MaizeGDB: The Maize Genetics and Genomics Database

Lisa Harper, Jack Gardiner, Carson Andorf, and Carolyn J. Lawrence

Abstract

MaizeGDB is the community database for biological information about the crop plant *Zea mays*. Genomic, genetic, sequence, gene product, functional characterization, literature reference, and person/organization contact information are among the datatypes stored at MaizeGDB. At the project's website (http://www.maizegdb.org) are custom interfaces enabling researchers to browse data and to seek out specific information matching explicit search criteria. In addition, pre-compiled reports are made available for particular types of data and bulletin boards are provided to facilitate communication and coordination among members of the community of maize geneticists.

Key words Maize, Database, Genetics, Genomics, Genome, Model organism database

1 Introduction

MaizeGDB is the repository for and interface to maize biological data. Many diverse types of public data can be found at MaizeGDB, such as: DNA sequence, genome assemblies, genetic maps, cytogenetic maps, haplotype maps, genetic mapping panels, genes and other loci, gene models, transcripts, ESTs, mutations, alleles, stocks, QTLs, SNPs, BACs, probes, gene products, proteins, pathway data, microarray data, expression atlas, RNAseq, microRNAs, insertion elements and stocks, images of mutants, images of gel patterns, references, people, organizations, tutorials, curated list of maize projects and resources, and community information such as history of the community. Formed in 2002 [1] by the fusion of two databases MaizeDB [2, 3] and ZmDB [4]; MaizeGDB is now a sequence centric one stop shop for maize biological data [5]. The MaizeGDB also provides a full featured Genome Browser which includes many custom tracks [5]. Data enters MaizeGDB in one of several ways: (1) Information from primary literature is entered via manual curation, (2) Data provided directly from maize researchers assisted by MaizeGDB, and (3) Data provided by other databases such as NCBI, Gramene, and PlantGDB.

MaizeGDB personnel work with generators of large datasets to get the data into a format that can be served and accessed at MaizeGDB. In addition to collecting, storing and making available maize data, workers at MaizeGDB also develop custom tools to interact efficiently with the truly enormous and complex datasets now available. "Big Data" is the catchphrase used to describe information sets that have extreme volume, variety, velocity, and complexity (reviewed at http://blogs.sap.com/analytics/2012/04/11/big-data-for-small-companies/). Traditional and even current data management applications are not designed to handle this data load. As it is used currently, "Big Data analysis" conveys the concept of extreme information management. Although large datasets are the aspect of Big Data management that is most obvious, large (and often fairly simplistic) datasets like genome-wide SNP datasets are actually the least interesting to serve and covey the least biologically relevant meaning. By comparison, the complexity aspect of Big Data management involves recapitulating biological meaning and is an aspect of the emerging discipline that MaizeGDB was a forerunner in addressing well. The depth and breadth of data at MaizeGDB continue to be a unique aspect of the data resource relative to other online repositories of biological information.

In addition to serving maize data and tools for interacting with the information, MaizeGDB personnel also provide services to the maize research community. Bulletin boards for news items, information of interest to cooperators, a curated list of maize projects and resources that focus on the scientific study of maize, an editorial board's recommended reading list, and educational outreach items are among the webpages made available through the MaizeGDB site (*see* Table 1). In addition, workers at MaizeGDB provide technical support for the Maize Genetics Executive Committee and the Annual Maize Genetics Conference.

Information about the history of MaizeGDB and the technical aspects of project's operation are described elsewhere [1, 5–8]. Reported here are the types of data and tools that are made available at MaizeGDB, some generalized search strategies that can be applied across various datatypes, and some specialized example usage cases. Mechanisms for adding data to the database also are described in detail.

2   Materials

This section lays out in detail the types of data stored at MaizeGDB.

2.1   Genomic Sequence Data

One item that relates directly to MaizeGDB's mission is to serve as the long term steward of whole genome sequence assemblies from any *Zea mays* subspecies and inbreds. We house or maintain current links to:

Table 1
Bulletin boards and static pages

Page title and web address	Content description
News column http://www.maizegdb.org/	News bulletins are displayed in the right margin. Older items are accessible through a link near the bottom
Tutorials http://outreach.maizegdb.org	Written and video tutorials that explain how to use various tools on the MaizeGDB website
Data Contribution "How To" Guide http://www.maizegdb.org/contribute_data	Displays sources of currently stored data and how researchers can contribute their own data
Editorial Board http://www.maizegdb.org/hot_new_papers	A list of noteworthy publications selected monthly by the MaizeGDB Editorial Board
Cooperators' page http://www.maizegdb.org/cooperators	Page of links to resources supporting the cooperative spirit shared among maize researchers
Maize Genetics Cooperation—Newsletter http://www.maizegdb.org/mnl	Makes accessible online copies of the MNL and provides information on how to receive hard copies
Maize Genetics Executive Committee http://www.maizegdb.org/mgec	Provides details on the membership, goals, function, and history of the MGEC
Maize Genetics Conference http://www.maizegdb.org/maize_meeting/	Centralized resource for current and past Maize Genetics Conferences which includes dates, locations, registration information, and abstract submission forms
Maize Research Projects List http://www.maizegdb.org/popcorn/search/	A curated list of maize projects and links to their respective project sites and resources generated
MaizeGDB/National Corn Growers Podcasts (http://www.maizegdb.org/)	An ongoing series of podcasts that educates MaizeGDB's stakeholders on the value of maize research and the need to make all the data publically available

1. Sequences from various maize inbred lines, including all versions of the B73 Reference Genome Assembly. Sequences from other inbred lines are shown in relationship to the inbred B73. These are accessible through MaizeGDB's GBrowse-based Genome Browser.

2. SNPs and flanking genomic sequences from different maize inbred lines, and access to larger databases such as Panzehhha (http://www.panzea.org/) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/).

3. BAC sequences from the minimum tiling path of the B73 Reference Genome Assembly, as well as contig information for both within BACs, and between BACs.

4. Molecular probe sequences and detection/amplification methods for RAPDs, ESTs, SSRs, RFLPs, AFLPs, overgos, and other genomic DNAs.

2.2	*Large Datasets*	With Next Generation Sequencing technologies, there has been a fundamental change in how maize researchers ask and address biological questions. It is now possible to generate large DNA/RNA sequence-based data sets at costs that were previously unthinkable. Because of this, these large datasets (often termed "Big Data" though, as mentioned above, large datasets are only one aspect of what constitutes Big Data) can contain millions, and even billions of data points. Access to these large data sets is currently problematic. Active research is ongoing in many fields to evolve methods of access and display to handle large datasets. Currently at MaizeGDB, various large datasets can be accessed through portals to other databases.

1. RNAseq primary data is stored at the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra), and served at MaizeGDB primarily as graphs of expression at each locus on the Genome Browser.

2. Access to billions of SNPs and flanking sequences from different maize inbred lines is currently available through Panzea and will eventually be accessible directly from MaizeGDB.

3. Proteomic data on the developing maize kernel made accessible through the Maize Proteome Project (http://maizeproteome.ucsd.edu/).

2.3	*Genetic Data*

1. Loci including (but not limited to) genes, chromosomal segments, centromeres, introns, probed sites, and quantitative trait loci (QTL).

2. Variations including known mutant and non-mutant alleles at a given locus, chromosomal structural variations, cytoplasmic variations, DNA polymorphisms, rearrangements, transpositions, etc.

3. QTL experiment environmental conditions, parental stocks, agronomic traits of interest, locus summaries, and raw data files.

4. High resolution genetic maps including nested association maps, hapmaps, and intermated population maps.

5. Over 2000 genetic maps, including both composite and individual maps, cytogenetic and cytological maps along with associated data including mapping panel descriptors, population size, and source information.

6. Seed stock (accession) descriptors consisting of a unique identifier (the stock name) and known synonyms, the stock source (e.g., an individual researcher's name or an organization name like the "Maize Genetics Cooperation—Stock Center"), and associated locus linkage group assignments, genotypic variations, karyotypic variations, phenotypes, and parental stock identifiers.

2.4   Gene Product and Functional Characterization Descriptions

1. Metabolic pathway data can be accessed through CornCyc, (http://corncyc.maizegdb.org, developed in collaboration with the Plant Metabolic Network, PMN) and MaizeCyc (http://maizecyc.maizegdb.org, developed in collaboration with Gramene).

2. Gene products with associated Enzyme Commission (EC) numbers, expression, induction conditions, subcellular localization data, metabolic pathway, known metabolic cofactors, mass (kDa), and links to loci that encode them.

3. Phenotypic descriptions that include trait descriptions and affected tissue types/organs (body parts) alongside mutant images.

2.5   Terms, Controlled Vocabularies, and Ontologies

1. Ontologies are hierarchically related controlled vocabularies that serve to enable communication across different databases and data sets. Ontology terms from many different established ontologies, such as the Gene Ontologies [9], the Plant Ontology, and the Trait Ontology [10], are assigned to data as appropriate. Where appropriate, phenotypes are described using Entity-Quality (EQ) statements which utilize the strengths of many different ontologies.

2. Terms and term definitions that describe stored data of various types.

3. Additional controlled vocabularies that are the set of terms that describe a given process or datatype. For example, terms of type "Developmental Stage" make up one controlled vocabulary.

2.6   Literature References and Person/ Organization Records

1. References from primary literature, the Maize Genetics Cooperation—Newsletter, and abstracts from the Annual Maize Genetics Conference; associated with virtually all other data types.

2. Contact information records for cooperators, authors, and organizations.

3   Methods

This section outlines the various ways to find and interact with data at MaizeGDB.

3.1   Interrogation Tools

Navigating data to find specific, useful pieces of information is not always a simple task. Learning to use the tools that will enable facile data navigation is, therefore, a good use of time. By learning the general methods for browsing and searching MaizeGDB, the time required to locate information will be decreased, allowing for more to be spent testing hypotheses at the bench. In each of the

following sections, descriptions for each type of search are provided and general techniques for efficiently and effectively navigating the MaizeGDB interface are described.

3.1.1 Interrogation Tools: Embedded Simple Search

The fastest and easiest way to navigate to data of interest at MaizeGDB is by using the search feature located on the right side of the horizontal green toolbar across every page. Clicking on "Search" will open the simple search box where you can enter your search term. If "all data" remains selected in the drop down menu, virtually all data will be searched simultaneously. Much faster searches can be done by selecting the data type from the dropdown menu. Press the button marked "Go!", or the enter or return button on your keyboard to start the search. Below, instructions are given to find the genetic position of the *bronze1* (*bz1*) locus, its gene model and how to visualize the gene model on the genome browser.

1. Go to http:// www.maizegdb.org.
2. Locate the green menu bar at the top of the page, and click "search" on the right hand side.
3. Read the note that appears in the popup window.
4. Specify criteria to locate records about *bronze1* by selecting "locus/loci" from the dropdown menu and by typing *bz1* into the field to the right. Click the button marked "Go!".
5. The locus page for *bz1* gene is first in the results list. Click the link to the *bz1* gene.
6. If a gene model has been linked to the bz1 locus, it will be present next to the ear of corn. Some gene models have yet to be linked to loci. On this page, scroll down to see the list of BACs, overgos, and other probes known to mark the *bz1* locus.
7. To see the *bz1* gene model on the Genome Browser, look in the "Overview" box. Scroll down to "Associated Gene Models", and click the "Genome Browser" link.
8. To download sequence, the GBrowse2 tools can be used: in the search box at the top of the view, select "download decorated FASTA file" or "Download Sequence File" from the pull down menu on the left side.

3.1.2 Interrogation Tools: Advanced Search of Data Centers

Questions asked by biologists are complex, so tools that query the database must enable complex queries to be made. MaizeGDB has grouped like data types into "data centers" (e.g., Genes/Gene models, Alleles and Polymorphisms, Expression, Images, Maps, Phenotypes, Sequences, References). This allows researchers to use custom queries that more efficiently search specific data types. For example, the search algorithms are different when searching for a reference, versus a sequence, versus a phenotype, so the advanced

search boxes allow different inputs. These custom searches also allow more logical display of the results.

Each of the Data Centers can be selected from the "Data Center" drop down menu on the green horizontal bar at the top of most pages. Often used Data Centers are also shown as a button in the center of the home page. Data Center names are linked to a page that explains the Data Center and makes available a Simple Search (similar in function to the search available through the search bar described above), an Advanced Search (if relevant, discussed more fully in this section), and a Discussion of the Data Type (written at a level comprehensible by the general public). The Expression Data Center is an exception: currently, this Data Center brings together many offsite tools to access and analysis expression data from various plant sites.

To demonstrate the use and functionality of the various Data Centers' Advanced Search tools, here are two examples of their use from two disparate Data Centers.

The Gene/Gene Models Data Center. In this Data Center's Advance search box, you can search for genes by name, type, known phenotypes, gene products, by chromosome or by a combination of these parameters. Below the Advanced search box is a box where you can search for genes and gene models by sequence using BLAST. Below that is a very useful box that allows users to download all gene between two genome coordinates, between coordinates of two markers or BACs, or download all genes on individual BACs. These search features are useful for researchers performing map based cloning. Next is a search box where you can enter a list of gene model names or transcript IDs (up to 8000), and retrieve their sequence. Lastly on this Data Center page are lists of useful links for accessing and downloading more gene and gene model information.

The Maps Data Center. There are over 2000 genetic maps of maize, and finding an appropriate map can be challenging for researchers. At the top of the Maps Data Center is a "Handy Reference" that describes the most commonly used genetic maps, and how they where constructed. On the Maps Data Center, you can search for genetic maps that contain markers of interest, that are from particular sources (individuals, companies and public institutions) or by mapping panel, or by a combination of these variables. Once a map has been chosen, information about that map can be found on its page. Below the advanced Search box, are links to the most commonly used genetic maps.

3.1.3 Interrogation Tools: Finding Projects and Resources

The *PrO*ject *P*ortal for corn (POPcorn) was developed as a single entry point for researchers to explore maize projects and resources that have been developed by maize researchers worldwide to advance maize research [11]. Currently, POPcorn contains 159 projects and 137 resources. Projects and resources are distinguished

from each other in that projects are generally knowledge driven with distinct deliverables and a specific endpoint. In contrast, resources, which often extend beyond the lifetime of the project, provide either biological stocks (either DNA based or seed) or software tools for navigating the data that has been developed by POPcorn projects. The central idea driving the development of a single access point for maize projects and resources was that not only would this save maize researchers time and effort in locating projects that might otherwise be over looked, but also to make sure that when the funding period ended, that valuable data and tools are maintained long term. Currently, POPcorn serves the maize community in three capacities: (1) POPcorn allows localization and utilization of community databases and large scale data sets with a single search. Curated POPcorn projects and resources can be searched by keyword, investigator, institution, category, and country. (2) POPcorn allows single or multiple DNA sequence searches (via BLAST) that access all POPcorn associated DNA sequence databases (~45 are currently represented) and returns a single, collated output for easy viewing. (3) POPcorn provides a mechanism for preserving raw data and associated annotations contained within POPcorn for migration to MaizeGDB for long-term storage. An important feature of POPcorn is that it is still actively curated even though project funding (NSF DBI 074804) has ended. MaizeGDB curators spend a few hours each month looking for new projects and resources while an automated utility checks URLs and send an email to alert curators that a particular link has become inactive. These curation efforts ensure that POPcorn will continue to be a relevant resource to the maize community.

3.1.4 Interrogation Tools: Accessing the Maize Community

The maize community spans a timeline of over 80 years and has a rich tradition of sharing community resources, both physical and intellectual. Not surprisingly, maize as both an applied and model research organism has benefitted from the cohesiveness of the maize community. One of MaizeGDB's primary objectives is to serve as a clearing-house for organizational information of interest to the maize community. Maize researchers can access the MaizeGBD community page from the menu bar on the home page. Contact information for maize researchers (Cooperators) as well as the Maize Executive Committee can be located. Information on the Annual Maize Genetics Conference, including direct links to registration and hotel accommodation's can be obtained. In addition, researchers can get directions and guidelines on contributing their data to MaizeGDB or becoming a community data curator. Information on the Maize Genetics Newsletter, current job listings, and the Maize Editorial Board's recommended readings are also available.

3.2 Analysis Tools

MaizeGDB provided open source and custom tools to allow users to drill down to the data the way they need to view it and use it.

3.2.1 Analysis Tools: BLAST

MaizeGDB created a powerful customizable BLAST tool which can be easily accessed from the home page through the BLAST pull down menu, by selecting BLAST in the Tools pull down menu, or by using the BLAST button in the center of the page. Using webservices, MaizeGDB BLAST can search locally stored sequences, all GenBank datasets (ESTs, GSS, HTGS, etc.), and many other sequence databases such as all gene model builds, all sequence assemblies, repeat databases, all known loci, microarray probes, transcription factors, and more. Selection of datasets to query is at the users' discretion. MaizeGDB BLAST will accept up to five input queries with a combined length of no more than 35,000 bp. Some web services may reject queries of this maximum size; if that happens the results page will show an error for that target host/data set. Longer query sequences or a large number of sequences can be aligned to the reference sequence assembly using another tool called ZeAlign (described below). MaizeGDB BLAST has four clearly delimited steps: Step 1: Input your sequences (Raw, FASTA, or GenBank IDs), and indicate the sequence is nucleotides or amino acids. Step 2: Select Datasets: MaizeGDB offers a wide variety of sequence datasets to search. This allows users to have a single place to BLAST where they can hit multiple databases at once. Each dataset is explained by hovering over the title of the dataset. Step 3: Select BLAST parameters: by choosing one of the preset options, or by modifying the advanced settings. An explanations for each option is available by hovering over the option title. Step 4: Select output type: MaizeGDB offers three possible outputs layouts; The standard BLAST text output (like the BLAST output at NCBI), The BLAST table output, and a expanded table output created by MaizeGDB for maximum integration with the website. Results can be displayed, or emailed, or both.

3.2.2 Analysis Tools: ZeAlign

Users that generate a large number of sequences can use the MaizeGDB developed ZeAlign tool to align their sequences to the current genome Assembly. This tool can also be used to remap older data to a newer genome assembly. ZeAlign is a BLAST tool, but unlike the regular MaizeGDB BLAST tool, ZeAlign allows up to 20,000 sequences to be aligned to a genome assembly. This tool is often used before researchers put their data on their own custom genome browser track which can be either private, or can be submitted to MaizeGDB for review and if approved, then for public viewing. ZeAlign can be reached by selecting it from the BLAST pull down menu on the green bar across the top of every page.

3.2.3 Analysis Tools: Genome Browser

MaizeGDB users can access five genome browsers covering three genotypes from the MaizeGDB home page: (1) MaizeGDB/B73, (2) Maizesequence.org/B73, (3) Genomaize/B73, (4) Phytozome/Mo17, (5) and QuerySequenceVisualizer/Palomero Toluqueño. The MaizeGDB Genome Browser uses a semi-customized version of GBrowse2 and is actively supported by MaizeGDB personnel.

The MaizeGDB genome browser allows access to archived sequence assembly versions (such as the original BAC-based sequence of B73 and B73 RefGen_V1), and fully supports the two most recent versions which currently are the B73_RefGen_V2 and B73 RefGen_V3 maize genome assemblies. This is critical to many MaizeGDB users as researchers often become vested in a particular version of the genome assembly. Genome version stability is essential for them to complete their research projects—especially those taking a positional cloning approach towards gene cloning. The current installation of GBrowse2 at MaizeGDB contains over 50 tracks that provide information on genome diversity, gene expression, gene models, genetic maps, DNA insertions, and repetitive elements. In addition to the standard GBrowse2 features such as the option to create snapshots and generate community and/or private tracks, MaizeGDB has created maize-specific documentation and tools tabs. The documentation tab introduces new users to various features in GBrowse2, whereas the maize specific tools tab gives users access to maize specific tools, such as BLAST, chromosomal bin, and incongruence viewers that have been embedded within the GBrowse2 environment.

3.2.4 Analysis Tools: Locus Lookup and Locus Pair Lookup

Not all genes in the maize genome have been identified, and not all in silico gene models have been linked with genes previously identified by classical genetics. Often researchers identify new genes based on mutant phenotype, genetically map them, and then want to identify their sequence. MaizeGDB provides the Locus Lookup and Locus Pair Lookup tools [12] to locate a genomic sequence interval where a gene of interest may reside, based on its position on the genetic map. This tool can be used if you have a genetic position for your gene of interest relative to one of several MGDB's composite genetic maps. The tool works by finding the nearest genetically mapped loci that flank the input locus and have an association with the B73 genome sequence, and returns those coordinates. Since many classically identified genes do have sequence coordinates, the Locus lookup tool goes through a hierarchical process:

Upon given the input coordinates, the tool:

1. checks if the locus is associated to any gene models and the coordinates for the gene model are returned, else

2. checks physical map coordinates to find out whether the locus is already placed. If so, the physically mapped locus coordinates are returned, else

3. checks the locus record at MaizeGDB to find out if any placed BACs are known to detect the locus and that BAC is returned within its genomic context, else

4. genetically mapped probes that are nearest 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.

The Locus Lookup tool will locate the genome region around one genetically mapped locus, while the Locus Pair Lookup allows the user to enter two genetically mapped loci and returns the genome sequence coordinates that include both loci and the region between them. The Locus lookup tool can be accessed from the home page by clicking the large "Locus Lookup" button, or by selecting this under the "Tools" drop down menu.

3.2.5 Analysis Tools: Cyc Databases

Metabolic pathways hosted at MaizeGDB can be accessed through either the expression or metabolic pathways quick links on the MaizeGDB homepage. Currently MaizeGDB hosts two independent, but complementary metabolic pathway viewers, MaizeCyc developed by Gramene [13, 14], and CornCyc, created by the Plant Metabolic Network which were both computationally inferred using either the Pathway Tools Software suite (MaizeCyc) or the Ensemble Enzyme Prediction Pipeline (E2P2. CornCyc). For MaizeCyc, MaizeGDB curators provided 772 literature-based GO term annotations, while for CornCyc, they provided curations on the auxin, brassinosteroid, and gibberellin pathways. The MaizeCyc and CornCyc respective pipelines identified similar numbers of enzymatic reactions and pathways. However, the two differ in that MaizeCyc contains a larger numbers of transporter proteins and reactions while CornCyc identifies over 4000 spliced protein variants. MaizeCyc and CornCyc were developed using different stringencies for their respective pipeline so it is not surprising their output differs with respect to the number of protein functions identified. Taken together, these two approaches are highly complementary and provide a more robust resource than either would alone: MaizeCyc identifies more enzymes and pathways but at lower accuracy, whereas CornCyc identifies fewer, but at a higher accuracy.

3.2.6 Analysis Tools: Expression Data Center

The expression data center at MaizeGDB houses both gene expression data as well as a collection of tools designed to facilitate its analysis. It can be reached either through the expression quick link on the MaizeGDB home page or from the data centers pull down menu at the top of the home page. The primary focus of the expression data center is to provide MaizeGDB users with access to a suite of complementary utilities that have been developed by external groups to leverage atlas style, genome-wide, gene expression data sets that have been generated by either high density microarrays or RNA sequencing. The gene expression tools accessible from MaizeGDB accommodate virtually any approach that a researcher might take towards leveraging expression data. The

maize eFP browser, developed by researchers at the University of Toronto [15] uses a pictogram approach to display the level of gene expression across a series of maize pictures representing over 60 maize tissues or cell types. It also allows the gene expression patterns of genes to be compared across tissues. This is particularly useful in comparing closely related homologs and identifying tissue specific expression patterns. The MapMan utility developed at the Max Planck Institute for Molecular Plant Physiology projects gene expression data on to a large collection of biochemical processes and metabolic pathways [16]. Users can compare and contrast two tissues or treatments across a range of metabolic pathways and quickly identify gene expression differences within specific steps in a biochemical pathway. MapMan is particularly well suited for control vs. experimental treatments such as mutant vs wild type or abiotic stress treatments where major shifts in metabolic pathways might be expected. The qTeller (QTL Teller) utility, which was developed by James Schnable and Mike Freeling at the University of California-Berkeley, allows users to view expression levels of genes within a user specified chromosomal interval. qTeller, which draws upon a large corpus of publically available gene expression data, also allows the expression levels of syntenic orthologs of rice, sorghum, Setaria, and Brachypodium to be compared which facilitates cross species comparisons of genes within a syntenic interval. Due to these unique features of QTeller, it is the expression analysis tool of choice for positional or QTL cloning projects.

3.2.7 Analysis Tools: Incongruency Tool

Maize has a large genome (about 2.7 gigabases [17]), many diverse lines with different DNA content, and the B73 reference sequence assembly is in flux. The B73 inbred line of maize was sequenced in a BAC by BAC approach (pubmed/19965430). First, a BAC library was made, then the identification of genetically mapped probes on each BAC allowed BACs to be aligned to the IBM genetic map, creating a high information content fingerprint (HICF) physical map. From this, a minimum tilling path of BACs was selected, and these BACs were sequenced, starting in 2005. Thus, the order of the sequences of each BAC is determined by the IBM genetic map. Since 2005, the density of markers per BAC (i.e., the complete sequence of the BAC) has increased, and the IBM genetic map has improved. The Incongruency tool allows users to check genomic regions for inconsistencies between the current genome assembly and the current IBM map (ISU Integrated IBM 2009 map). Where there is a discrepancy, it is difficult to determine which version is correct, however, the tool indicates a region of the genome assembly where the assembly itself needs improvement and serves as a warning to researchers to be careful with interpreting data in these regions. The Incongruency tool can be accessed from any page using in the TOOLS drop down menu.

Sometimes users want to see all the data available in a small region of the genome. While the Genome Browser is an excellent way to get data in a known sequence region, the Bin Viewer allows users to get data in a defined genetic region. Each maize chromosome is divided into 7–12 regions that are each about 20 centiMorgans [18]. To view all data in each bin, users can get to the Bin Viewer from the home page by clicking the centrally located Bin Viewer button, or selecting it from the "Tools" drop down menu. The Bin Viewer is a particular good way to visualize QTL data.

3.3   Ways to Add Data to the Database

MaizeGDB pulls sequence and other data from many sources. Other datasets are classified as one of three types: large datasets, small datasets, and notes. Large datasets are generally added to the database in bulk by members of the MaizeGDB Team, and are contributed by researchers directly. To contribute a large dataset to the project (or to find out whether the dataset you have generated constitutes a "large" or "small" dataset), use the feedback button at the top of any MaizeGDB page to make an inquiry.

Researchers can add "notes" to records. To add such a note, log in to the site using the "annotation" link displayed at the top right of any MaizeGDB page. Once logged in, click the "Add your own annotation to this record" link shown at the top of virtually all data displays. You may also add GO annotation in this way. Small datasets can be added to the database by researchers directly by way of the MaizeGDB Community Curation Tools. The method for adding a small dataset is explained below, using a newly published reference as the example usage case. The citation for our pretend reference is as follows:

Lawrence, CJ. (2005) How to use the reference curation module at MaizeGDB. *Plant Physiology* 9:3–4.

1. Click on the "annotation" link at the top of any MaizeGDB page. Click the link to "Create an Annotation Account" and fill out all information required. Be sure to check the box to become a MaizeGDB curator before clicking the submit button.

2. A confirmation email along with a Community Curation manual will be sent once the new account has been activated.

3. To begin adding data to the database, click on the link marked "tools" toward the top right of any MaizeGDB page.

4. Toward the bottom of this page click the link marked "Playground Community Curation Tools".

5. Log in using the newly created username and password.

6. Click the link toward the center of the page to download the curation tools' user manual for future reference.

7. In the left bar, click the link marked "Reference."

8. Fill in the title and select "article" as the reference type. (Because you are working at the "Playground Community

Curation Tools" feel free to make up pretend information for the purposes of this exercise.) When in doubt of what information to put into a given field, click on the buttons labeled with a question mark.

9. Fill in the year, volume, and pages information. For the "In Journal" field note the label "Lookup Field—Enter a Search String." Fill in the journal title.

10. Click the link beneath the "Author" heading to "Add Authors."

11. Half way down the page is a text field where the name of an author can be typed to locate a person record to associate with the new reference. For this example, type *Lawrence*.

12. Lawrence is the first (and only) author on this imaginary publication, so leave the dropdown menu with "Author" selected, and type the number **1** into the box labeled "Order." Press the "Submit & Continue" button.

13. Note that "Lawrence, CJ" is available in the dropdown menu. Select this item from the dropdown menu that has replaced the typing field for Author, scroll to the bottom of the page, and click the button labeled "Add to List of Authors."

14. Note that "Lawrence, CJ" now appears in the list of authors at the top of the page. Click the button marked "Author List Complete."

15. Scroll to the bottom of the page and press the button marked "Submit & Continue." Returned in place of the "In Journal" search string are available instances of matching journal names. Select the records "Plant Physiol".

16. Click the button at the bottom of the page marked "Insert into Database."

The newly created record enters a queue for approval by a worker at MaizeGDB. Once the record has been approved, it will become available through the MaizeGDB interface after the next database update. Other curation tool modules function similarly, and a detailed manual is available through the curation tools.

3.4 Leave Feedback   Personnel at MaizeGDB aim to be responsive to the questions and requests of people who use this resource. At the top of every page, on the horizontal green bar, is a feedback button. When used, it will record the page from which the feedback was initiated. Users can ask questions, make comments, and point out errors. MaizeGDB strongly encourages everyone to use the feedback button often!

3.5 Outreach   MaizeGDB provides written, video, and in-person tutorials on a broad range of topics all relating to interacting effectively with data at MaizeGDB [19, 20]. All tutorial and FAQ information can be found in the outreach section on the bottom left of the home page, Links to the MaizeGDB Facebook and Twitter pages are also there.

Acknowledgments

References

1. Lawrence CJ, Dong Q, Polacco ML, Seigfried TE, Brendel V (2004) MaizeGDB, the community database for maize genetics and genomics. Nucleic Acids Res 32:D393–D397

2. Polacco M, Coe E (1999) In: Letovsky SI (ed) Bioinformatics: databases and systems. Kluwer Academic Publishers, Norwell, MA, pp 151–162

3. Polacco M, Coe E, Fang Z, Hancock D, Sanchez-Villeda H, Schroeder S (2002) MaizeDB—a functional genomics perspective. Comp Funct Genomics 3(2):128–131

4. Dong Q, Roy L, Freeling M, Walbot V, Brendel V (2003) ZmDB, an integrated database for maize genome research. Nucleic Acids Res 31:244–247

5. Sen TZ, Andorf CM, Schaeffer ML, Harper LC, Sparks ME, Duvick J, Brendel VP, Cannon E, Campbell DA, Lawrence CJ (2009) MaizeGDB becomes 'sequence-centric'. Database. doi:10.1093/database/bap020

6. Lawrence CJ, Harper LC, Schaeffer ML, Sen TZ, Seigfried TE, Campbell DA (2008) MaizeGDB: the maize model organism database for basic, translational, and applied research. Int J Plant Genomics. doi:10.1155/2008/496957

7. Lawrence CJ, Schaeffer ML, Seigfried TE, Campbell DA, Harper LC (2007) MaizeGDB's new data types, resources and activities. Nucleic Acids Res 35(Database issue):D895–D900

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

9. Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. Genome Res 11:1425–1433

10. Bruskiewich R, Coe E, Jaiswal P, McCouch S, Polacco M, Stein L, Vincent L, Ware D (2002) The plant ontologytm consortium and plant ontologies. Comp Funct Genomics 3:137–142

11. Cannon EK, Birkett SM, Braun BL, Kodavali S, Jennewein DM, Yilmaz A, Antonescu V, Antonescu C, Harper LC, Gardiner JM, Schaeffer ML, Campbell DA, Andorf CM, Andorf D, Lisch D, Koch KE, McCarty DR, Quackenbush J, Grotewold E, Lushbough CM, Sen TZ, Lawrence CJ (2011) POPcorn: an online resource providing access to distributed and diverse maize project data. Int J Plant Genomics. doi:10.1155/2011/923035

12. Andorf CM, Lawrence CJ, Harper LC, Schaeffer ML, Campbell DA, Sen TZ (2010) The Locus Lookup tool at MaizeGDB: identification of genomic regions in maize by integrating sequence information with physical and genetic maps. Bioinformatics. doi:10.1093/bioinformatics/btp556

13. Monacoa MK, Senbc TZ, Dharmawardhanad PD, Rena L, Schaefferbe M, Naithanif S, Amarasinghed V, Thomasona J, Harperbg L, Gardinerch J, Cannonc EKS, Lawrencebc CJ, Wareab D, Jaiswal P (2013) Maize metabolic network construction and transcriptome analysis. Plant Genome. doi:10.3835/plantgenome2012.09.0025

14. Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, Amarasinghe V, Youens-Clark K, Thomason J, Preece J, Pasternak S, Olson A, Jiao Y, Lu Z, Bolser D, Kerhornou A, Staines D, Walts B, Wu G, D'Eustachio P, Haw R, Croft D, Kersey PJ, Stein L, Jaiswal P, Ware D (2013) Gramene 2013: comparative plant genomics resources. Nucleic Acids Res. doi:10.1093/nar/gkt1110

15. Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ (2007) An "electronic fluorescent pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS One 2(8), e718

16. Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37(6):914–939

17. Laurie DA, Bennett MD (1985) Nuclear DNA content in the genera *Zea* and *Sorghum*—intergenic, interspecific and intraspecific variation. Heredity 55:307–313

18. Gardiner JM, Coe EH, Melia-Hancock S, Hoisington DA, Chao S (1993) Development of a core RFLP map in maize using an immortalized F2 population. Genetics 134(3):917–930

19. Schaeffer ML, Harper LC, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Sen TZ, Lawrence CJ (2011) MaizeGDB: curation and outreach go hand-in-hand. Database (Oxford). doi:10.1093/database/bar022

20. Harper LC, Schaeffer ML, Thistle J, Gardiner JM, Andorf CM, Campbell DA, Cannon EK, Braun BL, Birkett SM, Lawrence CJ, Sen TZ (2011) The MaizeGDB Genome Browser tutorial: one example of database outreach to biologists via video. Database (Oxford). doi:10.1093/database/bar016