USDA United States Department of Agriculture

maizeGDB
MAIZE GENETICS AND GENOMICS DATABASE

December 2015:

Genome Assembly Stewardship.
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 fix and assembly issues (see Figure 1). It also has a set of standard operating procedures (SOPs) to record publically submitted assembly issues, inform researchers about progress on resolving the issues, and to release the 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. Current status: The B73 RefGen_v3 has been submitted to the GRC. Initially there were 92 critical errors that needed to be resolved to make the maize assembly compatible with the GRC data model. Ethy Cannon (MaizeGDB) resolved the critical errors, and adapted the GRC SOPs for use with the maize genome. The next maize assembly (B73 RefGen_v4) will also be submitted to the GRC tools (anticipated Jan. 2016) and this version will be used moving forward.

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 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). Each of these reference quality assemblies will be submitted to GenBank. 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.

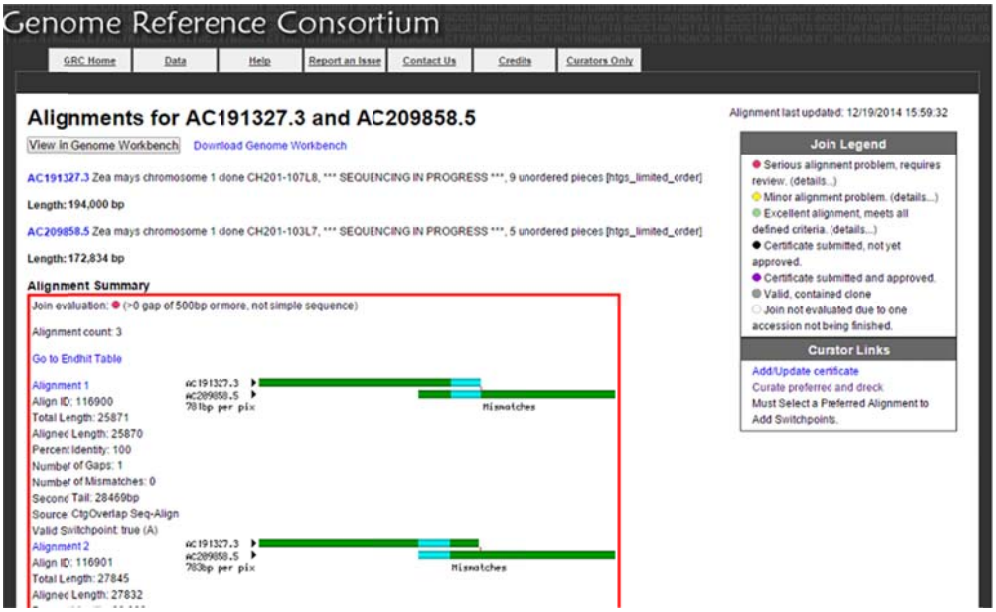

Figure 1: GRC for alignments between AC191327.3 and AC209858.5. It shows the length of the alignment, number of gaps, and the percent identity between the BACs, these are used to determine the quality of the alignment between the BACs.

Interface design.
This year 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 2). Figure 3 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.

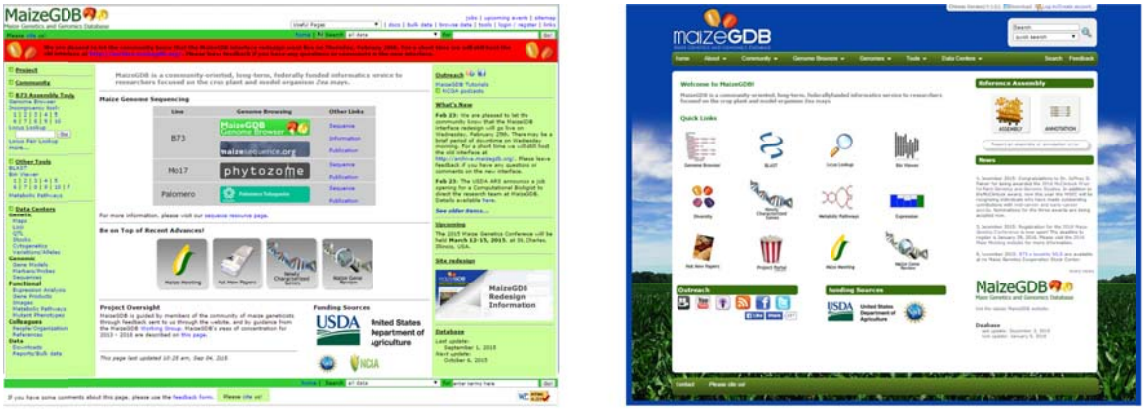

Figure 2: The MaizeGDB web interface before and after the interface redesign. On the left is the MaizeGDB look-and-feel from 2003 to 2014. On the right is the current web interface (released in March 2015). Webpages, tools, and data centers are organized in a menu within the header of each page.

Figure 3: MaizeGDB usage from 2011-2015. The release of the MaizeGDB interface redesign is labelled.

Breeder Tools and Resources to Visualize Diversity and Pedigree Relationships

The MaizeGDB Team prepared a survey to identify breeder needs for visualizing pedigrees, diversity data, and haplotypes, and distributed it to the maize community on behalf of the Maize Genetics Executive Committee (Summer 2015). We received 48 responses from researchers, of which more than half were self-identified as breeders. The researchers established their top priorities for visualization 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. 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 in part by this strong stakeholder input, MaizeGDB are 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, and 4) Visualizing genotypes from diversity datasets.

Genome Browser.

The MaizeGDB now has four versions of the B73 reference genome (BAC-based and B73 RefGen_v1 – 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 to 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.
- Genetic 2008 and IBM Neighbors maps are continually manually updated by MaizeGDB, and Ed Coe.

Genome Browser tracks.

New 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)

There are also over 20 new tracks (currently private) related to the W22, B104, and CML247 Genome Browsers.

Staff update.

Carson Andorf is now a Computation Biologist and the new lead scientist for MaizeGDB (filling the vacant vice-Lawrence-Dill position). John Portwood is a full-time scientific programmer and database administrator after serving over two years as a student programmer. New students hired through USDA 'Big-Data' funds include Brittney Dunfee (Genome stewardship, curator, social media), Mike Brumfield (interface development), David Schott (maize diversity tools), Ashley Enger (graphic design, multimedia), and Kyoung Tak Cho (predictive phenomics, PheWAS). Existing staff includes Taner Sen (Computational Biologist), Mary Schaeffer (Curator), Lisa Harper (Curator), Jack Gardiner (Curator), Ethy Cannon (Bioinformatics Engineer), and Bremen Braun (Interface Developer). In addition MaizeGDB has 5 vacant positions waiting to be filled. We are currently taking applications for a postdoctoral fellowship. Three positions should be advertised in the next few months: vice-Campbell (IT-specialist), software developer, and curator. An additional full-time IT-specialist should be filled later next year.

Outreach.
Tutorials were provided by Lisa Harper at the 57th annual Maize Genetics Conference hosted in St. Charles, IL. Currently Jack Gardiner and the NCGA have produced 8 podcasts on various topics about MaizeGDB and maize genetics. For more information see the home page at MaizeGDB. MaizeGDB continues to provide support for the Maize Genetics Conference (preparation of abstract booklet and program), and the Maize Genetics Executive Committee.

Publications.

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.

Harper, LC, Gardiner, JM, Andorf, CM, Lawrence, CJ. (2015) MaizeGDB: The Maize Genetics and Genomics Database. Plant Bioinformatics: Methods and Protocols.

Law M, Childs, KL, Campbell, MS, Stein, JC, Olson, AJ, Holt, C, Panchy, N, Lei, J, Jiao, D, Andorf, CM, Lawrence, CJ, Ware, D, Shiu, S, Sun, Y, Jiang, N, Yandell, M. (2015) Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. Plant physiology.

Acknowledgements.

Guidance is generously provided by the MaizeGDB Working group: A. Phillippy (Chair), A Barkan, Q Dong, D Jackson, T Lubberstedt, E Lyons, M Sachs (*ex officio*), M Settles, and N Springer; the Maize Genetics Executive Committee: N Springer (Chair), J Birchler, A Charcosset, P Chomet, S Flint-Garcia, S Hake, E Hiatt, S Kaeppler, K Newton, J Yu, R Sawers, P Schnable, and M Timmermans; the Maize Nomenclature Committee: M Sachs (Chair), T Brutnell, H Dooner, C Du, T Kellogg, M Schaeffer and P Stinard; and the MaizeGDB Editorial Board(2015). C Rasmussen, C Wang, M Gore, R Battaglia, W Song, M Facette.

We thank the USDA-ARS, the NSF, and the NCGA for funding.

Prepared by Mary Schaeffer, Carson Andorf and the MaizeGDB team.