MaizeGDB STATUS REPORT

UPDATES, ACTIVITIES, AND NEW INITIATIVES

Prepared by: The MaizeGDB Team
Bremen Braun, Ethalinda (Ethy) Cannon, Jack Gardiner, Lisa C. Harper, John Portwood,
Mary L. Schaeffer, Taner Z. Sen, and Carson M. Andorf

FEBRUARY 29, 2016

Contact: Carson Andorf
USDA-ARS
1027 Crop Genome Informatics Laboratory
Iowa State University
Ames, IA 50011
Email: Carson.Andorf@ars.usda.gov
URL: http://www.maizegdb.org
515-294-2019

Table of Contents

1 – Meeting agenda

Monday, February 29, 2016 *Times shown as Eastern/Central/Mountain/Pacific*

3pm/2pm/1pm/12pm – **Connect to the Web** (available 30 min earlier to enable early connection for anyone who wants to try it out)

3:00 pm (Eastern) Presentation: MaizeGDB's activities, accomplishments, and plans for the future

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

4:00 pm (Eastern) Working Group Executive Session
4:45 pm (Eastern) Working Group Summarizes for the MaizeGDB Team
Meeting adjourns

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

Current Working Group Membership
Alice Barkan, Qunfeng Dong, David Jackson, Thomas Lübberstedt, Eric Lyons, Adam Phillippy (chair), Marty Sachs (*ex-officio*), Mark Settles, and Nathan Springer.

2 – Executive Summary

Executive Summary

The long-term objectives of this project are to devise, synthesize, display, and provide access to maize genetics and genomics tools and data to enable the research community to investigate basic plant biology, accelerate the pace of genetic enhancement and breeding, and translate those findings into products that increase crop quality and production. The major challenge identified in the MaizeGDB's USDA-ARS 5-year project plan (2013-2018) was storing and providing useful access to a diversity of large-scale data types including, but not limited to, multiple whole-genome sequences and associated gene expression datasets. To address these challenges, we established five major objectives:

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

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

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

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

Objective 5: Facilitate the use of genomic and genetic data, information, and tools for germplasm improvement, thus empowering ARS scientists and partners to use a new generation of computational tools and resources.

The central goal of this project is to enable the maize genetics community and relevant informatics experts to identify and meet the intellectual, technological, and bioinformatics needs that currently limit the use of maize genetics and genomics information, both for crop improvement and as a model system. This will require not only stewardship and continued improvement of the assembly and annotation of the B73 reference genome, but also management of a variety of information becoming available as a result of the sequencing revolution. Data types that must be accommodated include genetic diversity data obtained via comparative sequencing of numerous maize haplotypes as well as myriad gene expression datasets, the availability of which will enable researchers to formulate and test functional hypotheses on a scale previously intractable. Flexible tools that enable easier, standardized means to conduct genetic and breeding analyses are required to interact with and make use of these datasets. To accomplish these tasks will require interactions with, and help from, researchers within and outside of the current maize research community (see Section 5 for more information about the 5-year project plan).

3 – Personnel list and project resources

Personnel

Federal

Carson Andorf, computational biologist and lead scientist, USDA-ARS in Ames, IA
Responsible for project management, coordination with outside groups, financial planning, and defining program direction in consultation with the MaizeGDB team.

Bremen Braun, interface developer, half-time, ORISE-DOE in Minneapolis, MN
Responsible for developing the MaizeGDB MVC (Model-View-Controller) templating language, the internal caching infrastructure and database, development on the gene model and stock pages, and a set of tools for breeders.

Lisa Harper, geneticist (curator & outreach coordinator), half-time, USDA-ARS in Albany, CA
Responsible for community outreach and education, video tutorials, Phenotypic Controlled Vocabulary, Editorial Board management, and literature curation.

John Portwood, computer programmer and database administrator, USDA-ARS in Ames, IA
Responsible for the relational databases which are the backbone of the MaizeGDB project. Lead developer on several projects as well as administering and updating the MaizeGDB Genome Browser.

Mary Schaeffer, geneticist (curator), USDA-ARS in Columbia, MO
Responsible for curation involving diversity phenotype data, metabolic pathways, genes/gene models, gene function annotations, maps, etc from published literature. A member of the Maize Nomenclature Committee and a co-editor of the Maize Newsletter. Mary is an excellent source of historical information on why certain data are represented in a particular way.

Taner Sen, computational biologist, USDA-ARS in Ames, IA
Responsible for MaizeCyc, CornCyc, and other systems biology and network analysis projects. Taner leads the efforts to create tools and resources to visualize diversity and pedigree relationships at MaizeGDB.

Iowa State University

Ethalinda (Ethy) Cannon, bioinformatics engineer, half-time, ISU in Ames, IA
Responsible for design and development of the original POPcorn website, the MaizeGDB Website Redesign, genome stewardship project implementation, managing the genome and annotations pages, and communication with groups providing data.

Jack Gardiner, curator, quarter-time, ISU employee located in Columbia, MO (previously at the University of Arizona in Tucson AZ) Responsible for identifying and recruiting gene expression, proteomics, physical and genetic mapping, and epigenetic datasets with a special emphasis on large data sets. Jack will lead the curation effort on implementing MaizeMine – a maize instance of the InterMine data warehousing software.

Michael Brumfield, undergraduate interface developer, ISU in Ames, IA
Responsible for web page development and debugging.

Brittney Dunfee, undergraduate curator, ISU in Ames, IA
Genome stewardship support, literature curation, and social media.

Kyoung Tak Cho, Ph.D. graduate student, ISU in Ames, IA
Responsible for development of MaizeDIG (a tool to integrate phenotypic images with genomic context), and predictive phenomics.

David Schott, undergraduate interface developer, ISU in Ames, IA
Responsible for development of maize diversity tools.

Ashley Enger, undergraduate graphic and multimedia designer, ISU in Ames, IA
Responsible for developing visual components of the MaizeGDB interface and editing tutorial videos.

Nancy Manchaanda, Ph.D. graduate student, ISU in Ames, IA
Responsible for development of tools to evaluate and structurally annotate maize genome assemblies.

Vacant positions (Federal)

IT-Specialist, full-time permanent
IT-Specialist, full-time permanent
IT-Specialist, half-time term
Geneticist, half-time permanent
Postdoctoral fellow

Funds: ~$990,000 per year in Ames

Current:
ARS (annual)
Ames IA: $966,416 federal in Ames
Covers salaries, hardware, travel, etc. for:
Carson Andorf, Ph.D.
Taner Sen, Ph.D.
Lisa Harper, Ph.D.
John Portwood
Ethy Cannon
Bremen Braun
and ISU students.

Columbia MO (federal): covers salary, hardware, travel, etc. for Mary Schaeffer, Ph.D.

Outside Funds (NSF Plant Genome Research Program) Genetic Networks Regulating Structure and Function of the Shoot Apical Meristem Mike Scanlon PI (Cornell University). Subcontract to Carolyn Lawrence, Iowa State University $1,937,366 total, $153,597 to CoPI Lawrence, Expires January 2018
Supports Jack Gardiner, Ph.D (0.25 FTE, 4 years).

4 – Past year's activities and accomplishments

Interface design.
This past year (March 2015) marked the formal release of the new interface. The archival copy remains accessible from the home-page (bottom right corner). The multi-year effort included reorganizing existing data, upgrading hardware and infrastructure, creating new tools, incorporating new data types (including diversity data, expression data, gene models, and metabolic pathways), and developing and deploying a modern interface (see Figure 1). Figure 2 shows usage statistics before and after the interface redesign. Design changes relating to germplasm include direct access to mutant stocks from gene record pages. Previously, access was only from alleles records, lookups at the stock query pages, and the COOP catalog. Stock pages have new views for pedigree and progeny information. In addition, trait-scores are accessible, as bulk downloads (at the Diversity center), and on individual trait and stock records. Germplasm ordering forms and outgoing links have been updated for the new release of GRIN-Global.

Additional tools in progress include tools to access large-scale maize diversity data and a platform to access flexible user specified queries to the MaizeGDB database (MaizeMine). The diversity tool will allow for SNP queries based on regions in the B73 genome and get alleles from over 17,000 public lines of maize. The initial dataset will be based on the GBS2.7 data from Panzea and will include lines from the following projects: Ames lines, NAM, IBM, and Maize-BREAD. The MaizeMine tool will be based on InterMine (http://intermine.org/), a 3rd party data warehouse software package. InterMine is highly indexed database system that integrates complex biological datasets and has a user-friendly interface to allow for quick bulk queries. InterMine has a data API's for programmatic access to data. This is a very customizable system and has been adapted for many other organisms.

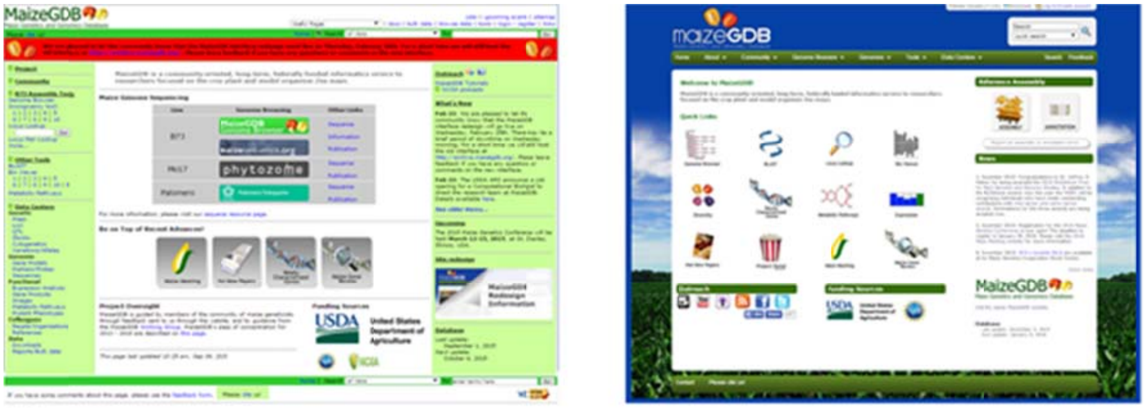


Figure 1: **The MaizeGDB web interface before and after the interface redesign**. On the left is the MaizeGDB look-and-feel from 2003 to 2014. On the right is the current web interface (released in March 2015). Webpages, tools, and data centers are organized in a menu within the header of each page.

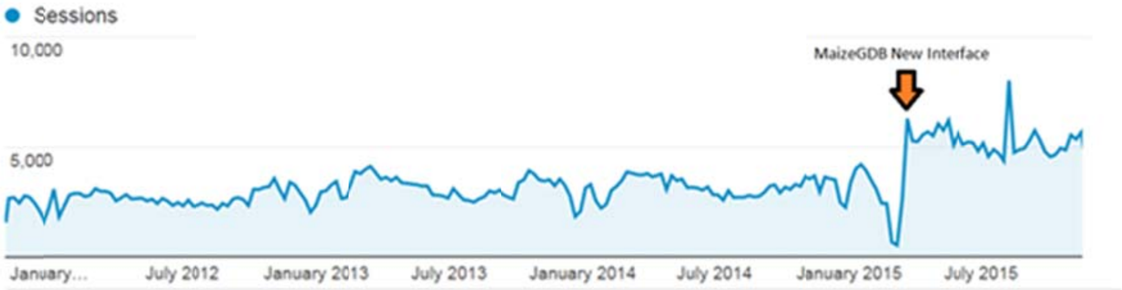


Figure 2: MaizeGDB usage from 2011-2015. The release of the MaizeGDB interface redesign is labelled.

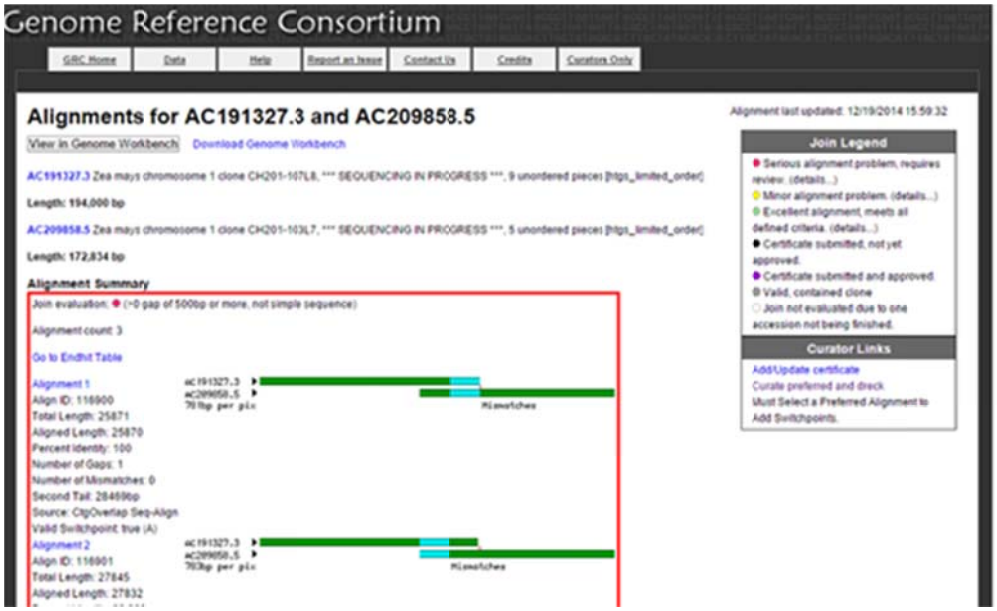
Maize genome stewardship and support for diverse maize genomes.
MaizeGDB has continued to collaborate with Doreen Ware (USDA, Gramene), Valerie Schneider (NCBI, Genome Reference Consortium (GRC)) and Kim Pruitt (NCBI, GenBank) to incorporate the B73 reference assembly into the GRC. The GRC has tools to visualize the quality of an assembly and to explore and fix assembly issues (see Figure 3). A set of standard operating procedures (SOPs) exist for handling publically submitted assembly issues, informing researchers about progress on resolving the issues, and for releasing resolved issues as patch assembly releases. An important aspect of a patch release is that it does not change assembly coordinates but does give the community access to improvements between major releases of the assembly.

The B73 RefGen_v3 assembly was successfully loaded into the GRC database, after reverse-engineering the GenBank BAC records to match the AGP file (which provides the BAC tiling path and indicates which portions of the BACs are used to create the pseudomolecules and scaffolds). An additional ~200 errors in the tiling path that were found by the GRC software were corrected. The new v4 assembly will also be loaded and the GRC tools will be used to improve the assembly and issue new patch and full releases.

A database for assembly and gene model issues reported by the community has been created at MaizeGDB. Issues are assessed by an assembly curator and resolved as possible through the GRC tools. Regions of the assembly and gene models with issues are reported on relevant MaizeGDB pages.

MaizeGDB has collected 547 assembly and annotation issues to date, from the literature and directly from the maize research community. Issues may be provided by any member of the research community through email and/or through a popup form at MaizeGDB, which can be found on the genome browser and the gene model search and record pages. These issues are loaded into a GRC-compliant issue tracker. The current set of issues will be used as test cases for the new B73 RefGen_v4 and any unresolved issues will become candidates for patch releases generated by the GRC tools.

Several additional reference-quality genomes will soon be available to the research community. MaizeGDB has worked closely with three of these projects: W22 (Tom Brutnell, Erik Vollbrecht, Hugo Dooner, Don McCarty, Charles Du), CML247 (Ed Buckler), and B104 (Kan Wang, Carolyn Lawrence-Dill, Carson Andorf). MaizeGDB is handling the GenBank submissions for each of these three reference quality assemblies. We expect that additional genome submissions will be done by the sequencing groups with support from MaizeGDB. MaizeGDB will provide a genomes page with consistent nomenclature, data downloads, quality statistics, genome browsers, BLAST tools, and general annotation pages. These pages are in progress and will be available in March 2016.



Alignments for AC191327.3 and AC209858.5

Alignment last updated: 12/19/2014 15:59:32

AC191327.3 Zea mays chromosome 1 clone CH201-107L8, *** SEQUENCING IN PROGRESS ***, 9 unordered pieces [htgs_limited_order]

Length: 194,000 bp

AC209858.5 Zea mays chromosome 1 clone CH201-103L7, *** SEQUENCING IN PROGRESS ***, 5 unordered pieces [htgs_limited_order]

Length: 172,834 bp

Alignment Summary

Join evaluation: ● (>0 gap of 500bp or more, not simple sequence)

Alignment count: 3

Alignment 1
Align ID: 116900
Total Length: 25871
Aligned Length: 25870
Percent Identity: 100
Number of Gaps: 1
Number of Mismatches: 0
Second Tail: 28469bp
Source: CtgOverlap Seq-Align
Valid Switchpoint: true (A)

Alignment 2
Align ID: 116901
Total Length: 27845
Aligned Length: 27832

Join Legend

- ● Serious alignment problem, requires review. (details...)
- ● Minor alignment problem. (details...)
- ● Excellent alignment, meets all defined criteria. (details...)
- ● Certificate submitted, not yet approved.
- ● Certificate submitted and approved.
- ● Valid, contained clone
- ○ Join not evaluated due to one accession not being finished.

Curator Links

Must Select a Preferred Alignment to Add Switchpoints.

Figure 3: GRC for alignments between AC191327.3 and AC209858.5. It shows the length of the alignment, number of gaps, and the percent identity between the BACs, these are used to determine the quality of the alignment between the BACs.

Breeder Tools and Resources to Visualize Diversity and Pedigree Relationships

In accordance to the Project Plan milestones created with the input from the maize community, the MaizeGDB Team prepared a survey draft to identify breeder needs for visualizing pedigrees, diversity data, and haplotypes. During the Maize Genetics Meeting in 2015, we sought input in person from about 10 maize researchers who are self-identified as breeders or work closely with them. Based on their feedback, we changed the wording and ordering of the questions, as well as possible answers to the questions.

The survey was then sent to the Maize Executive Committee (MGEC), and after a brief review, it was approved. The MaizeGDB team distributed it to the maize cooperators on behalf of the MGEC. The survey stayed open for 8 weeks. When we close the survey site in Fall 2015, we received 48 unique responses from maize researchers.

The survey was structured in three sections: 1) information about the survey respondent, 2) researchers' visualization needs, and 3) maize populations that need to be visualized:

1. *Respondents*. The 48 respondents show a broad distribution in their make-up: 55% were PIs and 23% were scientists/postdocs. The respondents were overwhelmingly from academia (64%), with 20% from government, and 16% from industry. 54% were self-identified as breeders, and 38% specified the focus of their research program as "breeding," and 33% as quantitative genetics, demonstrating a broad definition of breeding.
2. *Visualization needs*. The researchers established their highest priorities as: 1) SNPs in a region for a given list of lines, 2) haplotype analysis in a given list of lines, and 3) pedigree relationships.
3. *Populations*. The survey uncovered further that the following two populations are the most beneficial to visualize for researchers: 1) 3000 inbred lines from the paper of Romay et al. (Genome Biol, 14:R55, 2013), and 2) Expired PVP lines (Plant Variety Protection Act).

Driven by this strong stakeholder input, MaizeGDB is currently working in four areas:

1. Displaying immediate progenies of current stocks at the MaizeGDB Stock pages
2. Curating most recent ex-PVP lines in GRIN into the maize database and their display on the MaizeGDB Stock pages
3. Developing network views of pedigree relationships
4. Visualizing genotypes such as SNPs from diversity lines

Genome Browser.

The MaizeGDB now has four versions of the B73 reference genome (BAC-based, B73 RefGen_v1, B73 RefGen_v2, and RefGen_v3). We currently actively update and support B73 RefGen_v2 and RefGen_v3. The other instances are archived. We will be representing other genomes in the near future. We currently have development instances for W22, CML247, and B104. We anticipate having a browser for B73_RefGen_v4 (referred to above) in a few months. Listed below are new datasets and datasets available as tracks on the genome browser:

New data highlights.

- Additional DS-GFP and Uniform Mu insertion stocks added. Links to Vollbrecht insertion stocks altered to reflect availability from the COOP.
- Shoot apical meristem (SAM) trait data (best linear unbiased predictions) with accompanying longitudinal images were uploaded into MaizeGDB and connected to 10 new SAM terms and 1121 inbred lines.
- A set of B73-teostinte NILs entered into MaizeGDB, now accessible at COOP.
- Phenotype diversity data integrated last year for the NAM and IBM mapping panels, will soon include data for association panels such as the Goodman and Ames panels, along with definition of methods, environments and conditions used for traits. Standard terms, with computable accessions, are used from internationally accepted ontologies. Interfaces updated to access these from Diversity page, Stock and Trait records.
- Genetic 2008 and IBM Neighbors maps are continually manually updated by MaizeGDB, and Ed Coe.

New Genome Browser B73 RefGen_v3 Tracks:

- MAKER-P Gene Models, (Law et al. 2015)
- HapMapV3 (Bukowski 2015)
- NCBI Annotation Release 100, (NCBI)
- G4 Quadruplex Motifs (4 tracks): (Andorf et al. 2014)
- RNA-Seq Expression Atlas, Shawn Kaeppler (Stelpflug et al. 2015)
- (Private – waiting on publication) Phosphorylated Peptides from 33 Tissues, Justin Walley/Steve Briggs group
- (Private – waiting on publication) Non-modified Peptides from 33 Tissues, Justin Walley/Steve Briggs group
- Pan-genome Sequence Anchors, (Lu et al. 2015)
- Mo17 SNPs and Indels track from JGI
- Fluorescent protein tags from JCVI
- GBS v2.7 diversity data from Panzea.
- De novo transcript assemblies from JGI (Martin et al. 2014).
- Non-modified and Phosphorylated peptides (Walley et al. 2013).
- TSS transcription initiation sites, experimental, CAGE (“CAP analysis of gene expression”) Mejis-Guerra et al 2015 Plant Cell tpc.15.00630“ (in progress)
- Converted 25 tracks from v2 to v3.

New W22 Tracks (all currently private - waiting on publication):

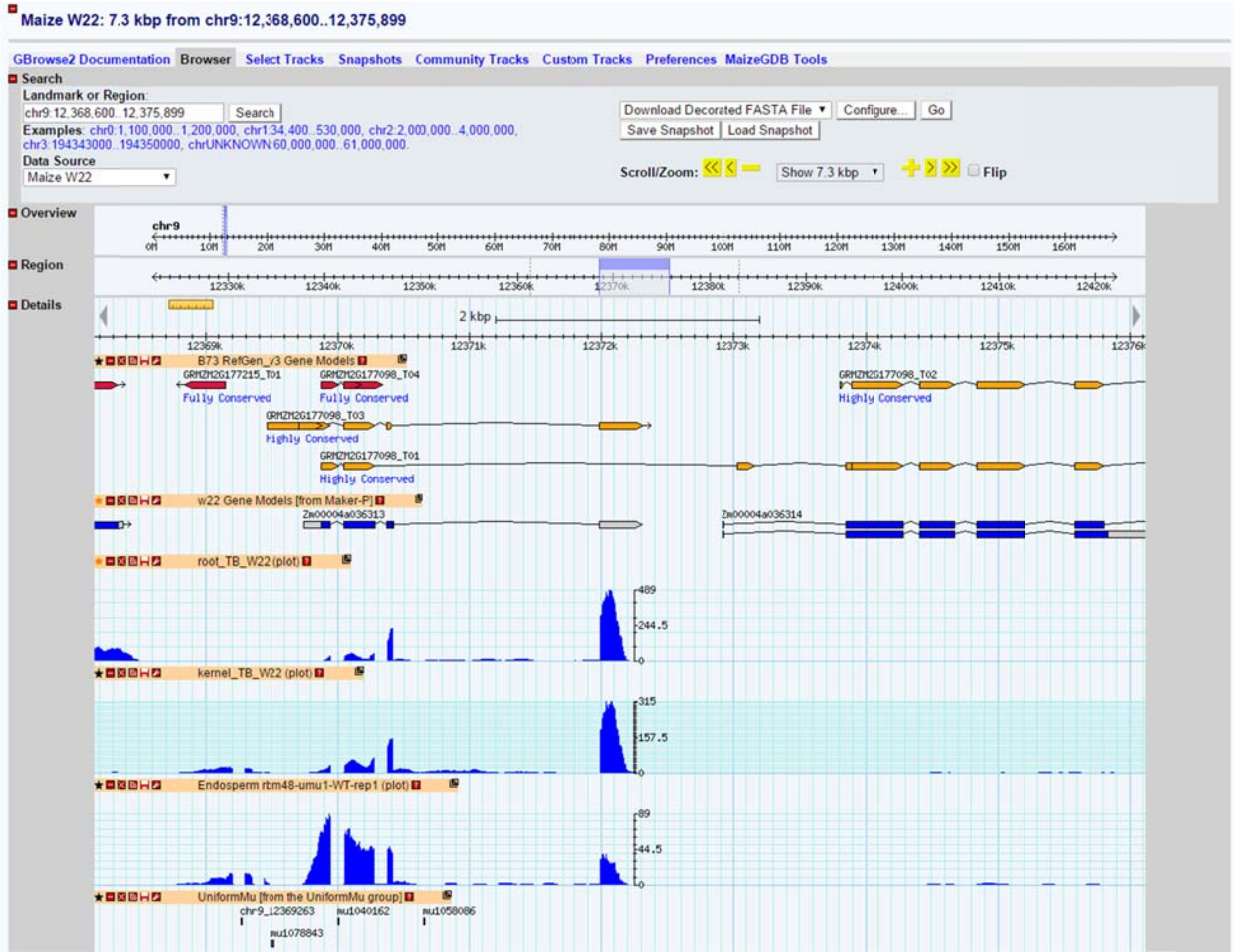
- DNA/GC Content, from W22 Group (assembly statistics)
- 6-frame translation, from W22 Group (assembly statistics)
- Bins, (Andorf)
- Core Bin Markers, (Andorf)
- Gaps, W22 Group (assembly statistics)
- B73 Maize G4v2 Motifs, (Andorf)
- W22 Maize G4v2 Motifs, (Bass and Andorf)
- W22 MAKER-P Gene Models, from W22 Group
- B73 RefGen_v3 Gene Models, (Wimalanathan and Vollbrecht)
- UniformMu Insertions, UniformMu group (McCarty and Koch)
- Ds Flanks, (AcDstagging.org)
- RNA-Seq reads: ear, kernel, shoot, root, endosperm, leaf, from W22 Group

New B104 Tracks (private):

- B104 Assembly, from B104 Group (Wang, Lawrence-Dill, Andorf)
- B104 MAKER-P Gene Models
- B73 RefGen_v3 Gene Models

New CML247 Tracks (private):

- CML247 MAKER-P Gene Models, CML247 Group (Buckler)
- B73 RefGen_v3 Gene Models, CML247 Group (Buckler)



Maize W22: 7.3 kbp from chr9:12,368,600..12,375,899

Landmark or Region: chr9:12,368,600..12,375,899

Examples: chr0:1,100,000..1,200,000, chr1:34,400..530,000, chr2:2,000,000..4,000,000, chr3:194343000..194350000, chrUNKNOWN:60,000,000..61,000,000.

Data Source: Maize W22

Scroll/Zoom: Show 7.3 kbp

Figure 4: **Snapshot of the MaizeGDB W22 Genome Browser near the gene: stc1 - sesquiterpene cyclase1**. This view shows the W22 gene model (Zm00004a036313), the mapped B73 RefGen_v3 gene model (GRMZM2G177098), RNA-Seq expression levels in root, kernel, and endosperm, and UniformMu insertion sites.

Metadata template for plant sequences

We are working with other groups, including CyVerse and the Genomics Standards Consortium to develop a plant specific metadata template to collect information about any sequence data. This template is based on the MIxS (Minimum Information about any(x) Sequence) [http://gensc.org/projects/mixs-gsc-project/] and a plant extension to MIxS that is currently under review, which is centered on the GenBank genome submission process, to ease submission to GenBank as well as collect metadata. Scripts for loading the metadata into a Chado database and pages to display the metadata are under development and expected to be available in March, 2016.

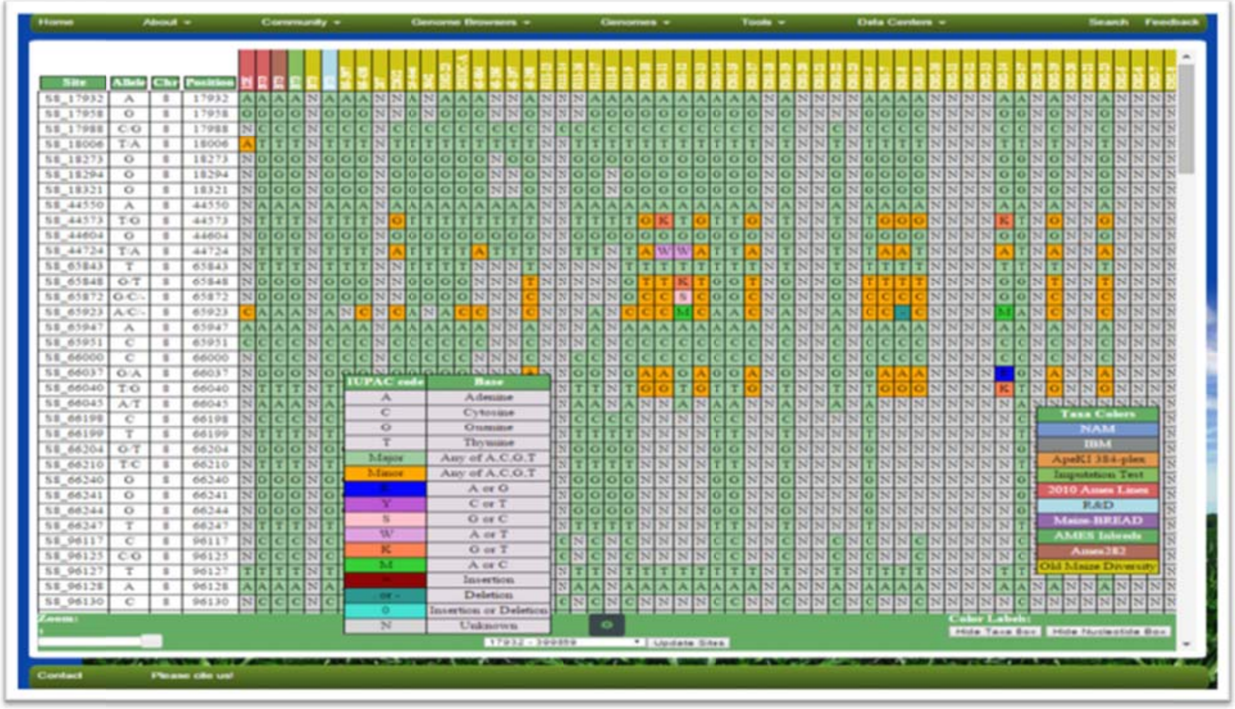
Data curation: Literature, which includes gene by gene annotation, and also large datasets

MaizeGDB literature curation has been focused on articles nominated by our six-member MaizeGDB Editorial Board. It has been expanded in this 5-year project cycle to include gene ontology (GO) annotation of experimentally documented gene functions. The GO annotation supports the BioCyc metabolism databases, and was also part of a text-mining research collaboration (BioCreative); it involves both undergraduate and senior curators. Literature was not comprehensively surveyed; instead the objective was to use literature as needed to document function.

Curation of large datasets included new genome browser tracks which involved close interaction between the curatorial and technical staff, and the data generating researchers to insure data are represented in the best way. Other large datasets included trait values from GWAS/QTL diversity data, where there are public genotypes. In these, metadata about traits (how measured, environments) was manually extracted from the literature. Other manual annotation used hierarchical dictionaries (ontologies) for phenotypes and traits. Curators have also worked to refine ontology content, and develop strategies to query data.

In-progress projects

The MaizeGDB Genotype Visualization Tool: One of the priorities set by the Breeders Tools and Resources survey we conducted in 2015 was displaying SNPs in a region for a given list of lines. To this end, the MaizeGDB Team is working towards developing a Web-based tool to help maize researchers to select a customized set of maize lines and a B73 genomic region, and visualize SNPs for that genomic region. The tool harnesses TASSEL APIs, and contains two versions of GBS data from ZeaGBSv2.7 from Panzea, raw and imputed. The raw data is 12 Gigabytes and imputed data is 5.5 Gigabytes. Both the data sets contain 955,690 SNPs at 17,280 lines in HDF5 format. The tool is currently being developed and will be made available in Summer 2016. Figure 5 shows a snapshot of the MaizeGDB Genotype Query Tool.



IUPAC code	Base
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
Major	Any of A,C,G,T
Minor	Any of A,C,G,T
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
+	Insertion
or -	Deletion
0	Insertion or Deletion
N	Unknown

Taxa Colors
NAM
IBM
ApeKI 384-plex
Imputation Test
2010 Ames Lines
RAD
Maize-BREAD
AMES Inbreds
Ames282
Old Maize Diversity

Figure 5: A snapshot of the MaizeGDB Genotype Visualization Tool. This tool provides visual displays for SNPs for thousands of taxa for a specified genomic region.

The MaizeGDB Pedigree Viewer: Another priority set by the Breeders Tools and Resources survey was to display pedigree relationships. In response to this stakeholder feedback, we are currently developing the MaizeGDB Pedigree Viewer. The Viewer is based on a pedigree network of 5487 maize lines that are currently available in the MAizeGDB Stock Pages. The viewer using cytoscape.js to display a network view (Figure 6). The tool will allow the user to apply a number of filters, select or upload their own breeding relationships, center a pedigree network on a maize line, and will display a shortest path between two maize lines on the pedigree network. The viewer is expected to be available in Summer 2016.

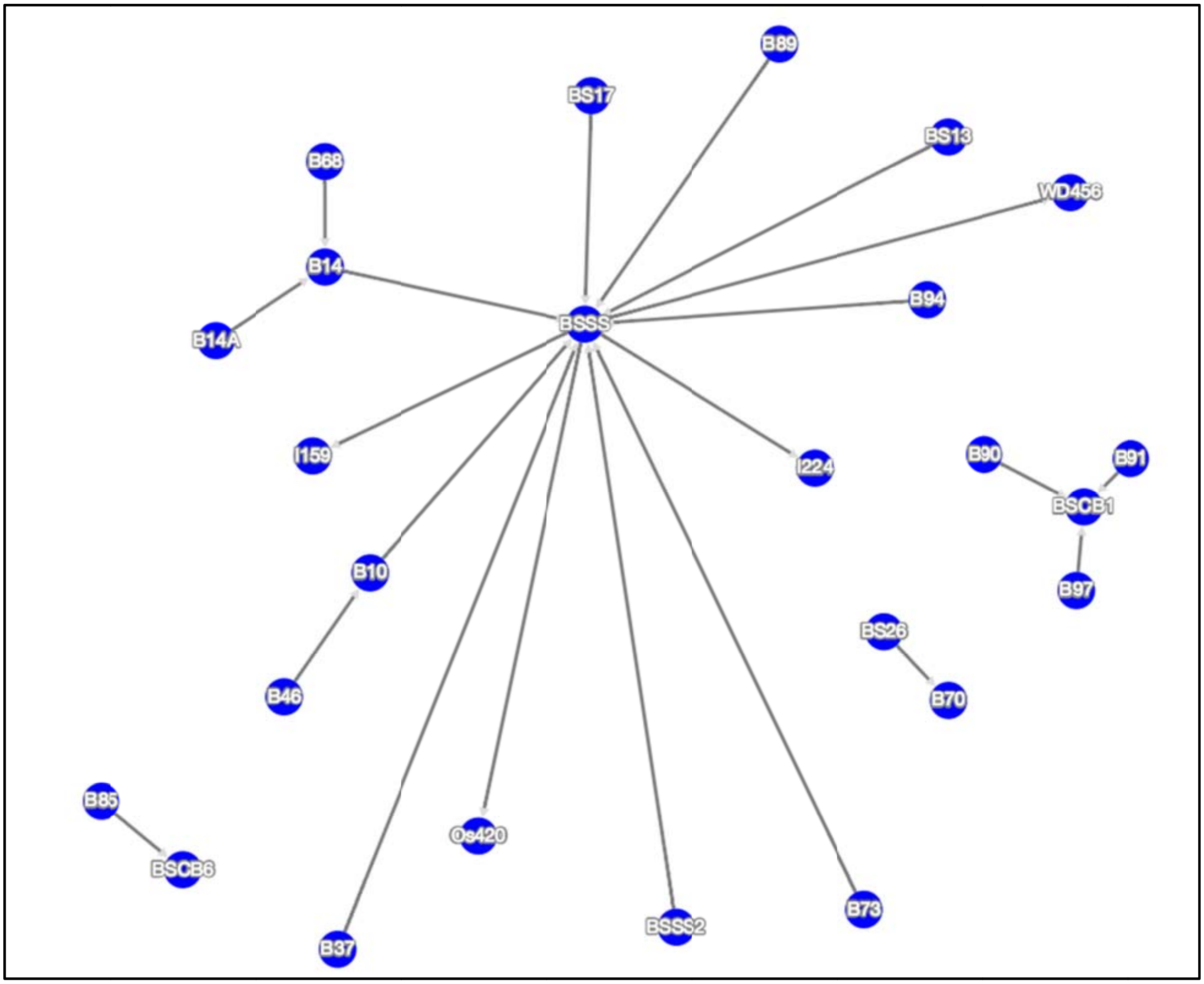
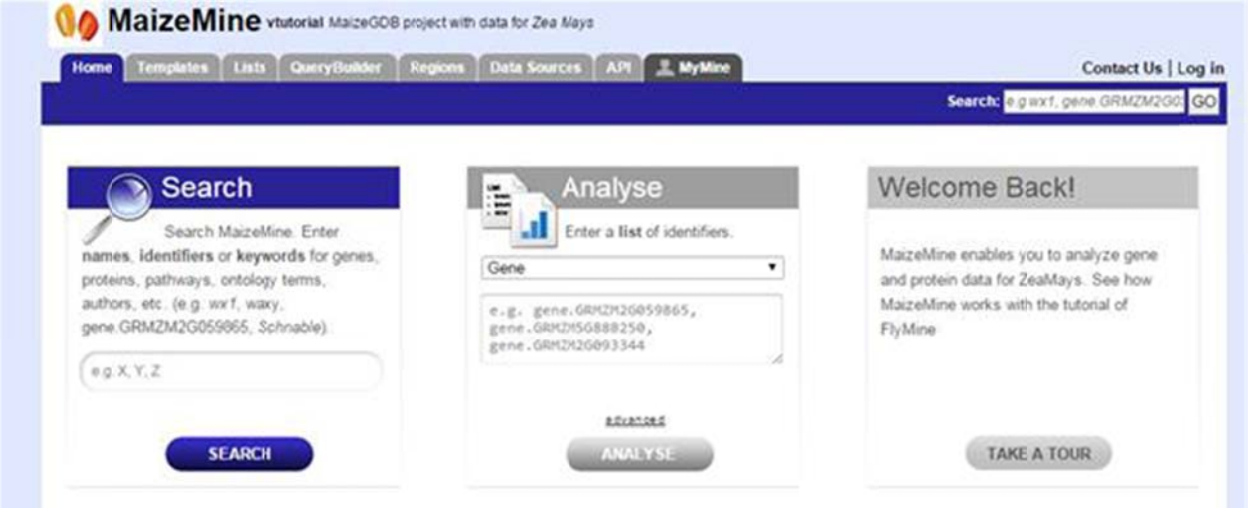


Figure 6. A snapshot of the MaizeGDB Pedigree Viewer. Maize lines originated from Iowa are shown as a pedigree network.

MaizeMine

As mentioned above, MaizeMine will be based on the 3rd party software, Intermine (http://intermine.org/). MaizeMine will address Objective 3 in the Project Plan to create a user-friendly interface that allows custom queries from complex biological data sets and deliver them in a variety of data formats (Figure 7). To achieve this objective, MaizeGDB has partnered with Dr. Chris Elsik, a computational biologist at the University of Missouri-Columbia. Chris has developed two previous instances of Intermine, Bovinemine and Hymenopteramine and as such has considerable expertise with the Intermine software package. Curator Jack Gardiner will work in the Elsik laboratory along with laboratory personnel and graduate students to develop a working instance in the coming year. The initial focus will be to get a functioning instance of MaizeMine installed and available from the MaizeGDB interface. In years two and three, Curator

Jack Gardiner will work both independently, and with MaizeGDB users, to develop custom query templates aided by computational assistance from Elsik laboratory personnel.



Search MaizeMine. Enter **names**, **identifiers** or **keywords** for genes, proteins, pathways, ontology terms, authors, etc. (e.g. wxf, waxy, gene.GRMZM2G059865, *Schnable*).

Enter a **list** of identifiers.

e.g. gene.GRMZM2G059865, gene.GRMZM5G888250, gene.GRMZM2G093344

MaizeMine enables you to analyze gene and protein data for ZeaMays. See how MaizeMine works with the tutorial of FlyMine

Figure 7. A snapshot of the homepage of MaizeMine. MaizeMine is based on InterMine, a data warehouse with a user-friendly interface that allows custom queries from complex biological data sets and deliver them in a variety of data formats.

5 – Five Year Project Plan

The USDA requires us to write a five year plan in response to their 5 year mission statement. We are in year 3 of the current 2013 to 2018 project plan.
Objective 1

Goal: Support stewardship of maize genome sequences and forthcoming diverse maize sequences.

- Goal 1.a: Enlist the community of maize researchers in the genome assembly and annotation process to enable their contributions to and use of improved reference genome sequences in real-time.
- Goal 1.b: Deliver sequence-based representations of maize diversity, both with respect to the B73 reference genome and in the absence of homologous reference sequence.

Anticipated Products: Tool suite to enable reference genome assembly improvement, documentation of diversity alongside the reference genome assembly, and contribution of structural and functional genome annotation by researchers directly.

Collaborators:
Valerie Schneider (Key Collaborator)
Doreen Ware (Key Collaborator)
Kim Pruitt (Key Collaborator)
Tom Brutnell (Key Collaborator)
Ed Buckler (Key Collaborator)
Qi Sun (Key Collaborator)
Andrew Olson
Mark Yandell
Yinping Jiao
Carolyn Lawrence
Kan Wang

Accomplishments:

- The B73 RefGen_v3 assembly was successfully loaded into the GRC database at NCBI
- Approximately 200 errors in the tiling path that were found by the GRC software were corrected.
- A database for assembly and gene model issues reported by the community has been created at MaizeGDB.
- MaizeGDB has collected 547 assembly and annotation issues to date, from the literature and directly from the maize research community.
- Genome and general gene model pages have been created for reference quality-assemblies.
- BLAST and Genome Browser tools have been created for W22, B104, and CML247.

Objective 2

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

- Goal 2a. Enable researchers to access high-quality functional descriptions for maize gene products by documenting their potential involvement in particular biochemical and metabolic pathways.
- Goal 2b. Develop and deploy network-based data access and analysis tools that support predictive biological investigations routinely pursued by basic biologists and breeders.

Collaborators

Peifen Zhang
Julie Dickerson
Ed Buckler
Marty Sachs
Thomas Lubberstedt
Candice Gardner
Mark Millard
Rick Vierling
Peter Karp

Accomplishments

- MaizeCyc 2.2 was released in collaboration with Gramene and MaizeGDB.
- MaizeCyc 2.2 and CornCyc 4.0 Biopax files were generated
- CornCyc and MaizeGDB are being archived off-site in Missouri
- MaizeCyc GO terms were migrated to CornCyc for the proteins are present in CornCyc for a total of 175 annotations.
- Survey Genomic Tools for Breeders:
    - Draft Survey prepared
    - Feedback received from (among others):
        - MaizeGDB personnel
        - Paul Scott
        - Jode Edwards
        - Sherry Flint-Garcia
        - Ed Buckler
        - Bill Tracy
        - Richard Vierling
        - Mark Millard
        - Thomas Lubberstedt
        - Stephanie Coffman

Objective 3

Goal: Allow researchers access to larger sets of data

- Data: Generating sets of data
- Analysis: Extracting information based on a set of data

Collaborators

- Christine Elsik
- Jack Gardiner

Accomplishments

- A Beta-version of Maize Mine has been created.
- A collaboration has been established to create and annotate a production version of MaizeMine.
- Search tools at the Gene Model data center has been updated to improve access to sequence and functional annotations.
- MaizeGDB search tools have been improved for quicker and more in-depth querying of the MaizeGDB database.

Objective 4

Goal: Provide community support services, training and documentation, meeting coordination, and support for community elections and surveys.

Collaborators

- Maize Genetics Executive Committee (MGEC)
- Maize Genetics Conference Steering Committee (MGCSC)
- Maize cooperators

Accomplishments

- Provided support for the Annual Maize Genetics Conference (website, abstract book, workshops, technical support)
- Conducted annual MGEC elections
- Conducted maize community surveys (2014 and 2016)
- Managed maize cooperators email list, job board, and community calendar

Objective 5

Goal: Facilitate the use of genomic and genetic data, information, and tools for germplasm improvement, thus empowering ARS scientists and partners to use a new generation of computational tools and resources.

Collaborators

- Ramona Walls
- Panzea
- Gramene
- AgBioData steering committee

Accomplishments

- Initiation and guidance of the AgBioData working group, an International group of databases that manage agricultural data.
- Metadata for sequences: a plant extension for MIxS (Minimum Information about any(x) Sequence) presented to the Genomic Standards Consortium for review
- Hired 9 students over the past 2 years to work on a variety of projects including the following: maize diversity tools, beta-version of MaizeMine, MaizeDig (tool to link images to genomic regions), interface development, debugging the MaizeGDB website, resolving errors in the B73 tiling path, and literature curation.

6 – Presentations/Outreach

- Organized the AgBioDatabase outreach booth at Plant and Animal Genome Conference (PAG 2015-2016)
- Ethy Cannon presented at workshop at American Society of Plant Biologists (ASPB 2015)
- Lisa Harper presented on comparative phenomics at PAG, 2015
- Carson Andorf presented at the Phenotypic prediction: image acquisition and analysis workshop
- Posters at Maize Meeting, Plant and Animal Genome Conference, and Biocreative V
- Annual reports to Corn Germplasm Committee (Chicago)
- Mary Schaeffer presented at PAG Plant Metabolic Network Resources and Applications Workshop(2014), and the University of Missouri IPG 2014 Symposium Plant Protein Phosphorylation (talk and tutorial on use of CornCyc/PlantCyc/MetaCyc)
- Organized Plant Phenotype workshop at PAG
- Organized and lead AgBioDatabase Meeting at PAG, with over 25 databases participating (2015 and 2016)

7 – Peer-reviewed journal articles since January 2014

1. Andorf, CM, Cannon, EK, Portwood, JL, Gardiner, JM, Harper, LC, Schaeffer, ML, Braun, BL, Campbell, DA, Vinnakota, AG, Sribalusu, VV, Huerta, M, Cho, KT, Wimalanathan, K, Richter, JD, Mauch, ED, Rao, BS, Birkett, SM, Richter, JD, Sen, TZ, Lawrence, CJ. (2015) MaizeGDB 2015: New tools, data, and interface for the maize model organism database. Nucleic Acids Research doi: 10.1093/nar/gkv1007.
2. Harper, LC, Gardiner, JM, Andorf, CM, Lawrence, CJ. (2016) MaizeGDB: The Maize Genetics and Genomics Database. Plant Bioinformatics: Methods and Protocols. pp. 187-202.
3. Law M, Childs KL, Campbell MS, Stein JC, Holt C, Panchy N, Lei J, Achawanantakun R, Jiao D, Andorf CM, Lawrence CJ, Ware D, Shiu S, Sun Y, Jiang N, Yandell M. (2014) Automated update, revision and quality control of the Zea mays genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. Plant Physiol. 2015 Jan;167(1):25-39. doi: 10.1104/pp.114.245027. Epub 2014 Nov 10.
4. Oellrich A, Walls RL, Cannon EK, Cannon SB, Cooper L, Gardiner J, Gkoutos GV, Harper L, He M, Hoehndorf R, Jaiswal P, Kalberer SR, Lloyd JP, Meinke D, Menda N, Moore L, Nelson RT, Pujar A, Lawrence CJ, Huala E. 2015. An ontology approach to comparative phenomics in plants. Plant Methods 25;11:10.

5. Van Auken K, Schaeffer ML, McQuilton P, Laulederkind SJF, Li D, Wang S-W, Hayman GT,Tweedie S, Arighi C, DOne J, Muller HM, Sternberg PW, Mao Y, Wei CCH, and Lu Z. BC4GO: a full-text corpus for the BioCreative IV GO task. Database (Oxford). 2014. 2014. pii: bau074. doi: 10.1093/database/bau074
6. Walsh JR, Sen TZ, Dickerson JA. A computational platform to maintain and migrate manual functional annotations for BioCyc databases. BMC Syst. Biol. 8:115, 2014.
7. Andorf CM, Kopylov MS, Dobbs D, Koch KE, Stroupe ME, Lawrence CJ, and Bass HW. G-quadruplex motifs in maize (Zea mays L.) reside at specific sites in thousands of genes coupled to energy stress pathways, including hypoxia, low sugar, and nutrient deprivation. J Genet Genomics. 2014. Dec 20;41(12):627-47. doi: 10.1016/j.jgg.2014.10.004. Epub 2014 Nov 4.

8 – Charge to Working Group

The primary role of the Working Group is to help the MaizeGDB Team with strategic planning. This is especially important now, as we hope to hire or re-train personnel, and as we start preparing for the next USDA-ARS 5-year plan next year. We would greatly appreciate both your comments on what we have presented today, and your guidance on the following four charges. We would appreciate your written report by March 31, 2016. Thank you for your ongoing guidance and support!

Charge 1: Genome Assembly Stewardship. Below is a list of question. If you agree that these items are important, could you please provide a priority or ranking to help us set priorities?

1. Should we devote curation effort into patch and new assembly releases for B73? We now have a tool to collect assembly errors. We plan to continue to collect them, and show them, unedited, on a browser track. Should we spend effort to vet these and create patch assembly releases?
2. Should we continue to collect new assemblies and related metadata?
3. If so, to what extent should we integrate the data into MaizeGDB; eg, make gene models pages for every set of genes? Provide a genome browser? Provide a Cyc view?
4. What should MaizeGDB's role be in encouraging researchers to submit genome assemblies to GenBank? We plan to require that genomes served at MaizeGDB be submitted to GenBank and are developing a metadata collection template to ease the submission process. The template will be made freely available to researchers at MaizeGDB.
5. Should we map any data from B73 RefGen_v2/v3 to the v4 assembly? B73 RefGen_v2 had 58 tracks of data. Twenty-six of those were computationally mapped onto v3, while only 11 tracks were mapped directly to v3. Tracks are listed here http://www.maizegdb.org/gbrowse under the "select track" tab.

Charge 2: Big Data set identification, evaluation and incorporation. Below is a list of question. Can you offer advice and suggestions?

1. How do we triage Big Data sets reported in the literature? It is becoming difficult to decide which large data sets will be useful to the maize community. This includes data sets that could be accessible by a genome browser, incorporated into MaizeMine, or by special genotype-phenotype queries or other tools. Can you suggest new ways to evaluate large datasets? Ways to get community help to recommend data?

2. Our Project Plan Objective 2 is to incorporate experimentally confirmed functional genomic annotation including: GO annotation, phenotype/trait annotation, quantitative trait values, and Metabolic Pathway data. This requires extensive manual literature curation. Can you suggest better, faster, less labor intensive ways to accomplish this objective? Possibilities include using the Editorial board, automated literature annotation, work with journals to get authors to submit pre-publication, develop and send templates to authors post-publication, and more.
3. How valuable is comprehensive integration of public QTL and GWAS data at MaizeGDB? Currently we integrate into MaizeGDB a subset of available trait scores with metadata, but do not add researcher-defined QTL loci, either defined by a SNP or more loosely by a genetic region.

Charge 3: Tool Development. Below is a list of question. If you think these items are important, could you please provide a priority or ranking to help us make priorities:

1. Should we improve existing tools to visualize and access large-scale diversity data (genotype, SNP and GBS data) such as haplotype viewers?
2. Should we transition to the JBrowse genome browser to handle larger datasets?
3. Should we improve query tools that use the hierarchical nature of ontologies with regard to phenotypes, mutants and genes? For example, this tool would allow searches for mutant phenotype with parent term “leaf” to return phenotypes from all child terms (ligule, sheath, blade, margins, etc) as well as whole leaf phenotypes.
4. Should we update curation tools in collaboration with the Maize Genetics Stock Center? This would allow easier, faster manual literature curation.
5. Should we find ways to better integrate information about Mu and AC tagged sequences into all areas of MaizeGDB? For example, add available tagged alleles to gene model pages?

Charge 4: Future needs and expectations. Below is a list of expected needs in the next 5 years. Can you identify more, and can you comment on the importance of each of these?

1. Identifying and storing Genomes to Fields (G2F) type data (linking genotype and environment to phenotypes)
2. Finding a path to a pan genome infrastructure
3. Using large-scale semi-automated literature curation, such as annotation by publisher and authors check as part of proof-reading, etc.
4. Multiple maize genome comparison and display

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

1. When a user submits feedback they currently do not receive an automated confirmation that their ticket was accepted into our system.
2. Structure of Gene and certain data center pages can be revised to make key information more accessible.
3. Some of our tools that were originally developed for v2 do not have an equivalent version for v3/v4 (e.g. Incongruency and Locus Lookup).
4. Many of our search tools are not consistent in options and result formats
5. We have implemented a issue tracking system to collect internal and external feature requests, questions, tasks, and interface bugs. Over the last 2 years, we have resolved 768 issues. About 10 per month come from outside users and we give these the highest priority. Typically we resolve more issues per month than are created to make progress on the backlog (see Figure 8).

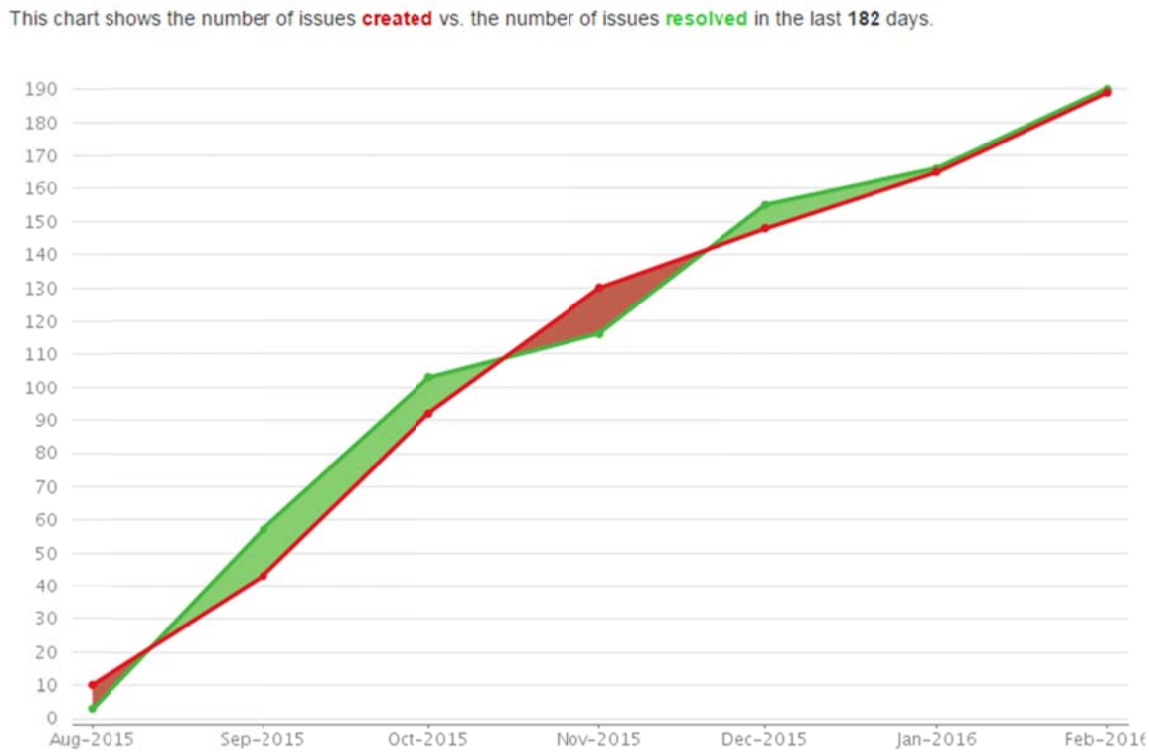


Figure 8: New versus resolved issues in the last 6 months. In the past six months, MaizeGDB has resolved a total of 190 issues and 189 new issues were created. Where the green line is above the red indicates progress against the backlog of remaining issues at MaizeGDB.