinsic feature of GBrowse, which is a community-based open source project that allows development of new modules via its customizable plug-in architecture.

(iii) Those surveyed commented on various Model Organism Databases (MODs) and their genome browsers. Sites using GBrowse [e.g. TAIR and FlyBase; (11) and (12)] received far fewer negative comments than sites using other mainstream software platforms.

As the MOD for maize, MaizeGDB strives to consolidate data from any group that makes information available. For this reason, the MaizeGDB BAC-based Genome Browser (http://bac.maizegdb.org) has been populated with many groups' data to create an integrated view of the maize genome within a single resource. Current contributors include the Maize Mapping Project

[MMP (13)], the Maize Genome Sequencing Consortium [MGSC via their MaizeSequence.org resource (6)], PlantGDB (14), the UniformMu group (15), the Department of Energy's Joint Genome Institute (http://www.phytozome.org/maize.php) and the Maize Assembled Gene Islands (MAGIs) resource (16). The MaizeGDB Genome Browser includes direct links to PlantGDB, MAGI and MaizeSequence.org from several 'tracks' allowing direct access to relevant data and tools available from those specialized resources.

The B73 MMP and MGSC outcomes serve as the basis for the Genome Browser's framework. The MGSC used a minimal tiling path of ~19 000 mapped bacterial artificial chromosome (BAC) clones derived from the MMP for BAC based genome sequencing. The focus has been to produce high-quality coverage of all identifiable gene-containing regions of the maize genome where only gene-containing regions are ordered, oriented and anchored to the physical and genetic maps of the maize genome (6).

For the MaizeGDB Genome Browser, genomic coordinates for elements that can be mapped to the B73 genome representation are provided by individual research groups, which creates a need for coordination in data release among the various contributors. To facilitate coordination, MaizeGDB, PlantGDB and MaizeSequence.org are currently working together to align features to the same GenBank (17) releases. Until this coordination is fully accomplished, genomic features displayed within the MaizeGDB Genome Browser may have been aligned by individual groups to different GenBank record releases, causing the elements on different tracks within a particular BAC to be out of order relative to each other. This is due to the fact that subsequences within each BAC can be reordered and reoriented as the records are updated in GenBank. Maize pseudomolecules have been computed (18) and will serve as the backbone that enables all groups to accomplish data coordination. A pseudomolecule based Genome Browser at MaizeGDB is currently under preparation.

In addition to data provided by contributors, two large-scale tracks have been calculated and are included in the MaizeGDB Genome Browser: a 'Sequenced FPC contig' track that clearly specifies regions that are not yet sequenced and a 'BIN' track that relies upon association of genetic markers to BACs for estimating maize bin boundaries (Figure 1). In the 'Sequenced FPC contig' track, sequenced regions are shown as boxes and unsequenced regions are shown as lines. These representations are based upon mapping the BACs to the FPC contigs (13) and demonstrate places where the minimal tiling path does not fully cover the contig. For the 'BIN' track, the

Figure 1. The MaizeGDB Genome Browser showing chromosome 5 from nucleotide position 5 110 700 to 11 637 499. The 'BIN' track shows bins 5.01 and 5.02, and the 'Sequenced FPC contig' track clearly displays regions within the FPC contigs that are not currently sequenced.

Database, Vol. 2009, Article ID bap020, doi:10.1093/database/bap020

90 genetic markers that delineate stretches of ~10–20 cM along each chromosome (19) were mapped to the genome. These bin boundary markers are called the 'core bin markers'. Because only a fraction of these core bin markers currently align with precision greater than a BAC, the displayed bin boundaries are only approximate. In some cases, there is no evidence (e.g. derived from sequence-based or hybridization-based methods) indicating bin boundary position. In such instances the corresponding bins are shown to extend between the flanking core bin markers. For example, although the core bin marker *umc5a* should delineate the boundary between bins 2.06 and 2.07 on chromosome 2, that marker sequence currently aligns only to a region on chromosome 7: the core bin marker for 2.07 is not known. This results in complete overlap of bins 2.06 and 2.07 within the browser.

Although the current view of the B73 genome is fluid and continues to change as more data become available, the addition of the genome browser to MaizeGDB enables researchers to use the emerging B73 sequence data in real time and allows them to visualize genomic elements, identify how these elements are positioned in the genome with respect to each other and relate the B73 maize physical map to the genetic maps by way of shared markers. As our view of the genome improves, the MaizeGDB Genome Browser's representation of the genome will track that progress so that researchers will be able to make use of available data as it emerges.

Genome browser integration with the existing MaizeGDB resource

The MaizeGDB Genome Browser provides integrated views among genomic regions, genetic markers and sequences based upon the B73 sequence assembly. To enhance the capabilities of the Genome Browser and to better integrate it with the existing MaizeGDB resource, images of relevant genomic regions have been integrated into various pages at MaizeGDB [e.g. expressed sequence tag (EST)], genome survey sequence (GSS) and molecular marker pages] where each image is linked into the appropriate regions within the Genome Browser. In addition, the MaizeGDB BLAST (22) tool outputs were upgraded to show genomic locations of hits visually and a Locus Lookup Tool (21) that integrates genetic data with genomic views was created.

The MaizeGDB BLAST tool, which is locally run and updated monthly, is accessible from various locations within MaizeGDB, including the top of the Genome Browser itself (alongside various other MaizeGDB tools). BLAST searches can be performed against BACs, ESTs, GSSs, overgos and other maize nucleotide sequences. Each BLAST hit with known genomic coordinates shows a snapshot of the appropriate genomic context, the e-value of the hit and a short description of the genomic element (Figure 2). Researchers can click on the image to access that region in the MaizeGDB Genome Browser. Detailed information about the BLAST hit along with nucleotide-level alignments of sequences can be reached by scrolling down the results page.

To more precisely map a BLAST hit to the genome, functionality has been added to allow researchers to upload BLAST hits as a separate track to the Genome Browser. By clicking a button on the results page (red arrow in Figure 2), a BLAST track becomes exclusively available to the machine that launched the BLAST search.

Maize is a species with a long genetic history, and over 1700 genetic maps are currently available via MaizeGDB. Because genomic coordinates are not available for all loci and some genetically mapped loci are not cloned, a mechanism for estimating the genomic location of such loci is needed for, e.g. researchers walking to genes. To meet this need, the Locus Lookup Tool (Figure 3) was implemented. The Locus Lookup Tool (available from the MaizeGDB front page, the Genome Browser and throughout MaizeGDB in other relevant locations) works by first checking physical map coordinates to find out whether the locus is already placed. If so, the physically mapped locus is highlighted in red in the appropriate genomic region. If not, the tool checks the locus record at MaizeGDB to find out if any sequenced BACs are known to detect the locus, and, if so, that BAC is returned within its genomic context. If not, genetically mapped probes that are nearest to the input locus are identified, the tool checks whether those probes have known genomic coordinates (working outward until appropriate probes are identified) and finally the region of the genome contained by the identified probes is reported with bounding probes shown in red.

It should be noted well that even though a locus may physically map to a known location within a BAC, because most B73 BAC sequences currently are concatenated sequence fragments of unknown order and orientation, the locus position on each individual BAC sequence may be anywhere on the BAC associated with a physically mapped probe. For this reason, a conservative window showing the entire BAC(s) to which the probe(s) have been physically mapped is displayed.

Expanded structural and genetic map sets integrated with BACs and the B73 genome sequence

As mentioned previously, MaizeGDB is the central archive for maize maps and documentation. MaizeGDB map pages include views of mapped markers and genotyping scores, updated files for sequence-based markers with assigned map locations, GenBank (17) accession numbers and FASTA-formatted sequences as bulk downloads (see, e.g. the UMC 98 'sequence view' for chromosome 1 at http://www.maizegdb.org/cgi-bin/displaymapwithaccessions.cgi?id=143431). To facilitate linking genetic maps to the B73 genome sequence, BAC sequence accessions are imported

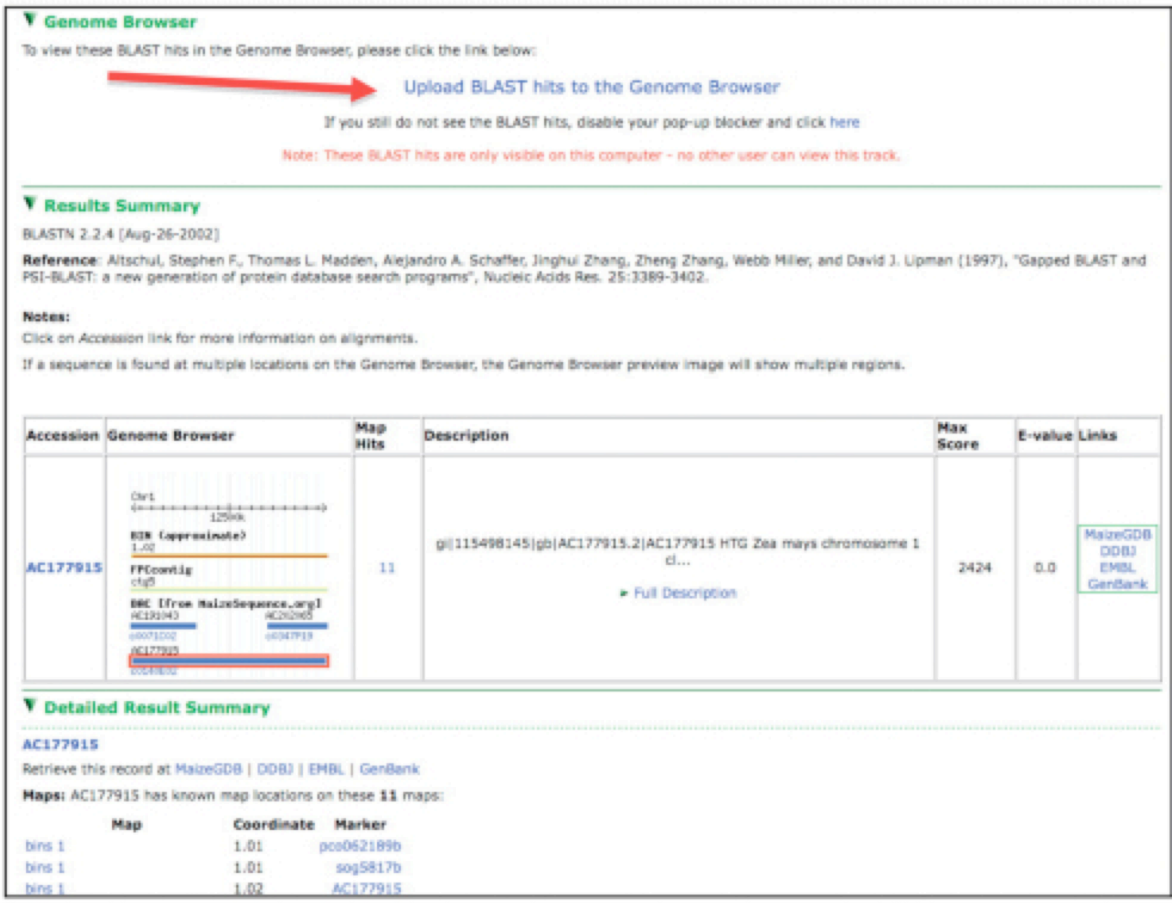

Figure 2. A view from the BLAST results interface. Note the link toward the top of the page allowing BLAST hits to be uploaded as a track in the Genome Browser, results table showing a thumbnail of the hit's genomic context and hit assignment to BACs via the molecular markers associated with the hit sequence.

monthly from GenBank, linked to MaizeGDB's molecular marker data and assigned a bin coordinate. The BAC bin values are based on coordinates provided by the MaizeSequence.org FTP site (http://ftp.maizesequence.org/current/fpc_report.txt) or, in the case of some 900 B73 BACs sequenced by other groups, computed locally to be consistent with MaizeSequence.org coordinates. There are currently 61 000 loci with some genetic map information. These include the 15 568 sequenced BACs and 27 870 (mostly EST-based) markers resolved at the level of a BAC on the IBM2 FPC0507 maps (2). New recombination-based maps, described below, contribute a large number of sequenced markers that can be used for integrating genetic maps with the genome sequence.

Newly available data include the new generation 'NAM' (Nested Associate Mapping) maps from the Maize Diversity-Based Genomics project (22). These are fully documented, with genotype data provided pre-publication. They constitute 27 map sets, based on high-throughput single nucleotide polymorphism (SNP) genotyping of some 1100 markers for nearly 5000 recombinant inbred lines (RILs). The maps are sequence based and closely tied by related overgos to the B73 genome sequence (23). The mapping stocks were designed to support candidate gene discovery for agronomic traits using a nested association mapping strategy (24–26). The RILs are available from the Maize Genetics Cooperation-Stock Center [MGCSC (27)]. Documentation includes the genotype scoring for all the lines, sequence accessions submitted to GenBank, sequences of the allele-specific interrogation primers for B73 and Mo17 and allele descriptors for each relevant SNP with nomenclature based on the style developed for the maize TILLING (28) project (e.g. http://www.maizegdb.org/cgi-bin/displaylocusrecord.cgi?id=978391). Links are provided to the source database as well as MGCSC.

Another addition is the Genetic 2008 map. It includes genes with experimentally confirmed gene products (1400) and/or phenotypes (380) that can be ordered to a resolution of 1–2 cM or better. The map was compiled manually from classical recombination data (29), along

Database, Vol. 2009, Article ID bap020, doi:10.1093/database/bap020

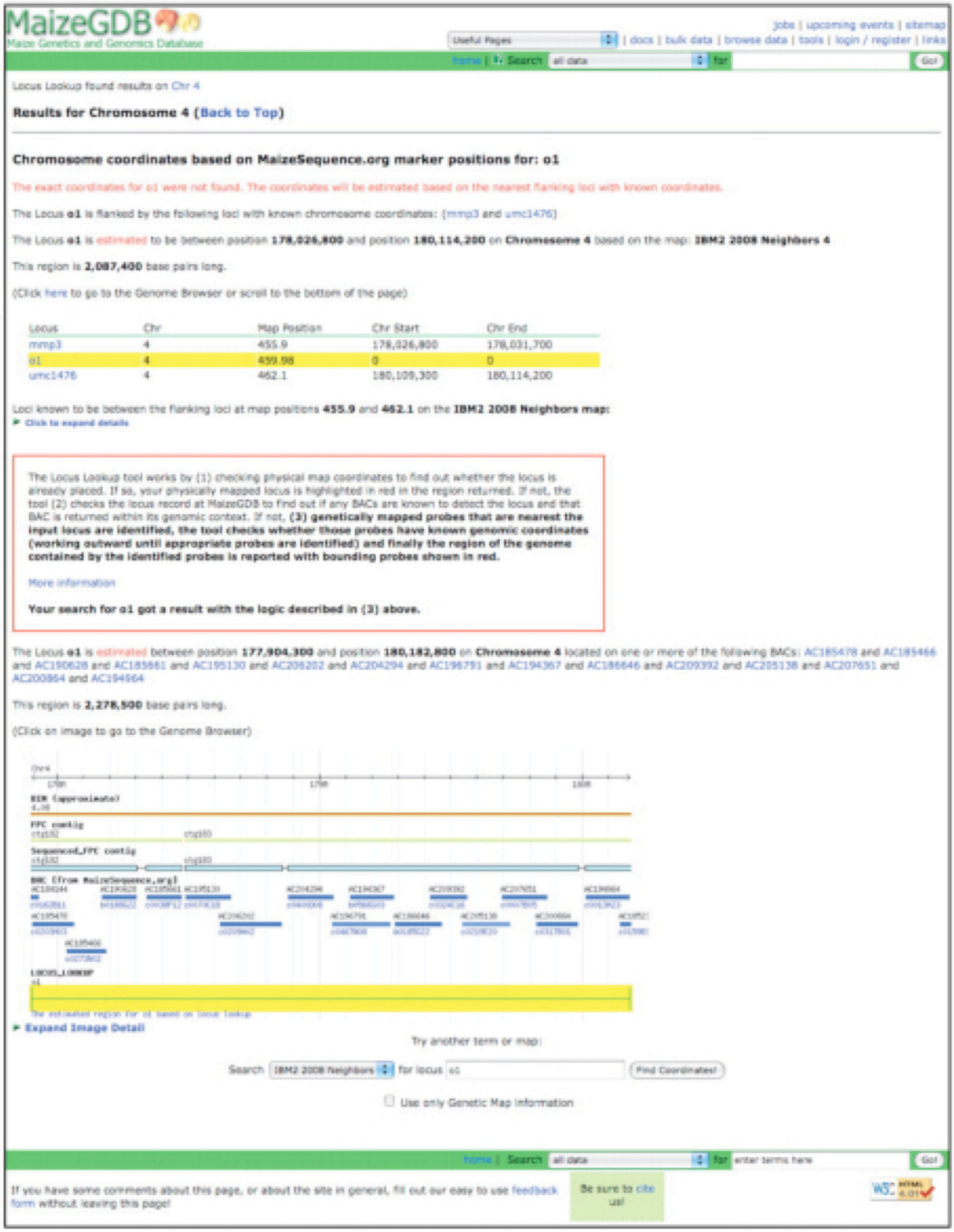

Figure 3. The result page for the approximate genomic coordinates for the *opaque endosperm1* (*o1*) locus when the Locus Lookup tool is used.

with sequence alignment to B73 BAC clones ordered by the maize FPC (13). New genes may be included by any cooperator directly by request.

The 2008 Neighbors Consensus maps incorporate the new genetic maps, including Genetic 2008. The current Neighbors maps include 15 904 loci, representing multiple marker technologies (6016 indels; 1388 SNPs; 1965 simple sequence repeats and 2153 overgos). Some 21% (3414) of these loci have been associated with the B73 genome BAC contigs by mappings based on sequence identity to markers/probes and/or high sequence similarity to markers/probes. This version of the Neighbors map differs from the previous iteration, IBM2 2005 (20), by including only markers ordered by recombination analysis and thus excludes most of the overgo and EST-based markers represented in the hybrid IBM2 FPC0507 map set (2).

Current endeavors

POPcorn—a PrOject Portal for corn

Maize researchers cannot easily leverage all available genetic and genomic data because the online locations of all resources are not easy to find and the sequence-indexed resources generated by individual projects must be searched independently. In addition, it is often the case that when a project's funding period ends, the generated data are lost because they are not moved to long-term repositories: these once-funded project sites degrade over time and sometimes disappear entirely. The MaizeGDB team aims to overcome these challenges in collaboration with the community of maize researchers through POPcorn (PrOject Portal for corn; http://www.maizegdb.org/popcorn), a needs-driven resource and data pipeline. POPcorn offers (i) a centralized web-accessible resource to search and browse ongoing maize genomics projects and will make available (ii) a single, stand-alone tool that makes use of web services and minimal data warehousing to enable researchers to carry out sequence searches at one location that return matches for all participating projects' related resources and (iii) a set of tools that enable collaborators to migrate their data to MaizeGDB at the project's conclusion. A functional version of the POPcorn resource that serves as a portal to research projects has just been released. Sequence search capabilities and data upload tools are planned for release in early 2010.

Community-driven maize genome annotation

The BAC-by-BAC maize genome sequencing effort is nearly complete with >17 000 Phase 1 BACs deposited in GenBank. The BACs are physically mapped (13). Although most BACs consists of several unordered sequence segments (or scaffolds) separated by unsequenced gaps, it is clear that sufficient sequence and map information is now available to derive gene models and assign them to classical genetic loci.

Maize stands out amongst the other plant genomes (e.g. *Populus*, Sorghum and *Brachypodium*) at similar annotation stages because of both the availability of rice and Arabidopsis as excellent model organisms to help the annotation effort and a large, well-organized research community that is ready to be engaged in the effort. Moreover, current annotation projects (including MaizeSequence.org, MAGI and PlantGDB) can be easily used as a springboard to complete community-driven annotation. For example, PlantGDB provides a daily updated table of new maize BACs from GenBank annotated with matching rice and Arabidopsis proteins (http://www.plantgdb.org/ZmGDB/DisplayGeneAnn.php). The table is sortable by location, rice gene product annotation or date and linked to genome context views with associated yrGATE community annotation tools. Thus, newly sequenced homologs of known rice or Arabidopsis genes are immediately accessible to the community for perusal and potential expert annotation.

In brief, the annotation process consists of: (i) a computational stage in which gene structures are predicted together with alignment evidence and (ii) subsequent rounds of manual annotation in which these models are improved. One set of computationally derived models for maize will be available from MaizeSequence.org (6). In addition, the EVidenceModeler (EVM) (31) meta-annotation system will be used to derive consensus gene structures. This program currently reconciles gene structures derived from a variety of computational gene finding tools and/or manual annotations according to a weight-based voting scheme, where weights are either set manually or learned using statistical training routines. Such systems have been shown to exhibit gene-finding competency commensurate with that of expert human annotators (32). As deployed for the maize community genome annotation project, EVM reconciles: gene structure predictions generated by the *ab initio* gene finders AUGUSTUS (33) and GeneMark.HMM (34), genes functionally supported by cDNA alignments as predicted by the GeneSeqer+MetWAMer system (35), as well as similarity data alignments generated by GenomeThreader [protein alignments (36)] and GeneSeqer [cDNA alignments (37)]. Elements in this panel of gene finders were selected largely based on their abilities to be retrained in a species-specific manner, enabling the generation of maize-specific parameter sets. It is anticipated that a more diverse collection of gene prediction software will be integrated into the pipeline over time. Once the maize genome has been computationally annotated using both *de novo* and comparative methods, instances of genes where cDNA and/or mRNA evidence are not congruent with the computationally derived annotation will be identified.

To facilitate community-driven maize genome annotation, the MaizeGDB and PlantGDB groups are working together. A brief overview of the annotation process is as follows. The gene models made available on MaizeSequence.org (6) and those derived from the EVM pipeline (described above) are evaluated for congruency with the most recent available data (EST and cDNA evidence that align to the gene). Next, a list of gene models that should be manually annotated is created using the xGDB GAEVAL (Gene Annotation EVALuation) system at PlantGDB (14) and made available to the community. Annotation project personnel contact researchers who may be working on particular genes in the list; the researcher annotates the gene directly using yrGATE, the xGDB genome annotation tool (38); and that annotation becomes available at both PlantGDB [via the maize xGDB browser called ZmGDB (39)] and MaizeGDB, pending approval by project personnel. Annotation project personnel will (i) conduct training in the use of yrGATE; (ii) alert members of the maize community to the need for curation of incongruent gene models that are of interest to them; (iii) encourage researchers to do the annotations (though this may seem minor, it is anticipated that this will require the majority of time and effort); and (iv) report the success of the project at conferences. The yrGATE training sessions will be conducted at the Annual Maize Genetics Conference and via three or more site visits per year to locations where many maize researchers are located.

It is anticipated that researchers may wish to annotate genes based upon: (i) genomic location; (ii) gene family membership; and (iii) involvement in particular biochemical pathways. To aid in identifying researchers and students who may wish to be involved in the genome annotation project, a query will be sent out to all maize cooperators asking for the contact information for at least one person per research group who will serve to annotate on behalf of the lab. Along with contact information, respondents will be asked for a list of locations/gene families/pathways of interest to the research group. Again, this effort in maize will benefit from the many resources now available for other model organisms and pan-species studies, e.g. http://www.kegg.com/, http://www.biocyc.org/, http://www.ebi.ac.uk/Tools/InterProScan/.

As outlined above, the MaizeGDB Team has chosen GBrowse to serve as the basis for the MaizeGDB Genome Browser. This is very helpful for working with PlantGDB's xGDB and yrGATE systems given that GBrowse supports Distributed Annotation System (DAS; http://www.biodas.org) for data transfer and ZmGDB has DAS capabilities. PlantGDB's ZmGDB resource will serve annotations as a 'Community Annotation Track' to the MaizeGDB Genome Browser via DAS, so researchers' annotations will be available in real time. In addition, Ensembl (the genome browser software underlying MaizeSequence.org and Gramene) also supports DAS for sharing annotation data, which will make the generated gene models readily available for various plant database sites to pick up freely.

As with all other data generated by PlantGDB and MaizeGDB, all annotation efforts will be immediately and comprehensively available to the community, via web browsers, via DAS and by download. For example, two of the authors (M.E.S. and V.P.B.) recently identified 1665 non-redundant full-length maize genes on 1463 unique BACs that are highly conserved with Sorghum proteins (36), and which are served via DAS at http://sunx4600uno.gdcb.iastate.edu:9002/das/Zm_to_Sb with ProServer software (40). These data are proffered to the maize community as a high quality gene set for use in training and assessing gene finding software tools, which can be accessed online at http://www.plantgdb.org/ZmGDB/DisplayZmToSb.php.

Outreach

Outreach continues to be an area of ongoing action at MaizeGDB. Since March 2007, tutorials have been taught at the University of Florida (Gainesville), the University of California (Berkeley), Stanford University (California), the USDA-ARS Plant Gene Expression Center (Albany, CA) and Iowa State University (Ames). In addition, MaizeGDB Team members are available for people to call or email with questions. Responding to feedback is a high priority, and the MaizeGDB team strives to be responsive to the suggestions and comments of maize research community members. In addition to live tutorials, online movie tutorials are also available with more currently in the making. These short movies demonstrate a specific topic of interest and are available online alongside other outreach materials at http://www.maizegdb.org/tutorial.

Acknowledgements

We thank the MGSC and MaizeSequence.org groups (especially Shiran Pasternak and Doreen Ware) for sharing the maize genome sequence data and their analyzed data sets with us and with the community prior to publication. We also thank the MaizeGDB Working Group (E. Buckler, K. Cone, M. Freeling, O. Hoekenga, A. Lamblin, K. McGinnis, L. Mueller, P. Schnable, M. Pop, T. Slezak, A. Sylvester, D. Ware, M. Sachs and V. Brendel) as well as the broad maize research community for their help and guidance. We also thank G. Davis, M. McMullen and E. Coe for advice on mappings.

Funding

U.S. Department of Agriculture-Agricultural Research Service; National Science Foundation (grant numbers DBI 0743804 and 0606909). Funding for open access

charge: U.S. Department of Agriculture-Agricultural Research Service.

Conflict of interest. None declared.

References

1. Lawrence,C.J., Dong,Q., Polacco,M.L. *et al.* (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.*, **32**, D393–D397.
2. Lawrence,C.J., Schaeffer,M.L., Seigfried,T.E. *et al.* (2007) MaizeGDB's new data types, resources and activities. *Nucleic Acids Res.*, **35**, D895–D900.
3. Lawrence,C.J., Harper,L.C., Schaeffer,M.L. *et al.* (2008) MaizeGDB: the maize model organism database for basic, translational, and applied research. *Int. J. Plant Genomics*, **2008**, 496957.
4. Maize Genetics Executive Committee. (2008) Allerton Report. *Maize Genetics Cooperation Newsletter*, **82**, 111–118.
5. Buckler,E.S., Gaut,B.S. and McMullen,M.D. (2006) Molecular and functional diversity of maize. *Curr. Opin. Plant Biol.*, **9**, 172–176.
6. Schnable,P.S., Ware,D., Fulton,R.S., *et al.* (2009) The B73 maize genome: complexity, diversity and dynamics. *Science*, **326**, 1112–1115.
7. Stein,L.D., Mungall,C., Shu,S. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
8. Hubbard,T.J., Aken,B.L., Ayling,S. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
9. Pan,X., Stein,L. and Brendel,V. (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, **21**, 3461–3468.
10. Faga,B. (2007) Installing and configuring CMap. In: *Curr. Protoc. Bioinformatics*, Chapter 9, Unit 9.8.
11. Swarbreck,D., Wilks,C., Lamesch,P. *et al.* (2008) The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
12. Tweedie,S., Ashburner,M., Falls,K. *et al.* (2009) FlyBase: enhancing Drosophila gene ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
13. Wei,F., Coe,E., Nelson,W. *et al.* Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genetics*, **3**, e123PMID: 17658954.
14. Duvick,J., Fu,A., Muppirala,U. *et al.* (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.*, **36**, D959–D965.
15. McCarty,D.g., Settles,A.M., Suzuki,M. *et al.* (2005) Steady-state transposon mutagenesis in inbred maize. *The Plant J.*, **44**, 52–61.
16. Fu,Y., Emrich,S.J., Guo,L. *et al.* (2005) Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. *Proc. Natl Acad. Sci. USA*, **102**, 12282–12287.
17. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J. *et al.* (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
18. Wei,F., Zhang,J., Zhou,S. *et al.* (2009) The Physical and Genetic Framework of the Maize B73 Genome. *PLoS Genetics*, **5**, e1000715.
19. Gardiner,J.M., Coe,E.H., Melia-Hancock,S. *et al.* (1993) Development of a core RFLP map in maize using an immortalized F_2 population. *Genetics*, **134**, 917–930.
20. Altschul,S.F., Madden,T.L., Schäffer,A.A *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–33402.
21. Andorf,C.M., Lawrence,C.J., Harper,L.C. *et al.* (2009) The Locus Lookup Tool at MaizeGDB: Identification of Genomic Regions in Maize by Integrating Sequence Information with Physical and Genetic Maps. *Genetics*, in press.
22. McMullen,M.D., Kresovich,S., Villeda,H.S. *et al.* (2009) Genetic properties of the maize nested association mapping population. *Science*, **325**, 737–740.
23. Gardiner,J., Schroeder,S., Polacco,M.L. *et al.* (2004) Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. *Plant Physiol.*, **134**, 1317–1326.
24. Stich,B., Möhring,J., Piepho,H.P. *et al.* (2008) Comparison of mixed-model approaches for association mapping. *Genetics*, **178**, 1745–1754.
25. Liu,K., Goodman,M.M., Muse,S. *et al.* (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics*, **165**, 2117–2128.
26. Canaran,P., Buckler,E.S., Glaubitz,J.C. *et al.* (2008) Panzea: an update on new content and features. *Nucleic Acids Res.*, **36**, D1041–D1043.
27. Scholl,R., Sachs,M. and Ware,D. (2003) Maintaining collections of mutants for plant functional genomics. In: Grotewold,E. (ed). *Plant Functional Genomics*, Totowa, NJ, 236, Vol. 236, pp. 311–326.
28. Till,B., Reynolds,S., Weil,C. *et al.* (2004) Discovery of induced point mutations in maize genes by TILLING. *BMC Plant Biol.*, **4**, 12.
29. Coe,E. (2008) Genetic maps 2007. *Maize Genet. Coop. News Lett.*, **82**, 87–102.
30. Schaeffer,M., Gerau,M., Sanchez-Villeda,H. (2007) Population Explosion In The IBM Neighborhood - IPC And New Genetic Maps. In Plant & Animal Genomes XV Conference 2007, W250. http://www.intl-pag.org/pag/15/abstracts/PAG15_W38_250.html (last accessed date 9 November 2009).
31. Haas,B., Salzberg,S., Zhu,W. *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*, **9**, R7.
32. Allen,J. and Salzberg,S. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.
33. Stanke,M., Diekhans,M., Baertsch,R. *et al.* (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.
34. Lomsadze,A., Ter-Hovhannisyan,V., Chernoff,Y. *et al.* (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.*, **33**, 6494–6506.
35. Sparks,M.E. and Brendel,V. (2008) MetWAMer: eukaryotic translation initiation site prediction. *BMC Bioinformatics*, **9**, 381.
36. Sparks,M.E. and Brendel,V. (2005) Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants. *Bioinformatics*, **21**, ii20–ii30.
37. Gremme,G., Brendel,V., Sparks,M.E. *et al.* (2005) Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Tech.*, **47**, 965–978.
38. Wilkerson,M.D., Schlueter,S.D. and Brendel,V. (2006) yrGATE: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes. *Genome Biol.*, **7**, R58.
39. Schlueter,S.D., Wilkerson,M.D., Dong,Q. *et al.* (2006) xGDB: open-source computational infrastructure for the integrated evaluation and analysis of genome features. *Genome Biol.*, **7**, R111.
40. Finn,R.D., Stalker,J.W., Jackson,D.K. *et al.* (2007) ProServer: a simple, extensible Perl DAS server. *Bioinformatics*, **23**, 1568–1570.

Copyedited by: SM

MANUSCRIPT CATEGORY: APPLICATIONS NOTE

BIOINFORMATICS APPLICATIONS NOTE

2009, pages 1–3
doi:10.1093/bioinformatics/btp556

Databases and ontologies

The locus lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps

Carson M. Andorf[1], Carolyn J. Lawrence[1,2], Lisa C. Harper[3,4], Mary L. Schaeffer[5,6], Darwin A. Campbell[1] and Taner Z. Sen[1,2,*]

[1]US Department of Agriculture – Agricultural Research Service, Corn Insects and Crop Genetics Research Unit, [2]Department of Genetics, Development and Cell Biology; Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA 50011, [3]US Department of Agriculture – Agricultural Research Service, Plant Gene Expression Center, 800 Buchanan Street, Albany, CA 94710, [4]Department of Molecular and Biology, University of California Berkeley, Berkeley, CA 94720, [5]US Department of Agriculture – Agricultural Research Service, Plant Genetics Research Unit and [6]Division of Plant Sciences, University of Missouri Columbia, Columbia, MO 65211, USA

Received and revised on August 7, 2009; accepted on September 11, 2009

Associate Editor: Alex Bateman

ABSTRACT

Summary: Methods to automatically integrate sequence information with physical and genetic maps are scarce. The Locus Lookup Tool enables researchers to define windows of genomic sequence likely to contain loci of interest where only genetic or physical mapping associations are reported. Using the Locus Lookup tool, researchers will be able to locate specific genes more efficiently that will ultimately help them develop a better maize plant. With the availability of the well-documented source code, the tool can be easily adapted to other biological systems.

Availability: The Locus Lookup tool is available on the web at http://maizegdb.org/cgi-bin/locus_lookup.cgi. It is implemented in PHP, Oracle and Apache, with all major browsers supported. Source code is freely available for download at http://ftp.maizegdb.org/open_source/locus_lookup/.

Contact: taner.sen@ars.usda.gov

Maize (*Zea mays* ssp. *mays*) has been an important model organism for nearly a century due to its large chromosomes, ease of making genetic crosses, and economic importance. Maize researchers have created many genetic recombination maps, and more than 1700 are available online via MaizeGDB, the maize model organism database (Lawrence *et al*., 2006, 2007; Sen *et al*., 2009). Other maize maps include physical, cytological [reviewed in Lawrence *et al*. (2006), and optical maps (Aston *et al*., 1999; Zhou *et al*., 2009)]. In addition to these maps, maize researchers now have access to genomic sequence of the B73 inbred line's genic regions (Schnable *et al*., 2009) and pseudomolecules representing the sequenced regions of the genome (Wei *et al*., 2009).

The availability of the B73 reference genome sequence opens up new research possibilities. Investigators can more easily narrow the genomic regions responsible for specific phenotypes with higher accuracy in a shorter time frame: 'walking to genes' has become 'running to genes'. Researchers can dissect gene structure and function faster, and more clearly define the relationship between genotype and phenotype.

Although sequence data are becoming increasingly available [especially with the advent of next-generation sequencing technologies (Simon *et al*., 2009)], the bioinformatic tools needed to integrate sequence with existing map information are insufficient, not only for maize but for many other research model species. Although these datatypes reside side-by-side in databases, integrating the information in a meaningful way is not always straightforward. In the absence of expensive manual curation by database personnel, researchers often resort to copying and pasting rows of data for loci of interest into spreadsheets and analyzing and curating the data by hand.

To overcome some aspects of these challenges, the Locus Lookup Tool was developed and deployed within the context of the GBrowse-based MaizeGDB Genome Browser (Sen *et al*., 2009; Stein *et al*., 2002). In overview, the Locus Lookup Tool takes the name(s) of (i) a single locus or (ii) two loci that define a region and returns a snapshot representing the likely genomic region containing the locus of interest.

Here is how the Locus Lookup tool works. When a single search term is entered, the Locus Lookup Tool (i) checks whether the locus/loci is/are already mapped to the B73 sequence, and if so, displays the genomic/nucleotide coordinates on the MaizeGDB Genome Browser. If not, then (ii) it checks whether a probe/molecular marker (e.g. a BAC) that recognizes/contains that locus has known genomic coordinates. If no such probes/molecular markers are returned, then the Locus Lookup Tool (iii) checks a user-specified genetic map for the nearest left and right features/markers neighboring the locus that are also placed on the B73 sequence in genomic coordinates. If such left and right features can be found, the Locus Lookup will return the coordinates between which the gene of interest is likely to reside. Note also that because the B73 sequence is BAC-based and BAC sequences consist of scaffolds as well as

*To whom correspondence should be addressed.

[15:28 1/10/2009 Bioinformatics-btp556.tex]

Page: 1 1–3

Copyedited by: SM MANUSCRIPT CATEGORY: APPLICATIONS NOTE

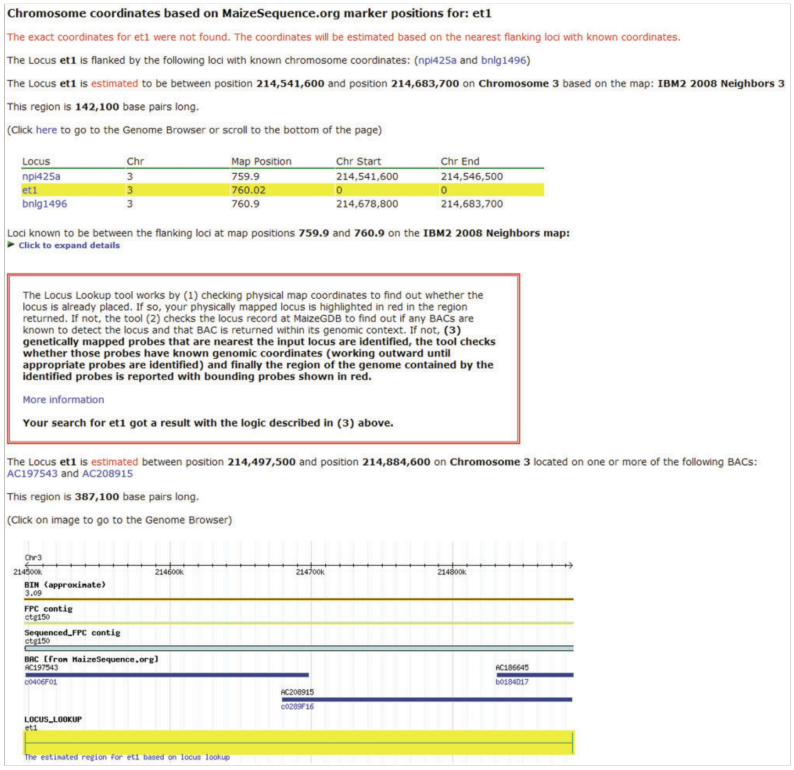

Fig. 1. The Locus Lookup view of the *etched 1* (et1) locus at MaizeGDB. At the top of the page, the estimated coordinates and the markers of known coordinates that are used for estimation are shown. In the middle of the page, a box with a bolded section specifies which method is used for estimation. At the bottom of the page, a view of the genomic region is shown. When the users click this view, they are lead to the MaizeGDB Genome Browser (with a custom Locus Lookup track), which enables them to harness the full capability of the MaizeGDB Genome Browser to address their research problems.

unordered and unoriented pieces, the returned snapshot contains the entire span of the left-most and right-most BACs in the region in case the locus of interest falls outside the window due to possible fragmentation of those BACs.

Because the precision of Locus Lookup Tool lies in the accuracy of both genetic coordinates on the selected genetic map and genomic coordinates on the B73 sequence, the imperfections in both are inherited. For the maize genetic maps, in particular consensus maps such as Neighbors and the Genetic 2008 that are composites of many different maps and experiments, genetic order and distance are only approximated. In addition, among different maize cultivars or lines, the order of certain genes along a chromosome varies in some cases (Fu and Dooner, 2002). Some of the current limitations (as of July, 2009) of the maize B73 sequence are: (i) it is not yet known how much sequence exists between physically mapped BAC contigs; (ii) the sequence of each BAC consists of scaffolds, as well as unordered

[15:28 1/10/2009 Bioinformatics-btp556.tex] Page: 2 1–3

Copyedited by: SM MANUSCRIPT CATEGORY: APPLICATIONS NOTE

and unoriented pieces separated by strings of 'N's; (iii) the order and/or orientation of BACs within a contig on the physical map may be wrong; and (iv) the order or orientation of sequence contigs may be wrong. In addition, as of this writing, the current version of the genome assembly (MaizeSequence.org's 3b.50 release) consists of 16 581 sequenced BACs, 1024 of which (i.e. 6%) are not assigned to a linkage group (chromosome). Sequences not assigned to a chromosome will be missed by the Locus Lookup Tool.

The Locus Lookup Tool at MaizeGDB can be reached a few different ways: (i) from the MaizeGDB homepage (http://www.maizegdb.org) in the left green margin, (ii) via selecting 'genome browser' in the dropdown to the left of the search box displayed at the top and bottom of every MaizeGDB page, and (iii) from the search box on the MaizeGDB Genome Browser (http://gbrowse.maizegdb.org). To do a Locus Lookup search, researchers enter the search term (locus name) and click on the search button. In the results page either the genomic coordinates are shown as a snapshot of the MaizeGDB Genome Browser or, if the position is based on flanking features, a genomic region will be specified and presented [see the example of *etched1* (et1) locus in Fig. 1]. Clicking the snapshot adds a custom track to the Genome Browser allowing the region to be viewed in its genomic context. In addition, if a genomic region bounded by two loci is desired, researchers can type 'locus1..locus2' into the search box to define the region bounded by the two loci using the same logic as above.

The logic and implementation for the Maize Locus Lookup Tool are not specific to maize and generally applicable to any research model organism and other model organism database groups. For example, SoyBase (Grant *et al*., 2009) and SGN (Mueller *et al*., 2005) are already planning to implement the Locus Lookup Tool's logic for the soybean and Solanaceae research communities, respectively.

ACKNOWLEDGEMENTS

The authors thank the maize community for their enthusiastic support and feedback, the MaizeGDB Working Group for their guidance, and the Maize Genome Sequencing Consortium for releasing the B73 genome as it is being sequenced.

Funding: US Department of Agriculture, Agricultural Research Service.

Conflict of Interest: none declared.

REFERENCES

Aston,C. *et al.* (1999) Optical mapping and its potential for large-scale sequencing projects. *Trends Biotechnol.*, **17**, 297–302.

Fu,H. and Dooner,H.K. (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl Acad. Sci. USA*, **99**, 9573–9578.

Grant,D. *et al.* (2009) SoyBase, The USDA-ARS Soybean Genome Database. http://soybase.org.

Lawrence,C.J. *et al.* (2006) Predicting chromosomal locations of genetically mapped loci in maize using the Morgan2McClintock Translator. *Genetics*, **172**, 2007–2009.

Lawrence,C.J. *et al.* (2007) MaizeGDB's new data types, resources and activities. *Nucleic Acids Res.*, **35**, D895–D900.

Mueller,L.A. *et al.* (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol.*, **138**, 1310–1317.

Schnable,P.S. *et al.* (2009) The B73 maize genome: complexity, diversity and dynamics. *Science*, submitted for publication.

Sen,T.Z. *et al.* (2009) MaizeGDB becomes 'sequence-centric'. *Database*, in press.

Simon,S.A. *et al.* (2009) Short-read sequencing technologies for transcriptional analyses. *Annu. Rev. Plant Biol.*, **60**, 305–333.

Stein,L.D. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Wei,F. *et al.* (2009) The physical and genetic framework of the B73 maize genome. submitted for publication.

Zhou,S. *et al.* (2009) B73 Optical map: a single molecule map of the maize genome, Maize Genetics Coop. *Newsletter*, **83**, 106–107.

[15:28 1/10/2009 Bioinformatics-btp556.tex] Page: 3 1–3

Accepted to DATABASE with minor revisions requested

Choosing a Genome Browser for a Model Organism Database (MOD): Surveying the Maize Community

Taner Z. Sen*[1,2], Lisa C. Harper[3,4], Mary L. Schaeffer[5,6], Trent E. Seigfried[1], Darwin A. Campbell[1], Carson M. Andorf[1], and Carolyn J. Lawrence[1,2]

1 USDA-ARS Corn Insects and Crop Genetics Research Unit, Ames, Iowa 50011, USA

2 Department of Genetics, Development and Cell Biology; Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa 50011, USA

3 USDA-ARS Plant Gene Expression Center, 800 Buchanan Street, Albany, CA 94710, USA

4 Department of Molecular and Biology, University of California Berkeley, Berkeley, CA 94720, USA

5 USDA-ARS Plant Genetics Research Unit, Columbia, MO 65211, USA

6 Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA

*corresponding author: Taner Z. Sen, taner.sen@ars.usda.gov, 1025 Crop Genome Informatics Laboratory, Iowa State University, Ames, IA 50011.

ABSTRACT

As the B73 maize genome sequencing project neared completion, MaizeGDB began an endeavor to integrate a graphical genome browser with its existing Web interface and database. To ensure that maize researchers would optimally benefit from the potential addition of a genome browser to the existing MaizeGDB resource, personnel at MaizeGDB surveyed researchers' needs. Collected data indicate that existing genome browsers for maize were inadequate and suggested that the implementation of a browser with quick interface and intuitive tools would meet most researchers' needs. Here we document the survey's outcomes, review functionalities of available genome browser software platforms with respect to the data collected, and offer our rationale for having chosen the GBrowse software suite for MaizeGDB. Because the genome as represented within the MaizeGDB Genome Browser is tied to detailed phenotypic data, molecular marker information, available stocks, etc., the MaizeGDB Genome Browser represents a novel mechanism by which the researchers can leverage maize sequence information toward crop improvement directly. The MaizeGDB Genome Browser can be reached at http://gbrowse.maizegdb.org/.

INTRODUCTION

A genome browser is to genomic sequence data as a Web browser is to the World Wide Web: both offer logical access to datastreams that are otherwise unintelligible. With the advent of new DNA sequencing technologies and the availability of copious amounts of sequence-based data from many species, genome browsers have been developed as a means for researchers to view, interact with, search through, and display sequenced genomes as well as to compare syntenic or similar regions of genomes among related species. Various genome browsers have been created over the years, each with particular strengths and weaknesses. Many provide independent solutions for integrating and visualizing sequence-based data alongside genetic and phenotypic information.

Community resources including Model Organism Databases (e.g., TAIR(1), FlyBase(2), etc.), Clade-Oriented Databases (e.g., Gramene(3), SGN(4), etc.), Automatic Annotation Shops (e.g., PlantGDB(5), TIGR(6,7), etc.), and others have a responsibility to provide timely access to sequence data well-integrated with existing traditional biological data. Determining how best to choose genome browser software to meet the needs of its users within the context of the group's maintenance capabilities is a major challenge for the groups working to build and maintain these community resources. Described here are the methodologies we used to determine which genome browser to implement at MaizeGDB(8-10), the MOD for maize.

The need for a Genome Browser at MaizeGDB

These are exciting times for maize. Not only is it a major production crop worldwide; a reference genome sequence for one inbred line, B73, has been released (www.maizesequence.org; (11)). As of August 2009, the minimum tiling path included 16,910 sequenced BAC and fosmid clones

and encompassed 2.12 Gb or 93% of the 2.3 Gb-long B73 genome(12). A pre-publication release of the B73 pseudomolecules is available through the Arizona Genomics Institute website. (http://www2.genome.arizona.edu/genomes/maize). Other whole genome sequences soon to become available include an ancient popcorn landrace, Palermo Toluqueño (Vielle-Calzada, personal communication) and outputs from a project to shotgun sequence the maize inbred line Mo17 (JGI-Joint Genome Institute). In addition, an extensive haplotype map is in progress for 27 lines of maize, enabling researchers to establish novel relations between genetic, physical, and diversity data (13,14). Other sequence-based resources include over 2 million public ESTs (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html) and a large number of genic sequences from gene-enriched libraries (15,16). Various research groups and consortia integrate large portions of these data sets, each in their own way. Examples include PlantGDB ((5); www.plantgdb.org), the Dana Farber (http://compbio.dfci.harvard.edu/tgi/tgipage.html; Lee et al 2005), MAGI (http://magi.plantgenomics.iastate.edu/;(17)), NCBI RefSeq(18), and Uniprot (www.unitprot.org; The UniProt Consortium 2009). Integration of the large data sets, at a single location, with the information about the position, orientation and sequence of genes, genetic markers, variations, and their association with phenotypic data would allow for a detailed understanding of the maize genome within its biological context, when presented as centrally accessible and simultaneously viewable.

At the completion of the Maize Sequencing Project, it is anticipated that genomic data and gene models will be transferred from the Maize Genome Sequencing Consortium's project database MaizeSequence.org to MaizeGDB (8-10) and Gramene (3) . As a federally funded, long-lived resource, MaizeGDB is tasked to serve the maize geneticists' and breeders' longitudinal data access and analysis needs. To accomplish these tasks, MaizeGDB primarily

relies on direct participation by members of the maize community including the Maize Genetics Executive Committee (MGEC; a group tasked to identify both the needs and the opportunities for maize genetics and to communicate this information to the broadest possible life science community), the MaizeGDB Working Group (a panel that offers guidance for MaizeGDB's continued development), and direct interaction with individual researchers from the maize community.

Currently, MaizeGDB stores information on: loci (genes and other genetically-defined genomic regions including QTLs), 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. Like many other MODs (such as TAIR (1), Oryzabase (19), and Soybase (20)), MaizeGDB incorporates and integrates newly generated genomic data into its existing database and develops tools to help visualize genome structure, gene models, functional data, and genetic variability. Toward this end, two groups directed MaizeGDB to evolve a more sequence-centric paradigm: the MaizeGDB Working Group (via their 2006 guidance document; see http://www.maizegdb.org/working_group.php) and maize principal investigators (in the 2007 Allerton Report that documents outcomes of a special two-day gathering of maize community with a focus on "The Future of Maize Genetics Planning for the Sequenced Genome Era"; see http://www.maizegdb.org/AllertonReport.doc). The time was right to carefully consider implementing a genome browser as a way to integrate genomic sequence features with the existing genetic and physical information at MaizeGDB.

At the time we began considering the implementation of a genome browser at MaizeGDB, various other resources already represented maize genomic sequence graphically via genome

browsers. Most notably, MaizeSequence.org, Gramene (which took down their genome browser when MaizeSequence.org was released), PlantGDB with its maize ZmGDB browser, and the Maize Assembled Gene Islands (MAGI) resource. A specific challenge for MaizeGDB was whether to follow the lead of the Maize Genome Sequencing Consortium and collaborate with that group to further develop MaizeSequence.org. This would be a very good use of funds in the short term given that both groups could collaborate to maintain a single maize genome browser. Another issue for consideration was the MaizeGDB team's charge to make decisions based upon input from the maize community as we are fortunate at MaizeGDB to be serving a cooperative community that communicates well. Therefore, we followed the following hierarchical strategy to gather the information needed to determine how to proceed with potentially implementing MaizeGDB Genome Browser:

1. Should MaizeGDB make a genome browser available at all? If researchers were happy with the existing options, implementing another resource would be a waste of time and resources.
2. If researchers wanted MaizeGDB to implement a genome browser, we needed to know:
    a. what they liked and did not like about available maize genome browsers and
    b. examples of workflows they would like to be able to carry out so that we could evaluate which software could best meet our stakeholders' needs.

With these ideas in mind, we approached the MGEC and MaizeGDB Working Group. These groups offered to survey the maize community on our behalf and worked with us to prepare a survey that aims to answer questions 1, 2a, and 2b.

Materials and Methods

Preparation of the survey

The MaizeGDB team prepared an initial draft survey, and sent it to the MaizeGDB Working Group and the MGEC for suggestions. The updated set of questions was considered by Dr. Patrick Armstrong in the Department of Psychology at Iowa State University who made recommendations on, e.g., how to eliminate potential bias by improving how the questions were worded or ordered. The final form of the survey can be found at http://www.maizegdb.org/browser_survey/ and in the *Supplementary Materials* section of this document.

In November of 2007 MaizeGDB personnel distributed via email a request by the MGEC for all "maize cooperators" (totaling 1,241 at that time) to take a survey regarding their use of online maize data resources with emphasis on browsing the available genome sequence. ("Maize cooperators" are a list of maize researchers maintained at MaizeGDB made up of attendees of maize meetings, researchers publishing frequently on maize, or people who specifically request to be considered a maize cooperator.) Each cooperator received a randomly generated unique key to ensure that each email recipient was only able to submit answers to the survey once.

The Number of Respondents. Among the 1,241 cooperators surveyed, 99 responded. This number is comparable to the number of participants to the last MGEC membership election where 234 of the 1,190 contacted cast a ballot. Because the Genome Browser Survey requested detailed answers to the researchers' needs and not every maize researcher would feel knowledgeable on genome browsers, this level of response to the survey exceeded our expectation.

RESULTS

The raw survey results can be found in the *Supplementary Material* section, as well as at http://www.maizegdb.org/browser_survey/analyze.php. Tabulated results are located at http://www.maizegdb.org/browser_survey/analyze-tab-delimited.php.

Time Spent Accessing Maize Data Online. 37% percent of the survey takers reported that they spend an hour or two each week online to access maize data. 39% percent spend between two and five hours. 15% spend more than five hours online to access maize data. Only 8% of the survey takers did not use online maize data resources.

Genome Browsers Used. 68% of the respondents reported that they use MaizeSequence.org and 66% use Gramene. A total of 76% use either MaizeSequence or Gramene. Although both sites use Ensembl (one genome browser software option; described in the Discussion section;(21)) as their genome browser, among the users of these websites, only a total of 35% of the all respondents acknowledged using Ensembl. This result shows that the users may not be aware of the underlying browser software that the various websites use.

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, among these users, only 31% acknowledged that they are using GBrowse (another genome browser software option; described in the Discussion section; (22)).

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 a genome browser that allows them to find data quickly and easily.

Desired Features. We thought that the "desired features" section of the survey might be very helpful to guide genome browser developers in the creation of new features. Survey respondents expressed interest to reach specific data using the most intuitive tools that require short learning time. They also reported a need for enhanced cross-referencing between different websites and called for downloadable data sets in various formats. In short, they want the most current data along with tools that are easy to learn and apply for their specific research needs. As expected, respondents want the minimized hassle and effort in reaching the needed maize data.

"Bad" Genome Browser Examples. We asked the respondents about what they do not like about the current genome browser examples in the hopes that this would give us indication as to which browsers or options to avoid. Among 29 comments left in "Bad genome browser examples", 19 of them cite either MaizeSequence.org or Gramene (66%), which use Ensembl as their 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 the discontent hints that real issues may lie with some features of Ensembl that need to be addressed by its developers. The respondents usually cited the perceived slowness of

the website as the major (and sometimes the only) problem. Another reported problem was, as one respondent says, "many, many noninituitive steps to get information".

Four Software Suites to Choose From

Although many genome browser software platforms exist, survey respondents were most familiar with Ensembl, GBrowse, the NCBI Map Viewer (23), the UCSC Genome Browser (24), and xGDB (25). Each genome browser is designed with a different focus (Table 1). Here, we provide a short review of some of their functionalities we considered in choosing a genome browser for MaizeGDB. Among these genome browsers, the NCBI Map Viewer is not downloadable to local machines, so it was not considered as a choice for the MaizeGDB Genome Browser. Because our users are extensively using the NCBI Map Viewer, however, we included it in our review for comparison.

(Table 1 here)

Ensembl. The Ensembl browser was originally developed to manage and display genomic data for the Ensembl project as a human genome browser (21). Its developers' current development focus is centered on supporting mammalian genomes. Ensembl especially shines in comparative genomics visualization and analysis. It provides a flexible framework that displays a wide variety of genomes (currently the Ensembl browser displays 48 genomes (21). A recent addition to Ensembl is the new multiple alignment pipeline that passes the data through 3 different programs (the Enredo-Pecan-Ortheus (EPO) pipeline (21,26,27) to obtain alignment results.

Ensembl's web interface combines many distinct, dynamically-generated views (e.g. genes, maps, contigs) to address different needs of the researchers. The framework is also integrated with multiple tools, including the similarity search tools BLAST and SSAHA, the retrieval software EnsMart, and the Distributed Annotation System (DAS) framework (28,29) for sharing and displaying distributed data sets on any publicly-available Ensembl instance (i.e., locally-installed software). Ensembl is designed to be portable—users with advanced programming skills can extend or modify Ensembl code through the Ensembl API (application programming interface), as it is a downloadable open-source package.

GBrowse. The Generic Model Organism Database Project (GMOD) (http://gmod.org) has the mission to build tools designed to serve the needs of model organism databases. One of the major and most popular tools developed by GMOD is the Generic Genome Browser (GBrowse) (22), an open-source web-based framework for displaying genomic annotations and features. Similar to other genome browsers, GBrowse allows the user to scroll and zoom within a genomic region, search for features based on name or keyword search, and customize feature tracks. A useful visual element in GBrowse is that each feature type can be represented by various customizable "glyph"s, which are essentially symbols that vary in shape, color, and size to represent genomic elements.

GBrowse was designed to be portable and extensible (i.e., its code is modifiable to add new capabilities). A developer can modify GBrowse at the following three different layers: the database layer, the data model layer, and the application layer. This flexibility allows the administrator to control how the data are stored, how the data are visualized, and how the user

interacts with the data. GBrowse is a downloadable, stand-alone, open source package and was designed to facilitate third-party plug-ins for data analysis and visualization. Some examples include plug-ins for calculating linkage disequilibrium, dumping data as GFF or FASTA, and facilitating the connection between GBrowse and Galaxy (30). Gbrowse can also be integrated with the comparative map viewer CMAP (31), the BioMart data mining system (32), and the TextPresso text mining tool (33). Some developers have even harnessed the GBrowse extensibility to create a web server for GBrowse that allows access without the hassle of local installation (34). Similar to the Ensemble browser, users can upload custom data (flat files or an URL) with ease through the DAS platform (28,29), which decentralizes data storage by allowing the display of third-party annotations. GBrowse is used by many MODs, including TAIR (1), WormBase (35), and Mouse Genome Informatics (MGI) (36), as well as CODs, such as SOL Genomics Network (4).

NCBI Map viewer. As a static repository, the National Center for Biotechnology Information (NCBI) strives to preserve the archives of large species-specific data sets for the scientific community. Its primary mission is to keep them up-to-date, searchable, and publicly available. NCBI accomplishes these herculean tasks in collaboration with many researchers and curators across species. NCBI also provides a range of tools for the visualization and analysis of genomes. Central to these tools is its genome browser, Map Viewer (23). Map Viewer is not designed to be customizable, but it is capable of visually representing maps and genomic elements and providing links to the webpages that include the most current and comprehensive data about these genomic elements.

One of the main disadvantages of Map Viewer is that it does not have the capability to be downloaded and installed to personal servers. It is specifically designed to work under the NCBI framework in accordance with the primary mission of serving to the public and the researchers as a static data repository.

The UCSC Genome Browser Database. The University of California Santa Cruz (UCSC) Genome Browser Database (24) started as part of the Human Genome Project (37) to make newly generated human genomic sequences publicly available. Although the UCSC genome browser remained focused on the human genome, its content over the years has extended to a cross-comparison platform of 19 vertebrate and 21 invertebrate species(24). The UCSC browser currently serves many tracks including evolutionary conservation track based on 28-species, variation and disease tracks, and mammalian gene collection tracks. Plant genomes are not included on the main UCSC site, any genome sequence, however, can be uploaded to a locally-installed instance of the UCSC browser. The browser is open-source; therefore customization by developers is possible. Also, the browser allows some customization in the form of "custom tracks" that might be uploaded to the UCSC website or to any available instance of the UCSC genome browser using the DAS framework. Note that similar to GBrowse and Ensemble, DAS tracks in the UCSC browser can be created temporarily on any instance that uses the DAS framework, but these tracks will only be privately available for the user who uploads them. It is also possible to use the DAS framework to create publicly available, permanent tracks to display data provided by third-party servers, but this requires access and administrative privileges to the main server that the genome browser is located.

xGDB. The eXtensible Genome Data Broker (xGDB) (25) is the genome browser developed by personnel working at PlantGDB (5) to facilitate their need for a system to manage, store, and display genomic evidence for 16 green plant genomes (including the maize genome). xGDB is a software package designed to view the outcomes of sequence analyses within a genomic context. xGDB can be customized for various individual research tasks and analysis needs. Other features of xGDB include search tools, online publishing, Web services, and third-party tool integration. The browser serves data through the DAS framework.

Choosing a Genome Browser

Choosing a genome browser to address the maize community presents a challenge given that several browsers (reviewed above) have different strengths and weaknesses. For example, one of the most popular genome browsers, Ensembl, provides the best tools for comparative genomics. In contrast, another popular genome browser, GBrowse, provides a wide range of tools for MODs, yet its tool repertoire for comparative genomics is not as rich as Ensembl's. Therefore, determining which software best suits the needs of maize geneticists is a task that requires a careful consideration.

Based upon results of the Genome Browser Survey, we chose GBrowse as the MaizeGDB Genome Browser for the following reasons:

1) Because maize researchers have a wide range of research interests, we decided to implement a genome browser that could be adapted to addresses general research questions. **UCSC**, **xGDB**, **Ensembl**, and **GBrowse** would all fit this need.

2) The **UCSC** genome browser is highly capable. However, one disadvantage of choosing it for the MaizeGDB Genome Browser is that no plant database currently uses the UCSC Genome Browser; TAIR, Soybase, and SOL Genomic Network (among others) use GBrowse. The availability of developers from plant databases, as well as from other MODs (e.g., Flybase, Mouse Genome Informatics), creates more opportunities for future collaboration to create similar solutions to respond to common challenges related to data integration and visualization.
3) **xGDB** is a downloadable open source package, but it is not in wide use yet: so far PlantGDB is the only site that uses **xGDB**, and it has a limited number of developers.
4) In the "Feature ranking", the three most desired features are chosen as: ease of use, visuals, and speed. The survey results indicate that cooperators do not consider **Ensembl** easy to use, and it is definitely perceived not as fast when compared to the other software available. Also, the desire to have cross-species comparison capability in a genome browser (where Ensembl shines) is only ranked 4th. Note that although not extensive as Ensembl's, Gbrowse has some cross-species tools already available (Synbrowse (38,39), CMap (31), and GBrowse_syn, which is included in the GBrowse 1.70 Release).
5) As indicated in the "Indispensable features", cooperators would like to see specific tool development in a 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 programming, tools development, and community involvement. The flexibility of tool development is intrinsic feature of GBrowse that allows customizable plug-in architecture as a community-based open source project. In the case of **Ensembl**, the code development is primarily done by a group in the United

Kingdom and *ad hoc* tool development is carried out by research groups for their specific needs. Because this tool development by databases is specific to a particular **Ensembl** version, the tools must be modified or re-written for each new version of **Ensembl**. This creates an issue with Ensembl as it requires more manpower and funding to adapt the code to new version of the genome browser. In the case of **xGDB**, the flexibility in code development is somewhat limited. Because this browser is not widely used, the number of independent developers working on **xGDB** is not comparable to the community of **GBrowse** developers.

6) MaizeSequence.org already provides maize genome sequence information using **Ensembl**. Providing this information using **GBrowse** and providing links to MaizeSequence.org would allow researchers to access different genome browsers for different applications and preferences. For example, when a cross-species comparison across many clades is necessary, **Ensembl** provides efficient solutions; however, when it comes to developing customizable visualization and analysis tools for maize-specific research problems, **GBrowse** stands out. Offering the availability of these two browsers to maize researchers will facilitate answering different research problems and will enhance agricultural research overall.

Although the MaizeGDB team devoted considerable time and effort to the decision of which genome browser to implement, we realize that the current selection of a specific genome browser may not matter in the long-term because the accelerating technology would certainly engender new and improved genome browsers that are not currently available to be adopted.

Implementing GBrowse

We started implementing the GBrowse-based MaizeGDB Genome Browser (described in detail in the MaizeGDB DATABASE paper (10) in February of 2008. We obtained maize data from various sources, including MaizeSequence.org, PlantGDB, and MAGI. We chose 5 people for guidance (from academia and industry in the U.S. and abroad) 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. The guidance and beta-testing groups provided many valuable inputs to improve our users' experience with the MaizeGDB Genome Browser. The MaizeGDB Genome Browser was released in December of 2008. We are still in the process of implementing the some of the guidance and beta testing groups' suggestions and continually work to better integrate the genome browser with existing data by creating new tools. One of these suggestions, provided to us by Dr. Sarah Hake, led to the creation and implementation of one of our most used tools in MaizeGDB: the Locus Lookup tool (40). This tool takes one or two loci as input and returns an approximate genomic region based on known physical and genetic associations, even in the case when the locus of interest is not yet mapped to the maize genome. The utility of the Locus Lookup tool is apparent especially for the genomes which are in the process of being sequenced.

FUNDING

This work is supported by the U.S. Department of Agriculture-Agricultural Research Service.

ACKNOWLEDGMENT

We thank USDA-ARS for its sustained funding, the members of the MaizeGDB Working Group (Volker Brendel Ed Buckler, Karen Cone, Mike Freeling, Owen Hoekenga, Lukas Mueller,

Marty Sachs, Pat Schnable, Tom Slezak, Anne Sylvester, and Doreen Ware), as well as the members of the MaizeGDB Executive Committee (Pat Schnable, Mary Alleman, Tom Brutnell, Sarah Hake, Jane Langdale, Jo Messing, Jean-Phillippe Vielle-Calzada, Anne Sylvester, William Tracy, Virginia Walbot, and Sue Wessler) for their direction, support, and inputs on this work. We also would like to thank Guidance Group (Peter Balint-Kurti, Sarah Hake, Damon Lisch, Mike Muszynski, and Virginia Walbot) and Beta-testers (Alain Charcosset, Olivier Dugas, James Estill, David Hessel, Damon Lisch, Mike Muszynski, Paul Scott, Virginia Walbot, Rachel Wang, and Cesar Alvarez-Mejia), without whose comments and suggestions we could not have created an implementation of the MaizeGDB Genome Browser customized to support our users' needs. We very much appreciate the useful comments by Dr. Patrick Armstrong. Last, but not least, we deeply appreciate and thank the maize community for their continuous support.

REFERENCES

1. Swarbreck, D., Wilks, C., Lamesch, P., *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, **36**, D1009-14.

2. Tweedie, S., Ashburner, M., Falls, K., *et al*. (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res*, **37**, D555-9.

3. Liang, C., Jaiswal, P., Hebbard, C., *et al*. (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res*, **36**, D947-53.

4. Mueller, L.A., Solow, T.H., Taylor, N., *et al*. (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol*, **138**, 1310-7.

5. Duvick, J., Fu, A., Muppirala, U., *et al*. (2008) PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res*, **36**, D959-65.

6. Childs, K.L., Hamilton, J.P., Zhu, W., *et al*. (2007) The TIGR Plant Transcript Assemblies database. *Nucleic Acids Res*, **35**, D846-51.

7. Chan, A.P., Pertea, G., Cheung, F., *et al*. (2006) The TIGR Maize Database. *Nucleic Acids Res*, **34**, D771-6.

8. Lawrence, C.J., Harper, L.C., Schaeffer, M.L., Sen, T.Z., Seigfried, T.E., Campbell, D.A. (2008) MaizeGDB: The Maize Model Organism Database for Basic, Translational, and Applied Research. *Int J Plant Genomics*, **2008**, 496957.

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

10. Sen, T.Z., Andorf, C.M., Schaeffer, M.L., *et al*. (2009) MaizeGDB becomes 'sequence-centric'. *Database*, **in press**.

11. Schnable, P.S., Ware, D., Fulton, B., Stein, J., Wei, F., al., e. (2009) The B73 maize genome: complexity, diversity and dynamics. *Science*, **in press**.

12. Wei, F., Zhang, J., Zhou, S., *et al*. (2009) The physical and genetic framework of the B73 maize genome. *PLoS Genetics*, **submitted**.

13. Buckler, E.S., Holland, J.B., Bradbury, P.J., *et al*. (2009) The genetic architecture of maize flowering time. *Science*, **325**, 714-8.

14. McMullen, M.D., Kresovich, S., Villeda, H.S., *et al*. (2009) Genetic properties of the maize nested association mapping population. *Science*, **325**, 737-40.

15. Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., *et al*. (2003) Maize genome sequencing by methylation filtration. *Science*, **302**, 2115-7.

16. Whitelaw, C.A., Barbazuk, W.B., Pertea, G., *et al*. (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science*, **302**, 2118-20.

17. Fu, H., Dooner, H.K. (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci U S A*, **99**, 9573-8.

18. Pruitt, K.D., Tatusova, T., Klimke, W., Maglott, D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*, **37**, D32-6.

19. Kurata, N., Yamazaki, Y. (2006) Oryzabase. An integrated biological and genome information database for rice. *Plant Physiol*, **140**, 12-7.

20. Grant, D., Nelson, R.T., Cannon, S.C., Shoemaker, R.C. (2009) SoyBase, The USDA-ARS Soybean Genome Database http://soybase.org.

21. Hubbard, T.J., Aken, B.L., Ayling, S., *et al*. (2009) Ensembl 2009. *Nucleic Acids Res*, **37**, D690-7.

22. Stein, L.D., Mungall, C., Shu, S., *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*, **12**, 1599-610.

23. Wolfsberg, T.G. (2007) Using the NCBI Map Viewer to browse genomic sequence data. *Curr Protoc Bioinformatics*, **Chapter 1**, Unit 1 5.

24. Karolchik, D., Kuhn, R.M., Baertsch, R., *et al*. (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, **36**, D773-9.

25. Schlueter, S.D., Wilkerson, M.D., Dong, Q., Brendel, V. (2006) xGDB: open-source computational infrastructure for the integrated evaluation and analysis of genome features. *Genome Biol*, **7**, R111.

26. Paten, B., Herrero, J., Beal, K., Fitzgerald, S., Birney, E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res*, **18**, 1814-28.

27. Paten, B., Herrero, J., Fitzgerald, S., *et al*. (2008) Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res*, **18**, 1829-43.

28. Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R., Stein, L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

29. Jenkinson, A.M., Albrecht, M., Birney, E., *et al*. (2008) Integrating biological data - the Distributed Annotation System. *BMC Bioinformatics*, **9 Suppl 8**, S3.

30. Giardine, B., Riemer, C., Hardison, R.C., *et al*. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, **15**, 1451-5.

31. Youens-Clark, K., Faga, B., Yap, I.V., Stein, L., Ware, D. (2009) CMap 1.01: A comparative mapping application for the Internet. *Bioinformatics*.

32. Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., Kasprzyk, A. (2009) BioMart Central Portal--unified access to biological data. *Nucleic Acids Res*, **37**, W23-7.

33. Muller, H.M., Kenny, E.E., Sternberg, P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, **2**, e309.

34. Podicheti, R., Gollapudi, R., Dong, Q. (2009) WebGBrowse--a web server for GBrowse. *Bioinformatics*, **25**, 1550-1.

35. Rogers, A., Antoshechkin, I., Bieri, T., *et al*. (2008) WormBase 2007. *Nucleic Acids Res*, **36**, D612-7.

36. Shaw, D.R. (2009) Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr Protoc Bioinformatics*, **Chapter 1**, Unit1 7.

37. Lander, E.S., Linton, L.M., Birren, B., *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.

38. Brendel, V., Kurtz, S., Pan, X. (2007) Visualization of syntenic relationships with SynBrowse. *Methods Mol Biol*, **396**, 153-63.

39. Pan, X., Stein, L., Brendel, V. (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*, **21**, 3461-8.

40. Andorf, C.M., Lawrence, C.J., Harper, L.C., Schaeffer, M.L., Campbell, D.A., Sen, T.Z. (2009) The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. *Bioinformatics*, **in press**.