December 2016: Report for the Corn Germplasm Committee
Prepared by Jack Gardiner, Carson Andorf, and the MaizeGDB team

GENOME ASSEMBLY STEWARDSHIP

A new B73 reference assembly-B73 RefGen_v4

In September, a new B73 Reference Assembly was released by the Ware laboratory (USDA-ARS) at Cold Spring Harbor and is now available through the MaizeGDB genome browser. This assembly, B73 RefGen_v4, is a de novo assembly and as such does not rely on prior assemblies v1-v3 that were based on the BAC physical map. This assembly utilized Pacific Biosciences Single Molecule Real Time (PacBio SMRT) sequencing technology. This represents a major leap forward in whole genome sequencing technology which will impact MaizeGDB's role in genome assembly stewardship. The B73 RefGen_v4 assembly increased DNA contig length 52-fold over v3. This significantly improved the organization intergenic spaces and centromeric regions and was done at a small fraction of the cost it took to produce prior maize genome assemblies. Because of this, we anticipate that submission of whole genome assemblies to MaizeGDB will increase dramatically over the next few years.

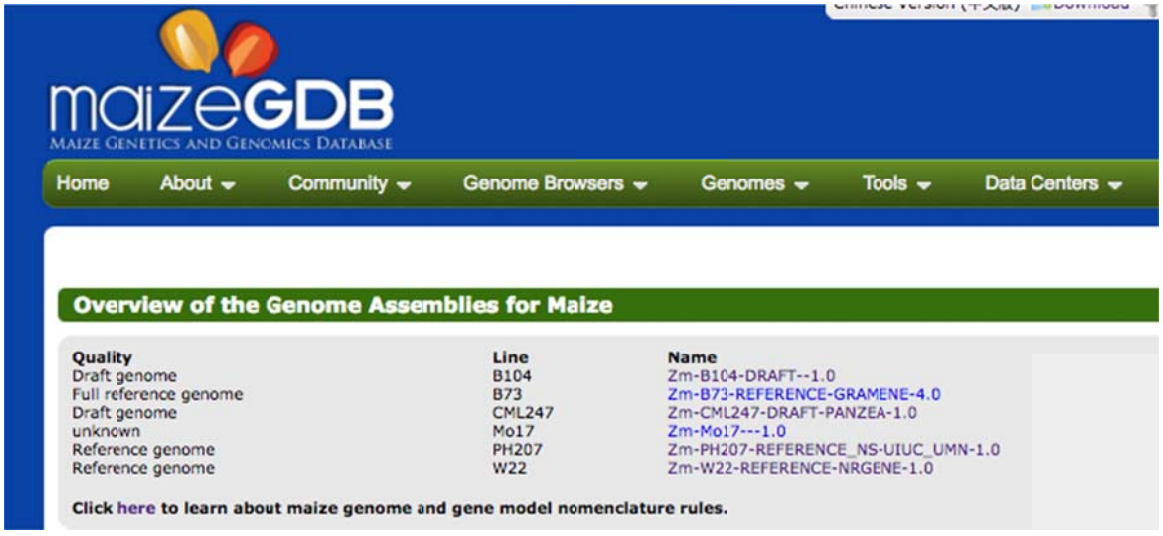

Figure 1: Whole genome assemblies submitted to MaizeGDB. Five whole genome assemblies are in the process of being submitted to MaizeGDB. Each submitted genome assembly will have its own genome browser and genome assembly page that will document standardized genome assembly information required by MaizeGDB.

MaizeGDB is preparing to display five additional genome assemblies

MaizeGDB is working with five additional groups to display their whole genome assemblies (Figure 1, http://www.maizegdb.org/genome/assemblies_overview). In addition, we have been contacted by several other groups who are in the early stages of developing additional whole genome assemblies. To ensure that all genome assemblies displayed at MaizeGDB can be fully leveraged for a wide variety of downstream data analysis applications, MaizeGDB has developed and is requiring standardized information for all whole genome assemblies. These required data standards, also known as metadata, allow full data recycling that gives life to the data beyond the original project. An example of the data collected for the B73 RefGen_v4 assembly can viewed here: http://www.maizegdb.org/genome/genome_assembly/Zm-B73-REFERENCE-GRAMENE-4.0 (See Figure 2). To accomplish this, MaizeGDB staff work closely with data generators to collect fully detailed genome assembly metadata that if often omitted from publications thereby rendering it essentially lost.

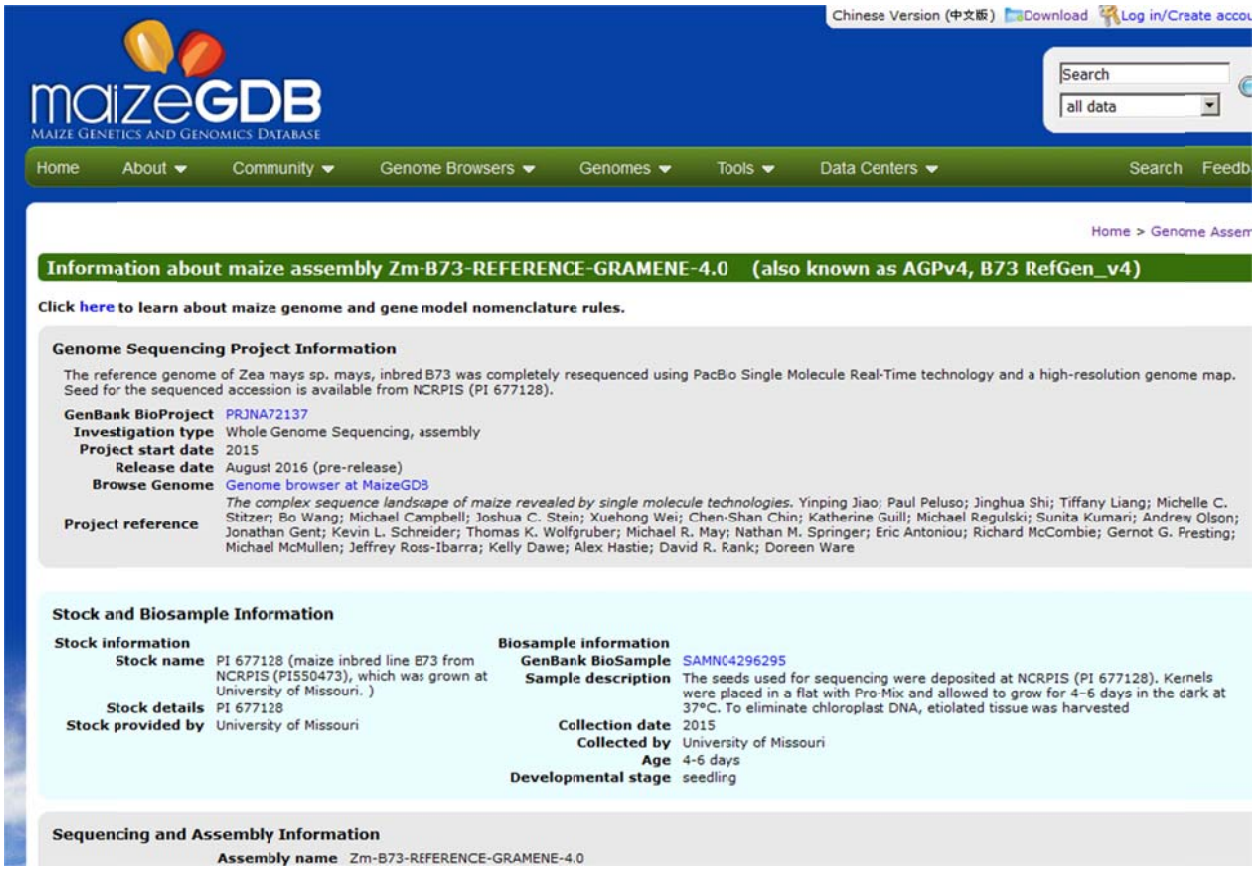

Figure 2: B73 RefGen_v4 genome page with metadata.

MaizeGDB has updated genome nomenclature rules

The addition of numerous whole genome assemblies and their associated gene models has necessitated a rethinking of the current nomenclature rules that are in effect for *Zea mays* and closely associated *Zea* species. It is important to get any new nomenclature rules in place sooner rather than later as non-standardized names create confusion in both the community and the literature. Indeed, the Rice and Arabidopsis have already addressed this issue by updating their nomenclature rules. MaizeGDB has expanded the current nomenclature rules (http://documents.maizegdb.org/nomenclature/maize_assembly_nomenclature_2016_update.pdf) and in doing so, has tried to take into account what has worked well for other model organism databases. The expanded nomenclature rules have been submitted to the maize nomenclature

committee (http://www.maizegdb.org/nomenclature#COMMITTEE) for review. These new rules attempt to take into account the many issues and complications that arise when hundreds of new whole genome assemblies and their associated gene models become available.

Development of SNP and pedigree viewers at MaizeGDB

In 2015, MaizeGDB prepared and distributed a survey to the maize community to identify visualization tools that would be of most use to the maize breeding community. The survey identified visualization of SNPs in a genomic region for a set of inbred lines, as one of the top priorities. The survey also identified two large populations as being of the most interest: 1) Ex Plant Variety Protection Act (PVP) lines and 2) 3000 inbred lines surveyed in the Romay et al. paper (Genome Biol, 14:R55, 2013). In the past year, staff at MaizeGDB have developed and released a SNP viewer (Fig 3, http://www.maizegdb.org/snpversity) which allows SNPs within a gene model or genomic interval to be visualized and downloaded. Development of the germplasm pedigree viewer is close to completion and is targeted for release early in 2017. A prerelease version can be viewed here: http://rho.maizegdb.org/breeders_toolbox?state=Iowa

Figure 3: The MaizeGDB SNP viewer. On the left is the options menu for gene model/genome interval and inbred line selection. Tool users can select from a variety of options to customize their search. On the right, SNP positions and respective alleles are visualized with links to relevant gene model pages.

Data highlights for 2016

- CornCyc 7.0 was released at MaizeGDB after being developed by the Plant Metabolic Network in collaboration with MaizeGDB.
- The maize proteomic expression atlas (Walley et al. 2016) is now fully available at MaizeGDB as both genome browser tracks and on the individual gene model pages.
- The MaizeGDB site and gene model pages have been reworked to accommodate multiple genome assemblies and gene model sets.
- A pre-release version of the maize B104 genome assembly is available that includes a genome browser and BLAST targets.
- A version of the maize PH207 genome assembly was released that includes a genome browser and BLAST targets.
- The W22 whole genome assembly was submitted to GenBank and will be available on the MaizeGDB genome browser soon.
- The integration of public germplasm that is being genotyped and phenotype into MaizeGDB, including Chinese NAM, and the AMEs panels, and Maize x teosinte NILs.

MaizeGDB outreach activities

Tutorials were provided to ~ 40 participants by outreach curator Lisa Harper at the 58 Annual Maize Genetics Conference in Jacksonville FL. This year, the tutorial took the form of a short introductory talk on data types and tools available at MaizeGDB. This was followed by problem sets, and working together as a group using tools and data resources available at MaizeGDB. In addition, MaizeGDB staff organized another workshop at the Maize Meeting with five groups giving presentations on their efforts to develop reference quality whole genome assemblies. Over 120 people attended this workshop. Lisa Harper also organizes the AgBioData working group that holds monthly conference calls to discuss challenges and solutions common to agricultural databases. The calls start with a presentation given by one the members followed by a group discussion. This October, Lisa was a course instructor at the Cold Spring Harbor workshop on Cereal Genomics where she presented several talks on available database resources and how to best utilize them. MaizeGDB, in collaboration with Gramene (www.gramene.org), also organizes the Agricultural Database booth at the Plant and Animal Genome (PAG) Meeting in San Diego where personnel from the various agricultural databases can present informational materials to PAG attendees and give hands on, in person tutorials. In 2016, over 20 databases participated in the Agricultural Database booth.

Staffing changes at MaizeGDB.

Carson Andorf continues to serve as the lead scientist for MaizeGDB. Curator Mary Schaeffer retired in April 2016 after more than 20 years of service to MaizeGDB. She now works part time (0.2 FTE) where she advises MaizeGDB on curation of QTLs, GWAS, and maize germplasm stocks. Computational biologist Taner Sen departed MaizeGDB to assume the role of lead scientist at Grain Genes in Albany, CA. Interface developer Bremen Braun has taken a position in private industry. Curator Jack Gardiner is now full time at MaizeGDB after transitioning from Carolyn Lawrence's group at Iowa State where he worked as a curator for the GXE project. He still works closely with the GXE group as their data is destined for MaizeGDB. He also works closely with Chris Elsik at the University of Missouri to develop an instance of MaizeMine. Recent Iowa State graduate Jesse Walsh has been recruited to fill the USDA postdoc position at MaizeGDB where he works on bioinformatics tool development. He currently is working on updating the metabolic pathway viewer to reflect the new B73 v4 assembly. Maggie Woodhouse has been recruited as the new MaizeGDB sequence curator to assist groups in submitting their genome assemblies to MaizeGDB. As noted above, Curator Lisa Harper manages outreach activities as well the Ag Bio Database group. Bioinformatics engineer Ethy Cannon continues to make improvements to the new MaizeGDB interface and is instrumental in developing the nomenclature guidelines and metadata standards. John Portwood continues as a full-time scientific programmer and database administrator and fulfills many of the responsibilities' of the vacant vice-Campbell position that is now under active recruitment. We hope to identify candidates for the vice-Campbell position by the end of the year. In addition, MaizeGDB has two additional vacant positions we hope to have advertised in the next few weeks: vice-Andorf (software developer) and vice-Sen (computational biologist). Current students hired through USDA 'Big-Data' funds include David Schott (maize diversity tools) and Kyoung Tak Cho (predictive phenomics, PheWAS).

Publications for 2016

Walsh, J, Schaeffer, M, Zhang, P, Rhee, S, Dickerson, J, Sen, T (2016) The Quality of Metabolic Pathway Resources Depends on Initial Enzymatic Function Assignments: A Case for Maize BMC Systems Biology 10:129 doi 10.1186/s12918-016-0369-0639-x

Sen, T, Braun, B, Schott, D, Portwood, J, Schaeffer, M, Harper, L, Gardiner, J, Cannon, E, Andorf, C (2016) Surveying the Maize Community for their Diversity and Pedigree Visualization Needs to Prioritize Tool Development and Curation. Submitted to: Database 10/26/16

Jones D, Zheng W, Huang S, Du C, Zhao X, Yennamalli R, Sen T, Nettleton D, Wurtele E, Li L. A Clade-Specific Arabidopsis Gene Connects Primary Metabolism and Senescence. Front Plant Sci. 2016 Jul 12;7:983. doi: 10.3389/fpls.2016.00983. eCollection 2016.

Andorf C, Cannon E, Portwood J, Gardiner J, Harper L, Schaeffer M, Braun B, Campbell D, Vinnakota A, Sribalusu V, Huerta M, Cho K, Wimalanathan K, Richter J, Mauch E, Rao B, Birkett S, Sen T, Lawrence-Dill C. MaizeGDB update: new tools, data and interface for the maize model organism database. Nucleic Acids Res. 2016 Jan 4;44(D1):D1195-201. doi: 10.1093/nar/gkv1007. Epub 2015 Oct 1.

Harper L, Gardiner J, Andorf C, Lawrence C. MaizeGDB: The Maize Genetics and Genomics Database. Methods Mol Biol. 2016;1374:187-202. doi: 10.1007/978-1-4939-3167-5_9.

Acknowledgements.

Maize GDB would like to acknowledge the guidance 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: J Birchler (Chair), P Chomet, N de Leon, S Flint-Garcia, S Hake, E Hiatt, S Kaeppler, K Koch, 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, and P Stinard; and the MaizeGDB Editorial Board (2016). C Rasmussen, A Ronceret, Y He, E Rodgers-Melnick, A L Eveland, M Facette.

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