MAIZE GENETICS AND GENOMICS DATABASE

# MaizeGDB STATUS REPORT

## *UPDATES, ACTIVITIES, AND NEW INITIATIVES*

USDA-ARS
Project No. 5030-21000-068 (Ames, IA)

Prepared by: The MaizeGDB Team

Carson Andorf, Ethalinda Cannon, Jack Gardiner, Lisa Harper, John Portwood,
Mary Schaeffer, and Margaret Woodhouse

**SEPTEMBER 26, 2018**

**Contact: C. 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** 3

**2 – MaizeGDB Mission and Progress Report** 4

**3 – Personnel list and formal collaborations** 10

**4 – Response to 2016 charge questions** 13

**5 – MaizeGDB 5-year Project Plan for USDA-ARS** 15

**6 – Recent Presentations and outreach** 18

**7 –Publications since March 2016** 19

## 1 – Meeting agenda

**Wednesday, September 26, 2016** *Times shown as Eastern/Central/Mountain/Pacific*
1pm/12pm/11am/10am – **Connect to the Web** (available 30 min earlier to enable early connection for anyone who wants to try it out). Online instructions will be sent by email.

12:00 pm (Central) Presentation: MaizeGDB's activities, accomplishments, and project plan

- MaizeGDB/Working Group Introductions (5 min)
- MaizeGDB 2017/2018 Accomplishments (15 min)
- MaizeGDB Project Plan (30 min)
  - Genome Stewardship and Pangenome (10 min)
  - Infrastructure to integrate genetic, genomic, and phenotypic data (10 min)
  - Curation (6 min)
  - Community Service (2 min)
  - External Collaboration (2 min)
- Charge Questions (10 min)

1:00 pm (Central) Working Group Executive Session
1:50 pm (Central) Working Group Summarizes for the MaizeGDB Team
Meeting adjourns (~2:00 pm Central)

### Working Group's Role

The Working Group is tasked with evaluating MaizeGDB's current status and recommending a course of action that will ensure 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.

### Current Working Group Membership

Alice Barkan, Qunfeng Dong, Candice Hirsch, David Jackson, Thomas Lübberstedt, Dorrie Main (chair), Marty Sachs (*ex-officio*), James Schnable, and Mark Settles.

## 2 – MaizeGDB Mission and Progress Report

**MaizeGDB is a community-oriented, long-term, federally funded informatics service to researchers focused on the crop plant and model organism Zea mays.**

For applied researchers to benefit from basic investigations, generated data must be made freely and easily accessible. MaizeGDB, the Maize Genetics and Genomics Database (http://www.maizegdb.org), is the research community's central repository for genetics and genomics information. The overall aim of our work is to create and maintain unified public resources that facilitate access to the outcomes of maize research. We provide reference genome stewardship within the context of extensive genomic diversity by adopting, developing, and deploying tools that enable community members to annotate and document updates to the genome and by developing and deploying genome visualization tools that enable user-friendly interaction with reference genomes and diversity data. In addition, we enable researchers to access data in a customized and flexible manner by deploying tools that facilitate direct interaction with the MaizeGDB database. Continued efforts to engage in education, outreach, and organizational needs of the maize research community involve the creation and deployment of tutorials, updating maize Cooperators on developments of interest to the community, and supporting the information technology needs of the Maize Genetics Executive Committee and Annual Maize Genetics Conference Steering Committee. To maximize tools, resources, and creation and enforcement of standards, we work with a group of 25 other Ag- related databases in the AgBioData Consortium.

**Progress Report:** Staff working at MaizeGDB provided tools and resources that make the maize genome sequence useful for investigative research and crop improvement. Genome sequences served include the latest version of the B73 representative genome (RefGen_v4), and the recently released genome sequence assemblies of 10 additional maize inbred lines. Genotype data for thousands of various other lines that represent the broad diversity represented by the *Zea* genus are also available as downloads or through a recently developed tool to visualize maize diversity (SNPversity; https://www.maizegdb.org/diversity). MaizeGDB focuses on curating high-quality, high-impact data sets for the maize research community. This includes the curation of over 75 datasets from the maize community that were added to the MaizeGDB genome browser. The current versions of the genome browser have over 150 datasets (tracks). In addition, MaizeGDB now supports 14 genome browsers for recently released maize genomes and over 90 data sets that can be used as targets for sequence similarity searches. This allows researchers to leverage research outcomes from many different sources and data types, all within the context of the maize reference assemblies. To enable a better understanding of how the genes in a plant define the potential phenotypes that will be observed in farmers' fields, we maintain a pathway view tool suite called CornCyc (https://www.maizegdb.org/metabolic_pathways/). Curation efforts have targeted high quality datasets and tools to support maize trait analysis, germplasm analysis, genetic studies, and breeding. MaizeGDB also hosts a wide range of data including recent support of new data types such as genome metadata, gene expression, protein analysis, and variation data. To improve access and visualization of data types several new tools have been implemented to: explore the variation and diversity in maize, download and compare how genes are expressed in the plant data, visualize pedigree data, link genes with phenotype images, and enable flexible user-specified queries to the MaizeGDB database. MaizeGDB also continues to be the community hub for maize research, coordinating activities and providing technical support to the maize research community. These resources provide long-term support, stability, and maintenance to maize research data and accelerate maize trait analysis, germplasm analysis, genetic studies, and breeding through better data access and utilization.

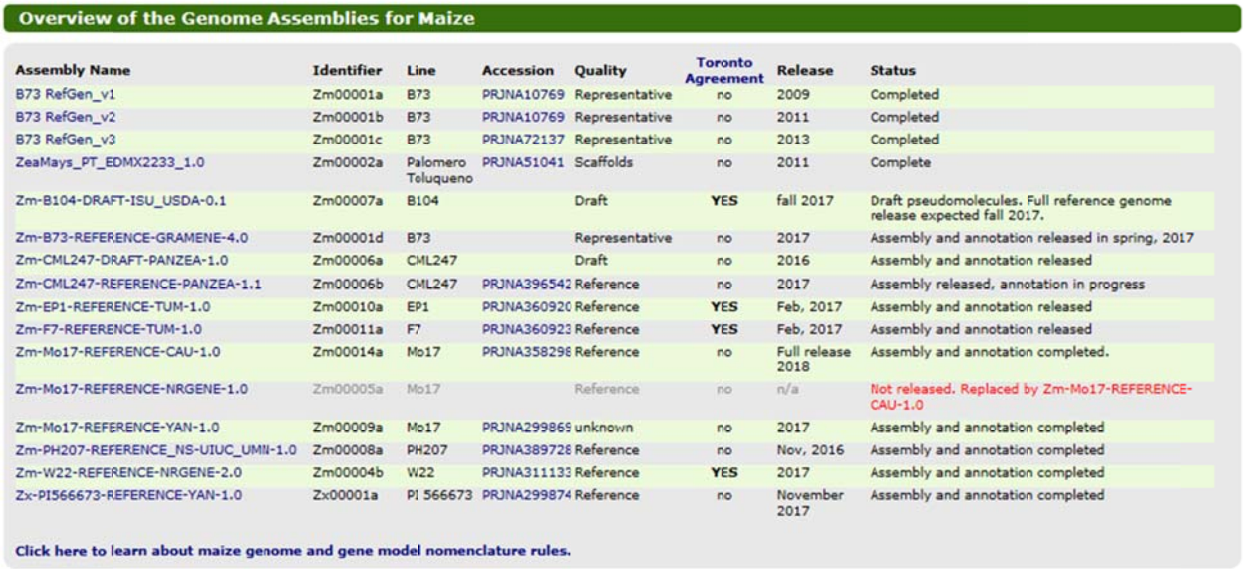
## Recent Accomplishments:

### Genome Assembly Stewardship

One of the main priorities at MaizeGDB is to provide genome assembly and annotation stewardship for the maize research community. We provide tools and resources that make the maize genome sequence useful for investigative research and crop improvement. MaizeGDB currently hosts information for multiple high-quality genome assemblies (including B104, B73, CML247, EP1, F7, Mo17 (Lai), Mo17 (Yan), PH207, and W22) (https://maizegdb.org/genome/assemblies_overview) and has integrated them with data held by MaizeGDB. This enables both exploring individual genomes, and comparing them in sets.

In anticipation of more genomes expected in the near future, MaizeGDB developed a set of minimum standards for hosting a new genome assembly, designed templates for collecting essential metadata related to the genome and assembly and annotation, enforced naming conventions set out by the maize nomenclature committee, and, after helping 5 groups submit complete assemblies to Genbank, we have created documentation to help submit genome assemblies to GenBank, and developed a pipeline for loading new assemblies. All of this enables comparative analysis.

In addition to bringing in new genome assemblies and providing the research community with a means of improvement, MaizeGDB will continue stewardship of the B73 genome assembly and annotation (https://maizegdb.org/assembly), which is expected to remain the representative reference maize genome assembly for the foreseeable future. Multiple, high-quality genome assemblies and annotations integrated with trait, phenotype, and germplasm data, will improve researchers' ability to conduct trait and germplasm analyses and to identify appropriate germplasm for breeding programs.



Overview of the Genome Assemblies for Maize

| Assembly Name | Identifier | Line | Accession | Quality | Toronto Agreement | Release | Status |
|---|---|---|---|---|---|---|---|
| B73 RefGen_v1 | Zm00001a | B73 | PRJNA10769 | Representative | no | 2009 | Completed |
| B73 RefGen_v2 | Zm00001b | B73 | PRJNA10769 | Representative | no | 2011 | Completed |
| B73 RefGen_v3 | Zm00001c | B73 | PRJNA72137 | Representative | no | 2013 | Completed |
| ZeaMays_PT_EDMX2233_1.0 | Zm00002a | Palomero Toluqueno | PRJNA51041 | Scaffolds | no | 2011 | Complete |
| Zm-B104-DRAFT-ISU_USDA-0.1 | Zm00007a | B104 |  | Draft | YES | fall 2017 | Draft pseudomolecules. Full reference genome release expected fall 2017. |
| Zm-B73-REFERENCE-GRAMENE-4.0 | Zm00001d | B73 |  | Representative | no | 2017 | Assembly and annotation released in spring, 2017 |
| Zm-CML247-DRAFT-PANZEA-1.0 | Zm00006a | CML247 |  | Draft | no | 2016 | Assembly and annotation released |
| Zm-CML247-REFERENCE-PANZEA-1.1 | Zm00006b | CML247 | PRJNA396542 | Reference | no | 2017 | Assembly released, annotation in progress |
| Zm-EP1-REFERENCE-TUM-1.0 | Zm00010a | EP1 | PRJNA360920 | Reference | YES | Feb, 2017 | Assembly and annotation released |
| Zm-F7-REFERENCE-TUM-1.0 | Zm00011a | F7 | PRJNA360923 | Reference | YES | Feb, 2017 | Assembly and annotation released |
| Zm-Mo17-REFERENCE-CAU-1.0 | Zm00014a | Mo17 | PRJNA358298 | Reference | no | Full release 2018 | Assembly and annotation completed. |
| Zm-Mo17-REFERENCE-NRGENE-1.0 | Zm00005a | Mo17 |  | Reference | no | n/a | Not released. Replaced by Zm-Mo17-REFERENCE-CAU-1.0 |
| Zm-Mo17-REFERENCE-YAN-1.0 | Zm00009a | Mo17 | PRJNA299869 | unknown | no | 2017 | Assembly and annotation completed |
| Zm-PH207-REFERENCE_NS-UIUC_UMN-1.0 | Zm00008a | PH207 | PRJNA389728 | Reference | no | Nov, 2016 | Assembly and annotation completed |
| Zm-W22-REFERENCE-NRGENE-2.0 | Zm00004b | W22 | PRJNA311133 | Reference | YES | 2017 | Assembly and annotation completed |
| Zx-PI566673-REFERENCE-YAN-1.0 | Zx00001a | PI 566673 | PRJNA299874 | Reference | no | November 2017 | Assembly and annotation completed |

Click here to learn about maize genome and gene model nomenclature rules.

**Genome Assemblies Overview.** Overview of the genome assemblies currently publicly listed at MaizeGDB. MaizeGDB hosts maize genomes with metadata, structural and functional annotations, genome browsers, and BLAST targets.

### MaizeMine, a Data Warehouse for Maize

In collaboration with Dr. Chris Elsik at the University of Missouri, MaizeGDB has developed MaizeMine, a working interoperability data warehouse based on the InterMine software package. MaizeMine accelerates genomic analysis by enabling researchers without scripting skills to create and export customized annotation datasets merged with their own research data for use in downstream analyses. MaizeMine integrates genomic sequences and gene annotations from the B73_RefGen_v3 and v4 genome assemblies, Gene Ontology, protein annotations (UniProt), protein families and domains (InterPro), homologs (Ensembl Compara), and pathways (CornCyc, KEGG, Plant Reactome). It also provides simple and sophisticated search tools, including a keyword search, built-in template queries with intuitive search menus, and a QueryBuilder tool for creating custom queries. The Regions search tool executes queries based on lists of genome coordinates, and supports both B73_RefGen_v3 and v4. The list tool allows users to upload identifiers to create custom datasets, perform list operations such as unions and intersections, and execute template queries with lists. When used with gene identifiers, the list tool automatically provides gene set enrichment for GO and pathways, with a choice of statistical parameters and background gene sets.

### SNPversity

SNPversity allows maize researchers to select a customized set of maize lines and a genomic region, and visualize DNA variations for that genomic region. The tool is loaded with datasets from Panzea that contain approximately 1 million regions of DNA variation for over 17,000 public maize lines. The SNPversity tool, along with the PedNet tool described below, has been built according to a survey conducted among maize researchers. The tool was upgraded to improve the flexibility of searches, produce more detailed results, and provide estimates on query runtimes. This tool has utility for maize geneticists trying to identify and clone genes of interest, relate genomic regions to phenotype, and understand the diversity in maize.  Genotype data for thousands of maize lines and individuals that represent the broad diversity represented by the *Zea* genus (i.e., maize and its near relatives) are also available as downloads through our Diversity Data center at https://www.maizegdb.org/diversity.

### Pedigree Viewer (PedNet)

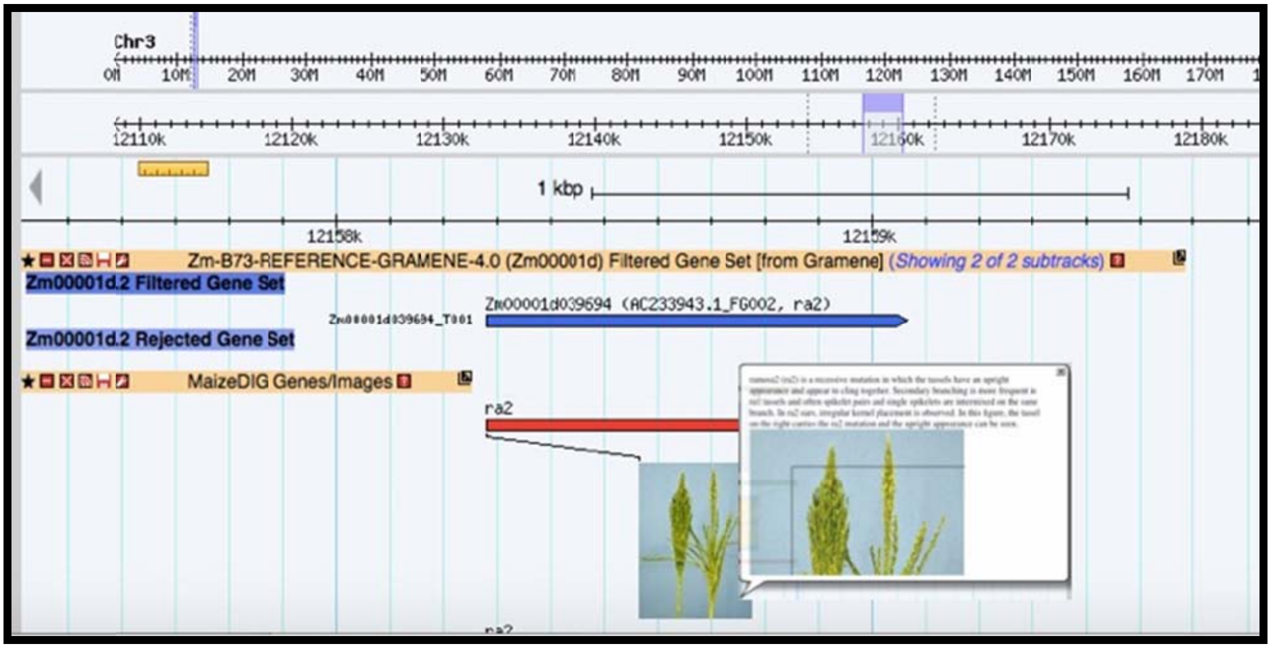
The Pedigree Viewer is a web-based pedigree viewer for maize. It allows users to build a network around a line, find the shortest path between two lines, build a network around the least common ancestor of two stocks, and filter pedigree networks by state, developer, source, and country. The PedNet viewer at MaizeGDB has a dynamically-generated pedigree network of 4,706 maize lines and 5,487 relationships. The tool has also been integrated into MaizeGDB's stock pages to show interactive images of pedigree trees. The Pedigree Viewer is customizable and allows users to upload their own data. This tool will help maize breeders identify appropriate maize lines to use in breeding projects which will improve available germplasm for researchers and farmers.

### qTeller

qTeller (http://www.qteller.com) is an RNA-seq processing pipeline and modular web interface that has been used to study quantitative expression variation in maize kernel row number, chitinase transcription during maize development , and other comparative expression analyses in maize, *Arabidopsis thaliana*, and *Brassica rapa*. qTeller draws RNA expression graphs for multiple RNA-seq datasets in real time for any gene of interest, and can be used to compare two genes at once by drawing a dot plot of relative expression for each RNA-seq dataset. MaizeGDB now hosts a version of qTeller (https://qteller.maizegdb.org) that currently has RNA-seq datasets from six different publications for B73, covering over 150 different tissues, conditions, and/or time points, and will soon host RNA-seq datasets for the inbred lines W22, Mo17, and others.

### MaizeDIG

MaizeDIG is an easy-to-use and easily extensible web-based resource to link genes, alleles, and genomes to phenotypes in images. MaizeDIG is based on the GMOD tool BioDIG, but has been enhanced to support multiple genomes and integrated with genome browsers to make tracks showing mutant phenotypes images within their genomic context. The ability to view images of mutant phenotypes at the genes' position on our genome browser is a new and unique feature at MaizeGDB. MaizeDIG allows for custom tagging of images to highlight regions related to the phenotypes and to curate and search by gene model, gene symbol, gene name, and allele. MaizeDIG has over 2,300 preloaded images available for 10 different genome browsers at MaizeGDB. Mutant phenotype images of 85 classical maize genes have been manually tagged to clarify and highlight phenotypes.


Chr3
0M 10M 20M 30M 40M 50M 60M 70M 80M 90M 100M 110M 120M 130M 140M 150M 160M 170M
12110k 12120k 12130k 12140k 12150k 12160k 12170k 12180k
1 kbp
12158k
12159k
Zm-B73-REFERENCE-GRAMENE-4.0 (Zm00001d) Filtered Gene Set [from Gramene] (Showing 2 of 2 subtracks)
Zm00001d.2 Filtered Gene Set
Zm00001d039694 (AC233943.1_FG002, ra2)
Zm00001d039694_T001
Zm00001d.2 Rejected Gene Set
MaizeDIG Genes/Images
ra2
ra2

ramosa2 (ra2) is a recessive mutation in which the tassels have an upright appearance and appear to cling together. Secondary branching is more frequent in ra2 tassels and often spikelet pairs and single spikelets are intermixed on the same branch. In ra2 ears, irregular kernel placement is observed. In this figure, the tassel on the right carries the ra2 mutation and the upright appearance can be seen.

**MaizeDIG data on the genome browser**: The screenshot shows what a mutant phenotype image of ra2 looks like after being tagged in MaizeDIG.

### AgBioData

MaizeGDB established AgBioData in 2015, and we continue to lead this now global effort. AgBioData is a consortium of agricultural biological databases and associated resources working together to ensure standards and best practices for acquisition, display and retrieval of genomic, genetic and breeding data. AgBioData has just completed our first work of establishing standards and best practices for: biocuration, ontology use and persistence, metadata database platforms, machine access to data, communication and sustainability (Database, 2018 coming soon). It is difficult for disparate groups to come to agreement, and thus this was a huge effort. As both generation of large data sets and the use of genomic databases increases exponentially, support for databases remain flat. Members of the AgBioData consortium are working together to build tools and increase interoperability so none of us re-invents the wheel. We are also promoting the FAIR data principles and hope to help researchers be good data stewards.

### Data Highlights for 2017-2018

- CornCyc versions 8.0/9.0 were released at MaizeGDB for the B73 _RefGen_v4 gene models after being developed by the Plant Metabolic Network in collaboration with MaizeGDB.
- Curated and loaded functional annotation in the form of GO and PO terms, collected from CornCyc, UniProt, and student-curated literature.
- The maize core bin markers have been converted to SNP-based markers based on pan- genome anchors. This will allow easy projection of core bin SNPs on to the numerous new whole genome assemblies being hosted at MaizeGDB.
- MaizeGDB has a new genome browser track for over 300,000 EMS point mutations and over 5,000 Indels contained in 1,942 sequence indexed maize lines that are available from the Chinese Academy of Agricultural Sciences.

- MaizeGDB curators, in collaboration with personnel at the NCBI, have selected over 300 RNA-seq datasets from over 90 B73 maize tissues for inclusion in the pipeline used to develop the NCBI RefSeq gene models for B73_RefGen_v4.
- MaizeGDB has created a new genome browser track for B73_RefGen_v3 and v4 for NCBI dbSNP's 58 million non-redundant SNPs with permanent RS#s. This is critical, as these non-redundant RS#s represent the best way for tracking SNPs across new whole genome assemblies.
- MaizeGDB has integrated multi-tissue RNA-seq expression data onto the B73_RefGen_v4 browser so that researchers can observe relative gene expression at a genomic region of interest.
- Student curators, guided by MaizeGDB curators, are systematically curating Maize Meeting abstracts to identify and document information on gene model function that would otherwise be missed by automated annotation algorithms.
- MaizeGDB in collaboration with the Schnable laboratory at ISU has uploaded over 1200 gel pictures for a survey of 24 inbred lines with ISU molecular markers. Information on molecular markers is one the most frequently requested bulk downloads at MaizeGDB.
- MaizeGDB worked with collaborators from the Plant Breeding, Technical University of Munich, Germany to make publicly available two whole genome sequences of the maize Flint lines EP1 and F7. The European Flint reference sequences provide additional diversity to other previously sequenced lines and complement the maize pan-genome.
- MaizeGDB worked with NRGene to submit four whole genome assemblies (EP1, F7, PH207, W22) to GenBank.
- MaizeGDB worked with collaborators at NRGene and the Chinese Agricultural University in Beijing, China to make the Lai lab version of Mo17 genome sequence publically available.
- MaizeGDB is working with collaborators at the Iowa State University and the ISU Transformation Facility to sequence, assemble, and annotate the maize line B104. New data includes 30X PacBio sequences. B104 is the inbred line used by the Transformation Facility to produce all of its primary transformants.
- 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.

## 3 – Personnel list and formal collaborations

### 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. Member (ex-officio) of the Maize Genetics Executive and Maize Genetics Conference Steering committees.

**Ethalinda (Ethy) Cannon**, Bioinformatics Engineer, half-time, USDA-ARS in Ames, IA
Responsible for creating metadata standards and templates for genomes and structural annotation, design and development of utilities to load, search, integrate, and download genome and gene model data, including functional annotation associated with gene models; assists with overall website and database development and support. Member of the maize nomenclature and AgBioData steering committees. .

**Lisa Harper**, geneticist (curator & outreach coordinator), half time, USDA-ARS in Albany, CA
Responsible for community outreach and education, video tutorials, Editorial Board management, literature curation, and identification of datasets of community-wide interest. Chair of the AgBioData consortium, member of the Maize Nomenclature Committee, and external advisor for the NSF Midwest Big Data Digital Agriculture Hub.

**John Portwood**, IT-Specialist, USDA-ARS in Ames, IA
Responsible for developing and administering the MaizeGDB interface, database, genome browsers, CornCyc, and various other tools. John maintains the hardware supporting our infrastructure, and is responsible for data security, backups, and disaster recovery. He serves as the abstract coordinator and webmaster for the annual maize genetics conference, and also administers elections and surveys for the maize community.

*Contract*

**Mary Schaeffer**, geneticist (curator), USDA-ARS (retired, 10% time) in Columbia, MO
Responsible for curation involving maps, loci, gene models, literature, QTL, etc. Formerly 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.

*Iowa State* University
**Margaret Woodhouse,** Scientist, full-time, ISU in Ames, IA
Responsible for genome curation, bioinformatics, and comparative genome analyses of sequenced maize genomes. Also identifies and maps RNA-seq data for qTeller, and assists in genome stewardship and tool development.

*University of Missouri*
**Jack Gardiner**, curator, full-time, Missouri University in Columbia, MO
Curator on location at Columbia, Missouri: Responsible for identifying and recruiting gene expression, proteomics, physical and genetic mapping, and epigenetic datasets with a special emphasis on large data sets. Jack leads the curation effort on implementing MaizeMine – a maize instance of the InterMine data warehousing software.

*Iowa State University Students*

**Kyoung Tak Cho**, graduate student, ISU in Ames, IA
Responsible for development of MaizeDIG (a tool to integrate phenotypic images to genomic context), and predictive phenomics. He is also developing machine learning approaches to predict gene expression and protein abundance in maize genes.

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

**Sagnik Banerjee,** graduate student, ISU in Ames, IA
Responsible for developing tools to globally analyze and correct maize gene annotation sets.

**Miranda Dietze,** undergraduate hourly curator, ISU in Ames, IA
Curates gene function from literature.

*Vacant positions (Federal):*

IT-Specialist, full-time permanent
Computational Biologist, full-time permanent

## Formal Collaborations

**Project title**: Development of MaizeMine, a Flexible Data Warehouse and Analysis System for Retrieval of Customized Datasets from the MaizeGDB Database
**Principal Investigators:** Carson Andorf, Christine Elsik (University of Missouri)
**MaizeGDB Personnel**: Jack Gardiner

**Objectives**: Continue development of MaizeMine, based on the InterMine data mining platform, to allow MaizeGDB users to query the maize database directly to retrieve large-scale, customized datasets in a variety of formats.

**Approach**: The InterMine software application has been used by Elsik laboratory personnel at the University of Missouri to develop MaizeMine for long-term use by MaizeGDB stakeholders. Core data sets have already been loaded into MaizeMine and additional high-value data sets will be loaded in the coming year. Anticipated benefits include providing the maize genetics and breeding communities with a powerful set of tools for parsing large-scale datasets to both identify and understand either genes or gene networks that underpin basic biological processes that contribute to crop productivity.

**Project title**: Maize pan-genomic tools and resources to explore the diversity of maize
**Principal Investigators:** Carson Andorf, Matthew Hufford (Iowa State University)
**MaizeGDB Personnel**: Margaret Woodhouse

**Objectives**: To use the upcoming maize genome assemblies from the maize nested association mapping (NAM) population founders along with other existing maize genome assemblies to curate the maize pan-genome and make maize assemblies, annotations, and metadata assessable to the maize research community through MaizeGDB. This work includes the creation, implementation, and updating of pan-genomic tools and resources incorporating representative cultivars of *Zea mays*, its wild relatives, and other related grasses.

**Approach**: The development and maintenance of a maize pan-genome will be done in collaboration with the maize community. Curation of the maize pan-genome and the upcoming genomes of maize and its wild relatives will include issuing genome and gene IDs in accordance with the Maize Nomenclature Committee, collecting statistics and metadata for each genome, helping sequencing labs to upload their genomes to NCBI and MaizeGDB, and performing comparative genomics analyses. Existing pan-genomic tools and resources will be updated to include data from the NAM founders. These tools/resources include the MaizeGDB genome and annotation pages, qTeller RNA expression visualization tool, genome browser tracks, and BLAST targets. Together, these data will provide an important comparative genomics resource for MaizeGDB stakeholders.

## 4 – Response to 2016 charge questions

Working Group response to Charge questions of 2016 (summarized) with MaizeGDB action to the responses in **bold**.
Accomplishments section:
https://documents.maizegdb.org/working_group/2016MaizeGDBWorkingGroupResponse.pdf

**Genome Assembly Stewardship**

- High-quality structural and functional annotation of reference genomes is made a top priority, though acknowledging that obtaining external funding for genome curation is extremely difficult
    - **MaizeGDB has included high-quality structural and functional annotation of reference genomes as a priority in the 5-year project plan.**
- Incentivizing and facilitating user improvements to the assemblies and annotation
    - **MaizeGDB has developed functionality for reporting assembly and annotation issues. Links to these forms are located in several locations on the MaizeGDB website.**
- Manual curation efforts should be limited to only a few of the highest quality genomes
    - **Manual curation has been limited to associating genes to gene models in the B73 genome. The same associations can be inferred based on orthologs among maize assemblies.**
- Taking ownership of these data will ease future assembly updates, allowing MaizeGDB to regenerate browser tracks when reference genomes are updated. Scripted regeneration of data tracks would be less prone to error than manually mapping old tracks to new assemblies.
    - **MaizeGDB has not acquired ownership of any of the maize genomes, but have worked closely with each of the sequencing groups.**
- Assisting users with NCBI submissions and enabling users to upload and browse their own custom tracks alongside public MaizeGDB data...several platforms currently provide similar services (e.g. Gramene, CyVerse), and MaizeGDB should talk with these groups to explore the reuse of existing tools.
    - **MaizeGDB has assisted several groups in uploading their genomes to NCBI including W22, PH207, EP1, and F7. MaizeGDB is also working with the NAM sequencing project to get the 25 founder lines submitted to NCBI.**

**Big data set identification, evaluation, and incorporation**

- Incorporation of new datasets should be prioritized based on semi-regular polling of an expert panel, stressing quality and importance over quantity.
    - **MaizeGDB has added additional ways to prioritize data curation in our 5-year project plan including consulting expert groups.**
- We believe that approaches such as automated literature curation are unlikely to succeed. Instead, we feel vacant curation positions should be staffed and community involvement incentivized.
    - **MaizeGDB has focused curation efforts on new maize genomes and targeted data sets that we feel will have high-impact (See data highlights in section 2).**
- We note that the community will be more likely to contribute if MaizeGDB provides some form of added value and appropriately maintains and archives the users' contributions.

- **Efforts have been made to create added value to the following data types: genomes, gene annotations, expression, diversity (SNP), images, and pedigree. In addition, we now give “badges” to community members who contribute.**

**Tool development**

- All development projects should be very carefully considered in terms of their cost versus benefit; freely available tools and services that may provide the desired functionality.
    - **We have considered free tools versus internally developed tools for a lot of our core resources including the genome browsers (GBrowse) pathway tools (Pathway Tools), expression visualization (qTeller), image curation (BioDIG), data management (InterMine), and sequence searches (BLAST).**
- MaizeGDB limits new development efforts, especially until full staffing is restored.
    - **MaizeGDB has limited internal development to the SNPversity and the Pedigree Viewer tools. We have focused on improving and integrating existing tools to be optimized for maize. New tool development is focused on R applications that perform a specific task and are easy to code, integrate, and maintain.**
- Efforts should be made to identify and improve the most critical and frequently used core services. This may include updating the genome browser to be more responsive and developing curation tools in collaboration with the Maize Genetics Stock Center to allow more efficient manual literature curation.
    - **MaizeGDB has put effort into improving the Genome Browser including upgrading the hardware. Due to lack of resources we have not updated the curation tools.**

**Future needs and expectations**

- Because of looming influx of new genomes and functional datasets generated by emerging sequencing technologies...we suggest adding more sequencing and genomics expertise to the team to help prepare for these new data.
    - **We hired Margaret Woodhouse, an associate scientist with Iowa State University, to lead efforts for genome curation and comparative genomics.**
- The maize pan-genome: As many whole genomes become available, the community will require interfaces for integrating data across the pan-genome, both to link between different copies of core genes and explore the diversity of accessory genes.
    - **We are monitoring developments in this field. We are in regular contact with Ed Buckler, Doreen Ware, and Paul Chomet (NRGene) about ways to represent and visualize a maize pan-genome.**
    - **We are working with the AgBioData Genome Nomenclature committee to develop at least a plant-wide standard genome and pan genome nomenclature to avoid/minimize confusion.**
- MaizeGDB to provide a centralized and reliable hub for the maize community with a focus on providing high-quality, reliable data and services.
    - **MaizeGDB continues to provide community support to the maize research community. We have also expanded these efforts with updated person pages to better recognize contributions, updated community calendar, digitized Maize Newsletter, and new workshops at the Maize Genetics Conference.**

## 5 – Five Year Project Plan

**ARS Process:**

***National Program 301 (Plant Genetic Resources, Genomics and Genetic Improvement)***

- ✓ **Website:**
    - https://www.ars.usda.gov/crop-production-and-protection/plant-genetic-resources-genomics-and-genetic-improvement/
- ✓ **Action Plan:**
    - https://www.ars.usda.gov/crop-production-and-protection/plant-genetic-resources-genomics-and-genetic-improvement/docs/action-plan-2018-2022/
- ✓ **Project Plan Title:** MaizeGDB: Enabling Access to Maize Breeding and Genomics Resources
- ✓ **Project Plan due to the ARS Office of Scientific Quality Review:** August 7, 2017
- ✓ **Project Plan certified:** March 23, 2018
- ✓ **Overall score (Ames):** 5.6 (of 8 possible). Minor revision.
- ✓ **Project start date:** April 10, 2018
- ✓ **Project end date:** April 9, 2023

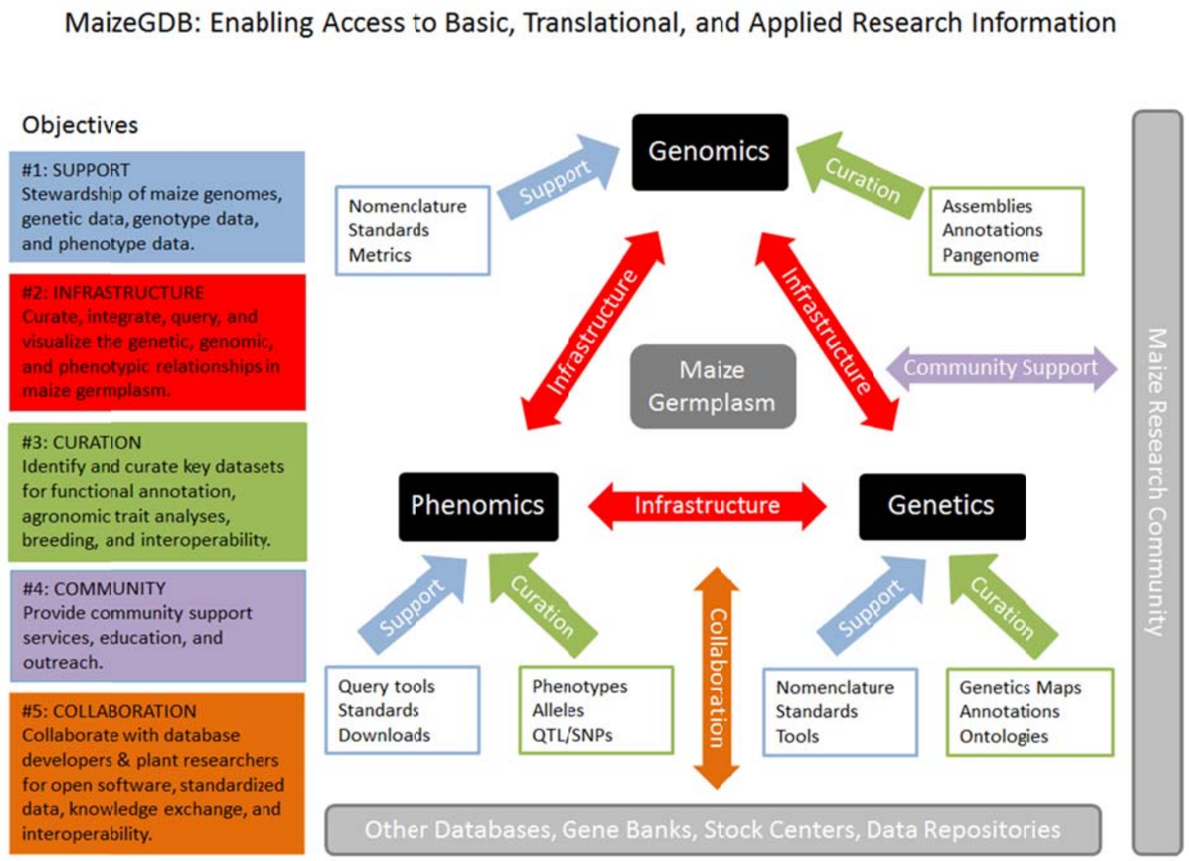
**OBJECTIVES**

The long-term objective of this project is to provide ongoing support and stability to the research community by providing a community database and informatics resource to store, query, integrate, analyze, and visualize maize genetics and genomics data. The MaizeGDB team performs non-hypothesis-driven research to fulfill these objectives. Our website, database, and underlying resources allow plant researchers to understand 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 to be met over the next five years will be storing, providing access to, and visualizing a variety of large-scale data types. Specifically, MaizeGDB will support and curate data that will highlight the genomic, genetic, and phenotypic relationships in maize and link it back to available maize germplasm. Some of the data types include, but are not limited to, multiple whole-genome sequences, pan-genomic data, association studies, expression data, and functional annotations. To address these challenges, we will implement the following objectives:

> **Objective 1:** Accelerate maize trait analysis, germplasm analysis, genetic studies, and breeding through stewardship of maize genomes, genetic data, genotype data, and phenotype data.
>
> **Objective 2:** Develop an infrastructure to curate, integrate, query, and visualize the genetic, genomic, and phenotypic relationships in maize germplasm.
>
> **Objective 3:** Identify and curate key datasets for benchmarking genomic discovery tools for the functional annotation of maize genomes, for agronomic trait analyses, for breeding (including genome editing), and for improving database interoperability.
>
> **Objective 4:** Provide community support services, training and documentation, meeting coordination, support for community elections and surveys, and support for the crop genome database community.

> **Objective 5:** Collaborate with database developers and plant researchers to develop improved methods and mechanisms for open, standardized data and knowledge exchange to enhance database utility and interoperability.

The central goal of this project is to allow the plant research community (with strong focus on the maize community) to utilize maize genetics and genomics data for both crop improvement and for basic research, using maize as a model system. The objectives to accomplish this goal can be broken down into five categories: support, infrastructure, curation, community, and collaboration. A flow-diagram showing the interconnectedness of these five categories is shown below. Support will be provided by the continued stewardship of the B73 maize reference assembly, other reference-quality assemblies, and their integration in a pan-genomic representation. In addition, the management of other data types including genotypes, phenotypes, gene expression, and association data will enable researchers to formulate and test functional hypotheses to understand the complexity and diversity of maize. The infrastructure component of our project will include tools, resources, and services to help integrate these datasets and build graphical interfaces to visualize the genetic, genomic, and phenotypic relationships in maize germplasm. Curation of high-quality and high-impact datasets has been the foundation of the MaizeGDB project since its inception. The project will balance the curation of traditionally important datasets that have long-term value and datasets with recent relevance based on current technologies and research trends. Focus will be on datasets that allow for high-quality functional annotation, agronomic trait analyses, breeding, and for cross-linking with other plant databases. The community outreach component of our plan is important to improve the utility and visibility of the MaizeGDB resources. This project has taken an active role in the maize research community by providing outreach and community support through tutorials/training/workshops, meeting coordination, community committee support, and community resource (e.g. job boards, calendars mailing list, etc.). The final component of the plan is collaboration. This includes both technology interoperability (cross-linking and data sharing) and coordination with other databases (data standards, tool co-development, and open collaborations).



MaizeGDB: Enabling Access to Basic, Translational, and Applied Research Information

Objectives

#1: SUPPORT
Stewardship of maize genomes, genetic data, genotype data, and phenotype data.

#2: INFRASTRUCTURE
Curate, integrate, query, and visualize the genetic, genomic, and phenotypic relationships in maize germplasm.

#3: CURATION
Identify and curate key datasets for functional annotation, agronomic trait analyses, breeding, and interoperability.

#4: COMMUNITY
Provide community support services, education, and outreach.

#5: COLLABORATION
Collaborate with database developers & plant researchers for open software, standardized data, knowledge exchange, and interoperability.

**MaizeGDB Project Flow Diagram**: The flow diagram shows how the project is divided into 5 interconnected categories, each corresponding to one of the 5 objectives: Support, Infrastructure, Curation, Community, and Collaboration. This example shows how these categories are applied to the genomics, phenomics, and genetics relationship in maize.

## 6 – Presentations/Outreach

2018

Andorf C (speaker). **MaizeGDB: How phenotype curation has co-evolved with genomic representations**. Short Talk at 5th Annual International Plant Phenotyping Symposium, Adelaide, Australia in 2018.

Harper L (speaker). **AgBioData consortium hits its stride**. Keynote speaker at NSF Big Data Hub: Digital Agriculture Community Meeting, Lincoln NE in 2018.

Andorf C (speaker). **Databases and Data Management in the Maize Research Community**. Short Talk at Maize Genetics Coordination Network Meeting, Madison, Wisconsin in 2018.

Cho KT (speaker), Portwood J, Harper LC, Gardiner JM, Lawrence CJ, Friedberg I, Andorf CM. **MaizeDIG: A mechanism for connecting gene models to phenotypes at MaizeGDB.** Short Talk at 26th Annual Conference Intelligent Systems for Molecular Biology (ISMB), Chicago, IL in 2018.

Wimalanathan K (speaker), Friedberg I, Andorf CM, Lawrence-Dill C. **Maize GO Annotation - Methods, Evaluation, and Review (maize-GAMER).** Short Talk at 6th Annual Conference Intelligent Systems for Molecular Biology (ISMB), Chicago, IL in 2018.

Cannon E (speaker) **How to be a good data citizen.** Short Talk at Community session at 60th Annual Maize Genetics Conference (MGC 2018), St. Malo, France.

Wimalanathan K (speaker), Friedberg I, Andorf CM, Lawrence-Dill C. **Maize - GO Annotation Methods Evaluation and Review (Maize-GAMER).** Short Talk at 60th Annual Maize Genetics Conference (MGC 2018), St. Malo, France.

Woodhouse M (speaker), Cannon EK, Andorf CM. **The Importance of Getting Genome Assemblies into Genbank, and How to Do It.** Short Talk during the Big Data: Manage Your Data Before Your Data Kills You Workshop at Plant and Animal Genome Conference, San Diego, CA in 2018.

Woodhouse M (speaker). **How the Maize Genome Database Helps Guide Data Management Best Practices in the Maize Community.** Short Talk during the Challenges and Opportunities in Plant Science Data Management - an International Workshop at Plant and Animal Genome Conference, San Diego, CA in 2018.

Elsik C (speaker). **MaizeMine: A Data Mining Warehouse for MaizeGDB.** Short Talk during the Digital Tools and Resources Session at Plant and Animal Genome Conference, San Diego, CA in 2018.

2017

Sen TZ (speaker), Braun BL, Schott DA, Portwood J, Schaeffer ML, Harper LC, Gardiner JM, Cannon EK, Andorf CM. (2017) **Surveying the Maize community for their diversity and pedigree visualization needs to prioritize tool development and curation.** Short Talk at the Biocuration 2017, Stanford, California, USA in 2017.

**MaizeGDB Outreach Activities**

MaizeGDB organized a series of six workshops at the Maize Genetics Conference in St. Louis, Missouri and four workshops at the Maize Genetics Conference in St. Malo, France. MaizeGDB team members Margaret Woodhouse, Jesse Walsh, and Lisa Harper led six of the workshops. Curator Lisa Harper and bioinformatic engineer Ethalinda Cannon also organize monthly seminars for the AgBioData consortium (https://www.agbiodata.org), which cover topics of importance to biocuration and biological databases. Lisa and Ethalinda were also among the organizers for an April 2017 meeting of AgBioData consortium in Salt Lake City, Utah. Over 45 scientists from 21 different institutions representing 32 agriculturally relevant databases and resources attended. This meeting covered topics in curation, metadata and persistence, database storage, data sharing using web services, policy, and communication. From these discussions and previous work, the AgBioData working group has published a white paper (Harper et al., 2018) on best practices in each of these areas. This will help enhance genetic, genomics and breeding research outcomes through standardization of practices and protocols across agricultural databases.

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 2017, over 20 databases participated in the Agricultural Database booth.

**Community Leadership Roles**

- USDA-ARS Executive Scientific Advisory Committee – C. Andorf
- USDA-ARS CERES High Performance Computing Policy Committee – J. Portwood
- USDA-ARS Database Committee - J. Portwood
- Maize Genetics Conference Steering Committee – C. Andorf (ex-officio)
- Maize Genetics Executive Committee – C. Andorf (ex-officio)
- AgBioDatabase Consortium Steering Committee – L. Harper (chair) and E. Cannon
- Maize Nomenclature Committee – L. Harper and E. Cannon

MaizeGDB provides technical support for the Maize Genetics Executive Committee, Maize Genetics Conference Steering Committee, the McClintock Prize for Plant Genetics and Genome Studies, and the Maize Community Awards.

**7 – Publications since March 2016**

2018

Alkhalifah N, Campbell DA, Falcon CM, Gardiner JM, Miller ND, Romay MC, Walls R, Walton R, Yeh CT, Bohn M, Bubert J, Buckler ES, Ciampitti I, Flint-Garcia S, Gore MA, Graham C, Hirsch C, Holland JB, Hooker D, Kaeppler S, Knoll J, Lauter N, Lee EC, Lorenz A, Lynch JP, Moose SP, Murray SC, Nelson R, Rocheford T, Rodriguez O, Schnable JC, Scully B, Smith M, Springer N, Thomison P, Tuinstra M, Wisser RJ, Xu W, Ertl D, Schnable PS, De Leon N, Spalding EP, Edwards J, Lawrence-Dill CJ. (2018) **Maize Genomes to Fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets.** BMC Res Notes. 2018 Jul 9;11(1):452. doi: 10.1186/s13104-018-3508-1.

Harper L, Campbell J, Cannon EK, Jung S, Main D, Poelchau M, Walls RL, Andorf CM, Arnaud E, Berardini, Birkett C, Cannon S, Carson J, Cooper L, Dunn N, Elsik C, Farmer A, Ficklin S, Grant D, Grau E, Hendon N, Hu Z, Humann J, Jaiswal P, Jonquet C, Laporte MA, Larmande P, Lazo G, McCarthy F, Menda N, Mungall C, Munoz-Torres M, Naithani S, Nelson R, Nesdill D, Park C, Reecy J, Reiser L, Sanderson LA, Sen TZ, Staton M, Subramaniam S, Karey Tello-Ruiz M, Unda V, Unni D, Wang L, Ware D, Wegrzyn J, Williams J, Woodhouse M. (2018) **AgBioData Consortium Recommendations for Sustainable Genomics and Genetics Databases for Agriculture.** Database. Volume 2018.

Reiser L, Harper L, Freeling M, Han B, Luan S (2018) **FAIR: A Call to Make Published Data More Findable, Accessible, Interoperable, and Reusable.** Molecular Plant, Vol. 11, Issue 9, p1105–1108.

Schott DA, Vinnakota AG, Portwood JL, Andorf CM, Sen TZ. (2018) **SNPversity: a web-based tool for visualizing diversity.** Database. Volume 2018.

Siegel ZD, Zhou N, Zarecor S, Lee N, Campbell DA, Andorf CM, Nettleton D, Lawrence-Dill CJ, Ganapathysubramanian B, Friedberg I, Kelly JW. (2018) **Crowdsourcing Image Analysis for Plant Phenomics to Generate Ground Truth Data for Machine Learning**. PLOS Computational Biology.

Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, Barbazuk WB, Bass HW, Baruch K, Ben-Zvi G, Buckler E, Bukowski R, Campbell MS, Cannon EKS, Chomet P, Dawe RK, Davenport R, Dooner HK, Du LH, Du C, Easterling KA, Gault CM, Guan JC, Hunter CT III, Jander G, Jiao YP, Koch KE, Kol G, Kollner TG, Kudo T, Li Q, Lu F, Mayfield-Jones D, Mei WB, McCarty DR, Noshay JM, Portwood JL, Ronen G, Settles AM, Shem-Tov D, Shi JH, Soifer I, Stein JC, Stitzer MC, Suzuki M, Vera DL, Vollbrecht E, Vrebalov JT, Ware DH, Wei S, Wimalanathan K, Woodhouse MHR, Xiong WW, Brutnell TP. (2018). **The maize W22 genome provides a foundation for functional genomics and transposon biology**. Nature Genetics. doi: 10.1038/s41588-018-0158-0.

Wimalanathan K, Friedberg I, Andorf CM, Lawrence-Dill CJ. (2018) **Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER).** Plant Direct. Vol 2, Issue 4. e00052.

2017

Odell SG, Lazo G., Woodhouse, MR, Hane DL, Sen TZ (2017) **The art of curation at a biological database: Principles and application**. Current Plant Biology. doi: 10.1016/j.cpb.2017.11.001.

Sen TZ, Braun BL, Schott DA, Portwood J, Schaeffer ML, Harper LC, Gardiner JM, Cannon EK, Andorf CM. (2017) **Surveying the Maize community for their diversity and pedigree visualization needs to prioritize tool development and curation**. Database, Vol 2017, Issue 1.

2016

Harper LC, Gardiner JM, Andorf CM, Lawrence CJ. (2016) **MaizeGDB: the maize genetics and genomics database.** Plant Bioinformatics; Humana Press, New York, NY, Pages 187-202.

Hoopen PT, Walls RL, Cannon EK, Cochrane G, Cole J, Johnston A, Karsch-Mizrachi I, Yilmaz P. (2016) **Plant specimen contextual data consensus**. Gigascience. 2016 Dec 1;5(1):1-4. doi: 10.1093/gigascience/giw002.

Walsh JR, Schaeffer ML, Zhang P, Rhee SY, Dickerson JA, Sen TZ. (2016) **The quality of metabolic pathway resources depends on initial enzymatic function assignments: a case for maize.** BMC Systems Biology 10:129